Distributed Monte Carlo simulation with large-scale Machine Learning : Bayesian Inference and Conformal Prediction

Vincent Plassier

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX063

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Distributed Monte Carlo simulation with application to large-scale Machine Learning: Bayesian Inference and Conformal Prediction

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'Ecole polytechnique

École doctorale n°574 Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 5 octobre 2023, par

VINCENT PLASSIER

Composition du Jury :

Gersende Fort Professor, Institut de Mathématiques de Toulouse	Présidente
Aurélien Bellet Chargé de recherche INRIA (HDR), University of Lille (CRIStAL)	Rapporteur
Gareth Roberts Professor, University of Warwick	Rapporteur
Sylvain Arlot Professor, Université Paris-Saclay	Examinateur
Christian Robert Professor, Université Paris Dauphine	Examinateur
Eric Moulines Professor, Ecole Polytechnique (CMAP)	Directeur de thèse
Alain Durmus Professor, Ecole Polytechnique (CMAP)	Co-directeur de thèse
Jean-Claude Belfiore Research engineer, Huawei Technologies France	Invité

À ma tante et à mon oncle,
Yvette et André

Remerciements

Je suis honoré et reconnaissant d'avoir achevé mon parcours doctoral. Je saisis cette occasion pour exprimer mes sincères remerciements à toutes les personnes qui m'ont aidé et soutenu tout au long de ce cheminement.

Tout d'abord, je tiens à exprimer ma profonde gratitude à mes directeurs de thèse Eric Moulines et Alain Durmus, dont l'expertise scientifique et l'esprit brillant ont été pour moi une source constante d'inspiration et de motivation durant cette période. Je tiens à saluer leur dévouement à la recherche. Un immense merci à Aymeric, pour son engagement scientifique, ainsi qu'à Maxim Panov pour nos fructueuses collaborations. Je suis très reconnaissant envers Johan Segers, pour sa bienveillance infinie.

Je salue les membres du jury, pour le temps consacré à l'évaluation des contributions. Je remercie particulièrement les rapporteurs Aurélien Bletter et Gareth Roberts, qui ont généreusement dédié leur temps à lire et à évaluer attentivement le manuscrit de ma thèse, et qui ont fourni des commentaires constructifs visant à améliorer la qualité de mon travail. Vos commentaires et suggestions perspicaces ont été d'une valeur inestimable, et je vous suis vivement reconnaissant. À tous les membres du jury qui étaient présents lors de la soutenance de mon doctorat. Pour les nommer: Sylvain Arlot et Christian Robert, ainsi que Gersande Fort, la présidente de mon jury. J'exprime ma sincère gratitude pour leur présence et l'intérêt qu'ils ont porté à ma recherche. Leurs questions stimulantes et leurs remarques pertinentes ont été très instructives, et je suis reconnaissant des efforts qu'ils ont déployés pour faire de cette soutenance une expérience précieuse d'apprentissage.

Je tiens à exprimer toute ma reconnaissance envers ma famille pour leur soutien indéfectible tout au long de mes études. En particulier, je remercie mon oncle et ma tante à qui la thèse est dédiée.

Je tiens à exprimer ma profonde gratitude à ceux qui m'ont soutenu tout au long de cette aventure. Un tout grand merci à Rémi, pour sa capacité à me motiver et pour ses innombrables conseils. Je remercie particulièrement Gilles pour nos échanges scientifiques ainsi que nos discussions passionnantes. Hamid pour les moments de détente passés autour d'une pizza Julia. Merci à Juliette, qui qui m'a soutenu lors de cette dernière année et qui m'a permis de garder le cap. Je remercie mes amis d'enfance Benjamin et Paul pour leur nature bienveillante.

Je n'oublie pas mes camarades du Centre Lagrange, avec qui j'ai partagé mon quotidien pendant cette thèse. Notamment, Mehdi pour l'aboutissement de notre projet commun, je lui souhaite le meilleur pour ses projets futurs. Je tiens à remercier les quatre fantastiques: Alex, Giaccomo, Sara, et Matthias, pour leur soutien et leur amitié. Une pensée pour Tom, l'ami des bulldogs, ainsi qu'à Louis pour sa tranquilité. Je suis très reconnaissant envers Pierre pour sa bonne humeur ainsi que ses bornes rivalisant de longueur avec les miennes. Je remercie Yazid pour les bons repas partagés ensembles, et je souhaite à Gabriel de continuer à s'épanouir dans la musique. Une mention spéciale à l'équipe zurichoise, Thomas pour sa sagesse et Valentin pour l'ambiance festive qu'il apporte au centre. Vous avez été une source de courage pendant cette thèse. Un grand merci aux assistantes, qui m'ont aidé quotidiennement, je pense particulièrement

à Yuqing, Xiaoyun et Ziwei. Bien que je n'aie pas pu vous dédier plus de lignes, je vous souhaite le meilleur pour la suite. Un grand merci à mes amis belges, en particulier ceux de Louvain-La-Neuve, avec qui j'ai partagé des moments inoubliables et des discussions passionnantes autour de sandwichs. Un grand merci à Dimitri, mon compagnon d'étude, pour ses précieux conseils et son soutien constant. Je remercie Louis Grenioux pour notre magnifique voyage à Hawaï. Je tiens à souligner la gentillesse d'Amin et de Naoufal, et je tiens à exprimer toute ma reconnaissance à Constantin pour l'aide que tu m'as apporté dans la dernière ligne droite. Une pensée pour Manon pour sa bienveillance, ainsi que pour Lisa, Leila, Antoine, Renaud, Mahmoud. Je vous remercie pour les bons moments passés ensemble. Une mention pour Anna Korba, pour sa précieuse aide. Je tenais également à remercier François Portier pour notre réalisation commune.

Publications

- **V. Plassier**, M. Vono, A. Durmus, and E. Moulines. DG-LMC: A turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs. In International Conference on Machine Learning, pp. 8577-8587. PMLR, 2021. [Plassier et al. (2021)]

- **V. Plassier**, M. Vono, A. Durmus, A. Dieuleveut, and E. Moulines. QLSD: Quantized Langevin Stochastic Dynamics for Bayesian federated learning. In International Conference on Artificial Intelligence and Statistics, pp. 6459-6500. PMLR, 2022. [Vono et al. (2022b)]

- **V. Plassier**, F. Portier, and J. Segers. Risk bounds when learning infinitely many response functions by ordinary linear regression. In Annales de l'Institut Henri Poincaré, et Statistiques. Institut Henri Poincaré, 2022. [Plassier et al. (2023c)]

- Jalalzai, H., E. Kadoche, R. Leluc, and **V. Plassier**. Membership inference attacks via adversarial examples. arXiv preprint arXiv:2207.13572. [Jalalzai et al. (2022)]

- **V. Plassier**, A. Durmus, and E. Moulines. Federated Averaging Langevin Dynamics: Toward a unified theory and new algorithms. International Conference on Artificial Intelligence and Statistics. PMLR, 2023. [Plassier et al. (2023b)]

- **V. Plassier**, M. Makni, A. Rubashevskii, E. Moulines and M. Panov. Conformal Prediction for Federated Uncertainty Quantification Under Label Shift. In International Conference on Machine Learning. PMLR, 2023. [Plassier et al. (2023a)]

- **V. Plassier**, N. Kotelevskii, A. Rubashevskii, F. Nosko, M. Velikanov, A. Fishkov, S. Horvath, M. Takac, E. Moulines and M. Panov. Efficient Conformal Prediction under Data Heterogeneity. [Preprint]

Abstract

Centralizing data is impractical or undesirable in many scenarios, especially when sensitive information is involved. In such cases, the need for alternative methods becomes evident. As large datasets are known to facilitate the learning of efficient models, distributed methods have emerged as a powerful tool to overcome the challenges posed by centralized data. Consequently, this thesis introduces innovative approaches to tackle large-scale Bayesian inference and uncertainty quantification, aiming to provide effective solutions in the context of distributed data environments. The federated Monte Carlo (MC) approaches allow multiple agents/nodes to conduct computations locally and securely, with a central server combining the results to obtain samples from the global posterior distribution. Bayesian posterior sampling techniques benefit from the incorporation of prior knowledge, leading to improved results. Additionally, the uncertainty associated with the parameters and the predictions are naturally quantified, which is crucial for decision-making. Especially with limited or noisy data, the ability to quantify uncertainty becomes even more essential.

The first part of this manuscript focuses on MC via Markov chains (MCMC) methods. In particular, we introduce two procedures, named `DG-LMC` and `FALD`, designed to target a global posterior distribution while ensuring scalability. Local agents are associated with a central server that aggregates information from each agent to generate samples from the posterior distribution. This approach minimizes the need to transmit large amounts of data across participating agents, making it especially advantageous in federated environments with limited bandwidth or low computational power. Considering the distributed nature of today's datasets, concerns about trust and confidence arise when transferring information to a central server. The proposed methods not only address practical applications but also extend existing learning algorithms to Bayesian inference problems. The proposed approach contributes to the development of more robust and efficient machine learning algorithms, and holds potential applications in various domains, including epidemiology and finance, where large-scale inference and data privacy are significant concerns. To demonstrate the effectiveness of the approach, real-world datasets are employed, and the results show the performance of federated MCMC simulation.

The second part of the thesis focuses on uncertainty management. Initially, we present the Bayesian approach, which involves defining a prior and a likelihood. To address bandwidth bottlenecks while efficiently generating samples, our proposed approach leverages compression operators. In the final part of this thesis, we introduce a novel frequentist FL method based on conformal predictions. Unlike other methods, our model-agnostic approach does not rely on specific model assumptions and can be applied to any underlying prediction model. Referred to as `DP-FedCP`, this method leverages quantile regression techniques to generate personalized prediction sets while maintaining robustness to outliers. The label shift between agents is addressed by determining quantiles based on importance weights. One crucial aspect of our approach is the preservation of differential privacy, it allows users to assess the confidence level of predictions and make informed decisions based on the associated level of uncertainty. By incorporating this privacy measure, we ensure safeguarding the user's sensitive information.

Résumé

Centraliser les données est indésirable dans de nombreux scénarios, notamment lorsque des informations sensibles sont traitées. Dans de tels cas, la nécessité de méthodes alternatives devient évidente. Puisqu'un grand nombre de données facilite l'apprentissage de modèles efficaces, les méthodes distribuées se sont imposées comme un outil puissant pour surmonter les défis de la centralisation des données. Cette thèse présente des approches innovantes dans les secteurs de l'inférence bayésienne à grande échelle et la quantification des incertitudes, avec pour but de fournir des solutions à la centralisation des données. Les approches de Monte Carlo fédéré permettent à plusieurs agents/nœuds d'effectuer des calculs localement et en toute sécurité, tandis qu'un serveur central combine les résultats obtenus pour échantillonner selon la posteriori globale. Ces techniques d'échantillonnage a posteriori bayésiennes bénéficient de l'incorporation des connaissances à travers la priori, ce qui conduit à des résultats améliorés. De plus, l'incertitude associée aux paramètres et aux prédictions est naturellement quantifiée, cette capacité étant d'autant plus nécessaire en présence d'un petit nombre de données ou de données bruitées.

La première partie de ce manuscrit se concentre sur les méthodes de Monte Carlo via les chaînes de Markov. En particulier, nous introduisons deux procédures, appelées `DG-LMC` et `FALD`, conçues pour cibler une distribution a posteriori tout en assurant la scalabilité. Chacune de ces méthodes reposent sur un serveur central pour orchestrer plusieurs entités locales. Celui-ci agrège l'information provenant de chaque agent afin de produire des solutions statistiques tout en limitant la quantité de données transférées. Cette approche réduit le nombre de communications entre participants, ce qui la rend particulièrement avantageuse dans les environnements fédérés avec une bande passante limitée. Étant donné la nature distribuée des ensembles de données d'aujourd'hui, des préoccupations concernant la confiance et la confidentialité se posent lors du transfert d'informations vers le serveur central. Les méthodes proposées non seulement abordent des applications pratiques, mais étendent également les algorithmes d'apprentissage existants aux problèmes d'inférence bayésienne. Les approches développées présentent des applications potentielles dans divers domaines, notamment l'épidémiologie et la finance, où l'inférence à grande échelle et la confidentialité des données sont des préoccupations majeures.

La deuxième partie de la thèse se concentre sur la gestion de l'incertitude. Initialement, nous présentons l'approche bayésienne, qui consiste à définir une a priori et une vraisemblance. Cette première méthode se base sur des opérateurs de compression afin de résoudre les problèmes de bande passante. Dans la dernière partie, nous introduisons une méthode fréquentiste basée sur les prédictions conformelles. Contrairement aux méthodes précédentes, cette approche fonctionne avec n'importe quel modèle prédictif. Nommée `DP-FedCP`, cette méthode utilise la technique de régression quantile pour générer des ensembles de prédictions personnalisés et robustes. En outre, elle aborde efficacement l'hétérogénéité entre agents via la détermination de quantiles basés sur des pondérations d'importance. Un aspect crucial de notre approche reste la préservation de la confidentialité, nous veillons à protéger les informations sensibles de chaque utilisateur.

Thesis outline and reading guide

Outline

This section provides an overview of the structure and content of the thesis, as well as the reading guide for a comprehensive understanding of the research. The thesis consists of an introductory part (Part I) followed by two main parts. Part II comprises two chapters that explore various distributed Monte Carlo sampling methods based on Markov Chain. Part III consists of two chapters that investigate the use of federated learning approaches for uncertainty quantification. Precisely, the thesis is organized into the following chapters:

Part I introduces the problem of federated learning and uncertainty quantification.

- Chapter 1 introduces the problem of federated learning and uncertainty quantification, outlining the main contributions of the thesis. This chapter presents the research questions, highlighting the significance of the study. It outlines the objectives, scope, and methodology of the research, providing a clear context for the subsequent chapters. The literature is reviewed critically, and we examine existing frameworks relevant to our topics. This chapter identifies gaps in the current knowledge and recalls the theoretical foundations upon which this study is built.

Part II presents the distributed Markov Chain Monte Carlo (MCMC) sampling methods. Both chapters proposed a distributed method based on local agents performing multiple local updates with a central server computing the consensus step.

- Chapter 2 investigates reliable large-scale Bayesian using distributed MCMC algorithms. The proposed methodology is designed to handle partitioned datasets stored within a master/slaves architecture. The scalability in high-dimensional settings through both synthetic and real data experiments is also demonstrated. This chapter is based on the conference paper Plassier et al. (2021).
- Chapter 3 develops one key direction of the thesis, addressing Bayesian inference in the context of federated learning. It introduces the Federated Averaging Langevin Dynamics (`FALD`) algorithm and proposes `VR-FALD`*, an enhanced version that utilizes control variates to correct client drift caused by statistical heterogeneity. Non-asymptotic bounds are established to showcase the effectiveness of `VR-FALD`* in mitigating the impact of statistical heterogeneity in federated learning benchmarks. This chapter is based on the conference paper Plassier et al. (2023b)

Part III highlights the proposed methodology for federated uncertainty quantification. Both chapters proposed a federated learning method that provides uncertainty quantification.

- Chapter 4 focuses on Bayesian inference in federated learning and introduces the Quantized Langevin Stochastic Dynamics algorithm, which addresses constraints such as privacy, communication overhead, and statistical heterogeneity. Variance reduction techniques are incorporated, leading to improved versions of

the algorithm. Both non-asymptotic and asymptotic convergence guarantees are provided, and the performance is demonstrated through various Bayesian Federated Learning benchmarks. This chapter is based on the conference paper Vono et al. (2022b).

- Chapter 5 addresses uncertainty quantification within the Federated Learning framework relying on conformal predictions. A novel federated conformal prediction method based on quantile regression is developed, taking into account privacy constraints and effectively handling the label shift between agents. The method provides theoretical guarantees for valid coverage of prediction sets while ensuring differential privacy, outperforming current competitors in extensive experimental studies. This chapter is based on the conference paper Plassier et al. (2023a).

Reading guide

Each chapter begins with a concise introduction, providing the necessary contextual information. For a quick overview of the contributions, readers are encouraged to focus on the summary of contributions in Chapter 1. While each chapter corresponds to an accepted conference article, some modifications have been made to improve readability and clarity. A chronological order has been established to facilitate the natural progression, but it is important to note that each chapter can be read independently.

Certain technical proofs have been omitted from the verbatim articles, and for a thorough grasp of all the details and proofs, please refer to the original articles. In any case, the main results are still presented, and most details, outcomes and primary proofs are included.

Contents

Notation

$:=$	Equal by definition
$\mathbb{N}, \mathbb{R}$	Sets of natural and real numbers
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\langle x, y \rangle$	Inner product of vectors $x, y \in \mathbb{R}^d$
$\|x\|_p$	ℓ_p-norm of vector $x \in \mathbb{R}^d$
$\|A\|$	Matrix norm induced $\|A\| = \sup\{\|Au\| : u \in \mathbb{R}^p, \|u\| = 1\}$
$\mathbb{R}^{n \times d}$	Set of real matrices of size $n \times d$
$\mathcal{S}_d(\mathbb{R})$	Set of real symmetric matrices of size $d \times d$
$\mathcal{S}_d^+(\mathbb{R}), \mathcal{S}_d^{++}(\mathbb{R})$	Set of real symmetric positive (semi)-definite matrices of size $d \times d$
I_d	Identity matrix of size $d \times d$
$A^\top$	Transpose of matrix A
$\mathrm{Tr}(A), det(A)$	Trace and Determinant of matrix A
$\lambda_{\min}(A), \lambda_{\max}(A)$	Smallest and Largest eigenvalue of matrix A
$A \otimes B$	Kronecker product of A and B
$vec(A)$	Vectorization of matrix A by stacking its columns
$\mathcal{B}(\mathcal{X})$	Borel σ-field on $\mathcal{X}$
$\mathbb{1}_E$	Characteristic function of set E
A^c	Complementary set of set A
$\mathbb{P}(\cdot)$	Probability of an event
$\mathbb{E}[\cdot]$	Expectation of a random variable
$\overset{\text{i.i.d.}}{\sim}$	Independent and Identically Distributed
$L_2(\pi)$	Set of square integrable functions with respect to measure π
$X \sim \pi$	Random variable X has distribution π
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance matrix Σ
∇f	Gradient function of $f : \mathbb{R}^d \to \mathbb{R}$
$\nabla^2 f$	Hessian matrix of $f : \mathbb{R}^d \to \mathbb{R}$

Part I

Introduction & Preliminaries

"Uncertainty is not a sign of weakness but a path to possibility."

Chapter 1

General Introduction, Motivations and Contributions

Contents

Machine learning and artificial intelligence (AI) have made great strides in the last two decades. These advances have been driven by the exponential growth of data and computational capabilities, benefitting from centralization to aggregate data in a single location with immense computational resources.

However, this fully centralized machine learning paradigm is increasingly at odds with real-world use cases due to both technological and societal reasons. On the technological side, centralized machine learning poses several challenges including (1) data processing bottlenecks, (2) inefficient utilization of communication resources, (3) coordination and synchronization issues that can lead to biased and incoherent models. At the societal level, transmitting data to centralized entities raises concerns about (1) privacy and exposure of individuals' private information, (2) ownership dilemmas, (3) centralization of power, and (4) objective disparities between individual agents at the network's edge and those of the centralized entity.

Recognizing these challenges, the machine learning community is now addressing the problems raised by networked agents. Depending on the context, an *agent* can either refer to an autonomous device equipped with local sensors and actuators, or an individual operating within a localized context supported by personal storage and computing facilities. Generally, this can be a company, a hospital, or a government agency. In any case, the technological trend is evident: as storage and computing capacity continues to increase at the agent level (referred to as the "edge" of the network), decentralizing computing tasks becomes increasingly appealing. It is crucial for the machine learning field to embrace this trend and adapt accordingly. Consequently, one of the significant challenges of our time is achieving learning in decentralized environments, accounting for distributed data sources, local computing resources, and heterogeneous goals.

1.1 Bayesian Inference in a nutshell

The Bayesian inference paradigm operates on the principle of treating parameters as random variables. Instead of "learning" parameters through the minimization of a loss function, the Bayesian approach infers a distribution, called the "posterior", over the parameters by applying the Bayes' rule. To obtain this posterior distribution, a "prior"

distribution that is independent of the data observations needs to be specified. While this may appear as an inconvenience, Bayesian inference treats all sources of uncertainty in the modeling process in a unified and consistent manner, requiring explicit assumptions and constraints. This in itself, is arguably an appealing feature of the paradigm. However, the most compelling aspect of the Bayesian approach is the automatic implementation of the "Occam's Razor". Within the Bayesian framework, there is a natural preference for simple models that sufficiently explain the data without unnecessary complexity.

The choice of the prior distribution $p(\theta)$ is the starting point of Bayesian learning. After observing $(z^{(1)}, \ldots, z^{(N)})$, this prior distribution is updated to a posterior distribution using Bayes' rule:

$$p(\theta \mid z^{(1)}, \ldots, z^{(N)}) = \frac{p(\theta)p(z^{(1)}, \ldots, z^{(N)} \mid \theta)}{p(z^{(1)}, \ldots, z^{(N)})} \propto p(\theta)\, \mathrm{L}(\theta; z^{(1)}, \ldots, z^{(N)})$$

The posterior distribution combines two components: (1) the likelihood, denoted as $\mathrm{L}(\theta; z^{(1)}, \ldots, z^{(N)})$, which encapsulates the information about the parameter θ derived from observations, and (2) the prior, which contains the information about θ derived from our background knowledge. Assuming independent observations, the likelihood can be expressed as follows:

$$\mathrm{L}(\theta; z^{(1)}, \ldots, z^{(N)}) = \prod_{i=1}^{N} p(z^{(i)}; \theta)$$

where $p(z^{(i)}; \theta)$ is the probability distribution function (pdf) of the observation for a given value of the parameter θ. To predict the value of a new observation z, a Bayesian approach integrates the predictive distribution over the different parameters with respect to the posterior distribution:

$$p(y \mid x, z^{(1)}, \ldots, z^{(N)}) = \int p(y \mid x, \theta)\, p(\theta \mid z^{(1)}, \ldots, z^{(N)})\, \mathrm{d}\theta. \tag{1.1}$$

The resulting predictive distribution, denoted as $p(y \mid x, z^{(1)}, \ldots, z^{(N)})$, is the outcome of Bayesian inference and serves various purposes based on user requirements. The ability to generate such a distribution is a fundamental advantage of the Bayesian approach. Computing the predictive distribution, as expressed in (1.1), lies at the core of Bayesian inference. Despite its apparent simplicity, it often poses significant computational challenges.

In the supervised learning setting, where $z^{(i)} = (x^{(i)}, y^{(i)})$, $y^{(i)}$ represents the response (or the dependent variable) while $x^{(i)}$ denotes the covariate (observation or features), and the likelihood function can be written as

$$\mathrm{L}\left(\theta; (x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\right) = \prod_{i=1}^{N} p(y^{(i)} \mid x^{(i)}, \theta).$$

1.1.1 Approximate Bayesian inference and MCMC

The Bayesian learning objective is to estimate label probabilities of new covariates. This involves finding predictive probabilities or making single-valued guesses. Both tasks require evaluating a function expressed as an expectation with respect to the

posterior distribution over the model's parameters. Writing the posterior probability density for the parameters as $p(\theta|\mathcal{D})$, with $\mathcal{D} = \{(x^{(k)}, y^{(k)})\}_{k=1}^{N}$, the expectation of $f(\theta)$ can be computed as follows:

$$\mathbb{E}[f(\theta) \mid \mathcal{D}] = \int f(\theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta.$$

Such expectations can be approximated by the Monte Carlo method, using samples drawn (approximately) from $p(\theta|\mathcal{D})$, the previous integral is approximated as:

$$\mathbb{E}[f(\theta) \mid \mathcal{D}] \approx \frac{1}{k} \sum_{j=0}^{k-1} f(\theta^{(j)})$$

While sampling directly from the posterior distribution $p(\cdot|\mathcal{D})$ is often computationally infeasible, it is still possible to generate an ergodic Markov chain with stationary distribution $p(\cdot \mid \mathcal{D})$.

Notably, while $p(\theta, \mathcal{D})$ has an explicit expression, the marginal distribution of the observations $p(\mathcal{D})$ is intractable. Let $K : (\Theta, \mathcal{B}(\Theta)) \to [0, 1]$ be a Markov kernel (Douc et al., 2018) and $\pi(\cdot|\mathcal{D})$ the posterior distribution (with probability density function $p(\theta|\mathcal{D})$). Assuming K admits $\pi(\cdot|\mathcal{D})$ as its unique invariant distribution, i.e., $\pi(\cdot|\mathcal{D})K = \pi(\cdot|\mathcal{D})$, where $\pi(A|\mathcal{D}) = \int K(\theta, A)\pi(\mathrm{d}\theta|\mathcal{D})$, for $A \in \mathcal{B}(\Theta)$. It is known that in such a case, K is ergodic; see e.g. Douc et al. (2018, Chapter 5). Consequently, sampling the new parameter θ_{k+1} at iteration $k \geq 0$ according to $K(\theta_k, \cdot)$ would generate samples targeting the posterior distribution $\pi(\cdot|\mathcal{D})$. In modern machine learning area, there is a demand for algorithms that perform well with high-dimensional parameters and scaled even for very large number of observations N. The computation of the likelihood poses a computational bottleneck. This challenge has led to significant research efforts over the past decade (see Welling and Teh (2011); Bardenet et al. (2017)). One possibility is the Euler-Maruyama approximation of the Langevin diffusion (Roberts and Tweedie, 1996) given by

$$\theta_{k+1} = \theta_k + \gamma \nabla \log p(\theta_k|\mathcal{D}) + \sqrt{2\gamma} Z_{k+1}, \tag{1.2}$$

where $(Z_k)_{k\geq 0}$ represents i.i.d. standard Gaussian noises, and $\gamma > 0$ denotes the time discretization step-size. The Langevin Monte Carlo technique defines a Markov chain with a transition kernel given by $K_\gamma(\theta, \mathsf{B})$ for $(\theta, \mathsf{B}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$, as follows:

$$K_\gamma(\theta, \mathsf{B}) = \frac{1}{(4\pi\gamma)^{d/2}} \int_{\mathsf{B}} \exp\left(-\frac{1}{4\gamma} \left\| \tilde{\theta} - \theta + \gamma \nabla \log p(\theta|\mathcal{D}) \right\|^2 \right) \mathrm{d}\tilde{\theta}.$$

At iteration $k \in \mathbb{N}$, the new parameter θ_{k+1} is sampled according to $K_\gamma(\theta_k, \cdot)$, which is equivalent to updating θ_k following (1.2). Under certain conditions on the step-size γ and the potential $U(\theta) = \log p(\theta|\mathcal{D})$ (refer to Dalalyan (2017b); Durmus and Moulines (2017) for further details), the distribution of $(\theta_k)_{k\in\mathbb{N}}$ converges to the stationary distribution $\pi\gamma$ (dependent on the step-size γ) as k goes to infinity. Additionally, note that π_γ approaches the target distribution π as the step-size γ tends to zero. Non-asymptotic bounds have been derived (in terms of total variation distance or Wasserstein distance) to analyze the impact of different parameters, such as the step-size γ, the number of samples N, the dimension of the parameter space d, and properties of the potential (e.g., log-concavity in the tails).

However, these methods require calculating the gradient of the log posterior at each iteration, which is computationally intensive—$\mathrm{O}(N)$ operations. When the dataset size is large, estimating the gradient over the entire dataset can be prohibitively expensive. To mitigate this computational cost, an unbiased estimate of the gradient can be computed using a subset $S_{k+1} \subset [N]$ of observations, known as a minibatch. This class of methods, which employ minibatches, is called Stochastic Gradient MCMC (SGMCMC). The cost per iteration for SGMCMC algorithms is $\mathrm{O}(b)$, where b is the minibatch size. In the SGLD method introduced by Welling and Teh (2011), the full gradient is replaced, for any $k \geq 0$, by

$$\widehat{\nabla \mathcal{L}(\theta_k, \mathcal{D})} := -\nabla \log p\left(\theta_k\right) - \frac{N}{b} \sum_{l \in S_{k+1}} \nabla \log p(y_l | x_l, \theta_k).$$

The convergence of this algorithm has been studied in terms of total variation and Wasserstein convergence bounds (see Dalalyan (2017b); Durmus and Moulines (2019); Durmus et al. (2019); Dalalyan and Karagulyan (2019)). Under certain assumptions of strong convexity and Lipschitz continuity on the potential and its Hessian, the number of iterations required to obtain a distribution that is ϵ-close to the target (in terms of total variation or Wasserstein distance) is shown to be $\mathrm{O}(d/\epsilon)$.

Research Question #1

How can we draw inspiration from optimization methods to design sampling algorithms? And how can we introduce control variates to improve accuracy?

While stochastic gradients are unbiased, they introduce additional noise to the Langevin scheme, which can negatively impact convergence speed. To address this issue, control variates are employed to reduce the variance of the SGMCMC gradient estimate. The Wasserstein bounds provided in Chatterji et al. (2018) illustrate a $\mathrm{O}(\sqrt{\gamma})$ improvement in asymptotic bias achieved. Furthermore, as demonstrated in Nagapetyan et al. (2017); Baker et al. (2019), the standard SGLD requires a minibatch size of $b = \mathrm{O}(N)$, whereas control variates only require a minibatch size of order $\mathrm{O}(1)$ to achieve similar performance. Numerous control variate-based algorithms have been proposed. One such algorithm is the fixed-point method by Brosse et al. (2018), which relies on control variates utilizing the minimum $\theta_\star = \arg\min \mathcal{L}(\cdot, \mathcal{D})$ of the loss function. At iteration $k \in \mathbb{N}$, the parameters θ_k are updated using the following estimator:

$$\widetilde{\nabla \mathcal{L}(\theta_k, \mathcal{D})} = \nabla \log p\left(\theta_k\right) - \nabla \log p(\theta^\star) + \frac{N}{b} \sum_{l \in S_{k+1}} \left\{ \nabla \log p(, y_l | x_l, \theta_k) - \nabla \log p(, y_l | x_l, \theta_\star) \right\}.$$

This new gradient estimate leads to improvements in the strongly convex case (Dubey et al., 2016) and Brosse et al. (2018) derive an upper bound in Wasserstein distance of order 2 between the distribution of the iterates $(\theta_k)_{k \in \mathbb{N}^*}$ and the Langevin diffusion. However, these control variates require the determination of the minimum $\theta_\star$ of the loss function, which is challenging to obtain in practice. Thus, Chatterji et al. (2018) propose an SVRG-Langevin variance reduction scheme based on the SVRG method (Johnson and Zhang, 2013). This method involves updating a reference point $\tilde{\theta}_k$ with a probability $q \in \left(0, 1\right]$. The gradient is then estimated using $\tilde{\theta}_k$, and the resulting

stochastic gradient is given by:

$$\nabla\overline{\mathcal{L}(\theta_k, \mathcal{D})} = \nabla\widehat{\mathcal{L}(\theta_k, \mathcal{D})} - \nabla\widehat{\mathcal{L}(\tilde{\theta}_k, \mathcal{D})} + \nabla\mathcal{L}(\tilde{\theta}_k, \mathcal{D}).$$

To reduce the variance of the stochastic gradient, we need to ensure that $\mathrm{Var}^{\mathcal{F}_k}(\nabla\overline{\mathcal{L}(\theta_k, \mathcal{D})}) < \mathrm{Var}^{\mathcal{F}_k}(\nabla\widehat{\mathcal{L}(\theta_k, \mathcal{D})})$, where $\mathrm{Var}^{\mathcal{F}_k}$ represents the variance conditioned on the random variables used up to the k-th iteration.

1.2 Distributed/Federated Learning

Training very high-dimensional models via loss minimization in a distributed/federated manner involves significant communication costs, which can become a major bottleneck and slow down training. Reducing communication costs has been identified as one of the major challenges of FL (Kairouz et al., 2021).

Two promising approaches have been proposed to address this challenge. The first approach is to have agents perform multiple optimization iterations locally before sending a model update to the central node. The second approach involves compressing the exchanged messages. While local updates have been used with some success in practice, they raise practical issues. Due to statistical heterogeneity, performing multiple steps can hinder convergence, as model updates target each agent's local minimizer (Li et al., 2019; Ro et al., 2021). This results in a tradeoff between communication cost and convergence (Wang et al., 2020b; Woodworth et al., 2020), and necessitates new algorithms to limit "client drift" (Karimireddy et al., 2020; Li et al., 2020b) (e.g., `SCAFFOLD`, `FED-PROX`).

Despite recent progress, developing new algorithms with theoretical guarantees in these domains remains a major challenge. These new algorithms significantly improve convergence when the target function is strongly convex and many local updates are performed. However, in the non-convex case (e.g., deep learning, latent-variable models), theoretical guarantees are essentially missing. Existing approaches often rely on Euclidean averaging of the model weights, which becomes inefficient when agents' models deviate significantly from the central model (Frankle et al., 2020). Initial attempts to improve the aggregation technique have been proposed, with two main approaches investigated: modifying the averaging scheme using optimal transport (Singh and Jaggi, 2020) or weight matching (Yurochkin et al., 2018; Wang et al., 2020a), which require significant computational effort, or using distillation (Lin et al., 2020; Sattler et al., 2020), but requires an additional public dataset, increased training overhead, but without theoretical guarantees.

Research Question #2

Can we leverage gradient compression schemes to sample from the posterior distribution?

Another approach to reduce communication costs is to decrease the number of bits in each message exchanged between agents and the central node. This is achieved through randomized lossy compression, often a mixture of sparsification and quantization. Biased compressors (e.g., `Top-`k) typically achieve higher compression ratios than

unbiased ones but can lead the algorithm to converge to spurious minima if directly applied (Karimireddy et al., 2019). The effect of bias can be mitigated by using error feedback methods, as advocated in Stich and Karimireddy (2019); Konečnỳ et al. (2016); Gorbunov et al. (2021); Hanzely and Richtárik (2020); Wang et al. (2021); Horváth and Richtarik (2020).

Random independent unbiased compressors perform better with increasing numbers of agents and are less sensitive to statistical heterogeneity. Several techniques have been proposed to develop such compression operators. However, for high-dimensional complex models, there is still a need for novel compression techniques. Specifically, the interactions between the distribution and structure of the compressed messages is still not clear; and the compression operator needs to be reconsidered for efficient high-dimensional compression methods. Additionally, the abundance of new compression techniques calls for more rigorous evaluation frameworks. These measures can be based on interpretable metrics, or small-scale. However, quantifying the impact of elementary algorithmic blocks on the overall performance of deep learning models is challenging.

1.2.1 Distributed/Federated Bayesian Learning

Distributed Monte Carlo methods aims to generate samples from the posterior distribution $p(\theta|\mathcal{D})$, but without exchanging observations between the workers and the central node. Each node has only access to its local dataset $\mathcal{D}_i$, and communicates with the central server to generate samples targeting the global posterior given by

$$\forall \theta \in \mathbb{R}^d, \quad \theta \sim p(\theta|\mathcal{D}) \propto p(\theta) \prod_{l=1}^{N} p(y_l|x_l,\theta) = \prod_{i=1}^{n} \left[p(\theta)^{1/n} p(\mathcal{D}_i|\theta) \right] = \prod_{i=1}^{n} p_i(\theta|\mathcal{D}_i).$$

The interest in distributed Bayesian inference has significantly grown over the past decade. In an early paper, Zinkevich et al. (2010) proposed running independent chains on each subset of the data, while periodically averaging the learned parameters. However, no clear theoretical convergence guarantee could be provided. Subsequently, sophisticated methods have been proposed to recombine local samples to approximate the desired global posterior (Neiswanger et al., 2014; Wang and Dunson, 2013; Minsker et al., 2014). Due to statistical heterogeneity, data imbalance, and noise, the local posteriors can differ significantly from each other. Better agents with more data might possess more accurate information on the parameter. Several alternative techniques have been proposed, they utilize the values from workers' chains to approximate the posterior distribution, each with its own benefits and drawbacks. Alternative techniques utilize the values from each chain and approximate posterior expectations, each with its own benefits and drawbacks. For instance, Neiswanger et al. (2014) propose Gaussian kernel density estimation (KDE), Wang and Dunson (2013) suggest Gaussian aggregation techniques based on local samples drawn from the Weierstrass transformation of the subposteriors (convolution of the subposteriors with Gaussian kernels), Minsker et al. (2014) develop a median posterior in a reproducing kernel Hilbert space (RKHS), and recombination of the samples using random partition trees (Wang et al., 2015).

In their seminal work, Scott et al. (2016) propose an exact algorithm for Gaussian subposteriors. They leverage the Bernstein-von Mises theorem (Van der Vaart, 2000) which states that under some conditions, if a unique parameter $\theta_\star = \arg\max \mathrm{L}(\cdot|\mathcal{D})$ exists, then the posterior tends to a normal distribution centered around $\theta_\star$ as the

number of observations increases. When $p_i(\cdot|\mathcal{D}_i)$ is the density of a Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$, the posterior distribution $p(\cdot|\mathcal{D}) \propto \prod_{i=1}^n p_i(\cdot|\mathcal{D}_i)$ is also Gaussian. It has a covariance $\Sigma = (\sum_{i=1} \Sigma_i^{-1})^{-1}$ and a mean $\mu = \Sigma \sum_{i=1}^n \Sigma_i^{-1}\mu_i$. Furthermore, if we draw n independent random variables $\theta_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, the weighted combination satisfies

$$\Sigma\left(\sum_{i=1}^{n} \Sigma_i^{-1}\theta_i\right) \sim \mathcal{N}\left(\mu, \Sigma\right). \tag{1.3}$$

Thus, combining these local draws $(\theta_i)_{i=1}^n$ according to (1.3) produces a sample distributed according to the target posterior distribution. However, it should be noted that this method lacks theoretical guarantees and performs well only for Gaussian or nearly Gaussian target distributions.

An alternative approach is to approximate the true posterior density using an estimate of the kernel density of the subposterior densities; see for example White et al. (2015). Neiswanger et al. (2014) propose an algorithm to sample according to $\widehat{p}_1 \times \cdots \times \widehat{p}_M$ instead of $\prod_{i\in[n]} p_i$. Each worker i independently samples parameters $\theta_i^1, \ldots, \theta_i^T$ from the subposterior $p_i(\cdot|\mathcal{D}_i)$, and the resulting samples are combined to derive the proxy $\widehat{p}_i$ given by

$$\widehat{p}_i(\theta) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{N}\left(\theta \mid \theta_t^i, h\mathsf{I}_d\right).$$

Here, $\mathcal{N}(\cdot|\cdot, h\mathsf{I}_d) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ denotes the Gaussian kernel with bandwidth parameter $h > 0$. Since the approximate subposterior $\widehat{p}_i$ is a Gaussian mixture, the product $\prod_{i\in[n]} \widehat{p}_i$ is also a Gaussian mixture contrary to $\prod_{i\in[n]} p_i$. Therefore, the second part of the algorithm involves sampling $\theta \in \mathbb{R}^d$ according to this Gaussian mixture. This method replaces the posterior sampling, from which it is difficult to sample, by mixture Gaussian sampling. However, this method has several drawbacks. Firstly, the parametric estimator can be asymptotically biased. Secondly, the number of samples required to achieve the same level of accuracy, exponentially depends on the dimension d due to the curse of dimensionality of the kernel density estimators. Moreover, for a multimodal posterior, the effect of averaging is not clear and can lead to mode collapse.

1.2.2 Bayesian inference methods using local steps on each client

Research Question #3

How to design efficient distributed sampling algorithms for high-dimensional models?

To address the question of designing efficient distributed sampling algorithms for high-dimensional models, Vono et al. (2022a); Rendell et al. (2020) have introduced a hierarchical Bayesian model that enables separate MCMC chains on each agent, client, or worker. As in the "embarrassingly parallel" approaches, the Global consensus Monte Carlo objective is to reduce the costs of communication latency. They employ a parameter relaxation method, which bears resemblance to the splitting technique used in optimization, such as the alternating direction method of multipliers (ADMM) (Boyd et al., 2011). In their approach, an auxiliary parameter z_i is associated with each agent/client/worker, which are assumed to be conditionally independent given the master's parameter θ. The algorithm targets an extended distribution $\pi_\rho(\theta)$ that depends

on a tolerance parameter $\rho > 0$. This distribution is defined as:

$$\pi_\rho(\theta) \propto p(\theta) \prod_{i=1}^{n} \left[K_i^{(\rho)}(\theta, z_i) p(z_i|\mathcal{D}_i) \right].$$

For $i \in [n]$, if $\lim_{\rho \to 0} K_i^{(\rho)}(\theta, z_i) = \mathbb{1}_\theta(z_i)$, then Scheffé's lemma (Scheffé, 1947) demonstrates that $(\pi_\rho)_{\rho>0}$ converges to π in total variation as $\rho \to 0$. The authors develop a Metropolis-within-Gibbs scheme that alternates between sampling the agent/client/worker parameters given the master's parameter, and sampling the master's parameter given the agent/client/worker parameters.

There exists a tradeoff between the bias, which requires $\rho \ll 1$, and the mixing time, which typically improves as the tolerance parameter increases. To obtain samples θ drawn according to π_ρ, the authors propose sampling the marginals separately. Each node independently samples z_i in parallel according to $\pi_\rho(z_i|\theta)$, while the central server samples $\theta \sim \pi_\rho(\theta|z_1, \ldots, z_n)$. In this procedure, communication is only necessary during aggregation steps, where an approximation of the full posterior is obtained using samples from the n chains. When exact sampling is not feasible in practice, Rendell et al. (2020) consider a Metropolis-Hastings scheme to sample from $\pi_\rho(z_i|\theta)$, while Vono et al. (2022a) suggest the use of a rejection sampling mechanism. Both approaches provide theoretical guarantees for their proposed schemes and prove that they admit π_ρ as a stationary distribution under mild assumptions.

Research Question #4

Can we provide theoretical guarantees for distributed sampling algorithms?

1.3 Federated Uncertainty Quantification

In the machine learning field, uncertainty quantification plays a critical role in decision-making processes. Traditional prediction models often provide point estimates without explicitly addressing the associated uncertainty, leaving decision-makers with limited insight into the reliability of the predictions. However, in many real-world applications, having an understanding of the uncertainty is essential for making informed decisions.

1.3.1 Bayesian uncertainty quantification and calibration

Over-parameterized deep models have shown the ability to memorize datasets even when the labels are completely randomized (Zhang et al., 2021). However, in many applications, especially those involving decision-making processes, overconfident predictions can be problematic (Amodei et al., 2016; Del Grosso et al., 2022). Therefore, uncertainty quantification is necessary to make reliable decisions (autonomous cars, health-related systems).

The importance of well-calibrated decisions is often emphasized as a means to mitigate the impact of rare but significant errors caused by poorly calibrated models; see Guo et al. (2017) for detailed calibration measures and Rahaman and Thiery (2021) for methods leading to better calibration. Deep learning methods are known to suffer from calibration issues, often producing overconfident estimates. These problems

become more pronounced in scenarios with limited data availability. While the calibration of probabilistic models has been extensively studied, calibrating extremely over-parameterized models in low-data regimes poses unique challenges. Frequentist learning proves effective in large data sets when the primary focus is accuracy but falls short in quantifying epistemic uncertainty due to limited data availability (Lakshminarayanan et al., 2017). Bayesian learning offers an alternative framework in which optimization is conducted on the distribution of model parameters, rather than a single vector of parameters as in frequentist learning.

Many Bayesian methods involve sampling weights to target specific distributions. However, assessing the quality of predictive uncertainty obtained through these methods presents a significant challenge. Metrics commonly used in the optimization community, such as accuracy or loss evaluation, are not well-suited to reflect how effectively an algorithm samples according to the posterior distribution. To evaluate the quality of predictive distributions obtained through classical methods, Wilson et al. (2021) compare various predictive distributions with a Hamiltonian ground truth distribution, which is known to be asymptotically accurate but computationally expensive for large neural network models.

1.3.2 Conformal predictions for uncertainty quantification

Conformal prediction experienced relatively recent developments in machine learning and statistics. These methods offer an attractive framework for uncertainty quantification. Unlike traditional approaches, conformal methods provide theoretical guarantees without any assumption except for exchangeability (Lei et al., 2013; Fontana et al., 2023). Conformal prediction leverages the concept of nonconformity scores to construct prediction intervals that capture the uncertainty associated to each prediction (Vovk et al., 1999; Shafer and Vovk, 2008; Balasubramanian et al., 2014). One of the key advantages is its model-agnostic nature. This flexibility allows practitioners to utilize conformal prediction as a powerful tool for uncertainty quantification in a wide range of applications. It can be applied to any underlying prediction model, including regression, classification, and more advanced machine learning algorithms. We will only detail here the main derivations of conformal split-prediction methods. As in the classical optimization framework, the training dataset $\mathcal{D}^{\text{train}}$ is used to learn the predictor $\widehat{f}$ while the calibration dataset $\mathcal{D}^{\text{cal}}$ is reserved for the confidence interval constructions.

At its core, conformal prediction introduces a notion of valid prediction sets by analyzing the distribution of non-conformity measures. These measures quantify the deviation between a new non-conformity score and the available non-conformity distribution estimate on the training data. Larger scores meaning worse agreement between x and y. These scores are often based on the predictor $\widehat{f}$ and can be therefore considered as post-processing; for specific choices of non-conformity functions we refer to Angelopoulos and Bates (2021) and references therein. As by hypothesis the data are i.i.d., the non-conformity scores must have the same distribution – which we note $P(V)$ – meaning that:

$$\forall (x, y) \in \mathcal{D}^{\text{cal}}, \qquad V(x, y) \overset{\text{i.i.d.}}{\sim} P(V).$$

Given a confidence threshold $\alpha \in (0, 1)$, the conformal prediction constructs a prediction set $\mathcal{C}_\alpha(x)$ for a new instance x by aggregating the non-conformity scores of the calibration data. From these non-conformity scores, a quantile is calculated. Denote

by N the number of calibration data and $q = \lceil (1-\alpha)(N+1) \rceil / (N+1)$, the $(1-\alpha)$-quantile corresponds to the qth largest value of $\{V(x,y)\}_{(x,y)\in\mathcal{D}^{\text{cal}}}$. The prediction set is ensured to contain the true label Y_{N+1} at the predefined confidence level $1-\alpha \in [0,1]$. Specifically, the prediction set is determined $\forall x \in \mathcal{X}$, by

$$\mathcal{C}_\alpha(x) = \left\{ y \in \mathcal{Y} \colon V(x,y) \leq Q_{1-\alpha} \left(\sum_{k=1}^{N} \frac{\delta_{V(X_k,Y_k)}}{N+1} + \frac{\delta_{V(X_{N+1},Y_{N+1})}}{N+1} \right) \right\}.$$

Under the exchangeability assumption on $\{(X_k, Y_k)\}_{k\in[N+1]}$, it is known (Papadopoulos et al., 2002; Tibshirani et al., 2019) that:

$$1-\alpha \leq \mathbb{P}\left(Y_{N+1} \in \mathcal{C}_\alpha(X_{N+1}) \mid \mathcal{D}^{(\text{train})} \right) \leq 1-\alpha + \frac{1}{N+1}.$$

Thus, the increase of the number of data allows the refinement of the prediction set. Indeed, the upper bound shows that this set becomes more and more informative when N increases. Note that $\mathcal{C}_\alpha(X_{N+1})$ cannot include too many possible outputs otherwise the upper bound would be relatively close to 1.

Research Question #5

How to adapt and customize these prediction sets to the federated case, and how can we keep theoretical guarantees despite shifts between local distributions?

Many research gaps remain in the federated conformal prediction framework, notably (1) regarding results on the quantiles federated computation, (2) ensuring valid coverage guarantees while (3) preserving privacy. However, only a few solutions have been proposed to address these challenges.

Research Question #6

How to efficiently calculate quantiles in a federated environment?

A natural approach to performing federated conformal prediction is to aggregate the quantiles of different agents. This is studied in Lu and Kalpathy-Cramer (2021); the authors suggest deriving prediction sets by averaging the local quantiles. However, this approach is not robust when dealing with heterogeneous data since global quantiles may not be suitable when at least one agent has limited data. For instance, if the threshold $\alpha \in (0,1)$ is taken such that $\alpha < (N^i+1)^{-1}$, then, the quantile for agent i becomes $Q_{1-\alpha}\{(N^i+1)^{-1}(\sum_{k=1}^{N^i} \delta_{V(X_k^i,Y_k^i)} + \delta_\infty)\} = \infty$. This demonstrates the lack of robustness of the quantile averaging approach, which results in problematic aggregations.

Research Question #7

How to generate prediction sets while preserving data confidentiality?

A more robust approach is developed by Humbert et al. (2023). The authors investigate the validity of a quantile of quantiles approach instead of using the average quantile. The theoretical study demonstrates its effectiveness for homogeneous datasets, however the study lacks mechanisms to handle data heterogeneity.

1.4 Summary of the contributions

Motivated by the research questions (RQ) previously mentioned, this thesis makes several contributions, which are outlined in detail in the following section. Each chapter focuses on a specific research direction, addressing the following key areas:

- ✦ Development of advanced distributed sampling methods targeting a global posterior distribution.
- ✦ Construction of efficient simulation methods for potentially high dimensional distributions, known up to some normalizing constant.
- ✦ Application of approximate inference methods for Bayesian deep learning.
- ✦ Derivation of federated uncertainty management methods based on conformal predictions.

Part II: Distributed Sampling & Langevin MC

• Chapter 2: *DG-LMC: A turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs* (RQ#3-RQ#4)

In this work, we propose an efficient sampling algorithm tailored for master/slave architectures. Our method specifically focuses on Bayesian inference from shared datasets $\{\mathcal{D}_i\}_{i=1}^n$ observed on n workers. We develop a procedure for approximating posterior distributions admitting a density given by

$$\pi(\theta|\mathcal{D}_{1:n}) \propto \prod_{i=1}^{n} \exp(-U_i(\theta)), \tag{1.4}$$

where the potential function $U_i \colon \mathbb{R}^{d_i} \to \mathbb{R}$ depends on the training set $\mathcal{D}_i$. The key idea of our novel methodology, called Distributed Gibbs using Langevin Monte Carlo (DG-LMC), consists in designing a joint distribution $\Pi_{\boldsymbol{\rho}}$ with auxiliary variables $z_1 \in \mathbb{R}^{d_1}, \ldots, z_n \in \mathbb{R}^{d_n}$ satisfying

$$\Pi_\rho(\mathcal{D}_{1:n}|z_{1:n},\theta) \propto \prod_{i=1}^{n} \Pi_\rho(\mathcal{D}_i|z_i), \qquad \Pi_\rho(z_{1:n}|\theta) = \prod_{i=1}^{n} \Pi_\rho(z_i|\theta), \tag{1.5}$$

where $\rho > 0$ is a tolerance parameter such that $\lim_{\rho\to 0} \Pi_{\boldsymbol{\rho}}(\theta|\mathcal{D}) = \pi(\theta|\mathcal{D})$. Working with Π_ρ has a significant advantage: the auxiliary variables $\{z_i\}_{i=1}^n$ are conditionally independent given θ. Consequently, utilizing (1.5) enables the following decomposition:

$$\begin{aligned}
\Pi_\rho(\theta|\mathcal{D}_{1:n}) &= \int \Pi_\rho(\theta, z_{1:n}|\mathcal{D}_{1:n})\,\mathrm{d}z_{1:n} \\
&= \frac{1}{\Pi_\rho(\mathcal{D}_{1:n})} \int \Pi_\rho(\theta, z_{1:n})\Pi_\rho(\mathcal{D}_{1:n}|\theta, z_{1:n})\,\mathrm{d}z_{1:n} \\
&= \frac{1}{\Pi_\rho(\mathcal{D}_{1:n})} \int \Pi_\rho(\theta) \prod_{i=1}^{n} \Big[\Pi_\rho(\mathcal{D}_i|z_i)\Pi_\rho(z_i|\theta)\Big]\,\mathrm{d}z_{1:n}.
\end{aligned}$$

By leveraging the Gibbs sampler, the distribution $\Pi_\rho(\theta, z_{1:n}|\mathcal{D}_{1:n})$ can be efficiently sampled in parallel without the need to transmit any data.

Contributions. The main contributions can be summarized as follows:

(1) We introduce a novel methodology called Distributed Gibbs using Langevin Monte Carlo (`DG-LMC`) in Section 2.2. This algorithm requires each worker to sample z_i from the conditional distribution $\Pi_\rho(z_i|\mathcal{D}_i, \theta)$ and to communicate this sample to the master node. Then, the central node sample θ according to $\Pi_\rho(\theta|z_{1:n})$ and sends back this parameter to every worker.

(2) Importantly, we present a comprehensive quantitative analysis of the induced bias and demonstrate explicit convergence results in Section 2.3. This represents our main contribution, and to the best of the authors' knowledge, this theoretical study is one of the most comprehensive among existing works that focus on distributed Bayesian machine learning with a master/slaves architecture. Specifically, we discuss the algorithm's complexity, the selection of hyperparameters, and offer practitioners simple guidelines for tuning them. Additionally, we conduct a thorough comparison of our method with existing approaches in Section 2.4.

(3) Finally, in Section 2.5, we demonstrate the advantages of the proposed sampler over popular and recent distributed MCMC algorithms through various numerical experiments.

Two main challenges remain: efficiently sampling from the conditional distribution $\Pi_{\boldsymbol{\rho}}(z_i|\theta, \mathcal{D}_i)$ for $i \in [n]$, and reducing frequent communication rounds with the master node. We address both issues using the Langevin Monte Carlo (LMC) algorithm to approximate sampling from $\Pi_{\boldsymbol{\rho}}(z_i|\theta, \mathcal{D}_i)$ (Rossky et al., 1978; Roberts and Tweedie, 1996). For $i \in [n]$, we introduce Π_ρ whose conditional densities given as follows:

$$\Pi_{\boldsymbol{\rho}}(z_i|\mathcal{D}_i, \theta) \propto \exp\left(-U_i(z_i) - \left\|z_i - \theta\right\|^2/(2\rho_i)\right),$$
$$\Pi_{\boldsymbol{\rho}}(\theta|z_{1:n}) = \mathcal{N}\left(\boldsymbol{\mu}(z_{1:n}), \mathbf{Q}^{-1}\right)$$

where the precision matrix $\mathbf{Q} = (\sum_{i=1}^n {\rho_i}^{-1})\mathbf{I_d}$ and the mean $\boldsymbol{\mu}(z_{1:n}) = \mathbf{Q}^{-1}\sum_{i=1}^n z_i/\rho_i$. When the tolerance parameter $\rho \to 0$, using (Scheffé, 1947) shows that this data augmentation scheme satisfies

$$\lim_{\rho\to 0} \Pi_{\boldsymbol{\rho}}(\theta|\mathcal{D}) = \lim_{\rho\to 0} \int \Pi_{\boldsymbol{\rho}}(\theta, z_{1:n})\mathrm{d}z_{1:n} = \pi(\theta|\mathcal{D}).$$

Based on the overdamped Langevin stochastic differential equation, at iteration k, we update the parameters as follows:

$$z_i^{(k+1)} = \left(1 - \frac{\gamma_i}{\rho_i}\right) z_i^{(k)} + \frac{\gamma_i}{\rho_i}\theta^{(k)} - \gamma_i \nabla U_i(z_i^{(k)}) + \sqrt{2\gamma_i}\xi_i^{(k)},$$
$$\theta^{(k+1)} = \boldsymbol{\mu}(z_{1:n}^{(k)}) + \mathbf{Q}^{-1/2}\xi_0^{(k)} \text{ during communication rounds else } \theta^{(k)},$$

where $\gamma_i > 0$ is a fixed step-size and $\{\xi_i^{(k)} : i \in [n], k \in \mathbb{N}\}$ is an i.i.d. sequence of standard Gaussian random variables. To mitigate communication costs, we allow each worker to perform $N_i \geq 1$ local LMC steps (Dieuleveut and Patel, 2019). Varying N_i across workers prevents `DG-LMC` from experiencing significant delays due to imbalanced worker response times (Ahn et al., 2014). We provide a detailed quantitative analysis of the bias and establish explicit non-asymptotic convergence results. Our analysis encompasses the complexity of `DG-LMC`, the selection of hyperparameters, and offers practitioners simple guidelines for tuning them. To the best of our knowledge, this theoretical study is one of the most comprehensive works on distributed Bayesian machine learning with a master/slave architecture.

Theorem 1.1 (Informal)**.** *Under some assumptions described in Chapter 2, there exist* $\kappa \in (0,1)$*,* $\gamma, \rho, C_0, C_1, C_2 > 0$ *such that for* $k \geq 0$*, the distribution* μ_k *of the sample* θ_k *satisfies*

$$W_2\Big(\mu_k, \pi(\cdot|\mathcal{D}_{1:n})\Big) \leq C_0(1-\kappa)^k + C_1\sqrt{d\gamma(\rho^2 + \gamma/\rho^2)} + C_2 d\rho.$$

- Chapter 3: *FALD: Federated Averaging Langevin Dynamics* (RQ#1)

In this chapter, we are interested in sampling from a target distribution π whose density can be decomposed as in (1.4). To address these issue, we propose an MCMC algorithm coined `FALD`, which combines the ideas of Stochastic Langevin Gradient Dynamics (SGLD) and Federated Averaging.

Contributions. The main contributions can be summarized as follows:

(1) We study a random loop version of the `FALD` algorithm proposed in Deng et al. (2021), and we establish non-asymptotic upper bounds in Wasserstein distance for strongly convex potentials U. An analysis of `FALD` was conducted in Deng et al. (2021, Theorem 5.7), however, the proof is plagued by an error; see Section 3.B.1.

(2) We give matching lower bounds to show that even with full batch gradients, `FALD` can be slower than `SGLD` due to client-drift.

(3) We propose a new method (`VR-FALD`*) that circumvents the shortcomings of `FALD`. This algorithm extends the Shifted Local-SVRG method of Gorbunov et al. (2021) to the Bayesian context. `VR-FALD`* combines the Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD) (Dubey et al., 2016) and adapts the bias reduction techniques from `SCAFFOLD` (Karimireddy et al., 2020).

(4) We derive theoretical guarantees for `VR-FALD`* which highlight its gradient variance reduction effect and ability to deal with data heterogeneity.

(5) The results are based on a general framework developed in the supplement, that encompasses a broad family of federated Bayes algorithms based on Langevin dynamics. This is the first unifying study among existing works on federated Bayesian inference.

(6) Finally, Section 3.4 illustrates our findings on classical FL benchmarks and provides a thorough comparison with existing FL Bayesian methods.

`FALD` algorithm samples from π while respecting a major constraint: each potential U_i and its gradient ∇U_i can only be computed by the i-th client. In this method, each client has a parameter θ_k^i which is updated locally while the global parameters θ_k^s is updated on the central server. At every round, the clients execute `SGLD` steps to update their local parameters

$$\tilde{\theta}_{k+1}^i = \theta_k^i - \gamma \nabla U^i(\theta_k^i) + \sqrt{2\gamma} Z_{k+1}^i,$$

where Z_{k+1}^i is a d-dimension Gaussian possibly correlated between clients. Each client sends $\tilde{\theta}_{k+1}^i$ to the central server with probability $p_c \in \big(0, 1\big]$ corresponding to the realization of a Bernoulli B_{k+1}. During communication rounds, the central server averages

the received parameters

$$\theta_{k+1}^s = (B_{k+1}/n) \sum_{i\in[n]} \tilde{\theta}_{k+1}^i + (1 - B_{k+1})\theta_k^s.$$

Then, this server parameter θ_{k+1}^s is returned to the local clients which update their local parameters θ_k^i following

$$\theta_{k+1}^i = B_{k+1}\theta_{k+1}^s + (1 - B_{k+1})\tilde{\theta}_{k+1}^i.$$

As stated in Theorem 1.2, the samples $\{\theta_k^s\}_{k\in\mathbb{N}}$ generated by the central server target the posterior distribution π. Further explanations on the convergence bounds of `FALD` are provided in Chapter 3. Although being theoretically sound, this method may suffer from high variance due to the stochastic gradients used during the local `SGLD` and the heterogeneity of the data, which hinders the convergence. More specifically, we show the impossibility for an algorithm not tackling heterogeneity to provide an asymptotic Wasserstein error below the discretization step-size $\mathrm{O}(\gamma)$. To solve this problem, we propose one alternative: `VR-FALD`* based on a combination of control variates and bias reduction techniques. Theoretical improvements are derived and experimental behaviors of our algorithms are provided.

Theorem 1.2 (Informal)**.** *Under assumptions described in Chapter 4, there exist* $\gamma_\star > 0$*, such that for* $\gamma \in (0, \gamma_\star)$*, there are* $\kappa \in (0, 1)$*,* $C_0, C_1, C_2, C_3 > 0$ *such that for* $k \geq 0$*, the distribution* μ_k *of the sample* θ_k *satisfies*

$$W_2^2\Big(\mu_k, \pi(\cdot|\mathcal{D}_{1:n})\Big) \leq (1-\kappa)^k C_0 + \gamma C_1 \mathbb{E}\Big(\textstyle\sum_{i=1}^n \widehat{U}_i(\theta_\star)\Big) + \frac{\gamma^2 C_2}{p_{\mathrm{c}}^2} \sum_{i=1}^n \|\nabla U_i\left(\theta_\star\right)\|^2 + \gamma^2 C_3.$$

Part III: Federated Uncertainty Quantification via Bayesian & Frequentist approaches

• Chapter 4: *`QLSD`: Quantized Langevin stochastic dynamics for Bayesian federated learning* (RQ#2)

Several works attempted to improve the efficiency of distributed/federated learning by reducing the communication cost. Some methods focused on quantizing each coordinate of the computed gradients (Alistarh et al., 2017), so that much fewer bits are needed to be transmitted. Aggressive quantization, such as the binary or ternary representation, has also been investigated. Other methods imposed sparsity onto gradients during communication, where only a small fraction of gradients gets exchanged across nodes in each iteration. The underlying ideas of these methods are basically to compress gradients, where each entry can be represented by much fewer bits than the original 32-bit floating-point number. Such compression introduces extra stochastic noises, i.e. quantization error, into the optimization process, and will slow down the convergence or even leads to divergence (Alistarh et al., 2017). The performance of these approaches relies on the tradeoff between the number of bits communicated per iteration and the quality of this information. Thus, aggressive schemes may only send one bit per coordinate (Bernstein et al., 2018; Tang et al., 2021) or used vector quantization (Leconte et al., 2021).

Contributions. The main contributions can be summarized as follows:

(1) We propose `QLSD`, a general MCMC algorithm specifically designed for Bayesian inference under the FL paradigm and two variance-reduced alternatives, especially tackling *heterogeneity*, *communication overhead* and *partial participation*.

(2) We provide a non-asymptotic convergence analysis of the proposed algorithms. The theoretical part highlights the impact of statistical heterogeneity measured by the discrepancy between local posterior distributions.

(3) We propose efficient mechanisms to mitigate the impact of statistical heterogeneity on convergence, either by using biased stochastic gradients or by introducing a *memory* mechanism that extends Horváth et al. (2022) to the Bayesian setting. In particular, we find that variance reduction indeed allows the proposed MCMC algorithm to converge towards the desired target posterior distribution when the number of observations becomes large.

(4) We illustrate the advantages of the proposed methods using several FL benchmarks. We show that the proposed methodology performs well compared to state-of-the-art Bayesian FL methods.

In this work, we extend these ideas to the Bayesian setting. We develop a novel federated Bayesian inference algorithm, called Quantized Langevin Stochastic Dynamics (`QLSD`) to address the communication bottleneck of distributed/federated algorithms. This framework incorporates the case of n clients, each owing a local potential $U_i : \mathbb{R}^d \to \mathbb{R}$ computed based on its local dataset $\mathcal{D}_i$. The agents perform Bayesian inference to target the posterior distribution proportional to $\exp(-\sum_{i=1}^n U_i)$ while respecting the federated learning constraints. Using an unbiased sequence $\{\mathcal{C}_k\}_{k\geq 1}$ of compression operators (Alistarh et al., 2017), these agents only communicate a quantized version of their stochastic gradient $\widehat{\nabla U_i}$ at each aggregation round. Then, the central server performs a Langevin dynamics step based on the received compressed gradients. The parameter θ_k is updated using the information of the participating clients $\mathcal{A}_{k+1}$:

$$\theta_{k+1} = \theta_k - \gamma \frac{n}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} \mathcal{C}_{k+1}(\widehat{\nabla U_i}(\theta_k)) + \sqrt{2\gamma} Z_{k+1}, \tag{1.6}$$

where Z_{k+1} is a standard Gaussian noise. Under assumptions stated in Theorem 4.5, the samples $\{\theta_k\}$ generated by (1.6) are approximately distributed according to $\prod_{i\in[n]} \exp(U_i)$. However, we illustrate theoretically and experimentally that this method suffers from heterogeneity and the use of a stochastic gradient $\widehat{\nabla U_i}$. To improve performance, we therefore introduce mechanisms leading to improved versions denoted `QLSD`* and `QLSD`$^{++}$. In the first version `QLSD`*, the stochastic gradient $\widehat{\nabla U_i}$ in (1.6) is replaced by the oracle $\widetilde{\nabla U_i}(\theta) = \widehat{\nabla U_i}(\theta) - \widehat{\nabla U_i}(\theta_\star)$, where $\theta_\star = \arg\min \sum_i U_i$; for more details, see the Langevin Fixed Point algorithm (Brosse et al., 2018). Interestingly, note that $\widetilde{\nabla U_i}$ is a biased estimate of ∇U_i since the expectation $\mathbb{E}[\widetilde{\nabla U_i}] \neq \nabla U_i$ in spite of $\mathbb{E}[\sum_i \widetilde{\nabla U_i}] = \sum_i \nabla U_i$. In Theorem 4.7, we derive asymptotic and non-asymptotic convergence guarantees for the proposed algorithm. However, obtaining the minimizer $\theta_\star$ is complicated in practical case scenario. Hence, we develop a last alternative coined `QLSD`$^{++}$ relying on the well-known SVRG technique (Johnson and Zhang, 2013) to reduce the noise introduced by the variance of the stochastic gradient combined with a memory mechanism to break down the heterogeneity problem (Horváth et al., 2022; Philippenko and Dieuleveut, 2020). Finally, we illustrate the performance of the proposed approach compared to various Bayesian federated learning benchmarks.

Furthermore, we numerically emphasize the compression benefits by achieving similar precision than classical methods with fewer bits.

Theorem 1.3 (Informal). *Under assumptions described in Chapter 4, there exist $\gamma_\star > 0$, such that for $\gamma \in (0, \gamma_\star)$, there are $\kappa \in (0,1)$, $C_0, C_1 > 0$ such that for $k \geq 0$, the distribution μ_k of the sample θ_k satisfies*

$$W_2^2\Big(\mu_k, \pi(\cdot|\mathcal{D}_{1:n})\Big) \leq (1-\kappa)^k C_0 + \gamma C_1.$$

• Chapter 5: *Conformal Prediction for Federated Uncertainty Quantification Under Label Shift* (RQ#5-RQ#6-RQ#7)

Accurate uncertainty quantification is crucial in modern machine learning applications. This is essential to develop reliable methods guaranteeing the validity of predictions. However, estimating valid prediction sets can be challenging in distributed settings, and this challenge is further exacerbated under label shift.

Contributions. The main contributions can be summarized as follows:

(1) We introduce a new method, `DP-FedCP`, to construct conformal prediction sets in a federated learning context that addresses label shift between agents; see Section 5.2. `DP-FedCP` is a federated learning algorithm based on federated computation of weighted quantiles of agent's non-conformity scores, where the weights reflect the label shift of each client with respect to the population. The quantiles are obtained by regularizing the pinball loss using Moreau-Yosida inf-convolution and a version of federated averaging procedure; see Section 5.3.

(2) We establish conformal prediction guarantees, ensuring the validity of the resulting prediction sets. Additionally, we provide differential private guarantees for `DP-FedCP`; see Section 5.4.

(3) We show that `DP-FedCP` provides valid confidence sets and outperforms standard approaches in a series of experiments on simulated data and image classification datasets; see Section 5.5.

Contrary to usual conformal methods, the `DP-FedCP` algorithm only computes the non-conformity scores on a subset of $\bar{N}$ calibration data. For example, $\bar{N} = \lfloor N/2 \rfloor$ when half of the calibration datapoints are used. One key mechanism of `DP-FedCP` consists in evaluating the discrepancy between the calibration and test distributions (P^{cal} and $P^\star$). Based on a Radon-Nikodym estimate of the likelihood ratio $\widehat{w}_y^\star = \mathrm{d}P_Y^\star / \mathrm{d}P_Y^{\mathrm{cal}}$, a valid prediction set can be obtained by weighting the non-conformity scores. Denote by $\{(X_k, Y_k)\}_{k\in[\bar{N}]}$ the calibration samples used to construct the prediction sets. For any $\mathbf{y} \in \mathcal{Y}$, we construct a family of weights $\{\widehat{p}_{y,\mathbf{y}}^\star\}_{y\in\mathcal{Y}}$ given by

$$\widehat{p}_{y,\mathbf{y}}^\star = \frac{\widehat{w}_y^\star}{\widehat{w}_{Y_{N^\star+1}^\star}^\star + \sum_{\ell=1}^{\bar{N}} \widehat{w}_{Y_\ell}^\star}.$$

Then using these weights, `DP-FedCP` leverages local non-conformity scores to derive personalized prediction sets for new datapoint $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star) \sim P^\star$, following

$$\begin{aligned} &\bar{\mu}_{\mathbf{y}}^\star = \widehat{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \textstyle\sum_{k=1}^{\bar{N}} \widehat{p}_{Y_k,\mathbf{y}}^\star \delta_{V_k}, \\ &\mathcal{C}_{\alpha,\bar{\mu}^\star}(X_{N^\star+1}^\star) = \Big\{\mathbf{y} \in \mathcal{Y} \colon V(X_{N^\star+1}^\star, \mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu}_{\mathbf{y}}^\star)\Big\}. \end{aligned} \tag{1.7}$$

Non-asymptotic bounds ensuring the validiy of these prediction sets are provided in Section 5.4. In particular, when the likelihood ratios are known, then the following result holds.

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| \leq \frac{6}{N} + \frac{36 + 6\log N}{N}\|\widehat{w}^\star\|^2_\infty + \frac{14\log N}{N}\sum\nolimits_{j:\ \frac{N^j}{12} < \log N} \sqrt{N^j},$$

where N^i corresponds to the calibration data owned by agent $i \in [n]$. The prediction set $\mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1})$ is generally intractable because determining the exact quantile $Q_{1-\alpha}(\bar{\mu}^\star_{\mathbf{y}})$ in a federated way is far from being straightforward. Actually, we develop a method solving this problem while ensuring that no attacker can determine with high confidence whether a particular individual's data is included in the dataset or not.

Part II

Distributed Sampling & Langevin MC

"The concrete is the abstract made familiar by use."

(**Paul Langevin**, La pensée et l'action, 1950)

Chapter 2

DG-LMC: Distributed Gradient Langevin Monte Carlo

Contents

Performing reliable Bayesian inference on a big data scale is becoming a keystone in the modern era of machine learning. A workhorse class of methods to achieve this task are Markov chain Monte Carlo (MCMC) algorithms and their design to handle distributed datasets has been the subject of many works. However, existing methods are not completely either reliable or computationally efficient. In this chapter, we propose to fill this gap in the case where the dataset is partitioned and stored on computing nodes within a cluster under a master/slaves architecture. We derive a user-friendly centralized distributed MCMC algorithm with provable scaling in high-dimensional settings. We illustrate the relevance of the proposed methodology on both synthetic and real data experiments.

2.1 Introduction

In the current machine learning era, data acquisition has seen significant progress due to rapid technological advances which now allow for more accurate, cheaper and faster data storage and collection. This data quest is motivated by modern machine learning techniques and algorithms which are now well-proven and have become common tools for data analysis. In most cases, the empirical success of these methods are based on a very large sample size (Bardenet et al., 2017; Bottou et al., 2018). This need for data is also theoretically justified by data probabilistic modelling which asserts that under appropriate conditions, the more data can be processed, the more accurate the inference can be performed. However, in recent years, several challenges have emerged regarding the use and access to data in mainstream machine learning methods. Indeed, first the amount of data is now so large that it has outpaced the increase in computation power of

computing resources (Verbraeken et al., 2020). Second, in many modern applications, data storage and/or use are not on a single machine but shared across several units (Raicu et al., 2006; Bernstein and Newcomer, 2009). Third, life privacy is becoming a prime concern for many users of machine learning applications who are therefore asking for methods preserving data anonymity (Shokri and Shmatikov, 2015; Abadi et al., 2016). Distributed machine learning aims at tackling these issues. One of its popular paradigms, referred to as data-parallel approach, is to consider that the training data are divided across multiple machines. Each of these units constitutes a worker node of a computing network and can perform a *local* inference based on the data it has access. Regarding the choice of the network, several options and frameworks have been considered. We focus here on the master/slaves architecture where the worker nodes communicate with each other through a device called the *master* node.

Under this framework, we are interested in carrying Bayesian inference about a parameter $\theta \in \mathbb{R}^d$ based on observed data $\{\mathbf{y}_k\}_{k=1}^n \in \mathsf{Y}^n$ (Robert, 2001). The dataset is assumed to be partitioned into S *shards* and stored on S machines among a collection of n worker nodes. The subset of observations associated to worker $i \in [n]$ is denoted by $\boldsymbol{y}_i$, with potentially $\boldsymbol{y}_i = \{\emptyset\}$ if $i \in [S+1:n]$, $n > S$. The posterior distribution of interest is assumed to admit a density w.r.t. the d-dimensional Lebesgue measure which factorizes across workers, *i.e.,*

$$\pi(\theta|\mathbf{y}_{1:n}) = Z_\pi^{-1} \prod_{i=1}^n \mathrm{e}^{-U_{\boldsymbol{y}_i}(\mathbf{A}_i\theta)}, \tag{2.1}$$

where $Z_\pi = \int_{\mathbb{R}^d} \prod_{i=1}^n \mathrm{e}^{-U_{\boldsymbol{y}_i}(\mathbf{A}_i\theta)} \mathrm{d}\theta$ is a normalization constant and $\mathbf{A}_i \in \mathbb{R}^{d_i \times d}$ are matrices that might act on the parameter of interest. For $i \in [n]$, the potential function $U_{\boldsymbol{y}_i} : \mathbb{R}^{d_i} \to \mathbb{R}$ is assumed to depend only on the subset of observations $\boldsymbol{y}_i$. Note that for $i \in [S+1:n]$, $n > S$, $U_{\boldsymbol{y}_i}$ does not depend on the data but only on the prior. For the sake of brevity, the dependency of π w.r.t. the observations $\{\boldsymbol{y}_i\}_{i=1}^n$ is notationally omitted and for $i \in [n]$, $U_{\boldsymbol{y}_i}$ is simply denoted by U_i.

To sample from π given by (2.1) in a distributed fashion, a large number of approximate methods have been proposed in the past ten years (Neiswanger et al., 2014; Ahn et al., 2014; Rabinovich et al., 2015; Scott et al., 2016; Nemeth and Sherlock, 2018; Chowdhury and Jermaine, 2018; Rendell et al., 2020). Despite multiple research lines, to the best of authors' knowledge, none of these proposals has been proven to be satisfactory. Indeed, the latter are not completely either computationally efficient in high-dimensional settings, reliable or theoretically grounded (Jordan et al., 2019).

This work is an attempt to fill this gap. To this purpose, we follow the data augmentation approach introduced in Vono et al. (2020) and referred to as asymptotically exact data augmentation (AXDA). Given a tolerance parameter $\boldsymbol{\rho}$, the main idea behind this methodology is to consider a joint distribution $\Pi_{\boldsymbol{\rho}}$ on the extended state space $\mathbb{R}^d \times \prod_{i=1}^n \mathbb{R}^{d_i}$ such that $\Pi_{\boldsymbol{\rho}}$ has a density w.r.t. the Lebesgue measure of the form $(\theta, z_{1:n}) \mapsto \prod_{i=1}^n \Pi_{\boldsymbol{\rho}}^i(\theta, z_i)$, with $\theta \in \mathbb{R}^d$ and $z_i \in \mathbb{R}^{d_i}$, $i \in [n]$. $\Pi_{\boldsymbol{\rho}}$ is carefully designed so that its marginal w.r.t. θ, denoted by $\pi_{\boldsymbol{\rho}}$, is a proxy of (2.1) for which quantitative approximation bounds can be derived and are controlled by $\boldsymbol{\rho}$. In addition, for any $i \in [n]$, $\Pi_{\boldsymbol{\rho}}^i(\theta, z_i)$ only depends on the data $\boldsymbol{y}_i$, and therefore plays a role similar to the local posterior $\pi^i(\theta) \propto \mathrm{e}^{-U_i(\mathbf{A}_i\theta)}$ in popular embarrassingly parallel approaches (Neiswanger et al., 2014; Scott et al., 2016). However, compared to this class of methods, AXDA does not seek for each worker to sample from $\Pi_{\boldsymbol{\rho}}^i$. Following a data augmentation strategy based on Gibbs sampling, AXDA instead requires each worker to

sample from the conditional distribution $\Pi_{\boldsymbol{\rho}}(z_i|\theta)$ and to communicate its sample to the master. $\Pi_{\boldsymbol{\rho}}$ is generally chosen such that sampling from $\Pi_{\boldsymbol{\rho}}(\theta|z_{1:n})$ is easy and does not require to access to the data. However, two main challenges remain: one has to sample efficiently from the conditional distribution $\Pi_{\boldsymbol{\rho}}(z_i|\theta)$ for $i \in [n]$ and avoid too frequent communication rounds on the master. Existing AXDA-based approaches unfortunately do not fulfill these important requirements (Vono et al., 2022a; Rendell et al., 2020). In this work, we leverage these issues by considering the use of the Langevin Monte Carlo (LMC) algorithm to approximately sample from $\Pi_{\boldsymbol{\rho}}(z_i|\theta)$ (Rossky et al., 1978; Roberts and Tweedie, 1996).

Our contributions are summarized in what follows.

(1) We introduce in Section 2.2 a new methodology called Distributed Gibbs using Langevin Monte Carlo (DG-LMC).

(2) Importantly, we provide in Section 2.3 a detailed quantitative analysis of the induced bias and show explicit convergence results. This stands for our main contribution and to the best of authors' knowledge, this theoretical study is one of the most complete among existing works which focused on distributed Bayesian machine learning with a master/slaves architecture. In particular, we discuss the complexity of our algorithm, the choice of hyperparameters, and provide practitioners with simple prescriptions to tune them. Further, we provide a thorough comparison of our method with existing approaches in Section 2.4.

(3) Finally, in Section 2.5, we show the benefits of the proposed sampler over popular and recent distributed MCMC algorithms on several numerical experiments.

Notations and conventions. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$. For $\ell \geq 1$, we refer to $\{1,\dots,n\}$ with the notation $[n]$ and for $i_1, i_2 \in \mathbb{N}$, $i_1 \leq i_2$, $\{i_1,\dots,i_2\}$ with the notation $[i_1 : i_2]$. For $0 \leq i < j$ and $(u_k; k \in \{i,\cdots,j\})$, we use the notation $u_{i:j}$ to refer to the vector $[u_i^\top,\cdots,u_j^\top]^\top$. We denote by $\mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$ the Gaussian distribution with mean vector $\mathbf{m}$ and covariance matrix $\boldsymbol{\Sigma}$. For a given matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we denote its smallest eigenvalue by $\lambda_{\min}(\mathbf{M})$. We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. We define the Wasserstein distance of order 2 for any probability measures μ,ν on $\mathbb{R}^d$ with finite 2-moment by $W_2(\mu,\nu) = (\inf_{\zeta\in\mathcal{T}(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\theta-\theta'\|^2 \mathrm{d}\zeta(\theta,\theta'))^{1/2}$, where $\mathcal{T}(\mu,\nu)$ is the set of transference plans of μ and ν.

2.2 Distributed Gibbs using Langevin Monte Carlo (DG-LMC)

In this section, we present the proposed methodology which is based on the AXDA statistical framework and the popular LMC algorithm.

AXDA relies on the decomposition of the target distribution π given in (2.1) to introduce an extended distribution which enjoys favorable properties for distributed computations. This distribution is defined on the state space $\mathbb{R}^d \times \mathsf{Z}$, where $\mathsf{Z} = \prod_{i=1}^n \mathbb{R}^{d_i}$, and admits a density w.r.t. the Lebesgue measure given, for any $\theta \in \mathbb{R}^d$, $z_{1:n} \in \mathsf{Z}$, by

$$\Pi_{\boldsymbol{\rho}}(\theta, z_{1:n}) \propto \prod_{i=1}^n \tilde{\Pi}^i_{\boldsymbol{\rho}}(\theta, z_i), \tag{2.2}$$

where $\tilde{\Pi}^i_{\boldsymbol{\rho}}(\theta, z_i) = \exp(-U_i(z_i) - \|z_i - \mathbf{A}_i\theta\|^2/2\rho_i)$ and $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^n \in \mathbb{R}^n_+$ is a sequence of positive tolerance parameters. Note that $\tilde{\Pi}^i_{\boldsymbol{\rho}}$ is not necessarily a probability density function. Actually, for $\Pi_{\boldsymbol{\rho}}$ to define a proper probability density, *i.e.* $\int_{\mathbb{R}^d \times \mathsf{Z}} \prod_{i=1}^n \tilde{\Pi}^i_{\boldsymbol{\rho}}(\theta, z_i) \mathrm{d}\theta \mathrm{d}z_{1:n} < \infty$, some conditions are required.

Assumption 2.1. *There exists $n' \in [n-1]$ such that the following conditions hold:* $\min_{i\in[n']} \inf_{z_i \in \mathbb{R}^{d_i}} U_i(z_i) > -\infty$, $\max_{i\in[n'+1:n]} \int_{\mathbb{R}^{d_i}} \mathrm{e}^{-U_i(z_i)} \mathrm{d}z_i < \infty$, *and* $\sum_{j=n'+1}^n \mathbf{A}_j^\top \mathbf{A}_j$ *is invertible.*

The next result shows that these mild assumptions are sufficient to guarantee that the extended model (2.2) is well-defined.

Proposition 2.2. *Assume Assumption 2.1. Then, for any $\boldsymbol{\rho} \in \mathbb{R}^n_+$, $\Pi_{\boldsymbol{\rho}}$ in (2.2) is a proper density.*

The data augmentation scheme (2.2) is approximate in the sense that the θ-marginal defined by

$$\pi_{\boldsymbol{\rho}}(\theta) = \int_{\mathsf{Z}} \Pi_{\boldsymbol{\rho}}(\theta, z_{1:n}) \mathrm{d}z_{1:n}, \tag{2.3}$$

coincides with (2.1) only in the limiting case $\max_{i\in[n]} \rho_i \downarrow 0$ (Scheffé, 1947). For a fixed $\boldsymbol{\rho}$, quantitative results on the induced bias in total variation distance can be found in Vono et al. (2022a). The main benefit of working with (2.2) is that conditionally upon θ, auxiliary variables $\{z_i\}_{i=1}^n$ are independent. Therefore, they can be sampled in parallel within a Gibbs sampler. For $i \in [n]$, the conditional density of z_i given θ writes

$$\Pi_{\boldsymbol{\rho}}(z_i|\theta) \propto \exp\Big(- U_i(z_i) - \frac{\|z_i - \mathbf{A}_i\theta\|^2}{2\rho_i} \Big). \tag{2.4}$$

On the other hand, the conditional distribution of θ given $z_{1:n}$ is a Gaussian distribution

$$\Pi_{\boldsymbol{\rho}}(\theta|z_{1:n}) = \mathcal{N}(\boldsymbol{\mu}(z_{1:n}), \mathbf{Q}^{-1}), \tag{2.5}$$

with precision matrix $\mathbf{Q} = \sum_{i=1}^n \mathbf{A}_i^\top \mathbf{A}_i/\rho_i$ and mean vector $\boldsymbol{\mu}(z_{1:n}) = \mathbf{Q}^{-1} \sum_{i=1}^n \mathbf{A}_i^\top z_i/\rho_i$. Under **H**2.1, note that $\mathbf{Q}$ is invertible and therefore this conditional Gaussian distribution is well-defined. Since sampling from high-dimensional Gaussian distributions can be performed efficiently, this Gibbs sampling scheme is interesting as long as sampling from (2.4) is cheap. Vono et al. (2022a) proposed the use of a rejection sampling step requiring to set $\rho_i = \mathrm{O}(1/d_i)$. When $d_i \gg 1$, this condition unfortunately leads to prohibitive computational costs and hence prevents its practical use for general Bayesian inference problems. Instead of sampling exactly from (2.4), Rendell et al. (2020) rather proposed to use Metropolis-Hastings algorithms. However, it is not clear whether this choice indeed leads to efficient sampling schemes.

To tackle these issues, we propose to build upon LMC to end up with a distributed MCMC algorithm which is both simple to implement, efficient and amenable to a theoretical study. LMC stands for a popular way to approximately generate samples from a given distribution based on the Euler-Maruyama discretization scheme of the overdamped Langevin stochastic differential equation (Roberts and Tweedie, 1996). At iteration t of the considered Gibbs sampling scheme and given a current parameter $\theta^{(t)}$, LMC applied to (2.4) considers, for $i \in [n]$, the recursion

$$z_i^{(t+1)} = \Big(1 - \frac{\gamma_i}{\rho_i}\Big) z_i^{(t)} + \frac{\gamma_i}{\rho_i} \mathbf{A}_i \theta^{(t)} - \gamma_i \nabla U_i\Big(z_i^{(t)}\Big) + \sqrt{2\gamma_i}\, \xi_i^{(t)}$$

Algorithm 2.1 Distributed Gibbs using LMC (DG-LMC)

Input: burn-in T_{bi}; for $i \in [n]$, tolerance parameters $\rho_i > 0$, step-sizes $\gamma_i \in (0, \rho_i/(1+\rho_i M_i)]$, local LMC steps $N_i \geq 1$.
Initialize $\theta^{(0)}$ and $z_{1:n}^{(0)}$.
for $t = 0$ **to** $T-1$ **do**
  // Sampling from $\Pi_{\boldsymbol{\rho}}(z_{1:n}|\theta)$
  for $i = 1$ **to** n // In parallel on the n workers **do**
    $u_i^{(0)} = z_i^{(t)}$
    for $k = 0$ **to** $N_i - 1$ // N_i local LMC steps **do**
      $\xi_i^{(k,t)} \sim \mathcal{N}\left(\mathbf{0}_{d_i}, \mathbf{I}_{d_i}\right)$
      $\mathbf{g}_i = \left(1 - \frac{\gamma_i}{\rho_i}\right) u_i^{(k)} + \frac{\gamma_i}{\rho_i} \mathbf{A}_i \theta^{(t)} - \gamma_i \nabla U_i\left(u_i^{(k)}\right)$
      $u_i^{(k+1)} = \mathbf{g}_i + \sqrt{2\gamma_i}\xi_i^{(k,t)}$ // See (2.4)
    $z_i^{(t+1)} = u_i^{(N_i)}$
  // Sampling from $\Pi_{\boldsymbol{\rho}}(\theta|z_{1:n})$
  $\theta^{(t+1)} \sim \mathcal{N}(\boldsymbol{\mu}(z_{1:n}^{(t+1)}), \mathbf{Q}^{-1})$ // See (2.5)
Output: samples $\{\theta^{(t)}\}_{t=T_{\text{bi}}-1}^{T}$.

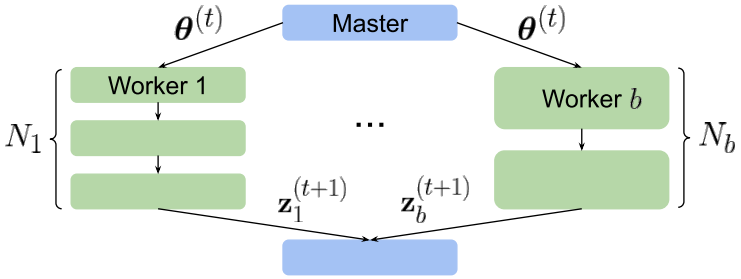


Figure 2.1 – Illustration of one global iteration of Algorithm 2.1. For each worker, the width of the green box represents the amount of time required to perform one LMC step.

where $\gamma_i > 0$ is a fixed step-size and $\{\xi_i^{(k)} : k \in \mathbb{N}\}_{i\in[n]}$ a sequence of independent and identically distributed (i.i.d.) d-dimensional standard Gaussian random variables. Only using a single step of LMC on each worker might incur important communication costs. To mitigate the latter while increasing the proportion of time spent on exploring the state-space, we instead allow each worker to perform N_i ($N_i \geq 1$) LMC steps (Dieuleveut and Patel, 2019; Rendell et al., 2020). Letting N_i varies across workers prevents Algorithm 2.1 to suffer from a significant block-by-the-slowest delay in cases where the response times of the workers are unbalanced (Ahn et al., 2014). The proposed algorithm, coined Distributed Gibbs using Langevin Monte Carlo (DG-LMC), is depicted in Algorithm 2.1 and illustrated in Figure 2.1.

2.3 Detailed analysis of DG-LMC

In this section, we derive quantitative bias and convergence results for DG-LMC and show that its mixing time only scales quadratically w.r.t. the dimension d. We also discuss the choice of hyperparameters and provide guidelines to tune them.

2.3.1 Non-Asymptotic Analysis

The scope of our analysis will focus on smooth and strongly log-concave target posterior distributions π. While these assumptions may be restrictive in practice, they allow for a detailed theoretical study of the proposed algorithm.

Assumption 2.3. *(i) For any $i \in [n]$, U_i is twice continuously differentiable and*

$$\sup_{z_i \in \mathbb{R}^{d_i}} \|\nabla^2 U_i(z_i)\| \leq M_i.$$

(ii) For any $i \in [n]$, U_i is m_i-strongly convex: there exists $m_i > 0$ such that

$$m_i \mathbf{I}_{d_i} \preceq \nabla^2 U_i.$$

Under these assumptions, it is shown in Lemma 2.24 in the Appendix that $-\log \pi$ is strongly convex with constant

$$m_U = \lambda_{\min}(\textstyle\sum_{i=1}^n m_i \mathbf{A}_i^\top \mathbf{A}_i). \tag{2.6}$$

Behind the use of LMC, the main motivation is to end up with a simple hybrid Gibbs sampler amenable to a non-asymptotic theoretical analysis based on previous works (Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019). In the following, this study is carried out using the Wasserstein distance of order 2.

Convergence Results

DG-LMC introduced in Algorithm 2.1 defines a homogeneous Markov chain $(V_t)_{t\in\mathbb{N}} = (\theta_t, Z_t)_{t\in\mathbb{N}}$ with realizations $(\theta^{(t)}, z_{1:n}^{(t)})_{t\in\mathbb{N}}$. We denote by $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ the Markov kernel associated with $(V_t)_{t\in\mathbb{N}}$. Since no Metropolis-Hastings step is used in combination with LMC, the proposed algorithm does not fall into the class of Metropolis-within-Gibbs samplers (Roberts and Rosenthal, 2006). Therefore, a first step is to show that $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ admits a unique invariant distribution and is geometrically ergodic. We proceed via an appropriate synchronous coupling which reduces the convergence analysis of $(V_t)_{t\in\mathbb{N}}$ to that of the marginal process $(Z_t)_{t\in\mathbb{N}}$. While the proof of the convergence of $(Z_t)_{t\in\mathbb{N}}$ shares some similarities with LMC (Durmus and Moulines, 2019), the analysis of $(Z_t)_{t\in\mathbb{N}}$ is much more involved and especially in the case $\max_{i\in[n]} N_i > 1$. We believe that the proof techniques we developed to show the next result can be useful to the study of other MCMC approaches based on LMC.

Proposition 2.4. *Assume Assumption 2.1-Assumption 2.3 and let $c > 0$ and $\boldsymbol{\gamma} = \{\gamma_i\}_{i=1}^n$ $\boldsymbol{N} = \{N_i\}_{i=1}^n$ satisfying $\max_{i\in[n]} \gamma_i \leq \bar{\gamma}$, $\min_{i\in[n]}\{N_i\gamma_i\}/\max_{i\in[n]}\{N_i\gamma_i\} \geq c$ and $\max_{i\in[n]}\{N_i\gamma_i\} \leq C_1$ where $\bar{\gamma}, C_1$ are explicit constants only depending on $(m_i, M_i, \rho_i)_{i\in[n]}$*[1,2].

[1]When $\boldsymbol{N} = \mathbf{1}_n$, $C_1 = \bar{\gamma} = 1/\max_{i\in[n]}\{M_i + \rho_i^{-1}\}$.

[2]When $\max_{i\in[n]} N_i > 1$, C_1 is of order $\min_{i\in[n]} \rho_i^2$ when $\max_{i\in[n]} \rho_i \to 0$, see Lemma 2.20 in the Appendix.

Then, there exists a probability measure $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ *such that* $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ *is invariant for* $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$*, there exists* $C_2 > 0$ *such that for any integer* $t \geq 0$ *and* $\mathbf{v} = (\theta, z) \in \mathbb{R}^d \times \mathsf{Z}$*, we have*

$$W_2(\delta_{\mathbf{v}} P^t_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}) \leq C_2 \cdot \left(1 - \min_{i\in[n]}\{N_i\gamma_i m_i\}/2\right)^t \cdot W_2(\delta_{\mathbf{v}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}).$$

Explicit expressions for C_1 *and* C_2 *are given in Proposition 2.21 in the Appendix. Finally, if* $\boldsymbol{N} = N\mathbf{1}_n$ *for* $N \geq 1$*, then* $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}} = \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\mathbf{1}_n}$.

We now discuss Proposition 2.4. If we set, for any $i \in [n]$, $N_i = 1$, the convergence rate in Proposition 2.4 becomes equal to $1 - \min_{i\in[n]}\{\gamma_i m_i\}/2$. In this specific case, we show in Proposition 2.13 that DG-LMC actually admits the tighter convergence rate $1 - \min_{i\in[n]}\{\gamma_i m_i\}$ which simply corresponds to the rate at which the slowest LMC conditional kernel converges. On the other hand, when $\max_{i\in[n]} N_i > 1$, the convergence of $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ towards $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ only holds if $\max_{i\in[n]}\{N_i\gamma_i\}$ is sufficiently small. This condition is necessary to ensure a contraction in W_2 and can be understood intuitively as follows in the case where $\boldsymbol{N} = N\mathbf{1}_n$ and $\boldsymbol{\gamma} = \gamma\mathbf{1}_n$. Given two vectors (θ_k, θ'_k) and an appropriate coupling (Z_{k+1}, Z'_{k+1}), we can show that $Z_{k+1} - Z'_{k+1}$ involves two competing terms: one keeping $Z_{k+1} - Z'_{k+1}$ close to $Z_k - Z'_k$ and another one driving $Z_{k+1} - Z'_{k+1}$ away from $\theta_k - \theta'_k$ (and therefore of $Z_k - Z'_k$) as N increases. This implies that N stands for a tradeoff and the product $N\gamma$ cannot be arbitrarily chosen. Finally, it is worth mentioning that the tolerance parameters $\{\rho_i\}_{i\in[n]}$ implicitly drive the convergence rate of DG-LMC. In the case $N_i = 1$, a sufficient condition on the step-sizes to ensure a contraction is $\gamma_i \leq 2/(M_i + m_i + 1/\rho_i)$. We can denote that the smaller ρ_i, the smaller γ_i and the slower the convergence.

Starting from the results of Proposition 2.4, we can analyze the convergence properties of DG-LMC. We specify our result to the case where we take for the specific initial distribution

$$\mu^\star_{\boldsymbol{\rho}} = \delta_{z^\star} \otimes \Pi_{\boldsymbol{\rho}}(\cdot|z^\star), \tag{2.7}$$

where $z^\star = ([\mathbf{A}_1\theta^\star]^\top, \cdots, [\mathbf{A}_n\theta^\star]^\top)^\top$, $\theta^\star = \arg\min\{-\log\pi\}$ and $\Pi_{\boldsymbol{\rho}}(\cdot|z^\star)$ is defined in (2.5). Note that sampling from $\mu^\star_{\boldsymbol{\rho}}$ is straightforward and simply consists in setting $z^{(0)} = z^\star$ and drawing $\theta^{(0)}$ from $\Pi_{\boldsymbol{\rho}}(\cdot \mid z^\star)$. For $t \geq 1$, we consider the marginal law of θ_t initialized at $\mathbf{v}^\star$ with distribution $\mu^\star_{\boldsymbol{\rho}}$ and denote it $\Gamma^t_{\mathbf{v}^\star}$. As mentioned previously, the proposed approach relies on two approximations which both come with some bias we need to control. This naturally brings us to consider the following inequality based on the triangular inequality and the definition of the Wasserstein distance:

$$W_2(\Gamma^t_{\mathbf{v}^\star}, \pi(\cdot|\mathcal{D})) \leq W_2(\mu^\star_{\boldsymbol{\rho}} P^t_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}) + W_2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho}}) + W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D})), \tag{2.8}$$

where $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$, $\Pi_{\boldsymbol{\rho}}$ and $\pi_{\boldsymbol{\rho}}$ are defined in Proposition 2.4, (2.2) and (2.3), respectively. In Proposition 2.22 in the Appendix, we provide an upper bound on the first term on the right-hand side based on Proposition 2.4. In the next section, we focus on controlling the last two terms on the right-hand side.

Quantitative Bounds on the Bias

The error term $W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D}))$ in (2.8) is related to the underlying AXDA framework which induces an approximate posterior representation $\pi_{\boldsymbol{\rho}}$. It can be controlled by the sequence of positive tolerance parameters $\{\rho_i\}_{i=1}^n$. By denoting $\bar{\rho} = \max_{i\in[n]} \rho_i$,

Table 2.1 – For the specific initialization $\mathbf{v}^\star$ with distribution $\mu^\star_{\boldsymbol{\rho}}$ given in (2.7), dependencies w.r.t. d and ε of the parameters involved in Algorithm 2.1 and of $t_{\text{mix}}(\varepsilon; \mathbf{v}^\star)$ to get a W_2-error of at most ε.

Assumptions		ρ_ε	γ_ε	N_ε	$t_{\text{mix}}(\varepsilon; \mathbf{v}^\star)$	Gradient evaluations
Assumption 2.1 Assumption 2.3	d	$\mathrm{O}(d^{-1})$	$\mathrm{O}(d^{-3})$	$\mathrm{O}(d)$	$\mathrm{O}(d^2 \log(d))$	$\mathrm{O}(d^3 \log(d))$
	ε	$\mathrm{O}(\varepsilon)$	$\mathrm{O}(\varepsilon^4)$	$\mathrm{O}(\varepsilon^{-2})$	$\mathrm{O}(\varepsilon^{-2} \lvert \log(\varepsilon) \rvert)$	$\mathrm{O}(\varepsilon^{-4} \lvert \log(\varepsilon) \rvert)$
Assumption 2.1 Assumption 2.7	d	$\mathrm{O}(d^{-1})$	$\mathrm{O}(d^{-2})$	$\mathrm{O}(1)$	$\mathrm{O}(d^2 \log(d))$	$\mathrm{O}(d^2 \log(d))$
Assumption 2.3	ε	$\mathrm{O}(\varepsilon)$	$\mathrm{O}(\varepsilon^2)$	$\mathrm{O}(1)$	$\mathrm{O}(\varepsilon^{-2} \lvert \log(\varepsilon) \rvert)$	$\mathrm{O}(\varepsilon^{-2} \lvert \log(\varepsilon) \rvert)$

Proposition 2.5 shows that this error can be quantitatively assessed and is of order $\mathrm{O}(\bar{\rho})$ for sufficiently small values of this parameter.

Proposition 2.5. *Assume Assumption 2.1, Assumption 2.3. In addition, let* $\mathbf{A} = [\mathbf{A}_1^\top, \dots, \mathbf{A}_n^\top]^\top$ *and denote* $\sigma_U^2 = \|\mathbf{A}^\top \mathbf{A}\| \max_{i \in [n]} \{M_i^2\}/m_U$, *where* m_U *is defined in (2.6). Then, for any* $\bar{\rho} \leq \sigma_U^2/12$,

$$W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D})) \leq \sqrt{2/m_U} \max(A_{\boldsymbol{\rho}}, B_{\boldsymbol{\rho}}),$$

where $A_{\boldsymbol{\rho}} = d\,\mathrm{O}(\bar{\rho})$ *and* $B_{\boldsymbol{\rho}} = d^{1/2}\,\mathrm{O}(\bar{\rho})$ *for* $\bar{\rho} \downarrow 0$. *Explicit expressions for* $A_{\boldsymbol{\rho}}, B_{\boldsymbol{\rho}}$ *are given in Section 2.C in the Appendix.*

In the case where π is Gaussian, the approximate distribution $\pi_{\boldsymbol{\rho}}$ admits an explicit expression and is Gaussian as well (e.g. when $n = 1$, the mean is the same and the covariance matrix is inflated by a factor $\rho \mathbf{I}_d$), see for instance Rendell et al. (2020, Section S2) and Vono et al. (2020, Section 5.1). Hence, an explicit expression for $W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D}))$ can be derived. Based on this result, we can check that the upper bound provided by Proposition 2.5 matches the same asymptotics as $\rho \to 0$ and $d \to \infty$.

The second source of approximation error is induced by the use of LMC within Algorithm 2.1 to target the conditional distribution $\Pi_{\boldsymbol{\rho}}(z_{1:n}|\theta)$ in (2.4). The stationary distribution of $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ whose existence is ensured in Proposition 2.4 differs from $\Pi_{\boldsymbol{\rho}}$. The associated bias is assessed quantitatively in Proposition 2.6.

Proposition 2.6. *Assume Assumption 2.1-Assumption 2.3. For any* $i \in [n]$, *define* $\tilde{M}_i = M_i + 1/\rho_i$ *and let* $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$, $\boldsymbol{N} \in (\mathbb{N}^*)^n$ *such that for any* $i \in [n]$,

$$\gamma_i \leq \frac{m_i}{40 \tilde{M}_i^2} \min_{i \in [n]} (m_i/\tilde{M}_i)^2 / \max_{i \in [n]} (m_i/\tilde{M}_i)^2, \tag{2.9}$$

$$N_i = \left\lfloor m_i \min_{i \in [n]} \{m_i/\tilde{M}_i\}^2 / (20 \gamma_i \tilde{M}_i^2 \max_{i \in [n]} \{m_i/\tilde{M}_i\}^2) \right\rfloor. \tag{2.10}$$

Then, we have

$$W_2^2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho}}) \leq C_3 \sum_{i=1}^n d_i \gamma_i \tilde{M}_i^2,$$

where $C_3 > 0$ *only depends on* $(m_i, M_i, \mathbf{A}_i, \rho_i)_{i=1}^n$ *and is explicitly given in Proposition 2.37 in the Appendix.*

With the notation $\bar{\gamma} = \max_{i\in[n]} \gamma_i$, Proposition 2.6 implies that $W_2(\Pi_{\boldsymbol{\rho}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}) \leq \mathrm{O}(\bar{\gamma}^{1/2})(\sum_{i=1}^n d_i)^{1/2}$ for $\bar{\gamma} \downarrow 0$. Note that this result is in line with Durmus and Moulines (2019, Corollary 7) and can be improved under further regularity assumptions on U, as shown below.

Assumption 2.7. *U is three times continuously differentiable and there exists $L_i > 0$ such that for all $z_i, z_i' \in \mathbb{R}^{d_i}$, $\|\nabla^2 U_i(z_i) - \nabla^2 U_i(z_i')\| \leq L_i \|z_i - z_i'\|$.*

Proposition 2.8. *Assume Assumption 2.1-Assumption 2.3-Assumption 2.7. For any $i \in [n]$, define $\tilde{M}_i = M_i + 1/\rho_i$ and let $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$, $\boldsymbol{N} \in (\mathbb{N}^*)^n$ such that for any $i \in [n]$, (2.9) and (2.10) hold. Then, we have*

$$W_2^2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho}}) \leq C_4 \sum_{i\in[n]} d_i \gamma_i (1/\tilde{M}_i^2 + \gamma_i \tilde{M}_i^2),$$

where $C_4 > 0$ only depends on $(m_i, M_i, L_i, \mathbf{A}_i, \rho_i)_{i=1}^n$ and is explicitly given in Proposition 2.41 in the Appendix.

Mixing Time with Explicit Dependencies

Based on explicit non-asymptotic bounds shown in Propositions 2.4, 2.5 and 2.6 and the decomposition (2.8), we are now able to analyze the scaling of Algorithm 2.1 in high dimension. Given a prescribed precision $\varepsilon > 0$ and an initial condition $\mathbf{v}^\star$ with distribution $\mu_{\boldsymbol{\rho}}^\star$ given in (2.7), we define the ε-mixing time associated to $\Gamma_{\mathbf{v}^\star}$ by

$$t_{\mathrm{mix}}(\varepsilon; \mathbf{v}^\star) = \min \left\{ t \in \mathbb{N} : W_2(\Gamma_{\mathbf{v}^\star}^t, \pi(\cdot|\mathcal{D})) \leq \varepsilon \right\}.$$

This quantity stands for the minimum number of DG-LMC iterations such that the θ-marginal distribution is at most at an ε W_2-distance from the initial target π. Under the condition that $n \max_{i\in[n]} d_i = \mathrm{O}(d)$ and by assuming for simplicity that for any $i \in [n]$, $m_i = m, M_i = M, L_i = L, \rho_i = \rho, \gamma_i = \gamma$ and $N_i = N$, Table 2.1 gathers the dependencies w.r.t. d and ε of the parameters involved in Algorithm 2.1 and of $t_{\mathrm{mix}}(\varepsilon; \mathbf{v}^\star)$ to get a W_2-error of at most ε. Note that the mixing time of Algorithm 2.1 scales at most quadratically (up to polylogarithmic factors) in the dimension. When Assumption 2.7 holds, we can see that the number of local iterations becomes independent of d and ε which leads to a total number of gradient evaluations with better dependencies w.r.t. to these quantities. Up to the authors' knowledge, these explicit results are the first among the centralized distributed MCMC literature and in particular give the dependency w.r.t. d and ε of the number of local LMC iterations on each worker. Overall, the proposed approach appears as a scalable and reliable alternative for high-dimensional and distributed Bayesian inference.

2.3.2 DG-LMC in Practice: Guidelines for Practitioners

We now discuss practical guidelines for setting the values of hyperparameters involved in Algorithm 2.1. Based on Proposition 2.4, we theoretically show an optimal choice of order $N_i\gamma_i \asymp m_i\rho_i^2/(\rho_i M_i + 1)^2$. Ideally, within the considered distributed setting, the optimal value for $(N_i, \gamma_i)_{i\in[n]}$ would boil down to optimize the value of $\max_{i\in[n]}\{N_i\gamma_i\}$ under the constraints derived in Proposition 2.4 combined with communication considerations. In particular, this would imply a comprehensive modelling of the communication costs including I/O bandwidths constraints. These optimization tasks fall outside

the scope of the present chapter, and therefore we let the search of optimal values for future works. Since our aim here is to provide practitioners with simple prescriptions, we rather focus on general rules involving tractable quantities.

Selection of $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$

From Durmus and Moulines (2017) and references therein, a simple sufficient condition on step-sizes $\boldsymbol{\gamma} = \{\gamma_i\}_{i=1}^n$ to guarantee the stability of LMC is $\gamma_i \leq \rho_i/(\rho_i M_i + 1)$ for $i \in [n]$. Both the values of γ_i and ρ_i are subject to a bias-variance tradeoff. More precisely, large values yield a Markov chain with small estimation variance but high asymptotic bias. Conversely, small values produce a Markov chain with small asymptotic bias but which requires many iterations to obtain a stable estimator. We propose to mitigate this tradeoff by setting γ_i to a reasonably large value, that is for $i \in [n]$, $\gamma_i \in [0.1\rho_i/(\rho_i M_i + 1), 0.5\rho_i/(\rho_i M_i + 1)]$. Since γ_i saturates to $1/M_i$ when $\rho_i \to \infty$, there is no computational advantage to choose very large values for ρ_i. Based on several numerical studies, we found that setting ρ_i of the order of $1/M_i$ was a good compromise between computational efficiency and asymptotic bias.

$\boldsymbol{N}$: A Trade-Off between Asymptotic Bias and Communication Overhead

In a similar vein, the choice of $\boldsymbol{N} = \{N_i\}_{i=1}^n$ also stands for a tradeoff but here between asymptotic accuracy and communication costs. Indeed, many local LMC iterations reduces the communication overhead but at the expense of a larger asymptotic bias since the master parameter is not updated enough. Ahn et al. (2014) proposed to tune the number of local iterations N_i on a given worker based on the amount of time needed to perform one local iteration, denoted here by τ_i. Given an average number of local iterations N_{avg}, the authors set $N_i = q_i n N_{\text{avg}}$ with $q_i = \tau_i^{-1} / \sum_{k=1}^n \tau_k^{-1}$ so that $n^{-1} \sum_{i=1}^n N_i = N_{\text{avg}}$. As mentioned by the aforementioned authors, this choice allows to keep the block-by-the-slowest delay small by letting fast workers perform more iterations in the same wall-clock time. Although they showed how to tune N_i w.r.t. communication considerations, they let the choice of N_{avg} to the practitioner. Here, we propose a simple guideline to set N_{avg} such that N_i stands for a good compromise between the amount of time spent on exploring the state-space and communication overhead. As highlighted in the discussion after Proposition 2.4, as γ_i becomes smaller, more local LMC iterations are required to sufficiently explore the latent space before the global consensus round on the master. Assuming for any $i \in [n]$ that γ_i has been chosen following our guidelines in Section 2.3.2, this suggests to set $N_{\text{avg}} = \lceil (1/n) \sum_{i \in [n]} \rho_i / (\gamma_i [\rho_i M_i + 1]) \rceil$.

2.4 Related work

As already mentioned in Section 2.1, hosts of contributions have focused on deriving distributed MCMC algorithms to sample from (2.1). This section briefly reviews the main existing research lines and draws a detailed comparison with the proposed methodology.

Table 2.2 – Synthetic overview of the main existing distributed MCMC methods under a master-slave architecture. The column *Exact* means that the Markov chain defined by the MCMC sampler admits (2.1) as invariant distribution. The column *Comm.* reports the communication frequency. A value of 1 means that the sampler communicates after every iteration. T stands for the total number of iterations and $N < T$ is a tunable parameter to mitigate communication costs. The acronym D-SGLD stands for distributed stochastic gradient Langevin dynamics.

Method	Type	Exact	Comm.	Bias bounds	Scaling
Wang and Dunson (2013)	one-shot	×	$1/T$	√	$\mathrm{O}(\mathrm{e}^d)$
Neiswanger et al. (2014)	one-shot	×	$1/T$	×	$\mathrm{O}(\mathrm{e}^d)$
Minsker et al. (2014)	one-shot	×	$1/T$	√	unknown
Srivastava et al. (2015)	one-shot	×	$1/T$	×	unknown
Wang et al. (2015)	one-shot	×	$1/T$	√	$\mathrm{O}(\mathrm{e}^d)$
Scott et al. (2016)	one-shot	×	$1/T$	×	unknown
Nemeth and Sherlock (2018)	one-shot	×	$1/T$	×	unknown
Jordan et al. (2019)	one-shot	×	$1/T$	√	unknown
Ahn et al. (2014)	D-SGLD	×	$1/N$	×	unknown
Chen et al. (2016)	D-SGLD	×	1	√	unknown
El Mekkaoui et al. (2021)	D-SGLD	×	$1/N$	√	unknown
Rabinovich et al. (2015)	g. consensus	×	$1/N$	×	unknown
Chowdhury and Jermaine (2018)	g. consensus	√	1	N/A	unknown
Rendell et al. (2020)	g. consensus	×	$1/N$	√	unknown
This chapter	g. consensus	×	$1/N$	√	$\mathrm{O}(d^2 \log(d))$

2.4.1 Existing distributed MCMC methods

Existing methodologies are mostly approximate and can be loosely speaking divided into three groups: *one-shot*, distributed stochastic gradient MCMC and *global consensus* approaches. To ease the understanding, a synthetic overview of their main characteristics is presented in Table 2.2.

One-shot approaches stand for communication-efficient schemes where workers and master only exchange information at the very beginning and the end of the sampling task; similarly to MapReduce schemes (Dean and Ghemawat, 2004). Most of these methods assume that the posterior density factorizes into a product of local posteriors and launch independent Markov chains across workers to target them. The local posterior samples are then combined through the master node using a single final aggregation step. This step turns to be the milestone of one-shot approaches and was the topic of multiple contributions Wang and Dunson (2013); Neiswanger et al. (2014); Minsker et al. (2014); Srivastava et al. (2015); Scott et al. (2016); Nemeth and Sherlock (2018). Unfortunately, the latter are either infeasible in high-dimensional settings or have been shown to yield inaccurate posterior representations empirically, if the posterior is not near-Gaussian, or if the local posteriors differ significantly Wang et al. (2015); Dai et al. (2019); Rendell et al. (2020). Alternative schemes have been recently proposed to tackle these issues but their theoretical scaling w.r.t. the dimension d is currently unknown (Jordan et al., 2019; Mesquita et al., 2020).

Albeit popular in the machine learning community, distributed stochastic gradient MCMC methods (Ahn et al., 2014) suffer from high variance when the dataset is large because of the use of stochastic gradients (Brosse et al., 2018). Some surrogates have been recently proposed to reduce this variance such as the use of *stale* or *conducive* gradients (Chen et al., 2016; El Mekkaoui et al., 2021). However, these variance reduction methods require an increasing number of workers for the former and come at the price of a prohibitive pre-processing step for the latter. In addition, it is currently unclear whether these methods are able to generate efficiently accurate samples from a given target distribution.

Contrary to aforementioned distributed MCMC approaches, global consensus methods periodically share information between workers by performing a consensus round between the master and the workers (Rabinovich et al., 2015; Chowdhury and Jermaine, 2018; Vono et al., 2019; Rendell et al., 2020). Again, they have been shown to perform well in practice, but their theoretical understanding is currently limited.

2.4.2 Comparison with the proposed methodology

Table 2.2 compares Algorithm 2.1 with existing approaches detailed previously. In addition to having a simple implementation and guidelines, it is worth noticing that DG-LMC appears to benefit from favorable convergence properties compared to the other considered methodologies.

We complement this comparison with an informal discussion on the computational and communication complexities of Algorithm 2.1. Recall that the dataset is assumed to be partitioned into S shards and stored on S workers among a collection of n computing nodes. Suppose that the s-th shard has size n_s, and let T be the number of total MCMC iterations and c_{com} the communication cost. In addition, denote by $c_{\text{eval}}^{(i)}$ the approximate wall-clock time required to evaluate U_i or its gradient. For the ease of exposition, we do not discuss the additional overhead due to bandwidth restrictions and assume similar computation costs, *i.e.*, $Nc_{\text{eval}} \simeq N_i c_{\text{eval}}^{(i)}$, to perform each local LMC step at each iteration of Algorithm 2.1. Under these assumptions, the total complexity of Algorithm 2.1 is $\mathrm{O}(T[2c_{\text{com}} + Nc_{\text{eval}}])$. Following the same reasoning, distributed stochastic gradient Langevin dynamics (D-SGLD) and one-shot approaches admit complexities of the order $\mathrm{O}(T[2c_{\text{com}} + Nc_{\text{eval}}n_{\text{mb}}/n_s])$ and $\mathrm{O}(Tc_{\text{eval}} + 2c_{\text{com}})$, respectively. The integer n_{mb} stands for the minibatch size used in D-SGLD. Despite their very low communication overhead, existing one-shot approaches are rarely reliable and therefore not necessarily efficient to sample from π given a prescribed computational budget, see Rendell et al. (2020) for a recent overview. D-SGLD seems to enjoy a lower complexity than Algorithm 2.1 when n_{mb} is small. Unfortunately, this choice comes with two main shortcomings: (i) a larger number of iterations T to achieve the same precision because of higher variance of gradient estimators, and (ii) a smaller amount of time spent on exploration compared to communication latency. By falling into the global consensus class of methods, the proposed methodology hence appears as a good compromise between one-shot and D-SGLD algorithms in terms of both computational complexity and accuracy. Section 2.5 will enhance the benefits of Algorithm 2.1 by showing experimentally better convergence properties and posterior approximation.

2.5 Experiments

This section compares numerically DG-LMC with the most popular and recent centralized distributed MCMC approaches namely D-SGLD and the global consensus Monte Carlo (GCMC) algorithm proposed in Rendell et al. (2020). Since all these approaches share the same communication latency, this feature is not discussed here.

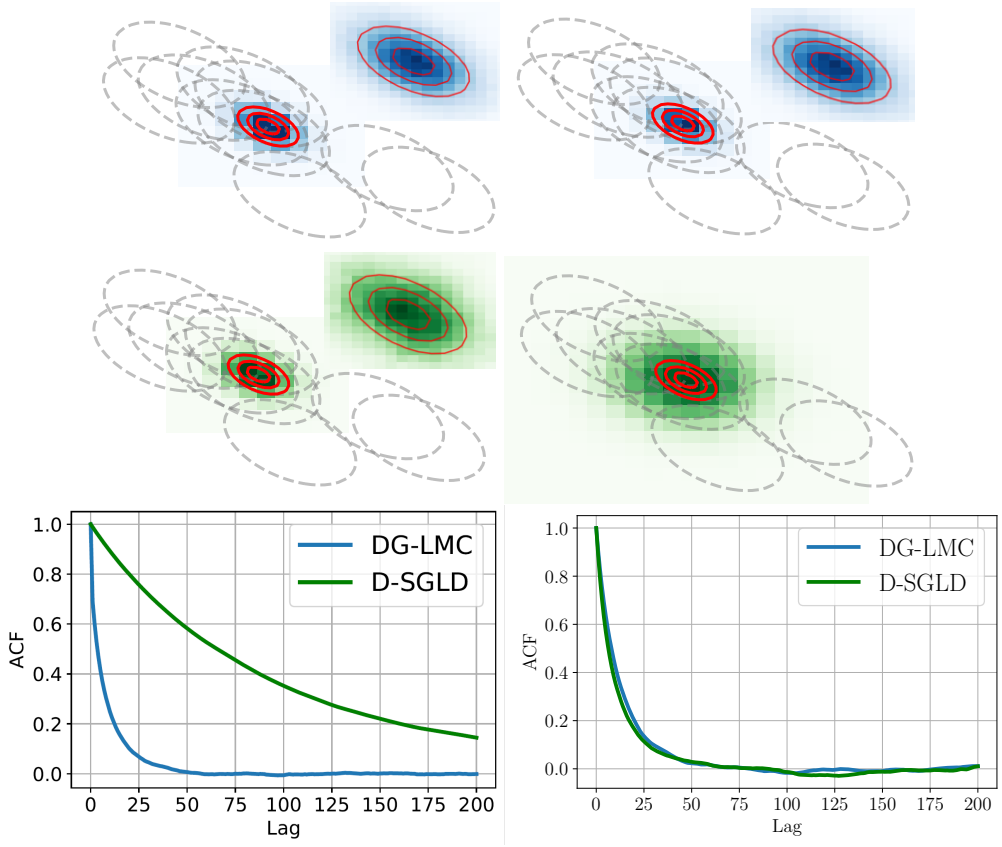


Figure 2.2 – Toy Gaussian experiment. (left) $N = 1$ local iterations and (right) $N = 10$. (top) DG-LMC, (middle) D-SGLD and (bottom) ACF comparison between DG-LMC and D-SGLD.

2.5.1 Toy Gaussian Example

In this toy example, we first illustrate the behavior of DG-LMC w.r.t. the number of local iterations which drives the communication overhead. We consider the conjugate Gaussian model $\pi(\theta|\mathbf{y}_{1:n}) \propto \mathcal{N}(\theta|\mathbf{0}_d, \boldsymbol{\Sigma}_\mathbf{0}) \prod_{i=1}^n \mathcal{N}(\mathbf{y}_i|\theta, \boldsymbol{\Sigma}_\mathbf{1})$, with positive definite matrices $\boldsymbol{\Sigma}_\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{1}$. We set $d = 2$, allocate $b = 20,000$ observations to a cluster made of $n = 10$ workers and compare DG-LMC with D-SGLD. Both MCMC algorithms have been run using the same number of local iterations N per worker and for a fixed budget of $T = 100,000$ iterations including a burn-in period equal to $T_{\text{bi}} = 10,000$. Regarding DG-LMC, we follow the guidelines in Section 2.3.2 and set for all $i \in [n]$, $\mathbf{A}_i = \mathbf{I}_d$, $\rho_i = 1/(5M_i)$ and $\gamma_i = 0.25\rho_i/(\rho_i M_i + 1)$. On the other hand, D-SGLD has been run with batch-size $b/(10n)$ and a step-size chosen such that the resulting posterior approximation is similar to that of DG-LMC for $N = 1$. Figure 2.2 depicts the results for $N = 1$ and $N = 10$ on the left and right columns, respectively. The top row (resp. middle row)

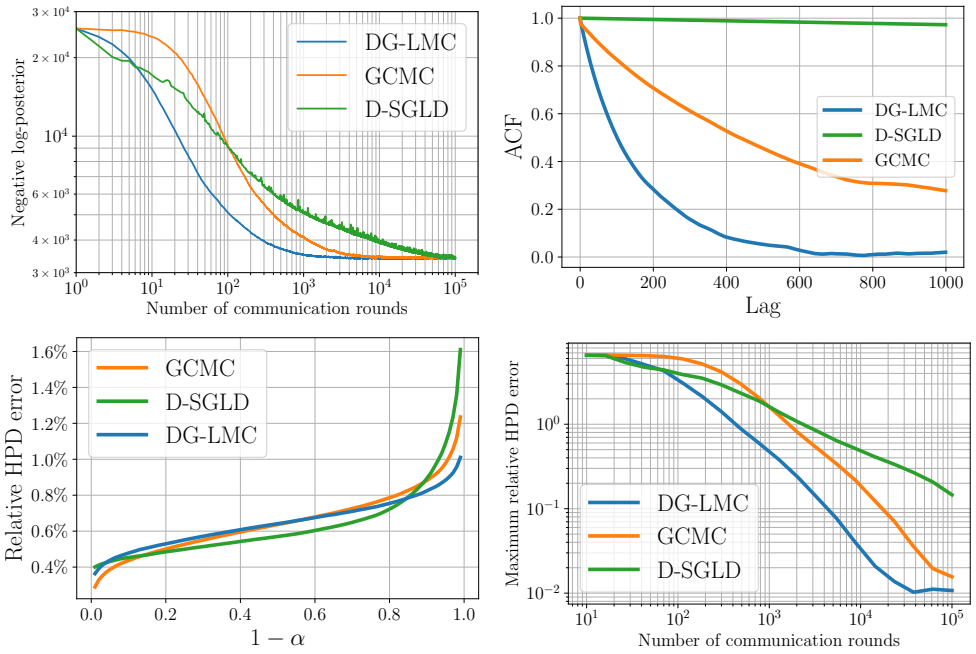


Figure 2.3 – Logistic regression. From left to right: negative log-posterior, ACF, HPD relative error after and during the sampling procedure.

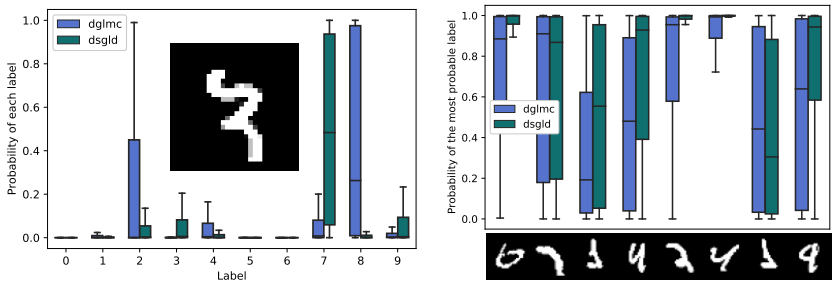


Figure 2.4 – Bayesian neural network. (left) probability of the most probable label for 8 examples and (right) probability of each label for a single example.

shows the contours of the n local posteriors in dashed gray, the contours of the target posterior in red and the 2D histogram built with DG-LMC (resp. D-SGLD) samples in blue (resp. green). When required, a zoomed version of these figures is depicted in the top right corner. It can be noted that DG-LMC exhibits better mixing properties while achieving similar performances as shown by the autocorrelation function (ACF) on the bottom row. Furthermore, its posterior approximation is robust to the choice of N in contrast to D-SGLD, which needs further tuning of its step-size to yield an accurate posterior representation. This feature is particularly important for distributed computations since N is directly related to communication costs and might often change depending upon the hardware architecture.

2.5.2 Bayesian Logistic Regression

This second experiment considers a more challenging problem namely Bayesian logistic regression. We use the *covtype*[3] dataset with $d = 54$ and containing $b = 581,012$ observations partitioned into $n = 16$ shards. We set $N = 10$, $T = 200,000$, $T_{\text{bi}} = T/10$ for all approaches, and again used the guidelines in Section 2.3.2 to tune DG-LMC. Under the Bayesian paradigm, we are interested in performing uncertainty quantification by estimating the highest posterior density (HPD) regions. For any $\alpha \in (0,1)$, define $\mathcal{C}_\alpha = \{\theta \in \mathbb{R}^d; -\log \pi(\theta|\mathbf{y}_{1:n}) \leq \eta_\alpha\}$ where $\eta_\alpha \in \mathbb{R}$ is chosen such that $\int_{\mathcal{C}_\alpha} \pi(\theta|\mathbf{y}_{1:n})\mathrm{d}\theta = 1-\alpha$. For the three approximate MCMC approaches, we computed the relative HPD error based on the scalar summary η_α, *i.e.* $|\eta_\alpha - \eta_\alpha^{\text{true}}|/\eta_\alpha^{\text{true}}$ where $\eta_\alpha^{\text{true}}$ has been estimated using the Metropolis adjusted Langevin algorithm. The parameters of GCMC and D-SGLD have been chosen such that all MCMC algorithms achieve similar HPD error. Figure 2.3 shows that this error is reasonable and of the order of 1%. Nonetheless, one can denote that DG-LMC achieves this precision level faster than GCMC and D-SGLD due to better mixing properties. This confirms that the proposed methodology is indeed efficient and reliable to perform Bayesian analyzes compared to existing popular methodologies.

2.5.3 Bayesian Neural Network

Up to now, both our theoretical and experimental results focused on the strongly log-concave scenario and showed that even in this case, DG-LMC appeared as a competitive alternative. In this last experiment, we propose to end the study of DG-LMC on an open note without ground truth by tackling the challenging sampling problem associated to Bayesian neural networks. We consider the MNIST training dataset consisting of $n = 60,000$ observations partitioned into $n = 50$ shards and such that for any $i \in [n]$ and $k \in [10]$, $\mathbb{P}(y_i = k|\theta, \mathbf{x}_i) = \beta_k$ where β_k is the k-th element of $\sigma(\sigma(\mathbf{x}_i^\top \mathbf{W}_1 + \mathbf{n}_1)\mathbf{W}_2 + \mathbf{n}_2)$, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{x}_i$ are covariates, and $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{n}_1$ and $\mathbf{n}_2$ are matrices of size 784×128, 128×10, 1×128 and 1×10, respectively. We set normal priors for each weight matrix and bias vector, $N = 10$ and ran DG-LMC with constant hyperparameters across workers $(\rho, \gamma) = (0.02, 0.005)$ and D-SGLD using a step-size of 10^{-5}. Exact MCMC approaches are too computationally costly to launch for this experiment and therefore no ground truth about the true posterior distribution is available. To this purpose, Figure 2.4 only compares the credibility regions associated to the posterior predictive distribution. Similarly to previous experiments, we found that D-SGLD was highly sensitive to hyperparameters choices (step-size and minibatch size). Except for a few testing examples, most of conclusions given by DG-LMC and D-SGLD regarding the predictive uncertainty coincide. In addition, posterior accuracies on the test set given by both algorithms are similar.

2.6 Conclusion

In this chapter, a simple algorithm coined DG-LMC has been introduced for distributed MCMC sampling. In addition, it has been established that this method inherits favorable convergence properties and numerical illustrations support our claims.

[3] www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets

2.A Proof of Proposition 2.2

Let $n' \in [n-1]$, $p' = \sum_{i=n'+1}^{n} d_i$ and consider

$$\begin{aligned} \mathbf{B}_{n'}^\top &= [\mathbf{A}_{n'+1}^\top/\rho_{n'+1}^{1/2} \cdots \mathbf{A}_n^\top/\rho_n^{1/2}] \in \mathbb{R}^{d\times p'}, \\ \bar{\mathbf{B}}_{n'} &= \mathbf{B}_{n'}^\top \mathbf{B}_{n'} = \sum_{i=n'+1}^{n} \{\mathbf{A}_i^\top \mathbf{A}_i/\rho_i\} \in \mathbb{R}^{d\times d}. \end{aligned} \tag{2.11}$$

Note that under Assumption 2.1, $\bar{\mathbf{B}}_{n'}$ is invertible. Indeed, it is a symmetric positive definite matrix since for any $\theta \in \mathbb{R}^d$, $\langle \bar{\mathbf{B}}_{n'}\theta, \theta\rangle \geq [\min_{i\in[n]} \rho_i^{-1}]\langle \sum_{i=n'+1}^{n} \mathbf{A}_i^\top \mathbf{A}_i \theta, \theta\rangle > 0$ using that $\sum_{i=n'+1}^{n} \mathbf{A}_i^\top \mathbf{A}_i$ is invertible. Define the orthogonal projection onto the range of $\mathbf{B}_{n'}$ and the diagonal matrix:

$$\mathbf{P}_{n'} = \mathbf{B}_{n'}\bar{\mathbf{B}}_{n'}^{-1}\mathbf{B}_{n'}^\top, \qquad \tilde{\mathbf{D}}_{n'} = \mathrm{diag}(\mathbf{I}_{d_{n'+1}}/\rho_{n'+1}, \ldots, \mathbf{I}_{d_n}/\rho_n). \tag{2.12}$$

2.A.1 Technical lemma

Lemma 2.9. *Assume Assumption 2.1. For any $(\theta, z_{n'+1:n}) \in \mathbb{R}^d \times \mathbb{R}^{p'}$, setting $z = z_{n'+1:n}$, we have*

$$\begin{aligned} \sum_{i=n'+1}^{n} \left\{\|z_i - \mathbf{A}_i\theta\|^2/\rho_i\right\} &= (\tilde{\mathbf{D}}_{n'}^{1/2} z)^\top \{\mathbf{I}_{p'} - \mathbf{P}_{n'}\}(\tilde{\mathbf{D}}_{n'}^{1/2} z) \\ &\quad + (\theta - \bar{\mathbf{B}}_{n'}^{-1}\mathbf{B}_{n'}^\top \tilde{\mathbf{D}}_{n'}^{1/2} z)^\top \bar{\mathbf{B}}_{n'}(\theta - \bar{\mathbf{B}}_{n'}^{-1}\mathbf{B}_{n'}^\top \tilde{\mathbf{D}}_{n'}^{1/2} z). \end{aligned}$$

Proof Setting $\mathbf{b} = \mathbf{B}_{n'}^\top \tilde{\mathbf{D}}_{n'}^{1/2} z$ and using the fact that $\bar{\mathbf{B}}_{n'}$ is symmetric, we have

$$\begin{aligned} \sum_{i=n'+1}^{n} \left\{\|z_i - \mathbf{A}_i\theta\|^2/\rho_i\right\} &= \theta^\top \bar{\mathbf{B}}_{n'}\theta - 2\theta^\top \mathbf{b} + \sum_{i=n'+1}^{n} \|z_i\|^2/\rho_i \\ &= \sum_{i=n'+1}^{n} \|z_i\|^2/\rho_i - \mathbf{b}^\top \bar{\mathbf{B}}_{n'}^{-1}\mathbf{b} + (\theta - \bar{\mathbf{B}}_{n'}^{-1}\mathbf{b})^\top \bar{\mathbf{B}}_{n'}(\theta - \bar{\mathbf{B}}_{n'}^{-1}\mathbf{b}). \end{aligned}$$

Using that $\mathbf{b}^\top \bar{\mathbf{B}}_{n'}^{-1}\mathbf{b} = (\tilde{\mathbf{D}}_{n'}^{1/2} z)^\top \mathbf{P}_{n'}(\tilde{\mathbf{D}}_{n'}^{1/2} z)$ and $\mathbf{P}_{n'}$ is a projection, $\mathbf{P}_{n'}^2 = \mathbf{P}_{n'}$ completes the proof. ∎

2.A.2 Proof of Proposition 2.2

Proposition 2.10. *Assume Assumption 2.1. Then, the function*

$$\psi : (\theta, z_{1:n}) \mapsto \prod_{i=1}^{n} \exp\{-U_i(z_i) - \|z_i - \mathbf{A}_i\theta\|^2/(2\rho_i)\}$$

is integrable on $\mathbb{R}^d \times \mathbb{R}^p$, where $p = \sum_{i=1}^{n} d_i$.

Proof Using Assumption 2.1 and the Fubini theorem, there exists $C_1 > 0$ such that:

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left[\prod_{i=1}^{n'} \int_{\mathbb{R}^{d_i}} \mathrm{e}^{-U_i(z_i)} \mathrm{e}^{-\frac{\left\|z_i - \mathbf{A}_i\theta\right\|^2}{2\rho_i}} \mathrm{d}z_i \cdot \prod_{j=n'+1}^{n} \int_{\mathbb{R}^{d_j}} \mathrm{e}^{-U_j(z_j)} \mathrm{e}^{-\frac{\left\|z_j - \mathbf{A}_j\theta\right\|^2}{2\rho_j}} \mathrm{d}z_j \right] \mathrm{d}\theta \\
&\quad \leq C_1 \int_{\mathbb{R}^d} \left[\prod_{i=1}^{n'} \int_{\mathbb{R}^{d_i}} \mathrm{e}^{-\frac{\left\|z_i - \mathbf{A}_i\theta\right\|^2}{2\rho_i}} \mathrm{d}z_i \cdot \prod_{j=n'+1}^{n} \int_{\mathbb{R}^{d_j}} \mathrm{e}^{-U_j(z_j)} \mathrm{e}^{-\frac{\left\|z_j - \mathbf{A}_j\theta\right\|^2}{2\rho_j}} \mathrm{d}z_j \right] \mathrm{d}\theta \\
&\leq C_1 \prod_{i=1}^{n'} (2\pi\rho_i)^{d_i/2} \int_{\mathbb{R}^d} \left[\prod_{j=n'+1}^{n} \int_{\mathbb{R}^{d_j}} \mathrm{e}^{-U_j(z_j)} \exp\left(- \left\| z_j - \mathbf{A}_j\theta \right\|^2 / (2\rho_j) \right) \mathrm{d}z_j \right] \mathrm{d}\theta \\
&= C_1 \prod_{i=1}^{n'} (2\pi\rho_i)^{d_i/2} \int_{\mathbb{R}^{d_{n'+1}}} \cdots \int_{\mathbb{R}^{d_n}} \left[\prod_{j=n'+1}^{n} \mathrm{e}^{-U_j(z_j)} \right] \left[\int_{\mathbb{R}^d} \prod_{j=n'+1}^{n} \mathrm{e}^{-\frac{\left\|z_j - \mathbf{A}_j\theta\right\|^2}{2\rho_j}} \mathrm{d}\theta \right] \mathrm{d}z_{n'+1:n}.
\end{aligned} \tag{2.13}$$

Using Lemma 2.9 and the fact that $\mathbf{I}_{p'} - \mathbf{P}_{n'}$ is positive definite, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^d} \prod_{j=n'+1}^{n} \exp\left(- \left\| z_j - \mathbf{A}_j\theta \right\|^2 / (2\rho_j) \right) \mathrm{d}\theta &= \exp\left(-(\tilde{\mathbf{D}}_{n'}^{1/2} z)^\top \{ \mathbf{I}_{p'} - \mathbf{P}_{n'} \} (\tilde{\mathbf{D}}_{n'}^{1/2} z)/2 \right) \\
&\quad \times \int_{\mathbb{R}^d} \exp\left(-(\theta - \bar{\mathbf{B}}_{n'}^{-1} \mathbf{B}_{n'}^\top \tilde{\mathbf{D}}_{n'}^{1/2} z)^\top \bar{\mathbf{B}}_{n'} (\theta - \bar{\mathbf{B}}_{n'}^{-1} \mathbf{B}_{n'}^\top \tilde{\mathbf{D}}_{n'}^{1/2} z)/2 \right) \mathrm{d}\theta \\
&\quad \leq \det\left(\bar{\mathbf{B}}_{n'} \right)^{-1/2} (2\pi)^{d/2}.
\end{aligned}$$

Then, the proof is completed by plugging this expression into (2.13) and using from Assumption 2.1 that $z_{n'+1:n} \mapsto \prod_{j=n'+1}^{n} \mathrm{e}^{-U_j(z_j)}$ is integrable. ■

2.B Proof of Proposition 2.4

This section aims at proving Proposition 2.4 in the main chapter. To ease the understanding, we dissociate the scenarios where $\max_{i\in[n]} N_i = 1$ and $\max_{i\in[n]} N_i > 1$. In addition, in all this section $\boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$ is assumed to be fixed.

2.B.1 Single local LMC iteration

In this section, we assume that a single LMC step is performed locally on each worker, that is $\max_{i\in[n]} N_i = 1$. For this, we introduce the conditional Markov transition kernel defined for any $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, $\theta \in \mathbb{R}^d$, $z = (z_1, \cdots, z_n) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$, and for $i \in [n]$, $\mathsf{B}_i \in \mathcal{B}(\mathbb{R}^{d_i})$, by

$$Q_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z, \mathsf{B}_1 \times \cdots \times \mathsf{B}_n | \theta) = \prod_{i=1}^{n} R_{\rho_i,\gamma_i}(z_i, \mathsf{B}_i | \theta),$$

where

$$R_{\rho_i,\gamma_i}(z_i,\mathsf{B}_i|\theta)=\int_{\mathsf{B}_i}\exp\left\{-\frac{1}{4\gamma_i}\left\|\tilde{z}_i-\left(1-\frac{\gamma_i}{\rho_i}\right)z_i-\frac{\gamma_i}{\rho_i}\mathbf{A}_i\theta+\gamma_i\nabla U_i(z_i)\right\|^2\right\}\frac{\mathrm{d}\tilde{z}_i}{(4\pi\gamma_i)^{d_i/2}}. \tag{2.14}$$

Recall that $p=\sum_{i=1}^n d_i$. The considered Gibbs sampler in Algorithm 2.1 defines a homogeneous Markov chain $X_\ell^\top=(\theta_\ell^\top,Z_\ell^\top)_{\ell\geq 1}$ where $Z_\ell^\top=([Z_\ell^1]^\top,\cdots,[Z_\ell^n]^\top)$. Indeed, it is easy to show that for any $\ell\in\mathbb{N}$ and measurable bounded function $f:\mathbb{R}^p\to\mathbb{R}_+$, $\mathbb{E}[f(Z_{\ell+1})|X_\ell]=\int_{\mathbb{R}^p}f(z)Q_{\boldsymbol{\rho},\boldsymbol{\gamma}}(Z_\ell,\mathrm{d}z|\theta_\ell)$ and therefore $(X_\ell)_{\ell\in\mathbb{N}}$ is associated with the Markov kernel defined, for any $\mathbf{x}^\top=(\theta^\top,z^\top)\in\mathbb{R}^d\times\mathbb{R}^p$ and $\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)$, $\mathsf{B}\in\mathcal{B}(\mathbb{R}^p)$, by

$$P_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathbf{x},\mathsf{A}\times\mathsf{B})=\int_{\mathsf{B}}Q_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z,\mathrm{d}\tilde{\mathbf{z}}|\theta)\int_{\mathsf{A}}\Pi_{\boldsymbol{\rho}}(\mathrm{d}\tilde{\theta}|\tilde{\mathbf{z}}), \tag{2.15}$$

where $\Pi_{\boldsymbol{\rho}}(\cdot|\tilde{\mathbf{z}})$ is defined in (2.5). Let $(\xi_\ell)_{\ell\geq 1}$ be a sequence of i.i.d. d-dimensional standard Gaussian random variables independent of the family of independent random variables $\{(\eta_\ell^i)_{\ell\geq 1}:i\in[n]\}$ where for any $i\in[n]$ and $\ell\geq 1$, η_ℓ^i is a d_i-dimensional standard Gaussian random variable. We define the stochastic processes $(X_\ell,\tilde{X}_\ell)_{\ell\geq 0}$ on $\mathbb{R}^p\times\mathbb{R}^p$ starting from $(X_0,\tilde{X}_0)=(\mathbf{x},\tilde{\mathbf{x}})=((\theta^\top,z^\top)^\top,(\tilde{\theta}^\top,\tilde{z}^\top)^\top)$ and following the recursion for $\ell\geq 0$,

$$X_{\ell+1}=(\theta_{\ell+1}^\top,Z_{\ell+1}^\top)^\top,\qquad \tilde{X}_{\ell+1}=(\tilde{\theta}_{\ell+1}^\top,\tilde{Z}_{\ell+1}^\top)^\top, \tag{2.16}$$

where $Z_{\ell+1}=([Z_{\ell+1}^1]^\top,\ldots,[Z_{\ell+1}^n]^\top)^\top$, $\tilde{Z}_{\ell+1}=([\tilde{Z}_{\ell+1}^1]^\top,\ldots,[\tilde{Z}_{\ell+1}^n]^\top)^\top$ are defined, for any $i\in[n]$, by

$$\begin{aligned}Z_{\ell+1}^i&=(1-\gamma_i/\rho_i)Z_\ell^i+(\gamma_i/\rho_i)\mathbf{A}_i\theta_\ell-\gamma_i\nabla U_i(Z_\ell^i)+\sqrt{2\gamma_i}\eta_{\ell+1}^i,\\ \tilde{Z}_{\ell+1}^i&=(1-\gamma_i/\rho_i)\tilde{Z}_\ell^i+(\gamma_i/\rho_i)\mathbf{A}_i\tilde{\theta}_\ell-\gamma_i\nabla U_i(\tilde{Z}_\ell^i)+\sqrt{2\gamma_i}\eta_{\ell+1}^i,\end{aligned} \tag{2.17}$$

and $\theta_{\ell+1},\tilde{\theta}_{\ell+1}$ by

$$\theta_{\ell+1}=\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}Z_{\ell+1}+\bar{\mathbf{B}}_0^{-1/2}\xi_{\ell+1},\qquad \tilde{\theta}_{\ell+1}=\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\tilde{Z}_{\ell+1}+\bar{\mathbf{B}}_0^{-1/2}\xi_{\ell+1}, \tag{2.18}$$

where $\bar{\mathbf{B}}_0$, $\mathbf{B}_0$ and $\tilde{\mathbf{D}}_0$ are given in (2.11) and (2.12), respectively. Note that X_ℓ and $\tilde{X}_\ell$ are distributed according to $\delta_{\mathbf{x}}P_{\boldsymbol{\rho},\boldsymbol{\gamma}}^\ell$ and $\delta_{\tilde{\mathbf{x}}}P_{\boldsymbol{\rho},\boldsymbol{\gamma}}^\ell$, respectively. Hence, by definition of the Wasserstein distance of order 2, it follows that

$$W_2(\delta_{\mathbf{x}}P_{\boldsymbol{\rho},\boldsymbol{\gamma}}^\ell,\delta_{\tilde{\mathbf{x}}}P_{\boldsymbol{\rho},\boldsymbol{\gamma}}^\ell)\leq\mathbb{E}\left[\|X_\ell-\tilde{X}_\ell\|^2\right]^{1/2}. \tag{2.19}$$

Thus, in this section we focus on upper bounding the squared norm $\|X_\ell-\tilde{X}_\ell\|$ from which we get an explicit bound on the Wasserstein distance thanks to the previous inequality.

Supporting lemmata

Note that Assumption 2.1 implies the invertibility of the matrix $\mathbf{B}_0$ defined in (2.11) since we have the existence of $n'\in[n-1]$, such that $\sum_{i=n'+1}^n\lambda_{\min}(\mathbf{A}_i^\top\mathbf{A}_i)/\rho_i>0$ and by the semi-positiveness of the symmetric matrices $\{\mathbf{A}_i^\top\mathbf{A}_i\}_{i\in[n]}$, we get that $\lambda_{\min}(\mathbf{B}_0)=\sum_{i=1}^n\lambda_{\min}(\mathbf{A}_i^\top\mathbf{A}_i)/\rho_i\geq\sum_{i=n'+1}^n\lambda_{\min}(\mathbf{A}_i^\top\mathbf{A}_i)/\rho_i$. To prove Proposition 2.4 in the case $\max_{i\in[n]}N_i=1$, we first upper bound (2.19) by building upon the following two technical lemmas.

Lemma 2.11. *Assume Assumption 2.1 and consider $(X_\ell, \tilde{X}_\ell)_{\ell\in\mathbb{N}}$ defined in (2.16). Then, for any $\ell \in \mathbb{N}$, it holds almost surely that*

$$\|X_{\ell+1} - \tilde{X}_{\ell+1}\|^2 \leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\|Z_{\ell+1} - \tilde{Z}_{\ell+1}\|^2.$$

Proof Let $\ell \geq 0$. By (2.18), we have $\theta_{\ell+1} - \tilde{\theta}_{\ell+1} = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}(Z_{\ell+1} - \tilde{Z}_{\ell+1})$ which implies that

$$\begin{aligned}\|X_{\ell+1} - \tilde{X}_{\ell+1}\|^2 &= \|\theta_{\ell+1} - \tilde{\theta}_{\ell+1}\|^2 + \|Z_{\ell+1} - \tilde{Z}_{\ell+1}\|^2 \\ &\leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\|Z_{\ell+1} - \tilde{Z}_{\ell+1}\|^2.\end{aligned}$$

■

For any $\mathbf{v} \in \mathbb{R}^n$, define the block diagonal matrix

$$\mathbf{D}_\mathbf{v} = \mathrm{diag}\left(v_1 \cdot \mathbf{I}_{d_1}, \ldots, v_n \cdot \mathbf{I}_{d_n}\right) \in \mathbb{R}^{p\times p} \tag{2.20}$$

and consider the following contraction factor:

$$\kappa_{\boldsymbol{\gamma}} = \max_{i\in[n]} \left\{|1 - \gamma_i m_i| \vee |1 - \gamma_i(M_i + 1/\rho_i)|\right\}. \tag{2.21}$$

Using this notation, the next result holds.

Lemma 2.12. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$. Then for any $\mathbf{x} = (z^\top, \theta^\top)^\top, \tilde{\mathbf{x}} = (\tilde{z}^\top, \tilde{\theta}^\top)^\top$, with $(\theta, \tilde{\theta}) \in (\mathbb{R}^d)^2$ and $(z, \tilde{z}) \in (\mathbb{R}^p)^2$, for any $\ell \geq 1$, we have*

$$\begin{aligned}W_2(\delta_\mathbf{x} P_{\boldsymbol{\rho},\boldsymbol{\gamma}}^\ell, \delta_{\tilde{\mathbf{x}}} P_{\boldsymbol{\rho},\boldsymbol{\gamma}}^\ell) \leq \kappa_{\boldsymbol{\gamma}}^{\ell-1} \cdot &\left((1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2) \cdot \frac{\max_{i\in[n]}\{\gamma_i\}}{\min_{i\in[n]}\{\gamma_i\}}\right)^{1/2} \\ &\times \left[\kappa_{\boldsymbol{\gamma}}\|z - \tilde{z}\| + \|\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\|\|\theta - \tilde{\theta}\|\right],\end{aligned}$$

where $\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}$ is defined as in (2.20) with $\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}} = (\gamma_1/\rho_1^{1/2}, \ldots, \gamma_n/\rho_n^{1/2})$, $\bar{\mathbf{B}}_0$, $\mathbf{B}_0$, $P_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ and $\kappa_{\boldsymbol{\gamma}}$ are given in (2.11), (2.15), (2.21), respectively.

Proof Consider $(X_k, \tilde{X}_k)_{k\in\mathbb{N}}$ defined in (2.16). By (2.19) and Lemma 2.11, we need to bound $(\|Z_k - \tilde{Z}_k\|)_{k\in\mathbb{N}}$. Let $\ell \in \mathbb{N}^*$. For any $i \in [n]$, we have by (2.17), that

$$Z_{\ell+1}^i - \tilde{Z}_{\ell+1}^i = \left(1 - \frac{\gamma_i}{\rho_i}\right)(Z_\ell^i - \tilde{Z}_\ell^i) + \frac{\gamma_i}{\rho_i}\mathbf{A}_i(\theta_\ell - \tilde{\theta}_\ell) - \gamma_i\left(\nabla U_i(Z_\ell^i) - \nabla U_i(\tilde{Z}_\ell^i)\right). \tag{2.22}$$

Since U_i is twice differentiable, we have

$$\nabla U_i(Z_\ell^i) - \nabla U_i(\tilde{Z}_\ell^i) = \int_0^1 \nabla^2 U_i(\tilde{Z}_\ell^i + t(Z_\ell^i - \tilde{Z}_\ell^i))\,\mathrm{d}t \cdot (Z_\ell^i - \tilde{Z}_\ell^i).$$

Using $\theta_\ell - \tilde{\theta}_\ell = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}(Z_\ell - \tilde{Z}_\ell)$, it follows that

$$Z_{\ell+1}^i - \tilde{Z}_{\ell+1}^i = \left(\left[1 - \frac{\gamma_i}{\rho_i}\right]\mathbf{I}_{d_i} - \gamma_i \int_0^1 \nabla^2 U_i(\tilde{Z}_\ell^i + t(Z_\ell^i - \tilde{Z}_\ell^i))\,\mathrm{d}t\right)(Z_\ell^i - \tilde{Z}_\ell^i)$$

$$+ \frac{\gamma_i}{\rho_i}\mathbf{A}_i\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}(Z_\ell - \tilde{Z}_\ell).$$

Consider the $p \times p$ block diagonal matrix defined by

$$\mathbf{D}_{U,\ell} = \mathrm{diag}\left(\gamma_1 \int_0^1 \nabla^2 U_1(\tilde{Z}_\ell^1 + t(Z_\ell^1 - \tilde{Z}_\ell^1))\,\mathrm{d}t, \cdots, \gamma_n \int_0^1 \nabla^2 U_n(\tilde{Z}_\ell^n + t(Z_\ell^n - \tilde{Z}_\ell^n))\,\mathrm{d}t\right).$$

With the projection matrix $\mathbf{P}_0$ defined in (2.12), the difference $Z_{\ell+1} - \tilde{Z}_{\ell+1}$ can be rewritten as

$$Z_{\ell+1} - \tilde{Z}_{\ell+1} = \left(\mathbf{I}_p - \mathbf{D}_{U,\ell} - \mathbf{D}_{\boldsymbol{\gamma}}^{1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}(\mathbf{I}_p - \mathbf{P}_0)\tilde{\mathbf{D}}_0^{1/2}\right)(Z_\ell - \tilde{Z}_\ell),$$

where $\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}$ is defined as in (2.20) with $\boldsymbol{\gamma}/\boldsymbol{\rho} = (\gamma_1/\rho_1, \ldots, \gamma_n/\rho_n)$. Since $\mathbf{D}_{U,\ell}$ commutes with $\mathbf{D}_{\boldsymbol{\gamma}}$ and $\mathbf{P}_0$ is an orthogonal projection matrix, using Assumption 2.3-(i)-(ii), we get

$$\begin{aligned}
&\|Z_{\ell+1} - \tilde{Z}_{\ell+1}\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} \\
&= \|\mathbf{D}_{\boldsymbol{\gamma}}^{-1/2}(\mathbf{D}_{\boldsymbol{\gamma}}^{1/2}\mathbf{D}_{\boldsymbol{\gamma}}^{-1/2} - \mathbf{D}_{\boldsymbol{\gamma}}^{1/2}\mathbf{D}_{U,\ell}\mathbf{D}_{\boldsymbol{\gamma}}^{-1/2} - \mathbf{D}_{\boldsymbol{\gamma}}^{1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}(\mathbf{I}_p - \mathbf{P}_0)\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{\gamma}}^{-1/2})(Z_\ell - \tilde{Z}_\ell)\| \\
&\leq \|\mathbf{I}_p - \mathbf{D}_{U,\ell} - \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\left(\mathbf{I}_p - \mathbf{P}_0\right)\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\|\|Z_\ell - \tilde{Z}_\ell\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}}.
\end{aligned}$$

Note that Assumption 2.1 and Assumption 2.3 and the fact that $\mathbf{P}_0$ is an orthogonal projector, so $\mathbf{0}_p \preccurlyeq \mathbf{I}_p - \mathbf{P}_0$, imply that

$$\begin{aligned}
&\mathrm{diag}(\{1 - \gamma_1(M_1 + 1/\rho_1)\}\mathbf{I}_{d_1}, \cdots, \{1 - \gamma_n(M_n + 1/\rho_n)\}\mathbf{I}_{d_n}) \\
&\qquad \preccurlyeq \mathbf{I}_p - \mathbf{D}_{U,\ell} - \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\left(\mathbf{I}_p - \mathbf{P}_0\right)\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2} \\
&\qquad\qquad \preccurlyeq \mathrm{diag}\left(\left\{1 - \gamma_1 m_1\right\}\mathbf{I}_{d_1}, \ldots, \left\{1 - \gamma_n m_n\right\}\mathbf{I}_{d_n}\right).
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\|Z_{\ell+1} - \tilde{Z}_{\ell+1}\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} &\leq \max_{i\in[n]}\left\{\max(|1 - \gamma_i m_i|, |1 - \gamma_i(M_i + 1/\rho_i)|)\right\}\|Z_\ell - \tilde{Z}_\ell\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} \\
&= \kappa_{\boldsymbol{\gamma}}\|Z_\ell - \tilde{Z}_\ell\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}}. \qquad (2.23)
\end{aligned}$$

An immediate induction shows, for any $\ell \geq 1$,

$$\|Z_\ell - \tilde{Z}_\ell\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} \leq \kappa_{\boldsymbol{\gamma}}^{\ell-1}\|Z_1 - \tilde{Z}_1\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}}. \qquad (2.24)$$

In addition, by (2.22), we have for any $i \in [n]$,

$$Z_1^i - \tilde{Z}_1^i = \left(1 - \frac{\gamma_i}{\rho_i}\right)(z_i - \tilde{z}_i) + \frac{\gamma_i}{\rho_i}\mathbf{A}_i(\theta - \tilde{\theta}) - \gamma_i(\nabla U_i(z_i) - \nabla U_i(\tilde{z}_i)).$$

It follows that $Z_1 - \tilde{Z}_1 = (\mathbf{I}_p - \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} - \mathbf{D}_{U,0})(z - \tilde{z}) + \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\tilde{\mathbf{D}}_0^{-1/2}\mathbf{B}_0(\theta - \tilde{\theta})$. Using the triangle inequality and Assumption 2.3 gives

$$\begin{aligned}
&\|Z_1 - \tilde{Z}_1\|_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} \leq (\min_{i\in[n]}\{\gamma_i\})^{-1/2}\|(\mathbf{I}_p - \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} - \mathbf{D}_{U,0})(z - \tilde{z}) + (\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0(\theta - \tilde{\theta})\| \\
&\leq (\min_{i\in[n]}\{\gamma_i\})^{-1/2}\left[\|\mathbf{I}_p - \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} - \mathbf{D}_{U,0}\|\|z - \tilde{z}\| + \|\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\|\|\theta - \tilde{\theta}\|\right]
\end{aligned}$$

$$\leq (\min_{i\in[n]}\{\gamma_i\})^{-1/2}\Big[\max_{i\in[n]}\{|1-\gamma_i(m_i+1/\rho_i)|,|1-\gamma_i(M_i+1/\rho_i)|\}\|z-\tilde{z}\| + \|\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\|\|\theta-\tilde{\theta}\|\Big]$$
$$\leq (\min_{i\in[n]}\{\gamma_i\})^{-1/2}\Big[\kappa_\gamma\|z-\tilde{z}\| + \|\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\|\|\theta-\tilde{\theta}\|\Big].$$

Combining (2.24) and the previous inequality and using Lemma 2.11, we get for $\ell \geq 1$,

$$\|X_\ell - \tilde{X}_\ell\|^2 \leq \kappa_{\boldsymbol{\gamma}}^{2(\ell-1)}\left(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2\right)\frac{\max_{i\in[n]}\{\gamma_i\}}{\min_{i\in[n]}\{\gamma_i\}} \times\left[\kappa_\gamma\|z-\tilde{z}\| + \|\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\|\|\theta-\tilde{\theta}\|\right]^2.$$

The proof is concluded by (2.19). ■

Specific case of Proposition 2.4

Based on the previous lemmata, we provide in what follows a specific instance of Proposition 2.4 in the scenario where $\max_{i\in[n]} N_i = 1$.

Proposition 2.13. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n]$, $\gamma_i \leq 2(m_i + M_i + 1/\rho_i)^{-1}$. Then, $P_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ defined in (2.15) admits a unique stationary distribution $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ and for any $\mathbf{x} = (z^\top, \theta^\top)^\top$ with $\theta \in \mathbb{R}^d$, $z \in \mathbb{R}^p$ and any $\ell \in \mathbb{N}^*$, we have*

$$W_2^2(\delta_{\mathbf{x}}P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq \Big(1-\min_{i\in[n]}\{\gamma_i m_i\}\Big)^{2(\ell-1)}\left((1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)\cdot\frac{\max_{i\in[n]}\{\gamma_i\}}{\min_{i\in[n]}\{\gamma_i\}}\right) \times\int_{\mathbb{R}^d\times\mathbb{R}^p}\left[(1-\min_{i\in[n]}\{\gamma_i m_i\})\|z-\tilde{z}\| + \|\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\|\|\theta-\tilde{\theta}\|\right]^2 \mathrm{d}\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\tilde{\mathbf{x}}),$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$, $P_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ are defined in (2.11) and (2.12).

Proof For any $i \in [n]$, note that the condition $0 < \gamma_i \leq 2(m_i + M_i + 1/\rho_i)^{-1}$ ensures that $\kappa_{\boldsymbol{\gamma}} = 1 - \min_{i\in[n]}\{\gamma_i m_i\} \in (0,1)$ and the proof follows from Lemma 2.12 combined with Douc et al. (2018, Lemma 20.3.2, Theorem 20.3.4). ■

2.B.2 Multiple local LMC iterations

In this section, we consider the general case $\max_{i\in[n]} N_i \geq 1$. For this, we introduce the conditional Markov transition kernel defined for any $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, $\boldsymbol{N} = (N_1, \ldots, N_n)$, $\theta \in \mathbb{R}^d$, $z = (z_1, \cdots, z_n) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$, for $i \in [n]$ and $\mathsf{B}_i \in \mathcal{B}(\mathbb{R}^{d_i})$, by

$$Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}\left(z, \mathsf{B}_1 \times \cdots \times \mathsf{B}_n|\theta\right) = \prod_{i=1}^n R^{N_i}_{\rho_i,\gamma_i}(z_i, \mathsf{B}_i|\theta), \tag{2.25}$$

where R_{ρ_i,γ_i} is defined by (2.14). Then, as in the case $\max_{i\in[n]} N_i = 1$, the Gibbs sampler presented in Algorithm 2.1 defines a homogeneous Markov chain $X_\ell^\top = (\theta_\ell^\top, Z_\ell^\top)_{\ell\geq 1}$ where $Z_\ell^\top = ([Z_\ell^1]^\top, \cdots, [Z_\ell^n]^\top)$. Indeed, it is easy to show that for any $\ell \in \mathbb{N}$ and measurable function $f : \mathbb{R}^p \to \mathbb{R}_+$, $\mathbb{E}[f(Z_{\ell+1})|X_\ell] = \int_{\mathbb{R}^p} f(z) Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(Z_\ell, \mathrm{d}z|\theta_\ell)$. Therefore, $(X_\ell)_{\ell\in\mathbb{N}}$ is associated with the Markov kernel defined, for any $\mathbf{x}^\top = (\theta^\top, z^\top)$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$, by

$$P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\mathbf{x}, \mathsf{A}\times\mathsf{B}) = \int_{\mathsf{B}} Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z, \mathrm{d}\tilde{\mathbf{z}}|\theta) \int_{\mathsf{A}} \Pi_{\boldsymbol{\rho}}(\mathrm{d}\tilde{\theta}|\tilde{\mathbf{z}}), \tag{2.26}$$

where $\Pi_{\boldsymbol{\rho}}(\cdot|\tilde{\mathbf{z}})$ is defined in (2.5). We now define a coupling between $\delta_{\mathbf{x}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ and $\delta_{\tilde{\mathbf{x}}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ for any $\ell \geq 1$ and $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^d \times \mathbb{R}^p$. Let $(\xi_\ell)_{\ell\geq 1}$ be a sequence of i.i.d. d-dimensional standard Gaussian random variables independent of the family of independent random variables $\{(\eta_\ell^i)_{\ell\geq 1} : i \in [n]\}$ where for any $i \in [n]$ and $\ell \geq 1$, η_ℓ^i is a d_i-dimensional standard Gaussian random variable. Define by induction the synchronous coupling $(\theta_\ell, Z_\ell)_{\ell\geq 0}, (\tilde{\theta}_\ell, \tilde{Z}_\ell)_{\ell\geq 0}$, for any $i \in [n]$ starting from $(\theta_0, Z_0) = \mathbf{x} = (\theta, z)$, $(\tilde{\theta}_0, \tilde{Z}_0) = \tilde{\mathbf{x}} = (\tilde{\theta}, \tilde{z})$ and for any $\ell \geq 0$ by

$$\begin{aligned} \tilde{Z}^i_{\ell+1} &= \tilde{Y}^{(i,\ell)}_{N_i}, \qquad & \tilde{\theta}_{\ell+1} &= \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} \tilde{Z}_{\ell+1} + \bar{\mathbf{B}}_0^{-1/2}\xi_{\ell+1}, \\ Z^i_{\ell+1} &= Y^{(i,\ell)}_{N_i}, \qquad & \theta_{\ell+1} &= \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} Z_{\ell+1} + \bar{\mathbf{B}}_0^{-1/2}\xi_{\ell+1}, \end{aligned} \tag{2.27}$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ are given by (2.11)-(2.12) and $\tilde{Y}_0^{(i,\ell)} = \tilde{Z}_\ell^i$, $Y_0^{(i,\ell)} = Z_\ell^i$, and for any $k \in \mathbb{N}$

$$\begin{aligned} \tilde{Y}^{(i,\ell)}_{k+1} &= \tilde{Y}^{(i,\ell)}_k - \gamma_i \nabla V_i(\tilde{Y}^{(i,\ell)}_k) + (\gamma_i/\rho_i)\mathbf{A}_i\tilde{\theta}_\ell + \sqrt{2\gamma_i}\eta^{(i,\ell)}_{k+1}, \\ Y^{(i,\ell)}_{k+1} &= Y^{(i,\ell)}_k - \gamma_i \nabla V_i(Y^{(i,\ell)}_k) + (\gamma_i/\rho_i)\mathbf{A}_i\theta_\ell + \sqrt{2\gamma_i}\eta^{(i,\ell)}_{k+1}, \end{aligned} \tag{2.28}$$

where, for any $z_i \in \mathbb{R}^{d_i}$, V_i is defined by

$$V_i(z_i) = U_i(z_i) + (2\rho_i)^{-1}\left\|z_i\right\|^2. \tag{2.29}$$

For any $\ell, k \in \mathbb{N}$ consider the $p \times p$ matrices defined by

$$\begin{aligned} \mathbf{H}^{(\ell)}_{U,k} = \mathrm{diag}\Bigg(&\gamma_1 \int_0^1 \nabla^2 U_1((1-s)Y^{(1,\ell)}_k + s\tilde{Y}^{(1,\ell)}_k)\,\mathrm{d}s, \\ &\ldots, \gamma_n \int_0^1 \nabla^2 U_n((1-s)Y^{(n,\ell)}_k + s\tilde{Y}^{(n,\ell)}_k)\,\mathrm{d}s\Bigg), \end{aligned}$$

$$\mathbf{J}(k) = \mathrm{diag}\left(\mathbb{1}_{[N_1]}(k+1)\cdot\mathbf{I}_{d_1}, \cdots, \mathbb{1}_{[N_n]}(k+1)\cdot\mathbf{I}_{d_n}\right), \tag{2.30}$$

$$\mathbf{C}^{(\ell)}_k = \mathbf{J}(k)(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} + \mathbf{H}^{(\ell)}_{U,k}), \tag{2.31}$$

$$\mathbf{M}^{(\ell)}_{k+1} = (\mathbf{I}_p - \mathbf{C}^{(\ell)}_0)^{-1}\ldots(\mathbf{I}_p - \mathbf{C}^{(\ell)}_k)^{-1}, \qquad \text{with } \mathbf{M}^{(\ell)}_0 = \mathbf{I}_p. \tag{2.32}$$

Under Assumption 2.3, we have $\|\mathbf{C}^{(\ell)}_k\| \leq \max_{i\in[n]}\{\gamma_i(M_i + 1/\rho_i)\}$, thus if we suppose that for any $i \in [n], 0 < \gamma_i < (M_i + 1/\rho_i)^{-1}$, the matrix $(\mathbf{I}_p - \mathbf{C}^{(\ell)}_k)$ is invertible. In addition, for any $\ell \in \mathbb{N}, k \geq \max_{i\in[n]}\{N_i\}, \mathbf{C}^{(\ell)}_k = \mathbf{0}_{p\times p}$, hence the sequence $(\mathbf{M}^{(\ell)}_k)_{k\in\mathbb{N}}$ is stationary, and we denote its limit by $\mathbf{M}^{(\ell)}_\infty$ which is equal to $\mathbf{M}^{(\ell)}_{\max_{i\in[n]}\{N_i\}}$.

Technical lemmata

Similarly to Lemma 2.11, the following result shows that it is enough to consider the marginal process $(Z_\ell, \tilde{Z}_\ell)_{\ell\geq 0}$ to control

$$W_2(\delta_{\mathbf{x}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \delta_{\tilde{\mathbf{x}}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}) \leq \mathbb{E}\left[\|X_\ell - \tilde{X}_\ell\|^2\right]^{1/2}. \tag{2.33}$$

Lemma 2.14. *Assume Assumption 2.1 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$. Then, for any $\ell \in \mathbb{N}$, the random variables $X_\ell = (\theta_\ell^\top, Z_\ell^\top)^\top, \tilde{X}_\ell = (\tilde{\theta}_\ell^\top, \tilde{Z}_\ell^\top)^\top$ defined in (2.27) satisfy*

$$\|\tilde{X}_{\ell+1} - X_{\ell+1}\|^2 \leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2,$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ are defined in (2.11)-(2.12).

Proof The proof is similar to the proof of Lemma 2.11 and is omitted. ■

To ease notation, for any $i \in [n]$, we consider all along this section the quantities

$$\tilde{m}_i = m_i + 1/\rho_i, \qquad \tilde{M}_i = M_i + 1/\rho_i. \tag{2.34}$$

The following lemma provides an explicit expression for $\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|$ with respect to $\|\tilde{Z}_\ell - Z_\ell\|$.

Lemma 2.15. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n], \gamma_i < \tilde{M}_i^{-1}$. Then, for any $\ell \geq 1$, we have*

$$\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \leq \left\| [\mathbf{M}_\infty^{(\ell)}]^{-1} + \sum_{k=0}^{\infty} [\mathbf{M}_\infty^{(\ell)}]^{-1} \mathbf{M}_{k+1}^{(\ell)} \mathbf{J}(k) \mathbf{D}_{\boldsymbol{N}}^{-1/2} \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2} \mathbf{P}_0 \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2} \mathbf{D}_{\boldsymbol{N}}^{1/2} \right\| \times \|\tilde{Z}_\ell - Z_\ell\|_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}}, \tag{2.35}$$

where $(\mathbf{M}_k^{(\ell)})_{k\in\mathbb{N}}$ is defined in (2.32), $(\tilde{Z}_k, Z_k)_{k\in\mathbb{N}}$ in (2.27), $\boldsymbol{N}\boldsymbol{\gamma} = (\gamma_1 N_1, \ldots, \gamma_n N_n)$ and $\boldsymbol{\gamma}/\boldsymbol{\rho} = (\gamma_1/\rho_1, \ldots, \gamma_n/\rho_n)$.

Proof Let $\ell \geq 1$. By (2.28), for any $i \in [n], k \in \mathbb{N}$, we obtain

$$\tilde{Y}_{k+1}^{(i,\ell)} - Y_{k+1}^{(i,\ell)} = \left(\mathbf{I}_{d_i} - \gamma_i \int_0^1 \nabla^2 V_i((1-s)Y_k^{(i,\ell)} + s\tilde{Y}_k^{(i,\ell)})\,\mathrm{d}s\right)(\tilde{Y}_k^{(i,\ell)} - Y_k^{(i,\ell)}) + (\gamma_i/\rho_i)\mathbf{A}_i(\tilde{\theta}_\ell - \theta_\ell).$$

Consider the process $((\tilde{\mathrm{Y}}_k^{(\ell)}, \mathrm{Y}_k^{(\ell)}) = \{\tilde{\mathrm{Y}}_k^{(i,\ell)}, \mathrm{Y}_k^{(i,\ell)}\}_{i=1}^n)_{k\in\mathbb{N}}$ with values in $\mathbb{R}^p \times \mathbb{R}^p$ defined for any $i \in [n]$, $k \geq 0$, by

$$\tilde{\mathrm{Y}}_k^{(i,\ell)} = \tilde{Y}_{\min(k,N_i)}^{(i,\ell)}, \qquad \mathrm{Y}_k^{(i,\ell)} = Y_{\min(k,N_i)}^{(i,\ell)}. \tag{2.36}$$

By (2.27), we have $\mathbf{A}_i(\tilde{\theta}_\ell - \theta_\ell) = \mathbf{A}_i\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}(\tilde{Z}_\ell - Z_\ell)$. Since

$$\mathbf{B}_0^\top = [\mathbf{A}_1^\top/\rho_1^{1/2} \cdots \mathbf{A}_n^\top/\rho_n^{1/2}]$$

and $\mathbf{P}_0 = \mathbf{B}_0\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top$ is the orthogonal projection matrix defined in (2.12), it follows that

$$\tilde{\mathrm{Y}}_{k+1}^{(\ell)} - \mathrm{Y}_{k+1}^{(\ell)} = \left(\mathbf{I}_p - \mathbf{C}_k^{(\ell)}\right)(\tilde{\mathrm{Y}}_k^{(\ell)} - \mathrm{Y}_k^{(\ell)}) + \mathbf{J}(k)\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{P}_0\tilde{\mathbf{D}}_0^{1/2}(\tilde{\mathrm{Y}}_0^{(\ell)} - \mathrm{Y}_0^{(\ell)}). \quad (2.37)$$

Since $\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}$ commutes with $\mathbf{C}_k^{(\ell)}$ and $\mathbf{J}(k)$, multiplying (2.37) by $\mathbf{M}_{k+1}^{(\ell)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}$, yields

$$\begin{aligned}\mathbf{M}_{k+1}^{(\ell)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_{k+1}^{(\ell)} - \mathrm{Y}_{k+1}^{(\ell)}) &= \mathbf{M}_k^{(\ell)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_k^{(\ell)} - \mathrm{Y}_k^{(\ell)}) \\ &\quad + \mathbf{M}_{k+1}^{(\ell)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\tilde{\mathbf{D}}_0^{1/2}(\tilde{\mathrm{Y}}_0^{(\ell)} - \mathrm{Y}_0^{(\ell)}). \quad (2.38)\end{aligned}$$

By definition of the processes in (2.27)-(2.28) and (2.36), we have for $k \geq \max_{i\in[n]}\{N_i\}$, $(\tilde{\mathrm{Y}}_k^{(\ell)}, \mathrm{Y}_k^{(\ell)}) = (\tilde{Z}_{\ell+1}, Z_{\ell+1})$ and $\mathbf{J}(k) = \mathbf{0}_{p\times p}$. Therefore, summing the previous equality (2.38) yields

$$\begin{aligned}\mathbf{M}_\infty^{(\ell)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{Z}_{\ell+1} - Z_{\ell+1}) = \Big[\mathbf{M}_0^{(\ell)} + \textstyle\sum_{k=0}^\infty \mathbf{M}_{k+1}^{(\ell)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\Big] \\ \times \mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_0^{(\ell)} - \mathrm{Y}_0^{(\ell)}).\end{aligned}$$

Multiplying this last equality by $[\mathbf{M}_\infty^{(\ell)}]^{-1}$ and applying the norm $\|\cdot\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}$ concludes the proof. ■

The three following lemmata aim at providing an explicit upper bound on (2.35). To this end, for $\ell, k \in \mathbb{N}$ and $i \in [n]$, consider $\mathbf{C}_k^{(i,\ell)}$ corresponding to the i-th diagonal block of $\mathbf{C}_k^{(\ell)}$ defined in (2.31), *i.e.*

$$\mathbf{C}_k^{(i,\ell)} = \mathbb{1}_{[N_i]}(k+1)\gamma_i\left\{\rho_i^{-1}\mathbf{I}_{d_i} + \int_0^1 \nabla^2 U_i((1-s)Y_k^{(i,\ell)} + s\tilde{Y}_k^{(i,\ell)})\,\mathrm{d}s\right\} \in \mathbb{R}^{d_i\times d_i}, \quad (2.39)$$

where, for any $\ell \in \mathbb{N}$ and $i \in [n]$, $(Y_k^{(i,\ell)}, \tilde{Y}_k^{(i,\ell)})_{k\in\mathbb{N}}$ is defined in (2.28). Thus, using the definition (2.32) of $\mathbf{M}_k^{(\ell)}$, we can write $[\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)}$ as a block-diagonal matrix $\mathrm{diag}(([\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)})^1, \ldots, ([\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)})^n)$ where for any $i \in [n]$, $([\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)})^i = \prod_{l=k}^{N_i-1}(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)}) \in \mathbb{R}^{d_i\times d_i}$.

Lemma 2.16. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{R}_+^*)^n$, $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n]$, $\gamma_i < \tilde{M}_i^{-1}$. Then, for any $i \in [n]$, $\ell \in \mathbb{N}$ and $k \in [N_i]$, we have*

$$\|([\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)})^i - \mathbf{I}_{d_i} - \textstyle\sum_{l=k}^\infty \mathbf{C}_l^{(i,\ell)}\| \leq \exp\{(N_i-k)\gamma_i\tilde{M}_i\} - 1 - (N_i-k)\gamma_i\tilde{M}_i,$$

where $\mathbf{M}_k^{(\ell)}, \tilde{M}_i$ are defined in (2.32), (2.34) respectively, and $\mathbf{M}_\infty^{(\ell)}$ is the limit of the stationnary sequence $(\mathbf{M}_k^{(\ell)})_{k\in\mathbb{N}}$.

Proof Let $\ell \in \mathbb{N}$, $i \in [n]$ and $k \in [N_i]$. The approximation error between $\prod_{l=k}^\infty(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)})$ and its linear approximation can be upper bounded as

$$\left\|\prod_{l=k}^\infty(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)}) - \mathbf{I}_{d_i} - \sum_{l=k}^\infty \mathbf{C}_l^{(i,\ell)}\right\| = \left\|\sum_{m=2}^\infty (-1)^m \sum_{k\leq l_1<\cdots<l_m} \mathbf{C}_{l_1}^{(i,\ell)}\cdots\mathbf{C}_{l_m}^{(i,\ell)}\right\|$$

$$\leq \sum_{m=2}^{\infty} \sum_{k \leq l_1 < \cdots < l_m} \|\mathbf{C}_{l_1}^{(i,\ell)}\| \cdots \|\mathbf{C}_{l_m}^{(i,\ell)}\| = \prod_{l=k}^{\infty} (1 + \|\mathbf{C}_l^{(i,\ell)}\|) - 1 - \sum_{l \geq k} \|\mathbf{C}_l^{(i,\ell)}\|$$
$$\leq \exp\left(\sum_{l=k}^{\infty} \|\mathbf{C}_l^{(i,\ell)}\| \right) - 1 - \sum_{l=k}^{\infty} \|\mathbf{C}_l^{(i,\ell)}\|,$$

where the products and the sums are well-defined since for any $l \geq N_i$, we have $\mathbf{C}_l^{(i,\ell)} = \mathbf{0}_{d_i}$. Finally, the proof is concluded using that $x \mapsto \exp(x) - 1 - x$ is increasing on $\mathbb{R}$ and for $l \in \mathbb{N}$, $\|\mathbf{C}_l^{(i,\ell)}\| \leq \gamma_i \tilde{M}_i \mathbb{1}_{[N_i]}(l+1)$ from Assumption 2.3-(i). ■

For any $\boldsymbol{N} = (N_1, \ldots, N_n) \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n) \in (\mathbb{R}_+^*)^n$, define the $p \times p$ block matrices

$$\mathbf{S}_1 = \mathrm{diag}(\{1 - N_1\gamma_1\tilde{M}_1\}\mathbf{I}_{d_1}, \cdots, \{1 - N_n\gamma_n\tilde{M}_n\}\mathbf{I}_{d_n}),$$
$$\mathbf{S}_2 = \mathbf{I}_p - \sum_{l=0}^{\infty} \mathbf{J}(l)\mathbf{H}_{U,l}^{(\ell)} - (\mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2}(\mathbf{I}_p - \mathbf{P}_0)(\mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2}, \tag{2.40}$$
$$\mathbf{S}_3 = \mathrm{diag}\left(\left\{1 - N_1\gamma_1 m_1\right\}\mathbf{I}_{d_1}, \ldots, \left\{1 - N_n\gamma_n m_n\right\}\mathbf{I}_{d_n} \right),$$

where for any $i \in [n]$, $\tilde{M}_i$ is defined in (2.34) and $\mathbf{P}_0, \mathbf{J}(l), \mathbf{H}_{U,l}^{(\ell)}$ are defined in (2.12), (2.87), (2.88), respectively.

Lemma 2.17. *Assume Assumption 2.1-Assumption 2.3. Then, for any $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$, we have*

$$\mathbf{S}_1 \preccurlyeq \mathbf{S}_2 \preccurlyeq \mathbf{S}_3.$$

As a result, under the additional assumption, for any $i \in [n]$, $\gamma_i N_i \leq 2/(m_i + M_i + 1/\rho_i)$, we get

$$\left\|\mathbf{S}_2\right\| \leq 1 - \min_{i \in [n]}\{N_i\gamma_i m_i\}. \tag{2.41}$$

Proof Since $\mathbf{P}_0$ is an orthogonal projection defined in (2.12), we have $\mathbf{P}_0 \preccurlyeq \mathbf{I}_p$, therefore we easily get

$$\mathbf{0}_{p\times p} \preccurlyeq (\mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2}(\mathbf{I}_p - \mathbf{P}_0)(\mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2} \preccurlyeq \mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}$$

and Assumption 2.3-(i)-(ii) imply

$$\mathrm{diag}(N_1\gamma_1 m_1\mathbf{I}_{d_1}, \cdots, N_n\gamma_n m_n\mathbf{I}_{d_n}) \preccurlyeq \sum_{l=0}^{\infty} \mathbf{J}(l)\mathbf{H}_{U,l}^{(\ell)} \preccurlyeq \mathrm{diag}(N_1\gamma_1 M_1\mathbf{I}_{d_1}, \cdots, N_n\gamma_n M_n\mathbf{I}_{d_n}).$$

Substracting these previous inequalities and adding $\mathbf{I}_p$ complete the first part of the proof. The additional condition, for any $i \in [n]$, $\gamma_i N_i \leq 2/(m_i + M_i + 1/\rho_i)$, ensures that $\mathbf{S}_1$ is definite-positive. Since $\mathbf{S}_1 \preceq \mathbf{S}_2$, we deduce that $\mathbf{S}_2$ is symmetric positive-definite as well. Then, $\left\|\mathbf{S}_2\right\|$ is equal to the largest eigenvalue of $\mathbf{S}_2$. The inequality $\mathbf{S}_2 \preceq \mathbf{S}_3$ concludes the second part of the proof. ■

For any $\boldsymbol{N} = (N_1, \ldots, N_n) \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n) \in (\mathbb{R}_+^*)^n$, define

$$r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} = \max_{i\in[n]}\{N_i\gamma_i/\rho_i\} \max_{i\in[n]}\{N_i\gamma_i\tilde{M}_i\}\left(1/2 + \max_{i\in[n]}\{N_i\gamma_i\tilde{M}_i\}\right)$$
$$+ 4\max_{i\in[n]}\{N_i\gamma_i\tilde{M}_i\}^2, \tag{2.42}$$

where $\tilde{M}_i$ is defined in (2.34).

Lemma 2.18. *Assume Assumption 2.1-Assumption 2.3. Let* $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ *such that, for any* $i \in [n]$, $N_i\gamma_i \le 2/(m_i + \tilde{M}_i)$ *and* $\gamma_i < \tilde{M}_i^{-1}$. *Then, for any* $\ell \in \mathbb{N}$, *we have*

$$\|[\mathbf{M}_\infty^{(\ell)}]^{-1} + \textstyle\sum_{k=0}^\infty [\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_{k+1}^{(\ell)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\| \le 1 - \min_{i\in[n]}\{N_i\gamma_i m_i\} + r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}},$$

where $\mathbf{P}_0$, $\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}$, $\mathbf{J}(k), \mathbf{M}_k^{(\ell)}$ *and* $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ *are defined in* (2.12), (2.30), (2.32) *and* (2.42), *respectively.*

Proof Let $\ell \in \mathbb{N}$. For any $k \in \mathbb{N}$, define

$$\begin{aligned}\mathbf{R}_k^{(\ell)} &= \prod_{l=k}^\infty (\mathbf{I}_p - \mathbf{C}_l^{(\ell)}) - \mathbf{I}_p + \sum_{l=k}^\infty \mathbf{C}_l^{(\ell)},\\ \mathbf{R}_k^{(i,\ell)} &= \prod_{l=k}^\infty (\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)}) - \mathbf{I}_{d_i} + \sum_{l=k}^\infty \mathbf{C}_l^{(i,\ell)}, \quad i \in [n],\end{aligned} \tag{2.43}$$

where $(\mathbf{C}_l^{(i,\ell)})_{l\in\mathbb{N}}$ is defined in (2.39) and remark that the products and the sums are well defined since for any $l \ge N_i$, we have $\mathbf{C}_l^{(i,\ell)} = \mathbf{0}_{d_i}$. By noting, for any $k \in [\max_{i\in[n]} N_i]$, that $[\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)} = \prod_{l=k}^\infty (\mathbf{I}_p - \mathbf{C}_l^{(\ell)})$, it follows that $[\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_k^{(\ell)} = \mathbf{I}_p - \sum_{l=k}^\infty \mathbf{C}_l^{(\ell)} + \mathbf{R}_k^{(\ell)}$. Since for any $i \in [n], l \ge N_i$, $\mathbf{R}_k^{(i,\ell)} = \mathbf{0}_{d_i}$, thus we have $\mathbf{J}(k)\mathbf{R}_{k+1}^{(\ell)} = \mathbf{R}_{k+1}^{(\ell)}$. In addition, using that $\mathbf{M}_0^{(\ell)} = \mathbf{I}_p$, $\mathbf{C}_l^{(\ell)} = \mathbf{J}(l)(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} + \mathbf{H}_{U,l}^{(\ell)})$, $\mathbf{D}_{\boldsymbol{N}} = \sum_{k=0}^\infty \mathbf{J}(k)$, $\mathbf{D}_{\boldsymbol{N}}\mathbf{C}_l^{(\ell)} = \mathbf{C}_l^{(\ell)}\mathbf{D}_{\boldsymbol{N}}$, we get

$$\begin{aligned}
&[\mathbf{M}_\infty^{(\ell)}]^{-1} + \sum_{k=0}^\infty [\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_{k+1}^{(\ell)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&= \mathbf{I}_p - \sum_{l=0}^\infty \mathbf{C}_l^{(\ell)} + \sum_{k=0}^\infty \mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&- \sum_{k=0}^\infty \sum_{l=k+1}^\infty \mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{C}_l^{(\ell)}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&+ \mathbf{R}_0^{(\ell)} + \sum_{k=0}^\infty \mathbf{R}_{k+1}^{(\ell)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&= \mathbf{I}_p - \sum_{l=0}^\infty \mathbf{J}(l)\mathbf{H}_{U,l}^{(\ell)} - \Big(\sum_{k=0}^\infty \mathbf{J}(k)\Big)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}(\mathbf{I}_p - \mathbf{P}_0)\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&- \sum_{l=1}^\infty \Big(\sum_{k=0}^{l-1} \mathbf{J}(k)\Big)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{C}_l^{(\ell)}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&+ \mathbf{R}_0^{(\ell)} + \sum_{k=0}^\infty \mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{R}_{k+1}^{(\ell)}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\\
&= \mathbf{S}_2 - \sum_{l=1}^\infty \Big(\sum_{k=0}^{l-1} \mathbf{J}(k)\Big)\mathbf{D}_{\boldsymbol{N}}^{-1}\mathbf{C}_l^{(\ell)}(\mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2}\mathbf{P}_0(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{D}_{\boldsymbol{N}})^{1/2}\\
&+ \mathbf{R}_0^{(\ell)} + \sum_{k=1}^\infty \mathbf{D}_{\boldsymbol{N}}^{-1}\mathbf{R}_k^{(\ell)}(\mathbf{D}_{\boldsymbol{N}}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2}\mathbf{P}_0(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{D}_{\boldsymbol{N}})^{1/2},
\end{aligned} \tag{2.44}$$

where $\mathbf{S}_2$ is defined in (2.40). We now bound the different terms of (2.44) separately. First, using (2.41), we have

$$\left\|\mathbf{S}_2\right\| \leq 1 - \min_{i\in[n]}\{N_i\gamma_i m_i\}.$$

By recalling $\mathbf{R}_0^{(\ell)}$ defined in (2.43), Lemma 2.16 shows that

$$\begin{aligned}
\|\mathbf{R}_0^{(\ell)}\| \leq \max_{i\in[n]} \|\mathbf{R}_0^{(i,\ell)}\| &= \max_{i\in[n]} \left\{ \left\| \prod_{l=0}^{\infty} \left(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)}\right) - \mathbf{I}_{d_i} - \sum_{l=0}^{\infty} \mathbf{C}_l^{(i,\ell)} \right\| \right\} && (2.45)\\
&\leq \max_{i\in[n]} \left\{ \exp\left(\sum_{l=0}^{\infty} \|\mathbf{C}_l^{(i,\ell)}\| \right) - 1 - \sum_{l=0}^{\infty} \|\mathbf{C}_l^{(i,\ell)}\| \right\} && (2.46)\\
&\leq \max_{i\in[n]} \left\{ \exp\{(N_i-1)\gamma_i\tilde{M}_i\} - 1 - (N_i-1)\gamma_i\tilde{M}_i \right\} && (2.47)\\
&\leq \max_{i\in[n]} \{((N_i-1)\gamma_i\tilde{M}_i)^2 \mathrm{e}^{(N_i-1)\gamma_i\tilde{M}_i}\}/2 && (2.48)\\
&\leq 4 \max_{i\in[n]} \{(N_i-1)\gamma_i\tilde{M}_i\}^2, && (2.49)
\end{aligned}$$

where, in the penultimate line, we used for any $t \geq 0$, that $\exp(t) - 1 - t \leq t^2 \exp(t)/2$. Regarding the second term of (2.44), using that $\mathbf{P}_0$ is an orthogonal projector, we get

$$\begin{aligned}
&\left\| \sum_{l=1}^{\infty} \left(\sum_{k=0}^{l-1} \mathbf{J}(k) \right) \mathbf{D}_{\boldsymbol{N}}^{-1} \mathbf{C}_l^{(\ell)} (\mathbf{D}_{\boldsymbol{N}} \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2} \mathbf{P}_0 (\mathbf{D}_{\boldsymbol{N}} \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2} \right\| \\
&\qquad \leq \max_{i\in[n]} \left(\frac{N_i\gamma_i}{\rho_i} \right) \left\| \sum_{l=1}^{\infty} \left(\sum_{k=0}^{l-1} \mathbf{J}(k) \right) \mathbf{D}_{\boldsymbol{N}}^{-1} \mathbf{C}_l^{(\ell)} \right\|.
\end{aligned}$$

Combining the following upper bound

$$\left\| \sum_{l=1}^{\infty} \left(\sum_{k=0}^{l-1} \mathbf{J}(k) \right) \mathbf{D}_{\boldsymbol{N}}^{-1} \mathbf{C}_l^{(\ell)} \right\| \leq \max_{i\in[n]} \left\{ \frac{1}{N_i} \sum_{l=1}^{\infty} l \|\mathbf{C}_l^{(i,\ell)}\| \right\}$$

with the fact, for any $i \in [n]$, that $\|\mathbf{C}_l^{(i,\ell)}\| \leq \gamma_i\tilde{M}_i \mathbb{1}_{[N_i]}(l+1)$, we get that

$$\begin{aligned}
&\left\| \sum_{l=1}^{\infty} \left(\sum_{k=0}^{l-1} \mathbf{J}(k) \right) \mathbf{D}_{\boldsymbol{N}}^{-1} \mathbf{C}_l^{(\ell)} (\mathbf{D}_{\boldsymbol{N}} \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2} \mathbf{P}_0 (\mathbf{D}_{\boldsymbol{N}} \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}})^{1/2} \right\| \\
&\qquad \leq \max_{i\in[n]} \left(\frac{N_i\gamma_i}{\rho_i} \right) \max_{i\in[n]} \left\{ \frac{N_i\gamma_i\tilde{M}_i}{2} \right\}. \qquad (2.50)
\end{aligned}$$

To upper bound the last term of (2.44), we start from the following inequality

$$\left\| \sum_{k=1}^{\infty} \mathbf{D}_{\boldsymbol{N}}^{-1} \mathbf{R}_k^{(\ell)} \right\| \leq \max_{i\in[n]} \left\{ \frac{1}{N_i} \sum_{k=1}^{N_i-1} \|\mathbf{R}_k^{(i,\ell)}\| \right\}.$$

Lemma 2.16 shows that for any $k \in [N_i - 1]$ and $i \in [n]$, $\|\mathbf{R}_k^{(i,\ell)}\| \leq \exp\{(N_i-k)\gamma_i\tilde{M}_i\} - 1 - (N_i-k)\gamma_i\tilde{M}_i$. Then, for any $i \in [n]$, we have

$$\frac{1}{N_i}\sum_{k=1}^{N_i-1}\|\mathbf{R}_k^{(i,\ell)}\| \le \frac{1}{N_i}\sum_{k=1}^{N_i-1}[\exp\{(N_i-k)\gamma_i\tilde{M}_i\}-1-(N_i-k)\gamma_i\tilde{M}_i]$$
$$\le (N_i\gamma_i\tilde{M}_i)^{-1}\int_0^{N_i\gamma_i\tilde{M}_i}(\mathrm{e}^t-1-t)\,\mathrm{d}t \le \frac{(N_i\gamma_i\tilde{M}_i)^2}{12}\left(\mathrm{e}^{N_i\gamma_i\tilde{M}_i}+1\right)$$
$$\le \max_{i\in[n]}\{(N_i\gamma_i\tilde{M}_i)^2\}, \quad (2.51)$$

where we have used $\mathrm{e}^2+1\le 12$. Plugging (2.51), (2.50), (2.49) into (2.41), we get

$$\left\|[\mathbf{M}_\infty^{(\ell)}]^{-1}+\sum_{k\in\mathbb{N}}[\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_{k+1}^{(\ell)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{P}_0\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{D}_{\boldsymbol{N}}^{1/2}\right\|$$
$$\le 1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}},$$

where $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ is defined in (2.42). ■

Lemma 2.19. *Assume Assumption 2.1-Assumption 2.3. Let $\boldsymbol{N}\in(\mathbb{N}^\star)^n, \boldsymbol{\gamma}\in(\mathbb{R}_+^*)^n$ such that, for any $i\in[n]$, $N_i\gamma_i\le 2/(m_i+\tilde{M}_i)$ and $\gamma_i<\tilde{M}_i^{-1}$. Then, for any $\mathbf{x}=(z^\top,\theta^\top)^\top, \tilde{\mathbf{x}}=(\tilde{z}^\top,\tilde{\theta}^\top)^\top\in\mathbb{R}^{p+d}$, with $(\theta,\tilde{\theta})\in(\mathbb{R}^d)^2, (z,\tilde{z})\in(\mathbb{R}^p)^2$ and any $\ell\ge 1$ we have*

$$W_2^2(\delta_{\tilde{\mathbf{x}}}P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}^\ell,\delta_{\mathbf{x}}P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}^\ell)\le(1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^{2\ell-2}(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)$$
$$\times\frac{\max_{i\in[n]}\{N_i\gamma_i\}}{\min_{i\in[n]}\{N_i\gamma_i\}}\left[\|[\mathbf{M}_\infty^{(0)}]^{-1}\|\|\tilde{z}-z\|+(\textstyle\sum_{i\in[n]}\|\mathbf{A}_i\|/\rho_i)\|\tilde{\theta}-\theta\|\right]^2,$$

where $\mathbf{B}_0,\bar{\mathbf{B}}_0,\tilde{\mathbf{D}}_0,P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}},\mathbf{M}_\infty^{(0)},r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ are defined in (2.11), (2.12), (2.26), (2.32), (2.42), respectively.

Proof Combining Lemma 2.15 and Lemma 2.18, we have for $\ell\ge 1$,

$$\|\tilde{Z}_{\ell+1}-Z_{\ell+1}\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}\le(1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})\|\tilde{Z}_\ell-Z_\ell\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}.$$

Thereby, for any $\ell\ge 1$, we obtain by induction

$$\|\tilde{Z}_\ell-Z_\ell\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}\le(1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^{\ell-1}\|\tilde{Z}_1-Z_1\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}. \quad (2.52)$$

Define the process $((\tilde{\mathrm{Y}}_k^{(0)},\mathrm{Y}_k^{(0)})=\{\tilde{\mathrm{Y}}_k^{(i,0)},\mathrm{Y}_k^{(i,0)}\}_{i=1}^n)_{k\in\mathbb{N}}$ with values in $\mathbb{R}^p\times\mathbb{R}^p$ defined for any $i\in[n]$, $k\ge 0$ by

$$\tilde{\mathrm{Y}}_k^{(i,0)}=\tilde{Y}_{\min(k,N_i)}^{(i,0)}, \qquad \mathrm{Y}_k^{(i,0)}=Y_{\min(k,N_i)}^{(i,0)}.$$

By (2.27), it follows that for any $i\in[n]$, $(\tilde{Z}_1^i,Z_1^i)=(\tilde{Y}_{N_i}^{(i,0)},Y_{N_i}^{(i,0)})$ where $(\tilde{Y}_0^{(i,0)},Y_0^{(i,0)})=(\tilde{Z}_0^i,Z_0^i)$. We get by (2.28) for $k\ge 0$,

$$\tilde{\mathrm{Y}}_{k+1}^{(0)}-\mathrm{Y}_{k+1}^{(0)}=(\mathbf{I}_p-\mathbf{C}_k^{(0)})(\tilde{\mathrm{Y}}_k^{(0)}-\mathrm{Y}_k^{(0)})+\mathbf{J}(k)\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0(\tilde{\theta}_0-\theta_0).$$

Hence, for $k\ge 0$, we obtain

$$\mathbf{M}_{k+1}^{(0)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_{k+1}^{(0)}-\mathrm{Y}_{k+1}^{(0)}) \\ =\mathbf{M}_{k}^{(0)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_{k}^{(0)}-\mathrm{Y}_{k}^{(0)})+\mathbf{M}_{k+1}^{(0)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{B}_0(\tilde{\theta}_0-\theta_0).$$

Summing the previous equality gives

$$\mathbf{M}_{\infty}^{(0)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_{\boldsymbol{N}}^{(0)}-\mathrm{Y}_{\boldsymbol{N}}^{(0)}) \\ =\mathbf{M}_{0}^{(0)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{\mathrm{Y}}_{0}^{(0)}-\mathrm{Y}_{0}^{(0)})+\sum_{k=0}^{\infty}\mathbf{M}_{k+1}^{(0)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{B}_0(\tilde{\theta}_0-\theta_0).$$

Multiplying by $[\mathbf{M}_{\infty}^{(0)}]^{-1}$ and using the fact that $(\theta_0,\mathrm{Y}_0^{(0)})=(\theta,z)$, $(\tilde{\theta}_0,\tilde{\mathrm{Y}}_0^{(0)})=(\tilde{\theta},\tilde{z})$, we get

$$\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{Z}_1-Z_1)=[\mathbf{M}_{\infty}^{(0)}]^{-1}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}(\tilde{z}-z)+\sum_{k=0}^{\infty}[\mathbf{M}_{\infty}^{(0)}]^{-1}\mathbf{M}_{k+1}^{(0)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{B}_0(\tilde{\theta}-\theta).$$

Plugging the result in (2.52) implies for any $\ell\geq 1$,

$$\|\tilde{Z}_\ell-Z_\ell\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}\leq(1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^{\ell-1} \\ \times\left[\|[\mathbf{M}_{\infty}^{(0)}]^{-1}\|\|\tilde{z}-z\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}+\left\|\sum_{k=0}^{\infty}[\mathbf{M}_{\infty}^{(0)}]^{-1}\mathbf{M}_{k+1}^{(0)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{B}_0\right\|\|\tilde{\theta}-\theta\|\right]. \tag{2.53}$$

By Assumption 2.3-(ii) and the definitions of $\mathbf{C}_l^{(0)},\mathbf{M}_k^{(0)}$ given in (2.31), (2.32), we have $\|\mathbf{I}_{d_i}-\mathbf{C}_l^{(i,0)}\|\leq 1-\gamma_i\tilde{m}_i$. As a result and since $([\mathbf{M}_{\infty}^{(0)}]^{-1}\mathbf{M}_k^{(0)})^i=\prod_{l=0}^{k-1}(\mathbf{I}_{d_i}-\mathbf{C}_l^{(i,0)})$, the triangle inequality implies

$$\begin{aligned}\left\|\sum_{k=0}^{\infty}[\mathbf{M}_{\infty}^{(0)}]^{-1}\mathbf{M}_{k+1}^{(0)}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2}\mathbf{B}_0\right\|&\leq\sum_{i\in[n]}\sqrt{\gamma_i/N_i}(\|\mathbf{A}_i\|/\rho_i)\sum_{k=1}^{N_i}\|([\mathbf{M}_{\infty}^{(0)}]^{-1}\mathbf{M}_k^{(0)})^i\| \\ &\leq\sum_{i\in[n]}\sqrt{\gamma_i/N_i}(\|\mathbf{A}_i\|/\rho_i)\sum_{k=0}^{N_i-1}(1-\gamma_i\tilde{m}_i)^k \\ &\leq\sum_{i\in[n]}\|\mathbf{A}_i\|\sqrt{N_i\gamma_i}/\rho_i.\end{aligned}$$

Plugging this result in (2.53), we get

$$\|\tilde{Z}_\ell-Z_\ell\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}\leq\Big(1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}\Big)^{\ell-1} \\ \times\left[\|[\mathbf{M}_{\infty}^{(0)}]^{-1}\|\|\tilde{z}-z\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}+\Big(\sum_{i\in[n]}\|\mathbf{A}_i\|\sqrt{N_i\gamma_i}/\rho_i\Big)\|\tilde{\theta}-\theta\|\right].$$

Finally, Lemma 2.14 gives

$$\|\tilde{X}_\ell-X_\ell\|^2\leq(1-\min_{i\in[n]}\{N_i\gamma_i m_i\}+r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^{2\ell-2}(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)$$

$$\times \frac{\max_{i\in[n]}\{N_i\gamma_i\}}{\min_{i\in[n]}\{N_i\gamma_i\}}\left[\|[\mathbf{M}_\infty^{(0)}]^{-1}\|\|\tilde{z}-z\| + \left(\sum_{i\in[n]}\|\mathbf{A}_i\|/\rho_i\right)\|\tilde{\theta}-\theta\|\right]^2.$$

Plugging this result into (2.33) concludes the proof. ■

The following result gives a condition on $\max_{i\in[n]}\{N_i\gamma_i\}$ to simplify the contrating term in Lemma 2.19 to $1-\min_{i\in[n]}\{N_i\gamma_i m_i\}/2$. To this end, define

$$A_0 = \max_{i\in[n]}\{\tilde{M}_i\}\max_{i\in[n]}\{1/\rho_i\}/2 + 4\max_{i\in[n]}\{\tilde{M}_i\}^2,$$
$$A_1 = \max_{i\in[n]}\{\tilde{M}_i\}^2\max_{i\in[n]}\{1/\rho_i\}.$$

Lemma 2.20. *Assume Assumption 2.1-Assumption 2.3 and let $c \in \mathbb{R}_+^*$, $\boldsymbol{N} \in (\mathbb{N}^\star)^n$, $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that*

$$\min_{i\in[n]}\{N_i\gamma_i\}/\max_{i\in[n]}\{N_i\gamma_i\} \geq c,$$
$$\max_{i\in[n]}\{N_i\gamma_i\} \leq \frac{c\min_{i\in[n]}\{m_i\}}{2A_0+\sqrt{2A_1 c\min_{i\in[n]}\{m_i\}}} \wedge \frac{2}{\max_{i\in[n]}\{m_i+M_i+1/\rho_i\}}. \tag{2.54}$$

Then, $1-\min_{i\in[n]}\{N_i\gamma_i m_i\} + r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} < 1-\min_{i\in[n]}\{N_i\gamma_i m_i\}/2 < 1$, where $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ is defined in (2.42).

Proof The proof is straightforward solving a second order polynomial inequality and using for any $a, b \in \mathbb{R}_+^*$, $a + \frac{b^2}{2a+b} \leq \sqrt{a^2+b^2}$. ■

Proof of Proposition 2.4

The next proposition quantifies the convergence of $(\delta_{\mathbf{x}}P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}})_{\ell\in\mathbb{N}}$ to a stationnary distribution $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Further, in the case $N_1 = \ldots = N_n$ we show the stationnary distribution $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ is equal to $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ derived in Proposition 2.13.

Proposition 2.21. *Assume Assumption 2.1-Assumption 2.3 and let $c > 0$ and $\boldsymbol{\gamma} = \{\gamma_i\}_{i=1}^n$, $\boldsymbol{N} \in (\mathbb{N}^\star)^n$ such that (2.54) is satisfied, for any $i \in [n]$, $N_i\gamma_i < 2/\max_{i\in[n]}\{m_i+\tilde{M}_i\}$ and $\gamma_i < \tilde{M}_i^{-1}$. Then, $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ defined in (2.26) admits a unique invariant probability measure $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$. In addition, for any $\mathbf{x} = (z^\top, \theta^\top)^\top$ with $(\theta, z) \in \mathbb{R}^d \times \mathbb{R}^p$, any integer $\ell \geq 1$, we have*

$$W_2^2(\delta_{\mathbf{x}}P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq (1-\min_{i\in[n]}\{N_i\gamma_i m_i\}/2)^{2\ell-2}\cdot(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)\frac{\max_{i\in[n]}\{N_i\gamma_i\}}{\min_{i\in[n]}\{N_i\gamma_i\}}$$
$$\times \int_{\mathbb{R}^d\times\mathbb{R}^p}\left[\|[\mathbf{M}_\infty^{(0)}]^{-1}\|\|\tilde{z}-z\| + \left(\sum_{i\in[n]}\|\mathbf{A}_i\|/\rho_i\right)\|\tilde{\theta}-\theta\|\right]^2 \mathrm{d}\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\tilde{\mathbf{x}}),$$

where $\mathbf{B}_0, \bar{\mathbf{B}}_0, \mathbf{M}_\infty^{(0)}$ are defined in (2.11), (2.32), respectively. Finally, if $\boldsymbol{N} = N(1,\ldots,1) = N\mathbf{1}_n$ for $N \geq 1$, then $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}} = \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\mathbf{1}_n}$.

Proof Note that under the conditions on $\boldsymbol{\gamma}$ and $\boldsymbol{N}$ stated in Proposition 2.21, Lemma 2.20 ensures that $1-\min_{i\in[n]}\{N_i\gamma_i m_i\}/2 < 1$. Then, from Lemma 2.19 and Douc et al. (2018, Lemma 20.3.2, Theorem 20.3.4), we deduce the existence and uniquness of a stationary distribution $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ for $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$. The proof is concluded by using the upper bound given in Lemma 2.19.

We now show the last statement and assume that $\boldsymbol{N} = N\mathbf{1}_n$, for $N \geq 1$. By Proposition 2.13, we have the existence and uniquness of a stationary distribution $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\mathbf{1}_n}$ which is invariant for $P_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ defined in (2.15). For ease of notation, we simply denote $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\mathbf{1}_n}$ by $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ We now show that $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ is also invariant for $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ defined in (2.26). Using the fact that $P_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ defined in (2.15) leaves $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ invariant from Proposition 2.13 and Fubini's theorem, we get for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$,

$$
\begin{aligned}
&\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}} P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\mathsf{A}\times\mathsf{B}) \\
&= \int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}((\tilde{\theta},\tilde{z}),(\mathrm{d}\theta,\mathrm{d}z)) \\
&= \int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\tilde{z},\mathrm{d}z|\tilde{\theta}) \Pi_{\boldsymbol{\rho}}(\mathrm{d}\theta|z) \\
&= \int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) \left[\prod_{i=1}^n R_{\rho_i,\gamma_i}^{N_i}(\tilde{z}_i,\mathrm{d}z_i|\tilde{\theta})\right] \Pi_{\boldsymbol{\rho}}(\mathrm{d}\theta|z) \\
&= \int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) \int_{\mathbb{R}^p} \left[\prod_{i=1}^n R_{\rho_i,\gamma_i}(\tilde{z}_i,\mathrm{d}\tilde{z}_i^{(1)}|\tilde{\theta})\right] \left[\prod_{i=1}^n R_{\rho_i,\gamma_i}^{N_i-1}(\tilde{z}_i^{(1)},\mathrm{d}z_i|\tilde{\theta})\right] \Pi_{\boldsymbol{\rho}}(\mathrm{d}\theta|z) \\
&= \int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \left[\int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) \left[\prod_{i=1}^n R_{\rho_i,\gamma_i}(\tilde{z}_i,\mathrm{d}\tilde{z}_i^{(1)}|\tilde{\theta})\right] \Pi_{\boldsymbol{\rho}}(\mathrm{d}\tilde{\theta}^{(1)}|\tilde{z}_i^{(1)})\right] \\
&\times \left[\prod_{i=1}^n R_{\rho_i,\gamma_i}^{N_i-1}(\tilde{z}_i^{(1)},\mathrm{d}z_i|\tilde{\theta})\right] \Pi_{\boldsymbol{\rho}}(\mathrm{d}\theta|z) \\
&= \int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta}^{(1)},\mathrm{d}\tilde{z}^{(1)}) \left[\prod_{i=1}^n R_{\rho_i,\gamma_i}^{N_i-1}(\tilde{z}_i^{(1)},\mathrm{d}z_i|\tilde{\theta}^{(1)})\right] \Pi_{\boldsymbol{\rho}}(\mathrm{d}\theta|z).
\end{aligned}
$$

Using a straightforward induction, we finally get

$$
\int_{\mathsf{A}\times\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}((\tilde{\theta},\tilde{z}),(\mathrm{d}\theta,\mathrm{d}z)) = \int_{\mathsf{A}\times\mathsf{B}} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\theta,\mathrm{d}z),
$$

which shows that $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ leaves $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ invariant. Since this stationary distribution is unique, we conclude that $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}} = \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$. ■

We specify our result to the case where we take a specific initial distribution. To define it, consider

$$
\mathbf{x}^\star = ([\theta^\star]^\top, [z^\star]^\top)^\top, \text{ where } \theta^\star = \arg\min\{-\log\pi\} \text{ and } z^\star = ([\mathbf{A}_1\theta^\star]^\top, \cdots, [\mathbf{A}_n\theta^\star]^\top)^\top.
$$

We define the probability measure

$$
\mu_{\boldsymbol{\rho}}^\star = \delta_{z^\star} \otimes \Pi_{\boldsymbol{\rho}}(\cdot|z^\star).
$$

Note that sampling from $\mu^\star_{\boldsymbol{\rho}}$ is straightforward and simply consists in setting $z_0 = z^\star$ and $\theta_0 = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} z_0 + \bar{\mathbf{B}}_0^{-1/2}\xi$, where ξ is a d-dimensional standard Gaussian random variable. We now specify our result when using $\mu^\star_{\boldsymbol{\rho}}$ as an initial distribution. Define the z-marginal under $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ by

$$\pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}} = \int_{\mathbb{R}^d} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\mathrm{d}\theta, z), \tag{2.55}$$

and the transition kernel of the Markov chain $\{Z_\ell\}_{\ell\geq 0}$, for all $z \in \mathbb{R}^p$ and $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$, by

$$P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z, \mathsf{B}) = \int_{\mathbb{R}^d} Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z, \mathsf{B}|\theta)\Pi_{\boldsymbol{\rho}}(\mathrm{d}\theta|z), \tag{2.56}$$

where $\Pi_{\boldsymbol{\rho}}(\cdot|\cdot)$ and $Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ are defined in (2.5) and (2.25), respectively.

Proposition 2.22. *Assume Assumption 2.1-Assumption 2.3 and let $c > 0$ and $\boldsymbol{\gamma} = \{\gamma_i\}_{i=1}^n$, $\boldsymbol{N} \in (\mathbb{N}^\star)^n$ such that (2.54) is satisfied, for any $i \in [n]$, $N_i\gamma_i < 2/\max_{i\in[n]}\{m_i + \tilde{M}_i\}$ and $\gamma_i < \tilde{M}_i^{-1}$. Then, for any integer $\ell \geq 1$, we have*

$$\begin{aligned}
&W_2(\mu^\star_{\boldsymbol{\rho}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}) \\
&\quad \leq 2^{1/2}(1 - \min_{i\in[n]}\{N_i\gamma_i m_i\}/2)^{\ell-1} \cdot (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)^{1/2} \max_{i\in[n]}\{N_i\gamma_i\}^{1/2} \\
&\qquad \times \left\{ \int_{\mathbb{R}^d} \|z_1 - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\mathrm{d}z_1) + \int_{\mathbb{R}^d} \|z_1 - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z^\star, \mathrm{d}z_1) \right\}^{1/2},
\end{aligned}$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ are defined in (2.11)-(2.12).

Proof Consider for $\ell \in \mathbb{N}^\star$, $X_\ell = (\theta_\ell^\top, Z_\ell^\top)^\top$, $\tilde{X}_\ell = (\tilde{\theta}_\ell^\top, \tilde{Z}_\ell^\top)^\top$ defined in (2.27) with X_0 distributed according to $\mu^\star_{\boldsymbol{\rho}}$ and $\tilde{X}_0$ distributed according to $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$. Combining Lemma 2.15, Lemma 2.18 and Lemma 2.20, we have for $\ell \geq 1$,

$$\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \leq (1 - \min_{i\in[n]}\{N_i\gamma_i m_i\}/2)\|\tilde{Z}_\ell - Z_\ell\|_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}}.$$

Thereby, for any $\ell \geq 1$, we obtain by induction

$$\|\tilde{Z}_\ell - Z_\ell\|_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \leq (1 - \min_{i\in[n]}\{N_i\gamma_i m_i\}/2)^{n-1}\|\tilde{Z}_1 - Z_1\|_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}}.$$

Using $\|\tilde{Z}_1 - Z_1\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \leq 2\|\tilde{Z}_1 - z^\star\|^2_{\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}} + 2\|Z_1 - z^\star\|^2_{\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}}$ combined with the definition of the Wasserstein distance and Lemma 2.14 give

$$\begin{aligned}
W_2(\mu^\star_{\boldsymbol{\rho}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) &\leq \mathbb{E}\left[\|\tilde{X}_\ell - X_\ell\|^2\right]^{1/2} \\
&\leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)^{1/2} \max_{i\in[n]}\{N_i\gamma_i\}^{1/2} \mathbb{E}\left[\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}}\right]^{1/2} \\
&\leq 2^{1/2}(1 - \min_{i\in[n]}\{N_i\gamma_i m_i\}/2)^{\ell-1}(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)^{1/2} \max_{i\in[n]}\{N_i\gamma_i\}^{1/2} \\
&\qquad \times \mathbb{E}\left[\|\tilde{Z}_1 - z^\star\|^2_{\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}} + \|Z_1 - z^\star\|^2_{\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}}\right]^{1/2}. \qquad (2.57)
\end{aligned}$$

Since $\tilde{X}_0$ is distributed according to the stationnary distribution $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$, $\tilde{X}_1$ also and therefore $\tilde{Z}_1$ is distributed according to $\pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$. Finally, by definition Z_1 has distribution $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z^\star,\cdot)$, therefore (2.57) completes the proof. ■

2.C Proof of Proposition 2.5

The proof of Proposition 3 stands for a generalization of Vono et al. (2022a, Proposition 6) which only considered the specific case $\rho_i = \rho^2$ for $i \in [n]$. This section is divided into two parts, the first gathers lemmas which allow us to upper bound the ξ^2-divergence between $\pi_{\boldsymbol{\rho}}$ and π. Then, in the second subsection, we combine these results to control the Wasserstein distance $W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D}))$ by showing that it is smaller than $\chi^2(\pi_{\boldsymbol{\rho}}|\pi(\cdot|\mathcal{D}))$. For any $\theta \in \mathbb{R}^d$ and $\boldsymbol{\rho} \in (\mathbb{R}^*_+)^n$, define

$$
\begin{aligned}
U_i^{\rho_i}(\mathbf{A}_i\theta) &= -\log\left(\int_{z_i\in\mathbb{R}^d} \exp\{-U_i(z_i) - \|z_i - \mathbf{A}_i\theta\|^2/(2\rho_i)\}\,\mathrm{d}z_i/(2\pi\rho_i)^{d_i/2}\right), \\
\overline{B}(\theta) &= \sum_{i=1}^n \rho_i\|\nabla U_i(\mathbf{A}_i\theta)\|^2/2 \\
\underline{B}(\theta) &= \sum_{i=1}^n \left\{\rho_i\|\nabla U_i(\mathbf{A}_i\theta)\|^2/[2(1+\rho_i M_i)] - d_i\log(1+\rho_i M_i)/2\right\}
\end{aligned}
\tag{2.58}
$$

and consider

$$
U(\theta) = \sum_{i\in[n]} U_i(\mathbf{A}_i\theta), \qquad\qquad U^{\boldsymbol{\rho}}(\theta) = \sum_{i\in[n]} U_i^{\rho_i}(\mathbf{A}_i\theta).
$$

2.C.1 Technical lemmata

We start this subsection by Lemma 2.23 which allow us to bound the ratio between the integrals defined by $\int_{\mathbb{R}^d}\exp\{-\sum_{i\in[n]} U_i^{\rho_i}(\mathbf{A}_i\theta)\}$ and $\int_{\mathbb{R}^d}\exp\{-\sum_{i\in[n]} U_i(\mathbf{A}_i\theta)\}\,\mathrm{d}\theta$.

Lemma 2.23. *Assume Assumption 2.1-Assumption 2.3-(i) and let $\boldsymbol{\rho} \in (\mathbb{R}^*_+)^n$. Then, we have $\underline{B}(\theta) \le U(\theta) - U^{\boldsymbol{\rho}}(\theta)$, for any $\theta \in \mathbb{R}^d$. If we assume in addition that for any $i \in [n]$, U_i is convex, we have $U(\theta) - U^{\boldsymbol{\rho}}(\theta) \le \overline{B}(\theta)$, for any $\theta \in \mathbb{R}^d$.*

Proof The proof follows from the same lines as in Vono et al. (2022a, Lemma 14). In what follows, we give it for the sake of completeness. First, note for any $\theta \in \mathbb{R}^d$ and $i \in [n]$,

$$
\begin{aligned}
&\exp\left\{U_i(\mathbf{A}_i\theta) - U_i^{\rho_i}(\mathbf{A}_i\theta)\right\} \\
&\qquad = \int_{\mathbb{R}^{d_i}} \exp\left(U_i(\mathbf{A}_i\theta) - U_i(z_i) - \|z_i - \mathbf{A}_i\theta\|^2/(2\rho_i)\right)\frac{\mathrm{d}z_i}{(2\pi\rho_i)^{d_i/2}}.
\end{aligned}
\tag{2.59}
$$

Using Assumption 2.3-(i), and a second order Taylor expansion, for any $\theta \in \mathbb{R}^d, i \in [n], z_i \in \mathbb{R}^{d_i}$, we have

$$
U_i(\mathbf{A}_i\theta) - U_i(z_i) \ge \nabla U_i(\mathbf{A}_i\theta)^\top(\mathbf{A}_i\theta - z_i) - M_i\|\mathbf{A}_i\theta - z_i\|^2/2.
$$

Hence, using (2.59), we have for any $\theta \in \mathbb{R}^d$ and $i \in [n]$,

$$\exp\left(\sum_{i=1}^n U_i(\mathbf{A}_i\theta) - U_i^{\rho_i}(\mathbf{A}_i\theta)\right) \geq \prod_{i=1}^n \exp\left(\frac{\rho_i}{2(1+\rho_i M_i)}\left\|\nabla U_i(\mathbf{A}_i\theta)\right\|^2\right)\left(1+\rho_i M_i\right)^{-d_i/2} = \exp(\underline{B}(\theta)).$$

Similarly, under the assumption that for any $i \in [n]$, U_i is convex, the proof for the upper bound follows from the same lines using, for any $i \in [n]$, $\theta \in \mathbb{R}^d$ and $z_i \in \mathbb{R}^{d_i}$, that

$$U_i(\mathbf{A}_i\theta) - U_i(z_i) \leq \nabla U_i(\mathbf{A}_i\theta)^\top(\mathbf{A}_i\theta - z_i).$$

■

Lemma 2.24. *Assume Assumption 2.1-Assumption 2.3. Then, U is m_U-strongly convex with*

$$m_U = \lambda_{\min}\left(\sum_{i=1}^n m_i \mathbf{A}_i^\top \mathbf{A}_i\right).$$

Proof Using by Assumption 2.3-(i) that for any $i \in [n]$, U_i is twice differentiable and by Assumption 2.3-(ii) the fact that for any $i \in [n]$, U_i is m_i-strongly convex, we have for any $\theta \in \mathbb{R}^d$

$$\nabla^2 U(\theta) = \sum_{i=1}^n \mathbf{A}_i^\top \nabla^2 U_i(\mathbf{A}_i\theta)\mathbf{A}_i \succeq \sum_{i=1}^n m_i \mathbf{A}_i^\top \mathbf{A}_i \succeq \lambda_{\min}\left(\sum_{i=1}^n m_i \mathbf{A}_i^\top \mathbf{A}_i\right)\mathbf{I}_d = m_U \mathbf{I}_d.$$

■

For any $\theta \in \mathbb{R}^d$, define

$$\beta(\theta) = \left(\sum_{i=1}^n \rho_i \left\|\nabla U_i(\mathbf{A}_i\theta)\right\|^2\right)^{1/2}. \tag{2.60}$$

Lemma 2.25. *Assume Assumption 2.3-(i) and let $\boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$. Then β is a Lipschitz function w.r.t. $\|\cdot\|$, with Lipschitz constant*

$$L_\beta = \lambda_{\max}\left(\sum_{i=1}^n \rho_i M_i^2 \mathbf{A}_i^\top \mathbf{A}_i\right)^{1/2}. \tag{2.61}$$

Proof For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have using $|(\sum_{i=1}^n a_i^2)^{1/2} - (\sum_{i=1}^n b_i^2)^{1/2}| \leq (\sum_{i=1}^n (a_i - b_i)^2)^{1/2}$, that

$$|\beta(\theta_1) - \beta(\theta_2)| \leq \left(\sum_{i=1}^n \rho_i \|\nabla U_i(\mathbf{A}_i\theta_1) - \nabla U_i(\mathbf{A}_i\theta_2)\|^2\right)^{1/2}$$

$$\leq \left(\sum_{i=1}^{n} \rho_i M_i^2 \|\mathbf{A}_i(\theta_1 - \theta_2)\|^2 \right)^{1/2},$$

which completes the proof. ■

Suppose Assumption 2.3-(ii) and for any $i \in [n]$, denote $\theta_i^\star$ a minimizer of $\theta \mapsto U_i(\mathbf{A}_i\theta)$.

Lemma 2.26. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$. Then for any $s < m_U/(12L_\beta^2)$, where L_β is defined in (2.61), we have*

$$\log \pi\left[\mathrm{e}^{s\{\beta^2 - \pi[\beta^2]\}}\right] \leq 8s^2 L_\beta^4/m_U^2 + 4s^2\{\pi[\beta]\}^2 L_\beta^2/m_U. \tag{2.62}$$

In addition,

$$\pi(\beta^2) \leq 2dL_\beta^2/m_U + 2\sum_{i=1}^{n} \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2. \tag{2.63}$$

Proof Using the decomposition

$$\beta^2(\theta) - \{\pi[\beta]\}^2 = (\beta(\theta) - \pi[\beta])^2 + 2\pi[\beta](\beta(\theta) - \pi[\beta])$$

and the Cauchy-Schwarz inequality imply, for any $s > 0$,

$$\pi\left[\mathrm{e}^{s\{\beta^2 - \{\pi[\beta]\}^2\}}\right] \leq \left\{\pi[\mathrm{e}^{2s\{\beta - \pi[\beta]\}^2}]\right\}^{1/2} \cdot \left\{\pi[\mathrm{e}^{4s\pi[\beta]\{\beta - \pi[\beta]\}}]\right\}^{1/2}. \tag{2.64}$$

The proof consists in bounding the two terms in the right-hand sided. Since $\beta : \mathbb{R}^d \to \mathbb{R}$ is L_β-Lipschitz by Lemma 2.25, for any $0 \leq s \leq m_U/(12L_\beta^2)$, using Vono et al. (2022a, Lemma 16) and Lemma 2.24 gives setting $\bar{\beta} = \beta - \pi[\beta]$, that

$$\pi\left[\exp\{2s(\bar{\beta}^2 - \pi[\bar{\beta}^2])\}\right] \leq \exp(16s^2 L_\beta^4/m_U^2). \tag{2.65}$$

In addition, using Bakry et al. (2013, Proposition 5.4.1), Lemma 2.25 and Lemma 2.24, we get for any $s \geq 0$,

$$\pi\left[\mathrm{e}^{4s\pi[\beta](\beta - \pi[\beta])}\right] \leq \mathrm{e}^{8s^2\{\pi[\beta]\}^2 L_\beta^2/m_U}.$$

Plugging this result and (2.65) into (2.64), we get

$$\pi\left[\mathrm{e}^{s\{\beta^2 - \{\pi[\beta]\}^2\}}\right] \leq \exp(s\pi(\bar{\beta}^2) + 8s^2 L_\beta^4/m_U^2 + 4s^2\{\pi[\beta]\}^2 L_\beta^2/m_U).$$

The proof of (2.62) follows using $\pi(\bar{\beta}^2) = \pi(\beta^2) - [\pi(\beta)]^2$ and rearranging terms.

Using the Young inequality, Assumption 2.3-(i),$\nabla U_i(\mathbf{A}_i\theta_i^\star) = 0$, $\nabla U(\theta^\star) = 0$, we have

$$\begin{aligned}
\pi(\beta^2) &= \int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} \rho_i \|\nabla U_i(\mathbf{A}_i\theta)\|^2 \right) \pi(\theta)\,\mathrm{d}\theta \\
&\leq 2\int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} \rho_i M_i^2 \|\mathbf{A}_i(\theta - \theta^\star)\|^2 \right) \pi(\theta)\,\mathrm{d}\theta + 2\sum_{i=1}^{n} \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2 \\
&\leq 2\lambda_{\max}\left(\sum_{i=1}^{n} \rho_i M_i^2 \mathbf{A}_i^\top \mathbf{A}_i \right) \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \pi(\theta)\,\mathrm{d}\theta + 2\sum_{i=1}^{n} \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2
\end{aligned}$$

$$\leq 2dL_\beta^2/m_U + 2\sum_{i=1}^n \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2,$$

where we have used $\pi[\|\theta - \theta^\star\|^2] \leq d/m_U$ by Durmus and Moulines (2019, Proposition 1 (ii)) and Lemma 2.24. ∎

Proposition 2.2 shows that $\pi_{\boldsymbol{\rho}}(\cdot) = \int_{\mathbb{R}^p} \Pi_{\boldsymbol{\rho}}(\cdot, z)\mathrm{d}z$ is well-defined and as such admits a finite normalizing constant. These two quantities are defined by

$$\mathrm{Z}_{\pi_{\boldsymbol{\rho}}} = \int_{\mathbb{R}^d} \exp\left\{-\sum_{i\in[n]} U_i^{\rho_i}(\mathbf{A}_i\theta)\right\}\mathrm{d}\theta, \quad \pi_{\boldsymbol{\rho}}(\cdot) = \exp\left\{-\sum_{i\in[n]} U_i^{\rho_i}(\mathbf{A}_i\cdot)\right\}/\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}. \tag{2.66}$$

Finally, note that the following quantity Z_π is a normalizing constant of π associated with the potential U, *i.e.* $\pi = \mathrm{e}^{-U}/\mathrm{Z}_\pi$,

$$\mathrm{Z}_\pi = \int_{\mathbb{R}^d} \exp\left\{-\sum_{i\in[n]} U_i(\mathbf{A}_i\theta)\right\}\mathrm{d}\theta. \tag{2.67}$$

Lemma 2.27. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$. Suppose in addition that $6L_\beta^2 \leq m_U$ where L_β is given in (2.61). Then, we have*

$$\log\left(\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}/\mathrm{Z}_\pi\right) \leq \left\{dL_\beta^2/m_U + \sum_{i=1}^n \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2\right\}(1 + 2L_\beta^2/m_U) + 2L_\beta^4/m_U^2.$$

Proof From the definitions (2.66) and (2.67), we have $\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}/\mathrm{Z}_\pi = \int_{\mathbb{R}^d} \pi(\theta)\exp\{\sum_{i=1}^n U_i(\mathbf{A}_i\theta) - U_i^{\rho_i}(\mathbf{A}_i\theta)\}\,\mathrm{d}\theta$. By Lemma 2.23, we obtain

$$\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}/\mathrm{Z}_\pi \leq \int_{\mathbb{R}^d} \pi(\theta)\exp(\overline{B}(\theta))\,\mathrm{d}\theta.$$

Note that $\overline{B} = \beta^2/2$ by (2.58)-(2.60), hence using that $6L_\beta^2 \leq m_U$, Lemma 2.26 applied with $s = 1/2$ shows that

$$\log\left(\int_{\mathbb{R}^d} \pi(\theta)\exp(\overline{B}(\theta))\,\mathrm{d}\theta\right) \leq \pi[\beta^2]/2 + 2L_\beta^4/m_U^2 + \{\pi[\beta]\}^2 L_\beta^2/m_U.$$

Using Lemma 2.26-(2.63) and $\pi[\beta] \leq \pi[\beta^2]$ concludes the proof. ∎

2.C.2 Proof of Proposition 2.5

Based on the technical lemmas derived in Section 2.C.1, we are now ready to bound the Wasserstein distance of order 2 between π and $\pi_{\boldsymbol{\rho}}$.

Proof [Proof of Proposition Proposition 2.5] Let $\boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$ such that $\max_{i\in[n]} \rho_i = \bar{\rho} \leq \sigma_U^2/12$, where $\sigma_U^2 = \|\mathbf{A}^\top\mathbf{A}\| \max_{i\in[n]}\{M_i^2\}/m_U$. Then, by definition of L_β (2.61), we get

$$12L_\beta^2 \leq m_U. \tag{2.68}$$

and Lemma 2.26 can be applied for $s = 1$ and Lemma 2.27 too. By Lemma 2.24, $U = -\log \pi$ is m_U-strongly convex therefore π satisfies a log-Sobolev inequality with constant m_U (Ledoux, 2001, Theorem 5.2). Finally, Otto and Villani (2000, Theorem 1) shows that π satisfies for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$W_2(\nu, \pi(\cdot|\mathcal{D})) \leq \sqrt{(2/m_U)\mathrm{KL}(\nu|\pi(\cdot|\mathcal{D}))} \leq \sqrt{(2/m_U)\chi^2(\nu|\pi(\cdot|\mathcal{D}))}, \tag{2.69}$$

where χ^2 is the chi-square divergence and where we have used for the last inequality that $\mathrm{KL}(\nu|\pi(\cdot|\mathcal{D})) \leq \chi^2(\nu|\pi(\cdot|\mathcal{D}))$ since for any $t > 0$, $\log(t) \leq t - 1$. We now bound $\chi^2(\pi_{\boldsymbol{\rho}}|\pi(\cdot|\mathcal{D}))$. By (2.66) and (2.67), for any $\theta \in \mathbb{R}^d$, consider the decomposition given by

$$\pi_{\boldsymbol{\rho}}(\theta)/\pi(\theta) - 1 = (\mathrm{Z}_\pi/\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}) \exp\left(\sum_{i=1}^n \left(U_i(\mathbf{A}_i\theta) - U_i^{\rho_i}(\mathbf{A}_i\theta)\right)\right) - 1. \tag{2.70}$$

In the sequel, we will both lower and upper bound (2.70) in order to upper bound $|1 - \pi_{\boldsymbol{\rho}}(\theta)/\pi(\theta)|$. Using the fact that for all $x \in \mathbb{R}, \exp(x) - 1 \geq x$, Lemmas 2.23 and 2.27 yield

$$\begin{aligned}\pi_{\boldsymbol{\rho}}(\theta)/\pi(\theta) - 1 &\geq \log\left(\mathrm{Z}_\pi/\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}\right) + \sum_{i=1}^n \left(U_i(\mathbf{A}_i\theta) - U_i^{\rho_i}(\mathbf{A}_i\theta)\right) \qquad (2.71)\\ &\geq -\left\{dL_\beta^2/m_U + \sum_{i=1}^n \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2\right\}(1 + 2L_\beta^2/m_U) - 2L_\beta^4/m_U^2 + \underline{B}(\theta) \geq -A_1,\end{aligned}$$

where

$$\begin{aligned}A_1 = \left\{dL_\beta^2/m_U + \sum_{i=1}^n \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2\right\}&(1 + 2L_\beta^2/m_U)\\ &+ 2L_\beta^4/m_U^2 + \sum_{i=1}^n (d_i/2)\log(1 + \rho_i M_i),\end{aligned}$$

where we have used in the last inequality that $\underline{B}(\theta) \geq -\sum_{i=1}^n (d_1/2)\log(1 + \rho_i M_i)$ by (2.58). In addition, by (2.66) and (2.67) we have

$$\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}/\mathrm{Z}_\pi = \int_{\mathbb{R}^d} \pi(\theta) \exp\{-\sum_{i=1}^n U_i^{\rho_i}(\mathbf{A}_i\theta)\}\,\mathrm{d}\theta / \int_{\mathbb{R}^d} \pi(\theta)\exp\{-\sum_{i=1}^n U_i(\mathbf{A}_i\theta)\}\,\mathrm{d}\theta,$$

which implies by Lemma 2.23 and Jensen inequality

$$\mathrm{Z}_{\pi_{\boldsymbol{\rho}}}/\mathrm{Z}_\pi \geq \int_{\mathbb{R}^d} \pi(\theta)\exp(\underline{B}(\theta))\,\mathrm{d}\theta \geq \exp(\pi[\underline{B}]).$$

It follows by (2.70) that $\pi_{\boldsymbol{\rho}}(\theta)/\pi(\theta) - 1 \leq \exp(\overline{B}(\theta) - \pi\left(\underline{B}\right)) - 1$. Combining this result and (2.71), it follows that the Pearson χ^2-divergence between π and $\pi_{\boldsymbol{\rho}}$ can be upper bounded as where

$$\chi^2(\pi_{\boldsymbol{\rho}}|\pi(\cdot|\mathcal{D})) \leq \max(A_1^2, A_2), \qquad A_2 = \int_{\mathbb{R}^d} \left(\exp(\overline{B}(\theta) - \pi\left(\underline{B}\right)) - 1\right)^2 \pi(\theta)\,\mathrm{d}\theta.$$

We now provide an explicit bound for A_2. First by Jensen inequality, we have $\pi(\exp(\overline{B})) \geq \exp(\pi(\overline{B}))$ which implies that $\exp(-\pi(\underline{B}))\pi[\exp(\overline{B})] \geq \prod_{i=1}^n (1+\rho_i M_i)^{d_i/2}$ by (2.58). Therefore, using that $\overline{B} = \beta^2/2$ by (2.58)-(2.60) and Lemma 2.26 with $s = 1$ since (2.68) holds, we get by (2.58),

$$
\begin{aligned}
A_2 &= \int_{\mathbb{R}^d} \left(\exp\left(\overline{B}(\theta) - \pi\left(\underline{B}\right)\right) - 1\right)^2 \pi(\theta)\,\mathrm{d}\theta \\
&= \exp\left(-2\pi\left(\underline{B}\right)\right)\pi\left[\exp\left(2\overline{B}\right)\right] - 2\exp\left(-\pi\left(\underline{B}\right)\right)\pi\left[\exp\left(\overline{B}\right)\right] + 1 \\
&\leq \prod_{i=1}^n (1+\rho_i M_i)^{d_i} \cdot \exp(-\pi\{\sum_{i=1}^n (\rho_i/(1+\rho_i M_i))\|\nabla U_i(\mathbf{A}_i\cdot)\|^2\})\pi\left[\exp(\beta^2)\right] \\
&\quad - 2\prod_{i=1}^n (1+\rho_i M_i)^{d_i/2} + 1 \\
&\leq \prod_{i=1}^n (1+\rho_i M_i)^{d_i} \cdot \exp(\pi\{\sum_{i=1}^n (\rho_i^2 M_i/(1+\rho_i M_i))\|\nabla U_i(\mathbf{A}_i\cdot)\|^2\}) \\
&\times \exp\left(8L_\beta^4/m_U^2 + 4\{2dL_\beta^2/m_U + 2\sum_{i=1}^n \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2\}L_\beta^2/m_U\right) \\
&\quad - 2\prod_{i=1}^n (1+\rho_i M_i)^{d_i/2} + 1,
\end{aligned}
\tag{2.72}
$$

where we have used for the last inequality that for $\theta \in \mathbb{R}^d$, $\beta(\theta)^2 - \sum_{i=1}^n (\rho_i/(1+\rho_i M_i))\|\nabla U_i(\mathbf{A}_i\theta)\|^2 = \sum_{i=1}^n (\rho_i^2 M_i/(1+\rho_i M_i))\|\nabla U_i(\mathbf{A}_i\theta)\|^2$, $\pi[\beta]^2 \leq \pi[\beta^2]$ by the Cauchy-Schwartz inequality and Lemma 2.26-(2.63). Similarly to the proof of Lemma 2.26-(2.63), by Assumption 2.3-(i), $\nabla U_i(\mathbf{A}_i\theta_i^\star) = 0$, $\nabla U(\theta^\star) = 0$, Durmus and Moulines (2019, Proposition 1 (ii)) and Lemma 2.24, we have

$$
\begin{aligned}
\pi\left[\sum_{i=1}^n (\rho_i^2 M_i/(1+\rho_i M_i))\|\nabla U_i(\mathbf{A}_i\cdot)\|^2\right] &\leq \pi\left[\sum_{i=1}^n \rho_i^2 M_i \|\nabla U_i(\mathbf{A}_i\cdot)\|^2\right] \\
&\leq 2d\lambda_{\max}\left(\sum_{i=1}^n \rho_i^2 M_i^3 \mathbf{A}_i^\top \mathbf{A}_i\right)/m_U + 2\sum_{i=1}^n \rho_i^2 M_i^3 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2.
\end{aligned}
$$

Therefore, we get by (2.72)

$$
\begin{aligned}
&A_2 \leq A_3 \\
&= \prod_{i=1}^n (1+\rho_i M_i)^{d_i} \exp\left(2d\lambda_{\max}\left(\sum_{i=1}^n \rho_i^2 M_i^3 \mathbf{A}_i^\top \mathbf{A}_i\right)/m_U + 2\sum_{i=1}^n \rho_i^2 M_i^3 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2\right) \\
&\exp\left(8L_\beta^4/m_U^2 + 8\left[dL_\beta^2/m_U + \sum_{i=1}^n \rho_i M_i^2 \|\mathbf{A}_i(\theta^\star - \theta_i^\star)\|^2\right]L_\beta^2/m_U\right) - 2\prod_{i=1}^n (1+\rho_i M_i)^{d_i/2} + 1.
\end{aligned}
\tag{2.73}
$$

It follows by (2.72) and (2.69) that

$$
W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D})) \leq \sqrt{(2/m_U)\max(A_1^2, A_3)}, \tag{2.74}
$$

where A_1 and A_3 are given by (2.71) and (2.73) respectively. Using that $L_\beta^2 = \mathrm{O}(\bar{\rho})$ and an expansion of the bound as $\bar{\rho} \to 0$ completes the proof. ■

2.D Proof of Proposition 2.6 and Proposition 2.8

As in Section 2.B, we assume in all this section that $\boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$ is fixed. For any $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n) \in (\mathbb{R}_+^*)^n$, we establish in this section explicit bounds on $W_2(\pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \pi_{\boldsymbol{\rho}})$ where $\pi_{\boldsymbol{\rho}}$ is given in (2.1) and $\pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ is the marginal distribution defined by

$$\pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\mathsf{A}) = \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(\mathsf{A} \times \mathbb{R}^p), \qquad \mathsf{A} \in \mathcal{B}(\mathbb{R}^d),$$

of the stationary probability measure $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ associated with the Markov chain $(Z_\ell, \theta_\ell)_{\ell \geq 0}$ defined in Algorithm 2.1. Note that in the case $\boldsymbol{N} = N(1, \dots, 1)$, this distribution is independent of N, see Proposition 2.21. To this purpose, we define an "ideal" dynamics from which we cannot sample but which converges geometrically towards $\Pi_{\boldsymbol{\rho}}$ under appropriate conditions. The corresponding ideal process will play the same role as the Langevin dynamics for the study of the unadjusted Langevin algorithm (Durmus and Moulines, 2019). This dynamics is defined as follows. Consider first for any $\theta \in \mathbb{R}^d$, $i \in [n]$, the stochastic differential equation (SDE) defined by

$$\mathrm{d}\tilde{Y}_t^{i,\theta} = -\nabla V_i(\tilde{Y}_t^{i,\theta})\,\mathrm{d}t - \rho_i^{-1}\mathbf{A}_i\theta + \sqrt{2}\,\mathrm{d}B_t^i, \tag{2.75}$$

where $(B_t^i)_{t \geq 0}$ is a d_i-dimensional Brownian motion and V_i is defined in (2.29). Note that under Assumption 2.3-(i), this SDE admits a unique strong solution (Revuz and Yor, 2013, Theorem (2.1) in Chapter IX). Denote for any $i \in [n]$, the Markov semigroup associated to (2.75) by $(\tilde{R}_{\rho_i,t}^i)_{t \geq 0}$ defined for any $\tilde{\mathbf{y}}_0^i \in \mathbb{R}^{d_i}$, $t \geq 0$ and $\mathsf{B}_i \in \mathcal{B}(\mathbb{R}^{d_i})$ by

$$\tilde{R}_{\rho_i,t}^i(\tilde{\mathbf{y}}_0^i, \mathsf{B}_i|\theta) = \mathbb{P}(\tilde{Y}_t^{i,\theta,\tilde{\mathbf{y}}_0^i} \in \mathsf{B}_i),$$

where $(\tilde{Y}_t^{i,\theta,\tilde{\mathbf{y}}_0^i})_{t \geq 0}$ is a solution of (2.75) with $\tilde{Y}_0^{i,\theta,\tilde{\mathbf{y}}_0^i} = \tilde{\mathbf{y}}_0^i$. For any bounded measurable function $f_i : \mathbb{R}^{d_i} \to \mathbb{R}_+$, Lemma 2.28 shows the measurability of the function $(\theta, \tilde{\mathbf{y}}_0^i) \mapsto \mathbb{E}[f_i(\tilde{Y}_t^{i,\theta,\tilde{\mathbf{y}}_0^i})]$ on $\mathbb{R}^d \times \mathbb{R}^{d_i}$ and therefore $\tilde{R}_{\rho_i,t}^i$ is a conditional Markov kernel.

Lemma 2.28. *For any bounded measurable function $f_i : \mathbb{R}^{d_i} \to \mathbb{R}_+$ and function f_i satisfying Assumption 2.3-(i), the mapping $(\tilde{\theta}_0, \tilde{\mathbf{y}}_0^i) \mapsto \mathbb{E}[f_i(\tilde{Y}_t^{i,\tilde{\theta}_0,\tilde{\mathbf{y}}_0^i})]$ is Borel measurable.*

Proof Consider the following stochastic differential equation

$$\begin{cases} \mathrm{d}\tilde{\theta}_t = \mathbf{0}_d, \\ \mathrm{d}\tilde{Y}_t^i = -\nabla V_i(\tilde{Y}_t^i)\,\mathrm{d}t - \rho_i^{-1}\mathbf{A}_i\tilde{\theta}_t + \sqrt{2}\,\mathrm{d}B_t^i. \end{cases}$$

Using Revuz and Yor (2013, Theorem (2.4) in Chapter IX), since U_i satisfies Assumption 2.3-(i), there exists a unique solution $(\tilde{X}_t^{\tilde{\mathbf{x}}})_{t \geq 0} = (\tilde{\theta}_t, \tilde{Y}_t^i)_{t \geq 0}$ with initial condition $\tilde{\mathbf{x}} = (\tilde{\theta}_0^\top, (\tilde{\mathbf{y}}_0^i)^\top)^\top \in \mathbb{R}^p$. Then, the proof follows from Revuz and Yor (2013, Theorem (1.9) in Chapter IX) and the fact that $\tilde{Y}_t^i$ is the unique solution of (2.75) with $\theta = \theta_0$. ∎

Define for any $\theta \in \mathbb{R}^d$, $z = (z_1^\top, \cdots, z_n^\top)^\top \in \mathbb{R}^p$, and for $i \in [n]$, $\mathsf{B}_i \in \mathcal{B}(\mathbb{R}^{d_i})$,

$$\tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}\left(z, \mathsf{B}_1 \times \cdots \times \mathsf{B}_n|\theta\right) = \prod_{i=1}^n \tilde{R}_{\rho_i,N_i\gamma_i}^i(z_i, \mathsf{B}_i|\theta),$$

and consider the Markov kernel defined, for any $\mathbf{x}^\top = (\theta^\top, z^\top)$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$, by

$$\tilde{P}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathbf{x}, \mathsf{A} \times \mathsf{B}) = \int_{\mathsf{B}} \tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z, \mathrm{d}\tilde{\mathbf{z}}|\theta) \int_{\mathsf{A}} \Pi_{\boldsymbol{\rho}}(\mathrm{d}\tilde{\theta}|\tilde{\mathbf{z}}) \,, \tag{2.76}$$

where $\Pi_{\boldsymbol{\rho}}(\cdot|\tilde{\mathbf{z}})$ is defined in (2.5). Note that $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ can be interpreted as a discretized version of $\tilde{P}_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ using the Euler-Maruyama scheme.

In the sequel, we first derive technical lemmata in Section 2.D.1 that are used to prove both Proposition 2.6 and Proposition 2.8. Based on these lemmata, we then prove each proposition in a dedicated section, namely Section 2.D.2 and Section 2.D.3.

2.D.1 Synchronous coupling and a first estimate

The main idea to prove Proposition 2.6 and Proposition 2.8 is to define $(X_\ell, \tilde{X}_\ell)_{\ell\in\mathbb{N}}$ such that for any $\ell \in \mathbb{N}$, $(X_\ell, \tilde{X}_\ell)$ is a coupling between $\delta_{\mathbf{x}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ defined in (2.26) and $\delta_{\tilde{\mathbf{x}}} \tilde{P}^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma}}$, and satisfies

$$\mathbb{E}\left[\|X_\ell - \tilde{X}_\ell\|^2\right] \leq c_1(\mathbf{x}, \tilde{\mathbf{x}}) \mathrm{e}^{-c_2 \min_{i\in[n]}\{\gamma_i m_i\}} + c_3 \gamma^\alpha,$$

where $c_2, c_3 > 0$ and $\alpha \in \{1, 2\}$ depending if Assumption 2.7 holds or not. Conditioning with respect to $(X_0, \tilde{X}_0)$ with distribution $\delta_{\mathbf{x}} \otimes \Pi_{\boldsymbol{\rho}}$, using the definition of the Wasserstein distance of order 2 and taking $n \to \infty$, we obtain

$$W_2(\pi_{\boldsymbol{\rho}}, \pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq W_2(\Pi_{\boldsymbol{\rho}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq \tilde{c}_3 \gamma^\alpha,$$

where $\tilde{c}_3 > 0$. We now provide the rigorous construction of $(X_\ell, \tilde{X}_\ell)_{\ell\in\mathbb{N}}$.

Let $\{(B_t^{(i,\ell)})_{t\geq 0} : i \in [n], \ell \in \mathbb{N}\}$ be independent random variables such that for any $i \in [n]$, the sequences $\{(B_t^{(i,\ell)})_{t\geq 0} : \ell \in \mathbb{N}\}$ are i.i.d. d_i-dimensional Brownian motions and let $(\xi_\ell)_{\ell\geq 0}$ be a sequence of i.i.d. standard d-dimensional Gaussian random variables independent of $\{(B_t^{(i,\ell)})_{t\geq 0} : i \in [n], \ell \in \mathbb{N}\}$. Consider the stochastic process $(\tilde{X}_\ell)_{\ell\geq 0}$ on $\mathbb{R}^d \times \mathbb{R}^p$ starting from $\tilde{X}_0$ distributed according to $\Pi_{\boldsymbol{\rho}}$ and defined by the recursion: for $\ell \in \mathbb{N}$, $i \in [n]$,

$$\tilde{X}_{\ell+1} = (\tilde{\theta}_{\ell+1}^\top, \tilde{Z}_{\ell+1}^\top)^\top, \quad \tilde{Z}^i_{\ell+1} = \tilde{Y}^{(i,\ell)}_{N_i\gamma_i}, \quad \tilde{\theta}_{\ell+1} = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} \tilde{Z}_{\ell+1} + \bar{\mathbf{B}}_0^{-1/2} \xi_{\ell+1}, \tag{2.77}$$

where $(\tilde{Y}_t^{(i,\ell)})_{t\geq 0}$, is a solution of (2.75) starting from $\tilde{Z}^i_\ell$ with parameter $\theta \leftarrow \theta_\ell$. Similarly to the process $(X_\ell)_{\ell\in\mathbb{N}}$ defined in Algorithm 2.1, the process $(\tilde{X}_\ell)_{\ell\in\mathbb{N}}$ defines a homogeneous Markov chain. Indeed, it is easy to show that for any $\ell \in \mathbb{N}$ and measurable function $f : \mathbb{R}^p \to \mathbb{R}_+$, $\mathbb{E}[f(\tilde{Z}_{\ell+1})|\tilde{X}_\ell] = \int_{\mathbb{R}^p} f(\tilde{z}) \tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\tilde{Z}_\ell, \mathrm{d}z|\tilde{\theta}_\ell)$ and therefore $(\tilde{X}_\ell)_{\ell\in\mathbb{N}}$ is associated with (2.76).

Proposition 2.29. *Assume Assumption 2.1-Assumption 2.3-(i), and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$. Then, the Markov kernel $\tilde{P}_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ defined in (2.76) admits $\Pi_{\boldsymbol{\rho}}$ as an invariant probability measure.*

Proof By property of the Langevin diffusion defined in (2.75), for all $\theta_0 \in \mathbb{R}^d$, the Markov kernel $\tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\cdot|\theta_0)$ admits $\Pi_{\boldsymbol{\rho}}(\cdot|\theta_0)$ as invariant measure, see *e.g.* Roberts and Tweedie (1996) or Kent (1978). Thus, for any $\theta_0 \in \mathbb{R}^d$ and $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$, we have

$$\int_{\mathsf{B}} \Pi_{\boldsymbol{\rho}}(z_1|\theta_0) \,\mathrm{d}z_1 = \int_{z_0\in\mathbb{R}^p} \tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z_0, \mathsf{B}|\theta_0) \Pi_{\boldsymbol{\rho}}(z_0|\theta_0) \,\mathrm{d}z_0. \tag{2.78}$$

Denote by $\pi^{\theta}_{\boldsymbol{\rho}}, \pi^{z}_{\boldsymbol{\rho}}$ the marginals under $\Pi_{\boldsymbol{\rho}}$: $\pi^{\theta}_{\boldsymbol{\rho}}(\mathsf{A}) = \Pi_{\boldsymbol{\rho}}(\mathsf{A} \times \mathbb{R}^p)$, $\pi^{z}_{\boldsymbol{\rho}}(\mathsf{B}) = \Pi_{\boldsymbol{\rho}}(\mathbb{R}^d \times \mathsf{B})$, for $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$, and consider the Markov chain $(\tilde{X}_\ell)_{\ell \in \mathbb{N}}$ defined in (2.77). For any measurable function $f : \mathbb{R}^{d+p} \to \mathbb{R}_+$, the Fubini-Tonelli theorem gives

$$
\begin{aligned}
\mathbb{E}[f(\tilde{X}_1)] &= \int_{\mathbb{R}^{d+p}} \int_{\mathbb{R}^{d+p}} f(\mathbf{x}_1) \Pi_{\boldsymbol{\rho}}(\theta_1 | z_1) \, \mathrm{d}\theta_1 \tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z_0, \mathrm{d}z_1 | \theta_0) \Pi_{\boldsymbol{\rho}}(\theta_0, z_0) \, \mathrm{d}\theta_0 \, \mathrm{d}z_0 \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^p} f(\mathbf{x}_1) \Pi_{\boldsymbol{\rho}}(\theta_1 | z_1) \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^p} \tilde{Q}_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z_0, \mathrm{d}z_1 | \theta_0) \Pi_{\boldsymbol{\rho}}(z_0 | \theta_0) \, \mathrm{d}z_0 \right] \pi^{\theta}_{\boldsymbol{\rho}}(\theta_0) \, \mathrm{d}\theta_0 \, \mathrm{d}\theta_1 \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^p} f(\mathbf{x}_1) \Pi_{\boldsymbol{\rho}}(\theta_1 | z_1) \left[\int_{\theta_0 \in \mathbb{R}^d} \Pi_{\boldsymbol{\rho}}(z_1 | \theta_0) \pi^{\theta}_{\boldsymbol{\rho}}(\theta_0) \, \mathrm{d}\theta_0 \right] \mathrm{d}z_1 \, \mathrm{d}\theta_1 \qquad (2.79) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^p} f(\mathbf{x}_1) \Pi_{\boldsymbol{\rho}}(\theta_1 | z_1) \pi^{z}_{\boldsymbol{\rho}}(z_1) \, \mathrm{d}z_1 \mathrm{d}\theta_1 \\
&= \int_{\mathbb{R}^{d+p}} f(\mathbf{x}_1) \Pi_{\boldsymbol{\rho}}(\theta_1, z_1) \, \mathrm{d}z_1 \mathrm{d}\theta_1 = \mathbb{E}[f(\tilde{X}_0)],
\end{aligned}
$$

where we have used (2.78) in (2.79). Therefore, X_1 has distribution $\Pi_{\boldsymbol{\rho}}$ and the Markov kernel $\tilde{P}_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ admits $\Pi_{\boldsymbol{\rho}}$ as a stationary distribution, which completes the proof. ■

Define by induction the synchronous coupling $(X_\ell = (\theta_\ell, Z_\ell))_{\ell \geq 0}, (\tilde{X}_\ell = (\tilde{\theta}_\ell, \tilde{Z}_\ell))_{\ell \geq 0}$, starting from $(\theta_0, Z_0) = (\theta, z)$, $(\tilde{\theta}_0, \tilde{Z}_0)$ distributed according to $\Pi_{\boldsymbol{\rho}}$, for any $i \in [n]$ and $\ell \geq 0$, as

$$
\begin{aligned}
\tilde{Z}^i_{\ell+1} &= \tilde{Y}^{(i,\ell)}_{N_i \gamma_i}, & \tilde{\theta}_{\ell+1} &= \bar{\mathbf{B}}_0^{-1} \mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} \tilde{Z}_{\ell+1} + \bar{\mathbf{B}}_0^{-1/2} \xi_{\ell+1}, \qquad (2.80) \\
Z^i_{\ell+1} &= Y^{(i,\ell)}_{N_i \gamma_i}, & \theta_{\ell+1} &= \bar{\mathbf{B}}_0^{-1} \mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} Z_{\ell+1} + \bar{\mathbf{B}}_0^{-1/2} \xi_{\ell+1},
\end{aligned}
$$

where we consider for any $i \in [n], k \in \mathbb{N}$, for $t \in [k\gamma_i, (k+1)\gamma_i)$

$$
\begin{aligned}
\tilde{Y}^{(i,\ell)}_t &= \tilde{Y}^{(i,\ell)}_{k\gamma_i} - \int_{k\gamma_i}^t \nabla V_i(\tilde{Y}^{(i,\ell)}_l) \, \mathrm{d}l + (t - k\gamma_i)(\rho_i)^{-1} \mathbf{A}_i \tilde{\theta}_\ell + 2^{1/2} (B^{(i,\ell)}_t - B^{(i,\ell)}_{k\gamma_i}), \\
Y^{(i,\ell)}_t &= Y^{(i,\ell)}_{k\gamma_i} - (t - k\gamma_i) \nabla V_i(Y^{(i,\ell)}_{k\gamma_i}) + (t - k\gamma_i)(\rho_i)^{-1} \mathbf{A}_i \theta_\ell + 2^{1/2} (B^{(i,\ell)}_t - B^{(i,\ell)}_{k\gamma_i}).
\end{aligned} \qquad (2.81)
$$

Let $\mathcal{G}_0 = \sigma(Z_0, \tilde{Z}_0, \theta_0, \tilde{\theta}_0)$, for any $\ell \in \mathbb{N}^\star$, let

$$
\mathcal{G}_\ell = \sigma\{(Z_0, \tilde{Z}_0, \theta_0, \tilde{\theta}_0), (B^{(i,k)}_t)_{t \geq 0} : i \in [n], k \leq \ell\}, \qquad (2.82)
$$

and for any $t \geq 0$, let $\mathcal{H}^{(\ell)}_t = \sigma(\{(B^{(i,\ell)}_s)_{s \leq t} : i \in [n]\})$, and

$$
\mathcal{F}^{(\ell)}_t \text{ the } \sigma\text{-field generated by } \mathcal{H}^{(\ell)}_t \text{ and } \mathcal{G}_{\ell-1}. \qquad (2.83)
$$

Note that X_ℓ and $\tilde{X}_\ell$ are distributed according to $\Pi_{\boldsymbol{\rho}} \tilde{P}^\ell_{\boldsymbol{\rho}}$ and $\delta_{\tilde{\mathbf{x}}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$, respectively. Hence, by definition of the Wasserstein distance of order 2, it follows since $\Pi_{\boldsymbol{\rho}} \tilde{P}^\ell_{\boldsymbol{\rho}} = \Pi_{\boldsymbol{\rho}}$ by Proposition 2.29 that

$$
W_2(\Pi_{\boldsymbol{\rho}}, \delta_{\mathbf{x}} P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}) \leq \mathbb{E}\left[\|X_\ell - \tilde{X}_\ell\|^2 \right]^{1/2}. \qquad (2.84)
$$

We start this section by a first estimate on $\mathbb{E}[\|X_\ell - \tilde{X}_\ell\|^2]^{1/2}$ and some technical results needed for the proof of Proposition 2.6 and Proposition 2.8. The following result holds regarding the process $(\tilde{Y}^{(i,\ell)}_t)_{t \in \mathbb{R}_+}$ defined, for any $i \in [n]$ and $\ell \in \mathbb{N}$, in (2.81).

Lemma 2.30. *Assume Assumption 2.1-Assumption 2.3. For $i \in [n], \ell \in \mathbb{N}$, denote by $z^i_{\ell,\star}$ the unique minimizer of $z_i \in \mathbb{R}^{d_i} \mapsto U_i(z_i) + \|z_i - \mathbf{A}_i\tilde{\theta}_\ell\|/(2\rho_i)$. Then, for any $i \in [n], k \in \mathbb{N}$ and $\ell \in \mathbb{N}$,*

$$\mathbb{E}^{\mathcal{G}_\ell}\left[\|\tilde{Y}^{(i,\ell)}_{k\gamma_i} - z^i_{\ell,\star}\|^2\right] \leq d_i/\tilde{m}_i, \tag{2.85}$$

where $\tilde{m}_i$ is defined in (2.34).

Proof Let $\ell \in \mathbb{N}$. By Durmus and Moulines (2019, Proposition 1), for $i \in [n]$ and $k \in \mathbb{N}$, we have

$$\mathbb{E}^{\mathcal{F}^{(\ell)}_{k\gamma_i}}\|\tilde{Y}^{(i,\ell)}_{k\gamma_i} - z^i_{\ell,\star}\|^2 \leq \|\tilde{Z}^i_\ell - z^i_{\ell,\star}\|^2 \mathrm{e}^{-2k\gamma_i\tilde{m}_i} + (d_i/\tilde{m}_i)(1 - \mathrm{e}^{-2k\gamma_i\tilde{m}_i}). \tag{2.86}$$

By (2.81), using Proposition 2.29 we get that $\tilde{X}_\ell$ has distribution $\Pi_{\boldsymbol{\rho}}$, therefore given $\tilde{\theta}_\ell$, $\tilde{Z}_\ell$ has distribution $\Pi_{\boldsymbol{\rho}}(\cdot|\tilde{\theta}_\ell)$. Then, using (2.86), Durmus and Moulines (2019, Proposition 1(ii)) combined with Assumption 2.3, and since $(\tilde{Z}^1_\ell, \dots, \tilde{Z}^n_\ell)$ are independent given $\tilde{\theta}_\ell$, we get the stated result. ■

Lemma 2.31. *Assume Assumption 2.1 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}^*_+)^n$. Then, for any $\ell \in \mathbb{N}$, the random variable $X_\ell = (\theta_\ell^\top, Z_\ell^\top)^\top, \tilde{X}_\ell = (\tilde{\theta}_\ell^\top, \tilde{Z}_\ell^\top)^\top$ defined in (2.80) satisfies*

$$\|\tilde{X}_{\ell+1} - X_{\ell+1}\|^2 \leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2,$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ are defined in (2.11)-(2.12).

Proof The proof is similar to the proof of Lemma 2.11 and is omitted. ■

For any $k, \ell \in \mathbb{N}, s \in \mathbb{R}_+$ consider the $p \times p$ matrices defined by

$$\mathbf{J}(k,s) = \mathrm{diag}\left(\mathbb{1}_{[N_1]}(k+1)\mathbb{1}_{[0,\gamma_1]}(s)\cdot\mathbf{I}_{d_1}, \cdots, \mathbb{1}_{[N_n]}(k+1)\mathbb{1}_{[0,\gamma_n]}(s)\cdot\mathbf{I}_{d_n}\right), \tag{2.87}$$

$$\mathbf{H}^{(\ell)}_{U,k} = \mathrm{diag}\Bigg(\gamma_1\int_0^1 \nabla^2 U_1((1-s)Y^{(1,\ell)}_{k\gamma_1} + s\tilde{Y}^{(1,\ell)}_{k\gamma_1})\,\mathrm{d}s, \tag{2.88}$$

$$\dots, \gamma_n\int_0^1 \nabla^2 U_n((1-s)Y^{(n,\ell)}_{k\gamma_n} + s\tilde{Y}^{(n,\ell)}_{k\gamma_n})\,\mathrm{d}s\Bigg),$$

$$\mathbf{C}^{(\ell)}_k = \mathbf{J}(k,0)(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} + \mathbf{H}^{(\ell)}_{U,k}), \tag{2.89}$$

$$\mathbf{M}^{(\ell)}_{k+1} = (\mathbf{I}_p - \mathbf{C}^{(\ell)}_0)^{-1}\dots(\mathbf{I}_p - \mathbf{C}^{(\ell)}_k)^{-1}, \qquad \text{with } \mathbf{M}^{(\ell)}_0 = \mathbf{I}_p. \tag{2.90}$$

Similarly to (2.28), for $\ell, k \in \mathbb{N}$ and $i \in [n]$, consider $\mathbf{C}^{(i,\ell)}_k$ corresponding to the i-th diagonal block of $\mathbf{C}^{(\ell)}_k$ defined in (2.89), *i.e.*

$$\mathbf{C}^{(i,\ell)}_k = \mathbb{1}_{[N_i]}(k+1)\gamma_i\left\{\rho_i^{-1}\mathbf{I}_{d_i} + \int_0^1 \nabla^2 U_i((1-s)Y^{(i,\ell)}_{k\gamma_i} + s\tilde{Y}^{(i,\ell)}_{k\gamma_i})\,\mathrm{d}s\right\} \in \mathbb{R}^{d_i\times d_i}, \tag{2.91}$$

where, for any $\ell \in \mathbb{N}$ and $i \in [n]$, $(Y^{(i,\ell)}_{k\gamma_i}, \tilde{Y}^{(i,\ell)}_{k\gamma_i})_{k\in\mathbb{N}}$ is defined in (2.81).

Lemma 2.32. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n], \gamma_i < 1/\tilde{M}_i$. Then, for any $\ell, k \in \mathbb{N}$, the matrix $(\mathbf{I}_p - \mathbf{C}_k^{(\ell)})$ is invertible and in addition, for any $i \in [n]$, we have*

$$\|\mathbf{I}_{d_i} - \mathbf{C}_k^{(i,\ell)}\| \leq 1 - \gamma_i \tilde{m}_i,$$

where $\mathbf{C}_k^{(i,\ell)}$ is defined in (2.91).

Proof Let $i \in [n], \ell, k \in \mathbb{N}$. By Assumption 2.3, we have $\|\nabla^2 U_i\| \leq M_i$ which implies by (2.91) that $\|\mathbf{C}_k^{(i,\ell)}\| \leq \gamma_i \tilde{M}_i$. Since $\gamma_i < 1/\tilde{M}_i$, the matrix $\mathbf{I}_p - \mathbf{C}_k^{(i,\ell)}$ is invertible and so is $\mathbf{I}_p - \mathbf{C}_k^{(\ell)}$. In addition, following the same lines as the proof of Lemma 2.17 implies $\|\mathbf{I}_{d_i} - \mathbf{C}_k^{(i,\ell)}\| \leq \max\{|1 - \gamma_i \tilde{m}_i|, |1 - \gamma_i \tilde{M}_i|\} = 1 - \gamma_i \tilde{m}_i$. ■

For any $\ell, k \in \mathbb{N}, i \in [n]$, if $\gamma_i \in (0, 1/\tilde{M}_i)$, Lemma 2.32 shows the invertibility of the matrices $\mathbf{I}_p - \mathbf{C}_k^{(\ell)}$. Therefore, $\mathbf{M}_\infty^{(\ell)}$ is invertible, and we can define

$$\mathrm{T}_1^{(\ell)} = [\mathbf{M}_\infty^{(\ell)}]^{-1} + \sum_{k=0}^{\infty} [\mathbf{M}_\infty^{(\ell)}]^{-1} \mathbf{M}_{k+1}^{(\ell)} \mathbf{J}(k,0) \mathbf{D}_{\boldsymbol{N}}^{-1/2} \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2} \mathbf{P}_0 \mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}^{1/2} \mathbf{D}_{\boldsymbol{N}}^{1/2}, \tag{2.92}$$

$$\mathrm{T}_2^{(\ell)} = \sum_{k=0}^{\infty} \left\{ [\mathbf{M}_\infty^{(\ell)}]^{-1} \mathbf{M}_{k+1}^{(\ell)} \mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2} \int_0^{+\infty} \mathbf{J}(k,l) [\nabla V(\tilde{Y}_{k\boldsymbol{\gamma}+l}^{(\ell)}) - \nabla V(\tilde{Y}_{k\boldsymbol{\gamma}}^{(\ell)})] \, \mathrm{d}l \right\}. \tag{2.93}$$

Using these matrices, we have the following result.

Lemma 2.33. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n], \gamma_i < 1/\tilde{M}_i$. Then, for any $\ell \geq 1$,*

$$\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2} (\tilde{Z}_{\ell+1} - Z_{\ell+1}) = \mathrm{T}_1^{(\ell)} (\tilde{Z}_\ell - Z_\ell) - \mathrm{T}_2^{(\ell)},$$

where $(Z_\ell, \tilde{Z}_\ell)_{\ell \in \mathbb{N}}$ is defined in (2.80) and $\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}} = \mathrm{diag}(N_1 \gamma_1 \mathbf{I}_{d_1}, \ldots, N_n \gamma_n \mathbf{I}_{d_n}) \in \mathbb{R}^{p \times p}$.

Proof Let $i \in [n]$ and $\ell \geq 1$. Recall that V_i is defined in (2.29) and for $z \in \mathbb{R}^p$, denote $V(z) = \sum_{i=1}^n V_i(z_i)$. For any $k \in \mathbb{N}$, we have

$$\nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)}) - \nabla V_i(Y_{k\gamma_i}^{(i,\ell)}) = \left[\int_0^1 \nabla^2 V_i((1-s) Y_{k\gamma_i}^{(i,\ell)} + s \tilde{Y}_{k\gamma_i}^{(i,\ell)}) \, \mathrm{d}s \right] (\tilde{Y}_{k\gamma_i}^{(i,\ell)} - Y_{k\gamma_i}^{(i,\ell)}).$$

For $k \geq 0$, it follows from (2.81) that

$$\begin{aligned} \tilde{Y}_{(k+1)\gamma_i}^{(i,\ell)} - Y_{(k+1)\gamma_i}^{(i,\ell)} = & \left(\mathbf{I}_{d_i} - \gamma_i \int_0^1 \nabla^2 V_i((1-s) Y_{k\gamma_i}^{(i,\ell)} + s \tilde{Y}_{k\gamma_i}^{(i,\ell)}) \, \mathrm{d}s \right) (\tilde{Y}_{k\gamma_i}^{(i,\ell)} - Y_{k\gamma_i}^{(i,\ell)}) \\ & - \int_0^{\gamma_i} \left[\nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)}) \right] \mathrm{d}l + (\gamma_i/\rho_i) \mathbf{A}_i (\tilde{\theta}_\ell - \theta_\ell). \end{aligned}$$

Consider the process $(\tilde{\mathrm{Y}}_t^{(\ell)}, \mathrm{Y}_t^{(\ell)})_{t \in \mathbb{R}_+}$ valued in $\mathbb{R}^p \times \mathbb{R}^p$ and defined for any $t \geq 0$ by

$$\tilde{\mathrm{Y}}_t^{(\ell)} = \tilde{Y}_{\min(t, N_i \gamma_i)}^{(\ell)}, \qquad \mathrm{Y}_t^{(\ell)} = Y_{\min(t, N_i \gamma_i)}^{(\ell)}. \tag{2.94}$$

The process (2.94) is continuous with respect to t and defined so that its component $(\tilde{\mathrm{Y}}_t^{(i,\ell)}, \mathrm{Y}_t^{(i,\ell)})$ equals $(\tilde{Y}_t^i, Y_t^i)$ for $t \leq N_i \gamma_i$ and is constant for $t > N_i \gamma_i$. For $l \geq 0$, we write $(\tilde{\mathrm{Y}}_{k\boldsymbol{\gamma}+l}^{(\ell)}, \mathrm{Y}_{k\boldsymbol{\gamma}+l}^{(\ell)}) = (\tilde{\mathrm{Y}}_{k\gamma_i+l}^{(i,\ell)}, \mathrm{Y}_{k\gamma_i+l}^{(i,\ell)})_{i \in [n]} \in \mathbb{R}^p \times \mathbb{R}^p$. Using the matrices defined in (2.90), for $k \in \mathbb{N}$, we obtain

$$\tilde{\mathrm{Y}}^{(\ell)}_{(k+1)\boldsymbol{\gamma}} - \mathrm{Y}^{(\ell)}_{(k+1)\boldsymbol{\gamma}} = (\mathbf{I}_p - \mathbf{C}^{(\ell)}_k)(\tilde{\mathrm{Y}}^{(\ell)}_{k\boldsymbol{\gamma}} - \mathrm{Y}^{(\ell)}_{k\boldsymbol{\gamma}}) - \int_0^\infty \mathbf{J}(k,l)\left[\nabla V(\tilde{\mathrm{Y}}^{(\ell)}_{k\boldsymbol{\gamma}+l}) - \nabla V(\tilde{\mathrm{Y}}^{(\ell)}_{k\boldsymbol{\gamma}})\right] \mathrm{d}l + \mathbf{J}(k,0)\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{P}_0\tilde{\mathbf{D}}_0^{1/2}(\tilde{\mathrm{Y}}^{(\ell)}_0 - \mathrm{Y}^{(\ell)}_0), \quad (2.95)$$

where $\mathbf{P}_0$ is defined in (2.12). Recall the matrix $\mathbf{M}^{(\ell)}_k$ defined in (2.90) with $\mathbf{M}^{(\ell)}_0 = \mathbf{I}_p$ and for $k \geq 1$, $\mathbf{M}^{(\ell)}_k = (\mathbf{I}_p - \mathbf{C}^{(\ell)}_0)^{-1} \dots (\mathbf{I}_p - \mathbf{C}^{(\ell)}_{k-1})^{-1}$. By multiplying (2.95) by $\mathbf{M}^{(\ell)}_{k+1}\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}$, we have

$$\begin{aligned}\mathbf{M}^{(\ell)}_{k+1}\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}(\tilde{\mathrm{Y}}^{(\ell)}_{(k+1)\boldsymbol{\gamma}} - \mathrm{Y}^{(\ell)}_{(k+1)\boldsymbol{\gamma}}) &= \mathbf{M}^{(\ell)}_k\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}(\tilde{\mathrm{Y}}^{(\ell)}_{k\boldsymbol{\gamma}} - \mathrm{Y}^{(\ell)}_{k\boldsymbol{\gamma}}) \\ &\quad - \mathbf{M}^{(\ell)}_{k+1}\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}\int_0^\infty \mathbf{J}(k,l)\left[\nabla V(\tilde{Y}^{(\ell)}_{k\boldsymbol{\gamma}+l}) - \nabla V(\tilde{Y}^{(\ell)}_{k\boldsymbol{\gamma}})\right] \mathrm{d}l \\ &\quad + \mathbf{M}^{(\ell)}_{k+1}\mathbf{J}(k,0)\mathbf{D}^{-1/2}_{\boldsymbol{N}}\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0\tilde{\mathbf{D}}_0^{1/2}(\tilde{\mathrm{Y}}^{(\ell)}_0 - \mathrm{Y}^{(\ell)}_0).\end{aligned}$$

By (2.94) and (2.80), we have for $t \geq \max_{i\in[n]}\{\gamma_i N_i\}$, $(\tilde{Z}_{\ell+1}, Z_{\ell+1}) = (\tilde{\mathrm{Y}}_t, \mathrm{Y}_t)$. Therefore, summing the previous expression over k, we get

$$\begin{aligned}\mathbf{M}^{(\ell)}_\infty\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}(\tilde{Z}_{\ell+1} - Z_{\ell+1}) &= -\sum_{k=0}^\infty \mathbf{M}^{(\ell)}_{k+1}\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}\int_0^\infty \mathbf{J}(k,l)[\nabla V(\tilde{Y}^{(\ell)}_{k\boldsymbol{\gamma}+l}) - \nabla V(\tilde{Y}^{(\ell)}_{k\boldsymbol{\gamma}})]\,\mathrm{d}l \\ &\quad + \left[\mathbf{M}^{(\ell)}_0 + \sum_{k=0}^\infty \mathbf{M}^{(\ell)}_{k+1}\mathbf{J}(k,0)\mathbf{D}^{-1/2}_{\boldsymbol{N}}\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{D}^{1/2}_{\boldsymbol{N}}\right]\mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}\cdot(\tilde{Z}_\ell - Z_\ell).\end{aligned}$$

By Lemma 2.32, $\mathbf{M}^{(\ell)}_\infty$ is invertible, and the proof is concluded by multiplying the previous equality by $[\mathbf{M}^{(\ell)}_\infty]^{-1}$. ∎

Based on Lemma 2.33, we have the following relation between $\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2$ and $\|\tilde{Z}_\ell - Z_\ell\|^2$.

Lemma 2.34. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}^*_+)^n$ such that, for any $i \in [n], \gamma_i < 1/\tilde{M}_i$. Then, for any $\epsilon > 0$ and $\ell \geq 1$,*

$$\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \leq (1+2\epsilon)\|\mathrm{T}^{(\ell)}_1\|^2\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} + (1 + 1/\{2\epsilon\})\|\mathrm{T}^{(\ell)}_2\|^2,$$

where $(Z_\ell, \tilde{Z}_\ell)_{\ell\in\mathbb{N}}$ is defined in (2.80) and $\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}} = \mathrm{diag}(N_1\gamma_1\mathbf{I}_{d_1}, \dots, N_n\gamma_n\mathbf{I}_{d_n}) \in \mathbb{R}^{p\times p}$.

Proof The proof follows from Lemma 2.33 and by using the fact that for $a, b \in \mathbb{R}^p, \epsilon > 0$ we have $2\langle a, b\rangle \leq 2\epsilon\|a\|^2 + (1/\{2\epsilon\})\|b\|^2$. ∎

Similarly to Lemma 2.18, we have the following result regarding the contracting term.

Lemma 2.35. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}^*_+)^n$ such that, for any $i \in [n]$, $\gamma_i < 1/\tilde{M}_i$ and $N_i\gamma_i \leq 2/(m_i + \tilde{M}_i)$. Then, for any $\ell \geq 0$, we have*

$$\|\mathrm{T}^{(\ell)}_1\| \leq 1 - \min_{i\in[n]}\{N_i\gamma_i m_i\} + r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}},$$

where $\mathrm{T}^{(\ell)}_1$ and $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ are defined in (2.92) and (2.42), respectively.

Proof The proof is similar to the proof of Lemma 2.18 and therefore is omitted. ■

In the next lemma, we upper bound the coefficient $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ defined in (2.42). For this, we explicit a choice of $\boldsymbol{N}$ that we denote $\boldsymbol{N}^\star = (N_1^\star(\gamma_1), \dots, N_n^\star(\gamma_n)) \in (\mathbb{N}^\star)^n$ defined for any $i \in [n]$, any $\gamma_i > 0$, by

$$N_i^\star(\gamma_i) = \left\lfloor m_i \min_{i\in[n]}\{m_i/\tilde{M}_i\}^2 / \left(20\gamma_i \tilde{M}_i^2 \max_{i\in[n]}\{m_i/\tilde{M}_i\}^2\right) \right\rfloor, \tag{2.96}$$

where $\tilde{M}_i = M_i + 1/\rho_i$.

Lemma 2.36. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n]$,*

$$\gamma_i \le \frac{m_i}{40\tilde{M}_i^2} \left(\frac{\min_{i\in[n]}\{m_i/\tilde{M}_i\}}{\max_{i\in[n]}\{m_i/\tilde{M}_i\}} \right)^2 .$$

Then, for any $i \in [n]$, we have $N_i^\star(\gamma_i) \in \mathbb{N}^\star$ and

$$r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}^\star} < \min_{i\in[n]}\{N_i^\star(\gamma_i)\gamma_i m_i\}/2,$$

where $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}^\star}$ is defined in (2.42).

Proof The assumption on γ_i combined with the definition (2.96) of $N_i^\star(\gamma_i)$ imply $N_i^\star(\gamma_i) \ge 2$, using in addition $m_i \le M_i$, $\max_{i\in[n]}\{N_i^\star(\gamma_i)\gamma_i\tilde{M}_i \mathbb{1}_{N_i^\star(\gamma_i)>1}\} \le 1/20$ and

$$\begin{aligned}\frac{1}{20}\left(\frac{\min_{i\in[n]}\{m_i/\tilde{M}_i\}}{\max_{i\in[n]}\{m_i/\tilde{M}_i\}} \right)^2 \ge \frac{N_i^\star(\gamma_i)\gamma_i\tilde{M}_i^2}{m_i} &> \frac{1}{20}\left(\frac{\min_{i\in[n]}\{m_i/\tilde{M}_i\}}{\max_{i\in[n]}\{m_i/\tilde{M}_i\}} \right)^2 - \frac{\gamma_i\tilde{M}_i^2}{m_i} \\ &\ge \frac{1}{40}\left(\frac{\min_{i\in[n]}\{m_i/\tilde{M}_i\}}{\max_{i\in[n]}\{m_i/\tilde{M}_i\}} \right)^2 . \end{aligned} \tag{2.97}$$

Using the definition (2.42) of $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}^\star}$, we have $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}^\star} < 5\max_{i\in[n]}\{N_i^\star(\gamma_i)\gamma_i\tilde{M}_i\mathbb{1}_{N_i^\star(\gamma_i)>1}\}^2$. Thus, plugging (2.97) in the previous inequality gives

$$r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}^\star} \le \max_{i\in[n]}\{m_i/\tilde{M}_i\}^2 \max_{i\in[n]}\left\{ \frac{N_i^\star(\gamma_i)\gamma_i\tilde{M}_i^2}{m_i} \right\} < \frac{\min_{i\in[n]}\{m_i/\tilde{M}_i\}^4}{80\max_{i\in[n]}\{m_i/\tilde{M}_i\}^2}. \tag{2.98}$$

In addition, (2.97) also shows that

$$\frac{1}{40}\left(\frac{\min_{i\in[n]}\{m_i/\tilde{M}_i\}}{\max_{i\in[n]}\{m_i/\tilde{M}_i\}} \right)^2 \left(\frac{m_i}{\tilde{M}_i} \right)^2 \le N_i^\star(\gamma_i)\gamma_i m_i. \tag{2.99}$$

Therefore, combining (2.98) and (2.99) completes the proof. ■

2.D.2 Proof of Proposition 2.6

We first give the formal statement of Proposition 2.6. For this, consider for any $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$, $i \in [n]$,

$$N_i^\star(\gamma_i) = \left\lfloor m_i \min_{i\in[n]}\{m_i/\tilde{M}_i\}^2 / \left(20\gamma_i \tilde{M}_i^2 \max_{i\in[n]}\{m_i/\tilde{M}_i\}^2\right) \right\rfloor, \tag{2.100}$$

and denote $\boldsymbol{N}^\star = (N_1^\star(\gamma_1), \dots, N_n^\star(\gamma_n))$.

Proposition 2.37. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that for any $i \in [n]$, $\gamma_i \le m_i/40\tilde{M}_i^2(\min_{i\in[n]}\{m_i/\tilde{M}_i\}/\max_{i\in[n]}\{m_i/\tilde{M}_i\})^2$. Then, we have*

$$W_2^2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}^\star}, \Pi_{\boldsymbol{\rho}}) \le \frac{4(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\max_{i\in[n]}\{m_i/\tilde{M}_i^2\}}{5\min_{i\in[n]}\{m_i/\tilde{M}_i\}^2 \max_{i\in[n]}\{m_i/\tilde{M}_i\}^2} \times \sum_{i=1}^n d_i\gamma_i m_i(1+\gamma_i^2\tilde{M}_i^2/12 + \gamma_i\tilde{M}_i^2/(2\tilde{m}_i)),$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ are defined in (2.11)-(2.12), and for any $i \in [n]$, $\tilde{m}_i$, $\tilde{M}_i$ are defined in (2.34).

By Lemma 2.31 and Lemma 2.34, we can note that the proof of Proposition 2.37 boils down to derive an upper bound on $\|\mathrm{T}_2^{(\ell)}\|^2$ defined in (2.93) for $\ell \in \mathbb{N}$. The following lemma provides such a bound.

Lemma 2.38. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n]$, $\gamma_i < 1/\tilde{M}_i$. Then, for any $\ell \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\|\mathrm{T}_2^{(\ell)}\|^2\right] \le \sum_{i=1}^n d_i N_i \gamma_i^2 \tilde{M}_i^2 \left[1 + \gamma_i^2\tilde{M}_i^2/12 + \gamma_i\tilde{M}_i^2/(2\tilde{m}_i)\right],$$

where $\tilde{m}_i, \tilde{M}_i, \mathrm{T}_2^{(\ell)}$ are defined in (2.34) and (2.93), respectively.

Proof Let $\ell \in \mathbb{N}$. Using (2.87), we can write, for any $l \in \mathbb{R}_+$ and $k \in \mathbb{N}$, $\mathbf{J}(k,l)$ as a block-diagonal matrix $\mathrm{diag}(\mathbf{J}^1(k,l), \dots, \mathbf{J}^n(k,l))$ with $\mathbf{J}^i(k,l) = \mathbb{1}_{[N_i]}(k+1)\mathbb{1}_{[0,\gamma_i]}(s)\cdot\mathbf{I}_{d_i}$ for any $i \in [n]$. By (2.90) and using for any $k \in \mathbb{N}$, that $[\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_{k+1}^{(\ell)} = \prod_{l=k+1}^\infty(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)})$ is finite by (2.89), we have

$$\begin{aligned}\|\mathrm{T}_2^{(\ell)}\|^2 &= \left\|\sum_{k=0}^\infty [\mathbf{M}_\infty^{(\ell)}]^{-1}\mathbf{M}_{k+1}^{(\ell)}\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1/2}\int_0^\infty \mathbf{J}(k,l)\left[\nabla V(\tilde{Y}_{k\boldsymbol{\gamma}+l}^{(\ell)}) - \nabla V(\tilde{Y}_{k\boldsymbol{\gamma}}^{(\ell)})\right]\mathrm{d}l\right\|^2 \\ &= \sum_{i=1}^n \frac{1}{N_i\gamma_i}\left\|\sum_{k=0}^\infty\prod_{l=k+1}^\infty(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)})\int_0^{\gamma_i}\mathbf{J}^i(k,0)\left[\nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)})\right]\mathrm{d}l\right\|^2.\end{aligned} \tag{2.101}$$

Since for any $i \in [n]$, $k \ge N_i$ we have $\mathbf{J}^i(k,0) = \mathbf{C}_l^{(i,\ell)} = \mathbf{0}_{d_i\times d_i}$, (2.101) can be rewritten as

$$\|\mathrm{T}_2^{(\ell)}\|^2 = \sum_{i=1}^n \frac{1}{N_i\gamma_i}\left\|\sum_{k=0}^{N_i-1}\prod_{l=k+1}^{N_i-1}(\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)})\int_0^{\gamma_i}\mathbf{J}^i(k,0)\left[\nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)})\right]\mathrm{d}l\right\|^2,$$

and the Cauchy-Schwarz inequality gives

$$\|\mathrm{T}_2^{(\ell)}\|^2 \leq \sum_{i=1}^{n} \frac{1}{\gamma_i} \left(\sum_{k=0}^{N_i-1} \left\| \prod_{l=k+1}^{N_i-1} (\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)}) \right\|^2 \left\| \int_0^{\gamma_i} \left[\nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)}) \right] \mathrm{d}l \right\|^2 \right). \tag{2.102}$$

Since, for any $i \in [n]$, $\gamma_i \tilde{M}_i < 1$, we get using Lemma 2.32,

$$\left\| \prod_{l=k+1}^{N_i-1} (\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)}) \right\|^2 \leq \{1 - \gamma_i \tilde{m}_i\}^{2(N_i-k-1)}.$$

By combining (2.102) with the previous result and the Jensen inequality, we have

$$\|\mathrm{T}_2^{(\ell)}\|^2 \leq \sum_{i=1}^{n} \sum_{k=0}^{N_i-1} \{1 - \gamma_i \tilde{m}_i\}^{2(N_i-k-1)} \int_0^{\gamma_i} \left\| \nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)}) \right\|^2 \mathrm{d}l. \tag{2.103}$$

For $i \in [n]$, using Durmus and Moulines (2019, Lemma 21) applied to the potential $V_i^\theta : \mathbf{y}^i \mapsto U_i(\mathbf{y}^i) + \|\mathbf{y}^i - \mathbf{A}_i\theta\|^2/(2\rho_i)$ yields

$$\begin{aligned} \int_0^{\gamma_i} \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \|\nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)})\|^2 \,\mathrm{d}l &= \int_0^{\gamma_i} \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \|\nabla V_i^{\tilde{\theta}_\ell}(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i^{\tilde{\theta}_\ell}(\tilde{Y}_{k\gamma_i}^{(i,\ell)})\|^2 \,\mathrm{d}l \\ &\leq \gamma_i^2 \tilde{M}_i^2 \left[d_i + d_i \gamma_i^2 \tilde{M}_i^2/12 + (\gamma_i \tilde{M}_i^2/2) \|\tilde{Y}_{k\gamma_i}^{(i,\ell)} - z_{\ell,\star}^i\|^2 \right], \end{aligned} \tag{2.104}$$

where $z_{\ell,\star}^i = \arg\min_{z_i \in \mathbb{R}^{d_i}} V_i^{\tilde{\theta}_\ell}(z_i)$.

By (2.104), (2.85), Lemma 2.30 and since $\max_{i\in[n]} \gamma_i \tilde{m}_i < 1$, we get

$$\begin{aligned} \sum_{i=1}^{n} \sum_{k=0}^{N_i-1} \{1 - \gamma_i \tilde{m}_i\}^{2(N_i-k-1)} \int_0^{\gamma_i} \mathbb{E} \|\nabla V_i(\tilde{Y}_{k\gamma_i+l}^{(i,\ell)}) - \nabla V_i(\tilde{Y}_{k\gamma_i}^{(i,\ell)})\|^2 \,\mathrm{d}l \\ \leq \sum_{i=1}^{n} d_i N_i \gamma_i^2 \tilde{M}_i^2 [1 + \gamma_i^2 \tilde{M}_i^2/12 + \gamma_i \tilde{M}_i^2/(2\tilde{m}_i)]. \end{aligned}$$

Combining this result with (2.103) completes the proof. ∎

We can now combine Lemma 2.38 and Lemma 2.35 with Lemma 2.34 to get the following bound.

Lemma 2.39. *Assume Assumption 2.1-Assumption 2.3 and let* $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ *such that, for any* $i \in [n]$, $\gamma_i < 1/\tilde{M}_i$, $N_i\gamma_i \leq 2/(m_i + \tilde{M}_i)$. *Suppose in addition* $\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} = \min_{i\in[n]}\{N_i\gamma_i m_i\} - r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} \in (0,1)$, *where* $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ *is defined in (2.42). Then, for* $\ell \geq 1$, *we have*

$$\begin{aligned} \mathbb{E}\left[\|\tilde{Z}_\ell - Z_\ell\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}^2 \right] &\leq (1 - \kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} + \kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}^2/2)^{2(\ell-1)} \mathbb{E}\left[\|\tilde{Z}_1 - Z_1\|_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}^2 \right] \\ &\quad + 2\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}^{-2} \sum_{i=1}^{n} d_i N_i \gamma_i^2 \tilde{M}_i^2 \left(1 + \frac{\gamma_i^2 \tilde{M}_i^2}{12} + \frac{\gamma_i \tilde{M}_i^2}{2\tilde{m}_i} \right), \end{aligned}$$

where, for any $i \in [n]$, $\tilde{M}_i$ *and* $\tilde{m}_i$ *are defined in (2.34).*

Proof Taking expectation in Lemma 2.34, we get for any $\ell \in \mathbb{N}, \epsilon > 0$ that

$$\mathbb{E}\left[\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \leq (1+2\epsilon)\mathbb{E}\left[\|\mathrm{T}_1^{(\ell)}\|^2\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] + (1+1/\{2\epsilon\})\mathbb{E}\left[\|\mathrm{T}_2^{(\ell)}\|^2\right],$$

where $\mathrm{T}_1^{(\ell)}$ and $\mathrm{T}_2^{(\ell)}$ are defined in (2.92) and (2.93), respectively. To ease notation, denote $\mathrm{B} = \sum_{i=1}^n d_i N_i \gamma_i^2 \tilde{M}_i^2 (1 + \gamma_i^2 \tilde{M}_i^2/12 + \gamma_i \tilde{M}_i^2/(2\tilde{m}_i))$. Using Lemma 2.38, we obtain for any $\ell \in \mathbb{N}, \epsilon > 0$

$$\mathbb{E}\left[\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \leq (1+2\epsilon)\mathbb{E}\left[\|\mathrm{T}_1^{(\ell)}\|^2\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] + (1+1/\{2\epsilon\})\mathrm{B}. \quad (2.105)$$

In addition, Lemma 2.35 implies that $\|\mathrm{T}_1^{(\ell)}\|^2 \leq (1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2$ almost surely. Therefore, taking $\epsilon = (1-[1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}]^2)/(4[1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}]^2)$, (2.105) yields for any $\ell \geq 0$,

$$\mathbb{E}\left[\left\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\right\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \leq \frac{1+(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2}{2}\mathbb{E}\left[\left\|\tilde{Z}_\ell - Z_\ell\right\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] + \frac{1+(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2}{1-(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2}\mathrm{B}.$$

An easy induction implies for any $\ell \geq 1$,

$$\mathbb{E}\left[\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \leq \left(\frac{1+(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2}{2}\right)^{\ell-1}\mathbb{E}\left[\|\tilde{Z}_1 - Z_1\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] + 2\frac{1+(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2}{(1-(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2)^2}\mathrm{B}. \quad (2.106)$$

Since $\kappa^2_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} = (\min_{i\in[n]}\{N_i\gamma_i m_i\} + r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2$ and using $\kappa^2_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} \leq 1$, we obtain

$$\begin{aligned}(1+(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2)/2 &= 1 - \kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} + \kappa^2_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}/2,\\ (1+(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2)/(1-(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^2)^2 &\leq \kappa^{-2}_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}.\end{aligned}$$

Combining these inequalities with (2.106) and (2.105) completes the proof. ∎

Lemma 2.40. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ such that, for any $i \in [n]$, $\gamma_i < 1/\tilde{M}_i$, $N_i\gamma_i \leq 2/(m_i+\tilde{M}_i)$ and $\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} = \min_{i\in[n]}\{N_i\gamma_i m_i\} - r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} \in (0,1)$, where $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ is defined in (2.42). Then, for any $\mathbf{x} \in \mathbb{R}^{d+p}$ and $\ell \geq 1$, we have*

$$\begin{aligned}&W_2^2(\delta_{\mathbf{x}}P^\ell_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho}})\\ &\leq (1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}} + \kappa^2_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}/2)^{2(\ell-1)}(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)\max_{i\in[n]}\{N_i\gamma_i\}\mathbb{E}\left[\|\tilde{Z}_1 - Z_1\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right]\\ &\quad + \frac{2(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|^2)\max_{i\in[n]}\{N_i\gamma_i\}}{\kappa^2_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}}\sum_{i=1}^n d_i N_i\gamma_i^2\tilde{M}_i^2[1+\gamma_i^2\tilde{M}_i^2/12 + \gamma_i\tilde{M}_i^2/(2\tilde{m}_i)],\end{aligned}$$

where $\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ are defined in (2.11)-(2.12), $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ is defined in (2.26), $(\tilde{Z}_\ell, Z_\ell)_{\ell\in\mathbb{N}}$ is defined in (2.80) and for any $i \in [n]$, $\tilde{M}_i$, $\tilde{m}_i$ are defined in (2.34).

Proof By Lemma 2.39, we have the following upper bound for $\ell \geq 1$,

$$\mathbb{E}\left[\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \leq (1 - \kappa_{\gamma,\rho,N} + \kappa^2_{\gamma,\rho,N}/2)^{2(\ell-1)}\mathbb{E}\left[\|\tilde{Z}_1 - Z_1\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] + 2\kappa^{-2}_{\gamma,\rho,N}\sum_{i=1}^n d_i N_i \gamma_i^2 \tilde{M}_i^2\left(1 + \frac{\gamma_i^2\tilde{M}_i^2}{12} + \frac{\gamma_i\tilde{M}_i^2}{2\tilde{m}_i}\right).$$

Using (2.80), Lemma 2.31, combined with the previous inequality, we get for any $\ell \geq 1, \mathbf{x} \in \mathbb{R}^{d+p}$,

$$\begin{aligned}
&W_2^2(\Pi_\rho, \delta_{\mathbf{x}} P^\ell_{\rho,\gamma,N}) \leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\mathbb{E}\left[\|\tilde{Z}_\ell - Z_\ell\|^2\right] \\
&\leq (1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\max_{i\in[n]}\{N_i\gamma_i\}\mathbb{E}\left[\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \\
&\leq (1 - \kappa_{\gamma,\rho,N} + \kappa^2_{\gamma,\rho,N}/2)^{2(\ell-1)}(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\max_{i\in[n]}\{N_i\gamma_i\}\mathbb{E}\left[\|\tilde{Z}_1 - Z_1\|^2_{\mathbf{D}^{-1}_{N\gamma}}\right] \\
&+ \frac{2(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\max_{i\in[n]}\{N_i\gamma_i\}}{\kappa^2_{\gamma,\rho,N}}\sum_{i=1}^n d_i N_i \gamma_i^2 \tilde{M}_i^2\left(1 + \frac{\gamma_i^2\tilde{M}_i^2}{12} + \frac{\gamma_i\tilde{M}_i^2}{2\tilde{m}_i}\right).
\end{aligned}$$

Hence, the stated result. ■

Proof of Proposition 2.6/Proposition 2.37. **Proof** Since for any $i \in [n]$, we know that

$$\gamma_i \leq m_i/40\tilde{M}_i^2(\min_{i\in[n]}\{m_i/\tilde{M}_i\}/\max_{i\in[n]}\{m_i/\tilde{M}_i\})^2,$$

setting

$$N_i^\star(\gamma_i) = \left\lfloor m_i \min_{i\in[n]}\{m_i/\tilde{M}_i\}^2/\left(20\gamma_i\tilde{M}_i^2\max_{i\in[n]}\{m_i/\tilde{M}_i\}^2\right)\right\rfloor$$

implies $\kappa_{\gamma,\rho,N^\star} \in (0,1)$ by Lemma 2.36. Thereby, letting n tend towards infinity in Lemma 2.40 and using Proposition 2.21 conclude the proof. ■

2.D.3 Proof of Proposition 2.8

We first give the formal statement of Proposition 2.8.

Proposition 2.41. *Assume Assumption 2.1-Assumption 2.3-Assumption 2.7 and let* $\rho \in (\mathbb{R}_+^*)^n, \gamma \in (\mathbb{R}_+^*)^n$ *such that for any* $i \in [n]$,

$$\gamma_i \leq m_i/(40\tilde{M}_i^2)(\min_{i\in[n]}\{m_i/\tilde{M}_i\}/\max_{i\in[n]}\{m_i/\tilde{M}_i\})^2.$$

Then, we have

$$W_2^2(\Pi_{\rho,\gamma,N^\star}, \Pi_\rho) \leq 4(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)\frac{\max_{i\in[n]}\{m_i/\tilde{M}_i^2\}}{\min_{i\in[n]}\{m_i/\tilde{M}_i\}^2}\mathcal{R}^\star(\gamma),$$

where setting $\mathfrak{f}_i = m_i/(20\tilde{M}_i)$,

$$\mathcal{R}^\star(\gamma) = \sum_{i=1}^n \left\{ d_i\gamma_i^2\tilde{M}_i^2 + \frac{d_i\gamma_i^2\mathfrak{f}_i}{\tilde{M}_i}\left(d_iL_i^2 + \frac{\tilde{M}_i^4}{\tilde{m}_i}\right) + d_i\gamma_i\tilde{M}_i\mathfrak{f}_i^3(1 + \mathfrak{f}_i + \mathfrak{f}_i^2)\right\},$$

$\bar{\mathbf{B}}_0, \mathbf{B}_0, \tilde{\mathbf{D}}_0$ *are defined in* (2.11)-(2.12)*, and for any* $i \in [n]$*,* $\tilde{m}_i$*,* $\tilde{M}_i$ *are defined in* (2.34)*.*

We provide the proof of Proposition 2.8 in what follows. Similarly to Lemma 2.34 for the proof of Proposition 2.6, we derive an explicit relation between $\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|$ and $\|\tilde{Z}_\ell - Z_\ell\|$.

Lemma 2.42. *Assume Assumption 2.1-Assumption 2.3-Assumption 2.7 and let* $\boldsymbol{N} \in (\mathbb{N}^\star)^n, \boldsymbol{\gamma} \in (\mathbb{R}_+^*)^n$ *such that for any* $i \in [n]$*,* $N_i\gamma_i \le 2/(m_i + \tilde{M}_i)$ *and* $\gamma_i < 1/\tilde{M}_i$*. Then, for* $\ell \ge 1$*, we have*

$$\mathbb{E}\Big[\|\tilde{Z}_{\ell+1} - Z_{\ell+1}\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}}\Big]^{1/2} \le \Big(1 - \min_{i\in[n]}\{N_i\gamma_i m_i\} + r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}\Big)\mathbb{E}\Big[\|\tilde{Z}_\ell - Z_\ell\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}}\Big]^{1/2} + \mathcal{R}(\boldsymbol{\gamma}, \boldsymbol{N})^{1/2},$$

where

$$\mathcal{R}(\boldsymbol{\gamma}, \boldsymbol{N}) = \sum_{i=1}^n d_i N_i \gamma_i^3 (d_i L_i^2 + \tilde{M}_i^4/\tilde{m}_i) + \sum_{i=1}^n \Big(d_i\gamma_i^2\tilde{M}_i^2 + d_i N_i^3\gamma_i^4\tilde{M}_i^4\Big) + \sum_{i=1}^n d_i N_i^4\gamma_i^5\tilde{M}_i^5(1 + N_i\gamma_i\tilde{M}_i), \quad (2.107)$$

$(\tilde{Z}_\ell, Z_\ell)_{\ell\in\mathbb{N}}$ *is defined in* (2.80)*,* $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ *in* (2.42) *and for any* $i \in [n]$*,* $\tilde{m}_i$*,* $\tilde{M}_i$ *are defined in* (2.34)*.*

Proof Let $\ell \in \mathbb{N}$. For any $k \in \mathbb{N}$, recall that $\mathbf{M}_k^{(\ell)}$ is defined in (2.90) and invertible by Lemma 2.32. Define

$$w_\ell = \mathbf{D}^{-1/2}_{\boldsymbol{N}\boldsymbol{\gamma}}(\tilde{Z}_\ell - Z_\ell).$$

Under this notation, the result given in Lemma 2.33 can be rewritten as

$$w_{\ell+1} = \mathrm{T}_1^{(\ell)} w_\ell - \mathrm{T}_2^{(\ell)},$$

where $\mathrm{T}_1^{(\ell)}$ and $\mathrm{T}_2^{(\ell)}$ are defined in (2.92) and (2.93), respectively. By the Minkowsky inequality and using (2.82), we have

$$\mathbb{E}^{\mathcal{G}_\ell}\Big[\|w_{\ell+1}\|^2\Big]^{1/2} \le \mathbb{E}^{\mathcal{G}_\ell}\Big[\|\mathrm{T}_1^{(\ell)} w_\ell\|^2\Big]^{1/2} + \mathbb{E}^{\mathcal{G}_\ell}\Big[\|\mathrm{T}_2^{(\ell)}\|^2\Big]^{1/2}. \quad (2.108)$$

Since by Lemma 2.35,

$$\|\mathrm{T}_1^{(\ell)}\| \le 1 - \min_{i\in[n]}\{N_i\gamma_i m_i\} + r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}, \quad (2.109)$$

it remains to bound $\mathbb{E}^{\mathcal{G}_\ell}[\|\mathrm{T}_2^{(\ell)}\|^2]$ to complete the proof.

For any $i \in [n]$, recall the function $V_i^{\theta_\ell} : \mathbb{R}^{d_i} \to \mathbb{R}$ defined for any $\mathbf{y}^i \in \mathbb{R}^{d_i}$ by $V_i^{\theta_\ell}(\mathbf{y}^i) = U_i(\mathbf{y}^i) + \|\mathbf{y}^i - \mathbf{A}_i\theta_\ell\|^2/(2\rho_i)$. Let $k \in \mathbb{N}$, using the Itô formula, we have for $l \in [k\gamma_i, (k+1)\gamma_i)$,

$$\nabla V_i(\tilde{Y}^{(i,\ell)}_{k\gamma_i+l}) - \nabla V_i(\tilde{Y}^{(i,\ell)}_{k\gamma_i}) = \int_{k\gamma_i}^{k\gamma_i+l} \left\{ \nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \nabla V_i^{\theta_\ell}(\tilde{Y}_u) + \vec{\Delta}(\nabla V_i^{\theta_\ell})(\tilde{Y}^{(i,\ell)}_u) \right\} \mathrm{d}u + \sqrt{2} \int_{k\gamma_i}^{k\gamma_i+l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \, \mathrm{d}B^i_u. \quad (2.110)$$

For any $i \in [n], k \in \mathbb{N}$, define

$$a^{(i,\ell)}_{1,k} = \mathbb{1}_{[N_i]}(k+1)[\mathbf{M}^{(i,\ell)}_\infty]^{-1}\mathbf{M}^{(i,\ell)}_{k+1} \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \nabla V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \, \mathrm{d}u \, \mathrm{d}l,$$

$$a^{(i,\ell)}_{2,k} = \mathbb{1}_{[N_i]}(k+1)[\mathbf{M}^{(i,\ell)}_\infty]^{-1}\mathbf{M}^{(i,\ell)}_{k+1} \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \vec{\Delta}(\nabla V_i^{\theta_\ell})(\tilde{Y}^{(i,\ell)}_u) \, \mathrm{d}u \, \mathrm{d}l,$$

$$a^{(i,\ell)}_{3,k} = \sqrt{2}\mathbb{1}_{[N_i]}(k+1)[\mathbf{M}^{(i,\ell)}_\infty]^{-1}\mathbf{M}^{(i,\ell)}_{k+1} \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \, \mathrm{d}B^i_u \, \mathrm{d}l.$$

With these notations and by (2.110), we have

$$\begin{aligned} \|T^{(\ell)}_2\|^2 &= \sum_{i\in[n]} \frac{1}{N_i\gamma_i} \left\| \sum_{k\in\mathbb{N}} \{a^{(i,\ell)}_{1,k} + a^{(i,\ell)}_{2,k} + a^{(i,\ell)}_{3,k}\} \right\|^2 \\ &\le E_1 + E_2 + E_3, \end{aligned} \quad (2.111)$$

where for any $j \in [3]$, $E_j = 3\sum_{i\in[n]} \| \sum_{k=0}^{N_i-1} a^{(i,\ell)}_{j,k} \|^2/(N_i\gamma_i)$. We now bound $\{E_j\}_{j\in[3]}$.

Upper bound on E_1. For any $i \in [n], k \in \mathbb{N}$, recall that we have $\left[\mathbf{M}^{(i,\ell)}_\infty\right]^{-1}\mathbf{M}^{(i,\ell)}_{k+1} = \prod_{l=k+1}^\infty (\mathbf{I}_{d_i} + \mathbf{C}^{(i,\ell)}_l)$ where $\mathbf{C}^{(i,\ell)}_l$ is defined in (2.89). In addition, since we suppose for any $i \in [n]$, that $\gamma_i\tilde{M}_i < 1$, Lemma 2.32 implies

$$\left\| \prod_{l=k+1}^{N_i-1} (\mathbf{I}_{d_i} - \mathbf{C}^{(i,\ell)}_l) \right\|^2 \le \left\{1 - \gamma_i\tilde{m}_i\right\}^{2(N_i-k-1)}.$$

Combining this result with the Cauchy-Schwarz inequality, we obtain

$$\frac{1}{N_i} \left\| \sum_{k=0}^{N_i-1} a^{(i,\ell)}_{1,k} \right\|^2 \le \sum_{k=0}^{N_i-1} \left\| \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \nabla V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \, \mathrm{d}u \, \mathrm{d}l \right\|^2. \quad (2.112)$$

For $i \in [n]$, using the definition of $z^i_{\ell,\star} = \arg\min_{\mathbf{y}^i\in\mathbb{R}^{d_i}} V_i^{\theta_\ell}(\mathbf{y}^i) \in \mathbb{R}^{d_i}$, we have $\nabla V_i^{\theta_\ell}(z^i_{\ell,\star}) = \mathbf{0}_{d_i}$. Therefore, for $i \in [n], k \in \mathbb{N}$, conditioning with respect to $\mathcal{F}^{(\ell)}_{k\gamma_i}$ defined in (2.83) and using the $\tilde{M}_i$-Lipschitz property of $V_i^{\theta_\ell}$ by Assumption 2.3 gives

$$\begin{aligned} \mathbb{E}^{\mathcal{F}^{(\ell)}_{k\gamma_i}} \left[\|\nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \nabla V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u)\|^2 \right] &\le \tilde{M}_i^2 \mathbb{E}^{\mathcal{F}^{(\ell)}_{k\gamma_i}} \left[\|\nabla V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) - \nabla V_i^{\theta_\ell}(z^i_{\ell,\star})\|^2 \right] \\ &\le \tilde{M}_i^4 \mathbb{E}^{\mathcal{F}^{(\ell)}_{k\gamma_i}} \left[\|\tilde{Y}^{(i,\ell)}_u - z^i_{\ell,\star}\|^2 \right]. \end{aligned}$$

For any $i \in [n], k \in \mathbb{N}$, combining this result with the Jensen inequality yields

$$\mathbb{E}^{\mathcal{F}^{(\ell)}_{k\gamma_i}} \left[\left\| \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \nabla V_i^{\theta_\ell}(\tilde{Y}^{(i,\ell)}_u) \, \mathrm{d}u \, \mathrm{d}l \right\| \right]^2$$

$$\leq \gamma_i \int_0^{\gamma_i} l \int_{k\gamma_i}^{k\gamma_i+l} \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \left[\|\nabla^2 V_i^{\theta_\ell}(\tilde{Y}_u^{(i,\ell)}) \nabla V_i^{\theta_\ell}(\tilde{Y}_u^{(i,\ell)})\|^2 \right] \mathrm{d}u \, \mathrm{d}l$$
$$\leq \gamma_i \tilde{M}_i^4 \int_0^{\gamma_i} l \int_{k\gamma_i}^{k\gamma_i+l} \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \left[\|\tilde{Y}_u^{(i,\ell)} - z_{\ell,\star}^i\|^2 \right] \mathrm{d}u \, \mathrm{d}l. \tag{2.113}$$

By Lemma 2.30, we have for any $i \in [n]$, $u \in \mathbb{R}_+$,

$$\mathbb{E}^{\mathcal{G}_\ell} \left[\|\tilde{Y}_u^{(i,\ell)} - z_{\ell,\star}^i\|^2 \right] \leq d_i / \tilde{m}_i.$$

Injecting this result in (2.113) yields

$$\mathbb{E} \left[\int_0^{\gamma_i} l \int_{k\gamma_i}^{k\gamma_i+l} \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \left[\|\tilde{Y}_u^{(i,\ell)} - z_{\ell,\star}^i\|^2 \right] \mathrm{d}u \, \mathrm{d}l \right] \leq d_i \gamma_i^3 / (3\tilde{m}_i).$$

Finally, this inequality, (2.113) and (2.112), we get

$$\mathbb{E}\left[E_1\right] \leq \sum_{i=1}^n d_i N_i \gamma_i^3 \tilde{M}_i^4 / \tilde{m}_i. \tag{2.114}$$

Upper bound on E_2. Using the Cauchy-Schwarz inequality, we have

$$\frac{1}{N_i} \left\| \sum_{k=0}^{N_i-1} a_{2,k}^{(i,\ell)} \right\|^2 \leq \sum_{k=0}^{N_i-1} \left\| \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \vec{\Delta}(\nabla V_i^{\theta_\ell})(\tilde{Y}_u^{(i,\ell)}) \, \mathrm{d}u \, \mathrm{d}l \right\|^2.$$

By Assumption 2.7, we have for any $z_i \in \mathbb{R}^{d_i}$, $\|\vec{\Delta}(\nabla V_i^{\theta_\ell})(z_i)\|^2 \leq d_i^2 L_i^2$. Therefore, we obtain

$$\left\| \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \vec{\Delta}(\nabla V_i^{\theta_\ell})(\tilde{Y}_u^{(i,\ell)}) \, \mathrm{d}u \, \mathrm{d}l \right\|^2 \leq \gamma_i \int_0^{\gamma_i} l \int_{k\gamma_i}^{k\gamma_i+l} \|\vec{\Delta}(\nabla V_i^{\theta_\ell})(\tilde{Y}_u^{(i,\ell)})\|^2 \, \mathrm{d}u \, \mathrm{d}l$$
$$\leq d_i^2 \gamma_i^4 L_i^2 / 3.$$

Thus, we get

$$\mathbb{E}\left[E_2\right] \leq \sum_{i=1}^n d_i^2 N_i \gamma_i^3 L_i^2. \tag{2.115}$$

Upper bound on E_3. For any $i \in [n], k \in \mathbb{N}$, define

$$\Delta_{3,k}^{(i,\ell)} = \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i+l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}_u^{(i,\ell)}) \, \mathrm{d}B_u^i \, \mathrm{d}l.$$

Using for any $i \in [n], k \in \mathbb{N}$, $[\mathbf{M}_\infty^{(i,\ell)}]^{-1} \mathbf{M}_{k+1}^{(i,\ell)} = \mathbf{I}_{d_i} - \sum_{l=k+1}^\infty \mathbf{C}_l^{(i,\ell)} + \mathbf{R}_k^{(i,\ell)}$ where $\mathbf{R}_k^{(i,\ell)}$ is defined in (2.43), we have, for any $i \in [n], k \in \mathbb{N}$,

$$\left\| \sum_{k=0}^{N_i-1} a_{3,k}^{(i,\ell)} \right\|^2 = \left\| \sqrt{2} \sum_{k=0}^{N_i-1} \prod_{l=k+1}^{N_i} \left[\mathbf{I}_{d_i} - \mathbf{C}_l^{(i,\ell)} \right] \Delta_{3,k}^{(i,\ell)} \right\|^2$$
$$= 2 \sum_{k_1,k_2=0}^{N_i-1} \langle \mathbf{R}_{k_1}^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \mathbf{R}_{k_2}^{(i,\ell)} \Delta_{3,k_2}^{(i,\ell)} \rangle + 2 \sum_{k_1,k_2=0}^{N_i-1} \langle \Delta_{3,k_1}^{(i,\ell)}, \Delta_{3,k_2}^{(i,\ell)} \rangle$$

$$
\begin{aligned}
&+ 2 \sum_{k_1,k_2=0}^{N_i-1} \langle \sum_{l=k_1+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \sum_{l=k_2+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_2}^{(i,\ell)} \rangle \\
&- 4 \sum_{k_1,k_2=0}^{N_i-1} \langle \sum_{l=k_1+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \Delta_{3,k_2}^{(i,\ell)} \rangle + 4 \sum_{k_1,k_2=0}^{N_i-1} \langle \mathbf{R}_{k_1}^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \Delta_{3,k_2}^{(i,\ell)} \rangle \\
&- 4 \sum_{k_1,k_2=0}^{N_i-1} \langle \mathbf{R}_{k_1}^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \sum_{l=k_2+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_2}^{(i,\ell)} \rangle . \qquad (2.116)
\end{aligned}
$$

We now control the quantities which appear in (2.116). First, by Assumption 2.3, for any $i \in [n], \mathbf{x}^i, \mathbf{y}^i \in \mathbb{R}^{d_i}$, note that we have

$$
\|\nabla^2 V_i^{\theta_\ell}(\mathbf{x}^i)\mathbf{y}^i\| \le \tilde{M}_i \|\mathbf{y}^i\| .
$$

By the Jensen inequality and the Itô isometry, for any $k \in \mathbb{N}$, we get

$$
\begin{aligned}
\mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \left[\|\Delta_{3,k}^{(i,\ell)}\|^2 \right] &= \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \left[\left\| \int_0^{\gamma_i} \int_{k\gamma_i}^{k\gamma_i + l} \nabla^2 V_i^{\theta_\ell}(\tilde{Y}_u^{(i,\ell)}) \, \mathrm{d}B_u^i \, \mathrm{d}l \right\|^2 \right] \\
&\le \gamma_i \tilde{M}_i^2 \int_0^{\gamma_i} \mathbb{E}^{\mathcal{F}_{k\gamma_i}^{(\ell)}} \left[\left\| \int_{k\gamma_i}^{k\gamma_i + l} \mathrm{d}B_u^i \right\|^2 \right] \mathrm{d}l = d_i \gamma_i^3 \tilde{M}_i^2 / 2 . \qquad (2.117)
\end{aligned}
$$

In addition, since for $i \in [n]$, $(\int_0^t \nabla^2 V_i^{\theta_\ell}(\tilde{Y}_u^{(i,\ell)}) \, \mathrm{d}B_u^i)_{t \ge 0}$ is a $(\mathcal{F}_t^{(\ell)})_{t \ge 0}$-martingale, for $(k_1, k_2) \in \{0, \dots, N_i - 1\}^2$ such that $k_1 < k_2$, we obtain

$$
\mathbb{E}^{\mathcal{G}_\ell} \left[\left[\Delta_{3,k_1}^{(i,\ell)} \right]^\top \Delta_{3,k_2}^{(i,\ell)} \right] = \mathbb{E}^{\mathcal{G}_\ell} \left[\mathbb{E}^{\mathcal{F}_{k_2\gamma_i}^{(\ell)}} \left[\Delta_{3,k_1}^{(i,\ell)\top} \Delta_{3,k_2}^{(i,\ell)} \right] \right] = 0 .
$$

Therefore,

$$
\sum_{k_1,k_2=0}^{N_i-1} \mathbb{E}^{\mathcal{G}_\ell} \left[\langle \Delta_{3,k_1}^{(i,\ell)}, \Delta_{3,k_2}^{(i,\ell)} \rangle \right] = d_i N_i \gamma_i^3 \tilde{M}_i^2 / 2 .
$$

Second, since for any $i \in [n], l \in \mathbb{N}, \mathbf{C}_l^{(i,\ell)} \in \mathbb{R}^{d_i \times d_i}$ is symmetric positive semi-definite, we have

$$
\sum_{k_1,k_2=0}^{N_i-1} \langle \sum_{l=k_1+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \Delta_{3,k_2}^{(i,\ell)} \rangle = \left\langle \left\{ \sum_{l=1}^{N} C_l \right\} \sum_{k_1=0}^{l-1} \Delta_{3,k_1}, \sum_{k_1=0}^{l-1} \Delta_{3,k_2} \right\rangle \ge 0 .
$$

Third, using for any $i \in [n], l \in \mathbb{N}$, using $\|\mathbf{C}_l^{(i,\ell)}\| \le \gamma_i \tilde{M}_i$ by definition (2.89) and Assumption 2.3 and combining the Cauchy-Schwarz inequality with (2.117), for any $i \in [n], (k_1, k_2) \in \{0, \dots, N_i - 1\}^2$, we get

$$
\sum_{k_1,k_2=0}^{N_i-1} \mathbb{E}^{\mathcal{G}_\ell} \left[\langle \sum_{l=k_1+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \sum_{l=k_2+1}^{N_i} \mathbf{C}_l^{(i,\ell)} \Delta_{3,k_2}^{(i,\ell)} \rangle \right] \le d_i N_i^4 \gamma_i^5 \tilde{M}_i^4 / 8 .
$$

Using (2.117) again and Lemma 2.16, for $i \in [n]$, we obtain

$$
\sum_{k_1,k_2=0}^{N_i-1} \mathbb{E}^{\mathcal{G}_\ell} \left[\langle \mathbf{R}_{k_1}^{(i,\ell)} \Delta_{3,k_1}^{(i,\ell)}, \mathbf{R}_{k_2}^{(i,\ell)} \Delta_{3,k_2}^{(i,\ell)} \rangle \right] \le (d_i \gamma_i^3 \tilde{M}_i^2 / 2) \sum_{k_1,k_2=0}^{N_i-1} \mathbb{E} \left[\|\mathbf{R}_{k_1}^{(i,\ell)}\| \|\mathbf{R}_{k_2}^{(i,\ell)}\| \right]
$$

$$\leq (d_i\gamma_i^3\tilde{M}_i^2/2)\left\{\sum_{k=0}^{N_i-1}(\exp[(N_i-k)\gamma_i\tilde{M}_i]-1-[(N_i-k)\gamma_i\tilde{M}_i])\right\}^2$$
$$\leq (d_i\gamma_i^3\tilde{M}_i^2/2)\left\{(\tilde{M}_i\gamma_i)^{-1}\int_0^{N_i\gamma_i\tilde{M}_i}\{\mathrm{e}^t-1-t\}\mathrm{d}t\right\}^2$$
$$\leq \frac{(\mathrm{e}^{N_i\gamma_i\tilde{M}_i}+1)^2}{288}d_iN_i^6\gamma_i^7\tilde{M}_i^6.$$

Similarly, we get Moreover, using the Cauchy-Schwarz inequality, for any $i\in[n]$ we get

$$\sum_{k_1,k_2=0}^{N_i-1}\mathbb{E}[\langle\Delta_{k_1}^{(i,\ell)},\mathbf{R}_{k_2}^{(i,\ell)}\Delta_{k_2}^{(i,\ell)}\rangle]\leq\sum_{k_1,k_2=0}^{N_i-1}\mathbb{E}\left[\|\Delta_{k_1}^{(i,\ell)}\|\|\Delta_{k_2}^{(i,\ell)}\|\|\mathbf{R}_{k_2}^{(i,\ell)}\|\right]$$
$$\leq\frac{d_iN_i\gamma_i^3\tilde{M}_i^2}{24}(\mathrm{e}^{N_i\gamma_i\tilde{M}_i}+1)N_i^3\gamma_i^2\tilde{M}_i^2$$
$$\leq d_iN_i^4\gamma_i^5\tilde{M}_i^4\frac{\mathrm{e}^{N_i\gamma_i\tilde{M}_i}+1}{24}.$$

In addition, for any $i\in[n]$, we have also

$$\sum_{k_1,k_2=0}^{N_i-1}\mathbb{E}\left[\langle\mathbf{R}_{k_1}^{(i,\ell)}\Delta_{3,k_1}^{(i,\ell)},\sum_{l=k_2+1}^{N_i}\mathbf{C}_l^{(i,\ell)}\Delta_{3,k_2}^{(i,\ell)}\rangle\right]\leq d_iN_i^5\gamma_i^6\tilde{M}_i^5\frac{\mathrm{e}^{N_i\gamma_i\tilde{M}_i}+1}{24}.$$

For any $i\in[n],k\in\mathbb{N}$, regrouping the previous results and using that $N_i\gamma_i\tilde{M}_i\leq 2$ give

$$\mathbb{E}[E_3]\leq\sum_{i=1}^n\{d_iN_i\gamma_i^2\tilde{M}_i^2+d_iN_i^3\gamma_i^4\tilde{M}_i^4\}+\sum_{i=1}^n d_iN_i^4\gamma_i^5\tilde{M}_i^5(1+N_i\gamma_i\tilde{M}_i). \tag{2.118}$$

Combination of our previous results. Injecting the three upper bounds (2.114), (2.115), (2.118) in (2.111), we get

$$\mathbb{E}\left[\|T_2^{(\ell)}\|^2\right]\leq\sum_{i=1}^n d_iN_i\gamma_i^3(d_iL_i^2+\tilde{M}_i^4/\tilde{m}_i)+\sum_{i=1}^n\{d_i\gamma_i^2\tilde{M}_i^2+d_iN_i^3\gamma_i^4\tilde{M}_i^4\}$$
$$+\sum_{i=1}^n d_iN_i^4\gamma_i^5\tilde{M}_i^5(1+N_i\gamma_i\tilde{M}_i). \tag{2.119}$$

Using the recursion defined in (2.108), and combining the upper bounds derived in (2.109) and (2.119) completes the proof. ■

Lemma 2.43. *Assume Assumption 2.1-Assumption 2.3-Assumption 2.7 and let* $\boldsymbol{N}\in(\mathbb{N}^\star)^n,\boldsymbol{\gamma}\in(\mathbb{R}_+^*)^n$ *such that for any* $i\in[n]$, $N_i\gamma_i\leq 2/(m_i+\tilde{M}_i)$, $\gamma_i<1/\tilde{M}_i$ *and* $\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}=\min_{i\in[n]}\{N_i\gamma_im_i\}-r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}\in(0,1)$, *where* $r_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}$ *is defined in (2.42). Then, for* $\ell\geq 1$, *we have*

$$\mathbb{E}\Big[\|\tilde{Z}_{\ell+1}-Z_{\ell+1}\|^2_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}\Big]^{1/2}\leq(1-\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}})^{\ell-1}\mathbb{E}\Big[\|\tilde{Z}_1-Z_1\|^2_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}}\Big]^{1/2}+\{\kappa_{\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{N}}\}^{-1}\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{N}),$$

where $\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{N})$ *is given in (2.107).*

Proof The proof follows from Lemma 2.42 combined with a straightforward induction. ■

Proof of Proposition 2.8/Proposition 2.41. **Proof** [Proof of Proposition 2.8/Proposition 2.41.] By Proposition 2.21 and Lemma 2.36, $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}^\star}$ converges in W_2 to $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}^\star}$. Therefore, using (2.84), Lemma 2.31 and Lemma 2.43 and taking $\ell \to +\infty$, we obtain

$$W_2^2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}^\star}, \Pi_{\boldsymbol{\rho}}) \leq 4(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2) \frac{\max_{i\in[n]}\{N_i^\star(\gamma_i)\gamma_i\}}{\min_{i\in[n]}\{N_i^\star(\gamma_i)\gamma_i m_i\}} \mathcal{R}(\boldsymbol{\gamma}, \boldsymbol{N}^\star), \quad (2.120)$$

where $\boldsymbol{N}^\star$ is defined in (2.100). By definition of $N_i^\star(\gamma_i)$, we have $\gamma_i \tilde{M}_i N_i^\star(\gamma_i) \leq \mathfrak{f}_i = m_i/(20\tilde{M}_i)$ which completes the proof upon using it in (2.120). ■

2.E Explicit mixing times

This section aims at providing mixing times for DG-LMC with explicit dependencies w.r.t. the dimension d and the prescribed precision ε. We specify our result to the case where for any $i \in [n]$, $m_i = m$, $M_i = M$, $\rho_i = \rho$, $\gamma_i = \gamma$, $N_i = N$ and for the specific initial distribution

$$\mu_{\boldsymbol{\rho}}^\star = \delta_{z^\star} \otimes \Pi_{\boldsymbol{\rho}}(\cdot|z^\star),$$

where

$$\mathbf{x}^\star = ([\theta^\star]^\top, [z^\star]^\top)^\top, \text{ where } \theta^\star = \arg\min\{-\log \pi\} \text{ and } z^\star = ([\mathbf{A}_1\theta^\star]^\top, \cdots, [\mathbf{A}_n\theta^\star]^\top)^\top.$$

Note that sampling from $\mu_{\boldsymbol{\rho}}^\star$ is straightforward and simply consists in setting $z_0 = z^\star$ and $\theta_0 = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} z_0 + \bar{\mathbf{B}}_0^{-1/2}\xi$, where ξ is a d-dimensional standard Gaussian random variable. Starting from this initialization, we consider the marginal law of θ_ℓ for $\ell \geq 1$ and denote it $\Gamma_{\mathbf{x}^\star}^\ell$. By Proposition 2.21, since for any $i \in [n]$, $N_i = N$, the stationary distribution associated to $P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ is $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}} = \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma},\mathbf{1}_n}$. We build upon the natural decomposition of the bias:

$$W_2(\Gamma_{\mathbf{x}^\star}^\ell, \pi(\cdot|\mathcal{D})) \leq W_2(\mu_{\boldsymbol{\rho}}^\star P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}^\ell, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) + W_2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}, \Pi_{\boldsymbol{\rho}}) + W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D})),$$

where $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$, $\Pi_{\boldsymbol{\rho}}$ and $\pi_{\boldsymbol{\rho}}$ are defined in Proposition 2.4, (2.2) and (2.3), respectively. The following subsections focus on deriving conditions on ℓ_ε, γ_ε, N_ε and ρ_ε to satisfy $W_2(\Gamma_{\mathbf{x}^\star}^{\ell_\varepsilon}, \pi(\cdot|\mathcal{D})) \leq \varepsilon$, where $\varepsilon > 0$.

2.E.1 Lower bound on the number of iterations ℓ_ε

In this section, we derive a lower bound on ℓ_ε such that $W_2(\mu_{\boldsymbol{\rho}}^\star P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}^{\ell_\varepsilon}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq \varepsilon/3$ following the result provided in Proposition 2.22. Recall that we define the z-marginal under $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ by

$$\pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}^z = \int_{\mathbb{R}^d} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\theta, z)\, \mathrm{d}\theta,$$

and the transition kernel of the Markov chain $\{Z_\ell\}_{\ell\geq 0}$, for all $z \in \mathbb{R}^p$ and $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$, by

$$P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z,\mathsf{B}) = \int_{\mathbb{R}^d} Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z,\mathsf{B}|\theta)\Pi_\rho(\theta|z)\,\mathrm{d}\theta,$$

where $\Pi_\rho(\cdot|z)$ and $Q_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ are defined in (2.5) and (2.25), respectively. In the case $\boldsymbol{N} = \mathbf{1}_n$, we simply denote $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ by $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$. We need to bound in Proposition 2.22 the factor

$$\left\{\int_{\mathbb{R}^d} \|z_1 - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}z_1) + \int_{\mathbb{R}^d} \|z_1 - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z^\star,\mathrm{d}z_1)\right\}^{1/2}.$$

Our next results provide such bounds.

Lemma 2.44. *Assume Assumption 2.1. Then, the transition kernel $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ leaves $\pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ invariant, that is $\pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}} = \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$, where $\pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ is defined by (2.55).*

Proof We have for any $\mathsf{B} \in \mathcal{B}(\mathbb{R}^p)$

$$\int_{\mathsf{B}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}z) = \int_{\mathsf{B}} \int_{\mathbb{R}^d} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\theta,\mathrm{d}z) = \int_{\mathsf{B}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}z) \int_{\mathbb{R}^d} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\theta|z).$$

Therefore, using the fact that $P_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ leaves $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ invariant from Proposition 2.13 and Fubini's theorem, we get

$$\begin{aligned}
\int_{\mathsf{B}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}z) &= \int_{\mathsf{B}} \int_{\mathbb{R}^d} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\theta,\mathrm{d}z) = \int_{\mathsf{B}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) P_{\boldsymbol{\rho},\boldsymbol{\gamma}}((\tilde{\theta},\tilde{z}),(\mathrm{d}\theta,\mathrm{d}z)) \\
&= \int_{\mathsf{B}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) Q_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\tilde{z},\mathrm{d}z|\tilde{\theta}) \Pi_{\boldsymbol{\rho}}(\theta|z)\,\mathrm{d}\theta \\
&= \int_{\mathsf{B}} \int_{\mathbb{R}^d\times\mathbb{R}^p} \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{\theta},\mathrm{d}\tilde{z}) Q_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\tilde{z},\mathrm{d}z|\tilde{\theta}) \int_{\mathbb{R}^d} \Pi_{\boldsymbol{\rho}}(\theta|z)\,\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{z}) P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\tilde{z},\mathsf{B}).
\end{aligned}$$

■

For any $i \in [n]$, let $\theta^\star_i$ a minimizer of $\theta \mapsto U_i(\mathbf{A}_i\theta)$, and define

$$u^\star = ([\mathbf{A}_1(\theta^\star - \theta^\star_1)]^\top, \cdots, [\mathbf{A}_n(\theta^\star - \theta^\star_n)]^\top)^\top$$

Lemma 2.45. *Assume Assumption 2.1-Assumption 2.3 and let $\boldsymbol{N} \in (\mathbb{N}^*)^n, \boldsymbol{\gamma}, \boldsymbol{\rho} \in (\mathbb{R}^*_+)^n$ such that, for any $i \in [n]$, $\gamma_i \leq 2/(m_i + M_i + 1/\rho_i)$ and denote $z^\star = ([\mathbf{A}_1\theta^\star]^\top, \cdots, [\mathbf{A}_n\theta^\star]^\top)^\top$. Then, for any $z \in \mathbb{R}^p$ and $\varepsilon > 0$,*

$$\begin{aligned}
\int_{\mathbb{R}^p} \|\tilde{z} - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z,\mathrm{d}\tilde{z}) \leq \min_{i\in[n]}\{N_i\}^{-1} \Big[&\kappa^2_{\boldsymbol{\gamma}}(1+2\varepsilon)\|z - z^\star\|^2_{\mathbf{D}^{-1}_\gamma} \\
&+ (1 + 1/(2\varepsilon)) \max_{i\in[n]}\{\gamma_i M_i^2\}\|u^\star\|^2 + \mathrm{Tr}(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0) + 2\sum_{i=1}^n d_i\Big],
\end{aligned}$$

where the transition kernel $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$ is defined in (2.56) with $\boldsymbol{N} = \mathbf{1}_n$.

Proof Let $\gamma_i \leq 2/(m_i + M_i + 1/\rho_i)$ for any $i \in [n]$, and ξ be a d-dimensional Gaussian random variable independent of $\{\eta^i : i \in [n]\}$ where for any $i \in [n]$, η^i is a d_i-dimensional Gaussian random variable. Take $z \in \mathbb{R}^p$ and let Z be the random variable distributed according to $\delta_z P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}$, and defined by

$$\theta = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} z + \bar{\mathbf{B}}_0^{-1/2}\xi,$$

and for any $i \in [n]$,

$$\begin{aligned}
Z^i &= \left(1 - \gamma_i/\rho_i\right) z_i - \gamma_i \nabla U_i(z_i) + \frac{\gamma_i}{\rho_i}\mathbf{A}_i\theta + \sqrt{2\gamma_i}\eta^i \\
&= \left(1 - \gamma_i/\rho_i\right) z_i - \gamma_i \nabla U_i(z_i) + \frac{\gamma_i}{\rho_i}\mathbf{A}_i\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} z + \frac{\gamma_i}{\rho_i}\mathbf{A}_i\bar{\mathbf{B}}_0^{-1/2}\xi + \sqrt{2\gamma_i}\eta^i \\
&= \left(1 - \gamma_i/\rho_i\right) z_i - \gamma_i[\nabla U_i(z_i) - \nabla U_i(\mathbf{A}_i\theta^\star)] - \gamma_i[\nabla U_i(\mathbf{A}_i\theta^\star) - \nabla U_i(\mathbf{A}_i\theta_i^\star)] \\
&+ \frac{\gamma_i}{\rho_i}\mathbf{A}_i\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} z + \frac{\gamma_i}{\rho_i}\mathbf{A}_i\bar{\mathbf{B}}_0^{-1/2}\xi + \sqrt{2\gamma_i}\eta^i.
\end{aligned}$$

Let

$$\begin{aligned}
\mathbf{D}^\star_U &= \operatorname{diag}\left(\gamma_1 \int_0^1 \nabla^2 U_1(z_1 + t(\mathbf{A}_1\theta^\star - z_1))\,\mathrm{d}t, \cdots, \gamma_n \int_0^1 \nabla^2 U_n(z_n + t(\mathbf{A}_n\theta^\star - z_n))\,\mathrm{d}t\right), \\
\tilde{\mathbf{D}}^\star_U &= \operatorname{diag}\left(\gamma_1 \int_0^1 \nabla^2 U_1(\mathbf{A}_1\theta^\star + t(\mathbf{A}_1\theta_1^\star - \mathbf{A}_1\theta^\star))\,\mathrm{d}t, \cdots, \gamma_n \int_0^1 \nabla^2 U_n(\mathbf{A}_n\theta^\star + t(\mathbf{A}_n\theta_n^\star - \mathbf{A}_n\theta^\star))\,\mathrm{d}t\right).
\end{aligned}$$

Since $\mathbf{P}_0\mathbf{D}_{\boldsymbol{\rho}}^{-1/2} z^\star = \mathbf{D}_{\boldsymbol{\rho}}^{-1/2} z^\star$, it follows that

$$\begin{aligned}
Z - z^\star = \left[\mathbf{I}_p - \mathbf{D}^\star_U - \mathbf{D}^{1/2}_{\boldsymbol{\gamma}}\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}(\mathbf{I}_p - \mathbf{P}_0)\mathbf{D}_{\boldsymbol{\rho}}^{-1/2}\right](z - z^\star) - \tilde{\mathbf{D}}^\star_U u^\star \\
+ \mathbf{D}^{1/2}_{\boldsymbol{\gamma}}\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{B}_0\bar{\mathbf{B}}_0^{-1/2}\xi + \mathbf{D}^{1/2}_{2\boldsymbol{\gamma}}\eta.
\end{aligned}$$

With the notation $\mathbf{H} = \mathbf{I}_p - \mathbf{D}^\star_U - \mathbf{D}^{1/2}_{\boldsymbol{\gamma}}\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}(\mathbf{I}_p - \mathbf{P}_0)\mathbf{D}_{\boldsymbol{\rho}}^{-1/2}$, (2.23), and using the fact that for any $\varepsilon > 0$, $a, b \in \mathbb{R}^d$, $|\langle a, b\rangle| \leq \varepsilon\|a\|^2 + (4\varepsilon)^{-1}\|b\|^2$, it follows, for any $z \in \mathbb{R}^p$, that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \|\tilde{z} - z^\star\|^2_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(z, \mathrm{d}\tilde{z}) \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^d} \left\|\mathbf{H}(z - z^\star) - \tilde{\mathbf{D}}^\star_U u^\star + \mathbf{D}^{1/2}_{\boldsymbol{\gamma}}\mathbf{D}^{1/2}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{B}_0\bar{\mathbf{B}}_0^{-1/2}\xi + \mathbf{D}^{1/2}_{2\boldsymbol{\gamma}}\boldsymbol{\eta}\right\|^2_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} \phi_d(\xi)\,\mathrm{d}\xi\phi_p(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} \\
&= \left\|\mathbf{H}(z - z^\star) - \tilde{\mathbf{D}}^\star_U u^\star\right\|^2_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} + \operatorname{Tr}(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0) + 2\sum_{i=1}^n d_i \\
&\leq \kappa^2_{\boldsymbol{\gamma}}\|z - z^\star\|^2_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} - 2\langle \mathbf{H}(z - z^\star), \tilde{\mathbf{D}}^\star_U u^\star\rangle_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} + \left\|\tilde{\mathbf{D}}^\star_U u^\star\right\|^2_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} + \operatorname{Tr}(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0) + 2\sum_{i=1}^n d_i \\
&\leq \kappa^2_{\boldsymbol{\gamma}}(1 + 2\varepsilon)\|z - z^\star\|^2_{\mathbf{D}_{\boldsymbol{\gamma}}^{-1}} + \left(1 + \frac{1}{2\varepsilon}\right)\max_{i\in[n]}\{\gamma_i M_i^2\}\|u^\star\|^2 + \operatorname{Tr}(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0) + 2\sum_{i=1}^n d_i.
\end{aligned}$$

■

Proposition 2.46. *Assume Assumption 2.1-Assumption 2.3 and let* $\boldsymbol{N} \in (\mathbb{N}^*)^n, \boldsymbol{\gamma}, \boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$ *such that, for any* $i \in [n]$, $\gamma_i \le 2/(m_i + M_i + 1/\rho_i)$. *Then, we have*

$$\int_{\mathbb{R}^d} \|z_1 - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}z_1) \le \min_{i\in[n]}\{N_i\}^{-1} \frac{2}{1-\kappa^2_{\boldsymbol{\gamma}}} \left(\frac{1+\kappa^2_{\boldsymbol{\gamma}}}{1-\kappa^2_{\boldsymbol{\gamma}}} \max_{i\in[n]}\{\gamma_i M_i^2\} \|u^\star\|^2 + \mathrm{Tr}(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0) + 2\sum_{i=1}^n d_i \right),$$

with $\kappa_{\boldsymbol{\gamma}}$ *defined in* (2.21).

Proof With the choice $\varepsilon = (1-\kappa^2_{\boldsymbol{\gamma}})/(4\kappa^2_{\boldsymbol{\gamma}})$ in Lemma 2.45 and using Lemma 2.44, we have

$$\int_{\mathbb{R}^p} \|\tilde{z} - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{\gamma}}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}\tilde{z}) \le \frac{\kappa^2_{\boldsymbol{\gamma}}+1}{2} \int_{\mathbb{R}^p} \|z - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{\gamma}}} \pi^z_{\boldsymbol{\rho},\boldsymbol{\gamma}}(\mathrm{d}z) + \frac{1+\kappa^2_{\boldsymbol{\gamma}}}{1-\kappa^2_{\boldsymbol{\gamma}}} \max_{i\in[n]}\{\gamma_i M_i^2\} \|u^\star\|^2 + \mathrm{Tr}(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}}\mathbf{P}_0) + 2\sum_{i=1}^n d_i.$$

Rearranging terms concludes the proof. ∎

Lemma 2.47. *Assume Assumption 2.1-Assumption 2.3 and let* $\boldsymbol{N} \in (\mathbb{N}^*)^n, \boldsymbol{\gamma}, \boldsymbol{\rho} \in (\mathbb{R}_+^*)^n$ *such that, for any* $i \in [n]$, $N_i\gamma_i \le 2/(m_i + M_i + 1/\rho_i), \gamma_i\tilde{M}_i < 1$ *and denote* $z^\star = ([\mathbf{A}_1\theta^\star]^\top, \cdots, [\mathbf{A}_n\theta^\star]^\top)^\top$. *Then, we have*

$$\int_{\mathbb{R}^p} \|\tilde{z} - z^\star\|^2_{\mathbf{D}^{-1}_{\boldsymbol{N}\boldsymbol{\gamma}}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z^\star, \mathrm{d}\tilde{z}) \le 2\sum_{i=1}^n \gamma_i N_i \left(1 + \mathrm{Tr}(\mathbf{P}_0)/\rho_i\right) + 4\sum_{i=1}^n d_i,$$

where the transition kernel $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}$ *is defined in* (2.56).

Proof Let $\{(\eta^i_k)_{k\ge1} : i \in [n]\}$ be independent random variables such that for any $i \in [n]$, the sequences $\{(\eta^i_k)_{k\ge1}\}$ are i.i.d. d_i-dimensional Brownian motions and let ξ a d-dimensional standard Gaussian random variable independent of $\{(\eta^i_k)_{k\ge1} : i \in [n]\}$. Consider the stochastic process $(Y_k)_{k\in\mathbb{N}}$ initialized for any $i \in [n]$ at $Y^i_0 = \mathbf{A}_i\theta^\star$ and defined, for any $i \in [n], k \in \mathbb{N}$, by

$$Y^i_{k+1} = Y^i_k - \gamma_i \nabla V_i(Y^i_k) + (\gamma_i/\rho_i)\mathbf{A}_i\theta + \sqrt{2\gamma_i}\eta^i_{k+1}, \tag{2.121}$$

where the potential $V_i = \mathbf{y}^i \mapsto U_i(\mathbf{y}^i) + \|\mathbf{y}^i\|^2/(2\rho_i)$ and

$$\theta = \bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2} z^\star + \bar{\mathbf{B}}_0^{-1/2}\xi. \tag{2.122}$$

In addition, we define the random variable $Z = (Z^1, \ldots, Z^n)$, for any $i \in [n]$, as

$$Z^i = Y^i_{N_i}.$$

By definition, note that Z is distributed according to $P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z^\star, \cdot)$. Define the process $(\mathrm{Y}_k = \{\mathrm{Y}^i_k\}_{i=1}^n)_{k\in\mathbb{N}}$ valued in $\mathbb{R}^p \times \mathbb{R}^p$ defined for any $i \in [n]$, $k \ge 0$ by

$$\mathrm{Y}^i_k = Y^i_{\min(k,N_i)},$$

and consider the following matrices defined, for any $k \in \mathbb{N}$, by

$$\begin{aligned}
\mathbf{H}_{U,k} &= \operatorname{diag}\Bigg(\gamma_1 \int_0^1 \nabla^2 U_1((1-s)Y_k^1 + sz^\star)\,\mathrm{d}s, \\
&\qquad\qquad \dots, \gamma_n \int_0^1 \nabla^2 U_n((1-s)Y_k^n + sz^\star)\,\mathrm{d}s\Bigg), \\
\mathbf{J}(k) &= \operatorname{diag}\left(\mathbb{1}_{[N_1]}(k+1)\cdot \mathbf{I}_{d_1}, \cdots, \mathbb{1}_{[N_n]}(k+1)\cdot \mathbf{I}_{d_n}\right), \\
\mathbf{C}_k &= \mathbf{J}(k)(\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{\rho}} + \mathbf{H}_{U,k}), \\
\mathbf{M}_{k+1} &= (\mathbf{I}_p - \mathbf{C}_0)^{-1}\dots(\mathbf{I}_p - \mathbf{C}_k)^{-1}, \qquad \text{with } \mathbf{M}_0 = \mathbf{I}_p.
\end{aligned} \tag{2.123}$$

Using these notations and (2.121), for any $k \in \mathbb{N}$, we get

$$\mathrm{Y}_{k+1} - z^\star = (\mathbf{I}_p - \mathbf{C}_k)(\mathrm{Y}_k - z^\star) + \mathbf{J}(k)\left(\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\theta - \mathbf{D}_{\boldsymbol{\gamma}}\nabla V(z^\star) + \mathbf{D}_{2\boldsymbol{\gamma}}^{1/2}\eta_{k+1}\right).$$

Multiplying the previous equality by $\mathbf{M}_{k+1}\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}$, we obtain, for $k \geq 0$,

$$\begin{aligned}
\mathbf{M}_{k+1}\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}(\mathrm{Y}_{k+1} - z^\star) &= \mathbf{M}_k\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}(\mathrm{Y}_k - z^\star) \\
&\quad + \mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}\left(\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\theta - \mathbf{D}_{\boldsymbol{\gamma}}\nabla V(z^\star) + \mathbf{D}_{2\boldsymbol{\gamma}}^{1/2}\eta_{k+1}\right).
\end{aligned}$$

Summing the previous equality over $k \in \mathbb{N}$ gives

$$\begin{aligned}
\mathbf{M}_\infty\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}(\mathrm{Y}_{\boldsymbol{N}} - z^\star) &= \mathbf{M}_0\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}(\mathrm{Y}_0 - z^\star) \\
&\quad + \sum_{k=0}^{\infty}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}\left(\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\theta - \mathbf{D}_{\boldsymbol{\gamma}}\nabla V(z^\star) + \mathbf{D}_{2\boldsymbol{\gamma}}^{1/2}\eta_{k+1}\right).
\end{aligned}$$

Multiplying the last equality by $[\mathbf{M}_\infty]^{-1}$ and using the fact that $\mathrm{Y}_0 = z^\star$, we get

$$\begin{aligned}
&\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}(Z - z_\star) \\
&\quad = \sum_{k=0}^{\infty}[\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}\left(\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\theta - \mathbf{D}_{\boldsymbol{\gamma}}\nabla V(z^\star) + \mathbf{D}_{2\boldsymbol{\gamma}}^{1/2}\eta_{k+1}\right).
\end{aligned} \tag{2.124}$$

Recall that $\mathbf{P}_0 = \mathbf{B}_0\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top$. Hence, by (2.122) and using $\mathbf{P}_0\mathbf{D}_{\boldsymbol{\rho}}^{-1/2}z^\star = \mathbf{D}_{\boldsymbol{\rho}}^{-1/2}z^\star$, we get

$$\mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\theta - \mathbf{D}_{\boldsymbol{\gamma}}\nabla V(z^\star) = \mathbf{D}_{\boldsymbol{\gamma}/\sqrt{\boldsymbol{\rho}}}\mathbf{B}_0\bar{\mathbf{B}}_0^{-1/2}\xi - \mathbf{D}_{\boldsymbol{\gamma}}\nabla U(z^\star).$$

Plugging this equality into (2.124) yields

$$\begin{aligned}
\mathbf{D}_{\boldsymbol{N\gamma}}^{-1/2}(Z - z^\star) &= -\sum_{k=0}^{\infty}[\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{N}}^{1/2}\nabla U(z^\star) \\
&\quad + \sum_{k=0}^{\infty}[\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{\gamma}/(\boldsymbol{N\rho})}^{1/2}\mathbf{B}_0\bar{\mathbf{B}}_0^{-1/2}\xi \\
&\quad + \sqrt{2}\sum_{k=0}^{\infty}[\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\eta_{k+1}.
\end{aligned} \tag{2.125}$$

Recall that $[\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1} = (([\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1})^1, \dots, ([\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1})^n)$ is a block-diagonal matrix where, for any $i \in [n]$, $([\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1})^i = \prod_{l=k+1}^{\infty}(\mathbf{I}_{d_i} - \mathbf{C}_l^i)$ where $\mathbf{C}_l^i$ is defined in (2.123). In addition, since we suppose for any $i \in [n]$, that $\gamma_i \tilde{M}_i < 1$, Lemma 2.32 implies

$$\left\| \prod_{l=k+1}^{N_i-1} (\mathbf{I}_{d_i} - \mathbf{C}_l^i) \right\|^2 \leq \left(1 - \gamma_i \tilde{m}_i\right)^{2(N_i-k-1)}.$$

We now upper bound separately each term on the right-hand side of (2.125). First, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left\| \sum_{k=0}^{\infty} [\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{\gamma}/\boldsymbol{N}}^{1/2} \right\|^2 &\leq \sum_{i=1}^{n} (\gamma_i/N_i) \left\| \sum_{k=0}^{\infty} ([\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1})^i \mathbf{J}^i(k) \right\|^2 \\
&\leq \sum_{i=1}^{n} (\gamma_i/N_i) \left\| \sum_{k=0}^{N_i-1} \prod_{l=k+1}^{N_i-1} \left(\mathbf{I}_{d_i} - \mathbf{C}_l^i\right) \right\|^2 \\
&\leq \sum_{i=1}^{n} \gamma_i \sum_{k=0}^{N_i-1} \left\| \prod_{l=k+1}^{N_i-1} \left(\mathbf{I}_{d_i} - \mathbf{C}_l^i\right) \right\|^2 \\
&\leq \sum_{i=1}^{n} \gamma_i \sum_{k=0}^{N_i-1} \left(1 - \gamma_i \tilde{m}_i\right)^{2(N_i-k-1)} \\
&\leq \sum_{i=1}^{n} N_i \gamma_i. \qquad (2.126)
\end{aligned}$$

Second, using the same techniques as for the above inequality, we obtain

$$\left\| \sum_{k=0}^{\infty} [\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{\gamma}/(\boldsymbol{N}\boldsymbol{\rho})}^{1/2}\mathbf{B}_0\bar{\mathbf{B}}_0^{-1/2}\xi \right\|^2 \leq \sum_{i=1}^{n} \frac{N_i\gamma_i}{\rho_i} \left\| \mathbf{B}_0\bar{\mathbf{B}}_0^{-1/2}\xi \right\|^2 \qquad (2.127)$$

Finally, the third term can be upper-bounded as

$$\mathbb{E}\left[\left\| \sqrt{2} \sum_{k=0}^{\infty} [\mathbf{M}_\infty]^{-1}\mathbf{M}_{k+1}\mathbf{J}(k)\mathbf{D}_{\boldsymbol{N}}^{-1/2}\eta_{k+1} \right\|^2 \right] \leq 2\sum_{i=1}^{n} d_i. \qquad (2.128)$$

Combining (2.125), (2.126), (2.127) and (2.128), we get

$$\int_{\mathbb{R}^p} \|\tilde{z} - z^\star\|^2_{\mathbf{D}_{\boldsymbol{N}\boldsymbol{\gamma}}^{-1}} P^z_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}(z^\star, \mathrm{d}\tilde{z}) \leq \sum_{i=1}^{n} \gamma_i N_i \left(1 + \mathrm{Tr}(\mathbf{P}_0)/\rho_i\right) + 2\sum_{i=1}^{n} d_i.$$

■

Given $\varepsilon > 0$, we are now ready to provide a condition on the number of iterations ℓ_ε to achieve $W_2(\mu^\star_{\boldsymbol{\rho}} P^{\ell_\varepsilon}_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq \epsilon/3$ in the case where for any $i \in [n]$, $m_i = m$, $M_i = M$, $\rho_i = \rho$, $\gamma_i = \gamma$ and $N_i = N$. Define

$$\mathsf{E}_0^2 = 18N\gamma(1+\|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top\tilde{\mathbf{D}}_0^{1/2}\|)\left[\frac{2}{N(1-\kappa_{\boldsymbol{\gamma}}^2)}\left(\frac{1+\kappa_{\boldsymbol{\gamma}}^2}{1-\kappa_{\boldsymbol{\gamma}}^2}\cdot\gamma M^2\|u^\star\|^2 + (\gamma/\rho)\mathrm{Tr}(\mathbf{P}_0) + 2\sum_{i=1}^n d_i\right) + 2b\gamma N\left(1+\mathrm{Tr}(\mathbf{P}_0)/\rho\right) + 4\sum_{i=1}^n d_i\right].$$

Theorem 2.48. *Assume Assumption 2.1-Assumption 2.3 and let* $\boldsymbol{N} = N\mathbf{1}_n, \boldsymbol{\gamma} = \gamma\mathbf{1}_n, \boldsymbol{\rho} = \rho\mathbf{1}_n$, $\rho > 0, \gamma > 0, N \geq 1$, *such that* $\gamma < 1/\tilde{M}$, $N\gamma < 2/(m+\tilde{M})$, *and* (2.54) *is satisfied. Then, for any* $\varepsilon > 0$, *any*

$$\ell_\varepsilon \geq 2\log(\mathsf{E}_0/\varepsilon)/(N\gamma m),$$

we have, $W_2(\mu_{\boldsymbol{\rho}}^\star P_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}^{\ell_\varepsilon}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq \varepsilon/3$.

Proof By some algebra and using $1/\log(1/(1-x)) \leq 1/x$ for $0 < x < 1$, the proof directly follows from Proposition 2.22 combined with Proposition 2.46 and Lemma 2.47. ■

2.E.2 Upper bound on the tolerance parameter $\boldsymbol{\rho}_\varepsilon$

Define

$$R_0 = 2\sigma_U^2\left(d\sigma_U^2 + \sum_{i=1}^n M_i^2\|\mathbf{A}_i(\theta^\star-\theta_i^\star)\|^2\right) + 2\sigma_U^4,$$

$$R_1 = d\sigma_U^2 + \sum_{i=1}^n M_i^2\|\mathbf{A}_i(\theta^\star-\theta_i^\star)\|^2 + \sum_{i=1}^n d_i M_i/2$$

$$R_2 = 2d\max_{i\in[n]}\{M_i\}\sigma_U^2 + 2\sum_{i=1}^n M_i^3\|\mathbf{A}_i(\theta^\star-\theta_i^\star)\|^2 + 8\sigma_U^4 + 8\sigma_U^2\left[2d\sigma_U^2 + 2\sum_{i=1}^n M_i^2\|\mathbf{A}_i(\theta^\star-\theta_i^\star)\|^2\right].$$

Recall that $\bar{\rho} = \max_{i\in[n]}\{\rho_i\}$. Then, the following result holds.

Lemma 2.49. *Assume Assumption 2.1-Assumption 2.3. For any* $\varepsilon > 0$, *let* $\boldsymbol{\rho}_\varepsilon \in (\mathbb{R}_+^*)^n$ *such that*

$$\bar{\rho}_\varepsilon \leq \frac{-R_1 + \sqrt{R_1^2 + 4R_0\varepsilon m_U^{1/2}/(3\sqrt{2})}}{2R_0} \wedge \frac{\varepsilon\sqrt{m_U}}{3\sqrt{2}\sqrt{R_2 + [R_2/(12\sigma_U^2) + \sum_{i=1}^n d_i M_i]^2}} \wedge \frac{1}{12\sigma_U^2} \wedge \frac{-\sum_{i=1}^n d_i M_i + \sqrt{(\sum_{i=1}^n d_i M_i)^2 + 6R_2}}{2R_2}.$$

Then, $W_2(\pi_{\boldsymbol{\rho}_\varepsilon}, \pi(\cdot|\mathcal{D})) \leq \varepsilon/3$.

Proof Let $\varepsilon > 0$. From (2.74), for any $\bar{\rho} \leq 1/(12\sigma_U^2)$, $W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D})) \leq \sqrt{\frac{2}{m_U}} \max(A_1, A_3^{1/2})$, where A_1, A_3 are defined in (2.71) and (2.73) respectively. This implies that $W_2(\pi_{\boldsymbol{\rho}}, \pi(\cdot|\mathcal{D})) \leq \varepsilon/3$ is verified if $\max(A_1, A_3^{1/2}) \leq \varepsilon\sqrt{m_U}/(3\sqrt{2})$. First, $A_1 \leq \varepsilon\sqrt{m_U}/(3\sqrt{2})$ holds if

$$\bar{\rho} \leq \frac{-R_1 + \sqrt{R_1^2 + 4R_0\varepsilon m_U^{1/2}/(3\sqrt{2})}}{2R_0} \wedge \frac{1}{12\sigma_U^2}. \tag{2.129}$$

We now focus on A_3. Using the fact that for any $x \in \mathbb{R}, \mathrm{e}^x \geq x + 1$, we have $2\prod_{i=1}^n (1 + \rho_i M_i)^{d_i} \geq 2 + \sum_{i=1}^n d_i \log(1 + \rho_i M_i)$ and therefore

$$A_3 \leq \exp\left(\bar{\rho}^2 R_2 + \sum_{i=1}^n d_i \log(1 + \rho_i M_i)\right) - 1 - \sum_{i=1}^n d_i \log(1 + \rho_i M_i).$$

Since $\sum_{i=1}^n d_i \log(1 + \rho_i M_i) \leq \bar{\rho} \sum_{i=1}^n d_i M_i$, $\bar{\rho}^2 R_2 + \sum_{i=1}^n d_i \log(1 + \rho_i M_i) \leq 3/2$ holds for

$$\bar{\rho} \leq \frac{-\sum_{i=1}^n d_i M_i + \sqrt{(\sum_{i=1}^n d_i M_i)^2 + 6R_2}}{2R_2}. \tag{2.130}$$

Since for any $x \leq 3/2, \mathrm{e}^x \leq 1 + x + x^2$ and using the fact that $\bar{\rho} \leq 1/(12\sigma_U^2)$, it follows that

$$A_3 \leq \bar{\rho}^2 R_2 + \left(\bar{\rho}^2 R_2 + \bar{\rho} \sum_{i=1}^n d_i M_i)\right)^2 \leq \bar{\rho}^2 \left[B_1 + \left(\frac{R_2}{12\sigma_U^2} + \sum_{i=1}^n d_i M_i\right)^2\right].$$

Hence, $A_3^{1/2} \leq \varepsilon\sqrt{m_U}/(3\sqrt{2})$ holds under (2.130) and

$$\bar{\rho} \leq \frac{\varepsilon\sqrt{m_U}}{3\sqrt{2}\sqrt{R_2 + \left(\frac{R_2}{12\sigma_U^2} + \sum_{i=1}^n d_i M_i\right)^2}}. \tag{2.131}$$

The proof is concluded by combining (2.129), (2.130) and (2.131). ■

2.E.3 Upper bound on the step-size $\boldsymbol{\gamma}_\varepsilon$ and number of local iteration $\boldsymbol{N}_\varepsilon$

Based on Proposition 2.37 or Proposition 2.41, we now determine an upper bound on $\boldsymbol{\gamma}_\varepsilon$ to ensure $W_2(\Pi_{\boldsymbol{\rho}}, \Pi_{\boldsymbol{\rho}, \boldsymbol{\gamma}_\varepsilon}) \leq \varepsilon/3$ in the case $\boldsymbol{N} = N\mathbf{1}_n, \boldsymbol{\gamma} = \gamma\mathbf{1}_n, \boldsymbol{\rho} = \rho\mathbf{1}_n$ where $\rho > 0, \gamma > 0, N \geq 1$. The following results hold depending if Assumption 2.7 is considered. Define

$$C_\rho = \frac{4\tilde{M}^2(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)}{5m},$$

$$C_0 = (\tilde{M}^2/2)\left[\tilde{M}/\tilde{m} + 1/6\right] \sum_{i=1}^n d_i, \quad C_1 = \sum_{i=1}^n d_i, \quad C_2 = \varepsilon^2/(9C_\rho).$$

Lemma 2.50. *Assume Assumption 2.1-Assumption 2.3 and let $\rho, \gamma > 0$ and $N \geq 1$. In addition, set $\boldsymbol{\rho} = \rho\mathbf{1}_n$, $\boldsymbol{\gamma}_\varepsilon = \gamma_\varepsilon \mathbf{1}_n$, $\boldsymbol{N}_\varepsilon = N_\varepsilon \mathbf{1}_n$ and $\varepsilon > 0$ such that*

$$\gamma_\varepsilon \leq \frac{-C_1 + \sqrt{C_1^2 + 4C_0C_2}}{2C_0} \wedge \frac{m}{40\tilde{M}^2}. \tag{2.132}$$

Then $W_2(\Pi_{\boldsymbol{\rho}}, \Pi_{\boldsymbol{\rho}, \boldsymbol{\gamma}_\varepsilon}) \leq \varepsilon/3$.

Proof Let $\varepsilon > 0$. By Proposition 2.37, note that $W_2^2(\Pi_{\boldsymbol{\rho}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}_\varepsilon,\boldsymbol{N}_\varepsilon}) \leq \varepsilon^2/9$ is satisfied if

$$C_0\gamma_\varepsilon^2 + C_1\gamma_\varepsilon \leq C_2.$$

Since this inequality is satisfied under the choice (2.132), we have $W_2(\Pi_{\boldsymbol{\rho}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}_\varepsilon,\boldsymbol{N}_\varepsilon}) \leq \varepsilon/3$. Eventually, using Proposition 2.21 shows that $\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}_\varepsilon,\boldsymbol{N}_\varepsilon} = \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}_\varepsilon}$. ■

In addition to the assumptions of Lemma 2.50, under Assumption 2.7 we get a more interesting mixing-time for γ. For any $\varepsilon \in \mathbb{R}_+^*, \rho \in \mathbb{R}_+^*$, define

$$C_{\boldsymbol{\rho}} = \frac{4(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)}{m}, \tag{2.133}$$

$$\bar{\gamma}_\varepsilon = \frac{\varepsilon}{3\tilde{M}\sqrt{C_{\boldsymbol{\rho}}\sum_{i\in[n]} d_i}} \wedge \frac{\varepsilon^2}{18C_{\boldsymbol{\rho}}\tilde{M}\mathfrak{f}^3\sum_{i\in[n]} d_i} \wedge \frac{\varepsilon\sqrt{\tilde{M}}}{3\sqrt{C_{\boldsymbol{\rho}}\mathfrak{f}\sum_{i\in[n]} d_i(d_iL_i^2 + \tilde{M}^4/\tilde{m})}} \wedge \frac{m}{40\tilde{M}^2}, \tag{2.134}$$

where $\mathfrak{f} = m/(20\tilde{M})$.

Lemma 2.51. *Assume Assumption 2.1-Assumption 2.3-Assumption 2.7. Then, for any $N \in \mathbb{N}^*, \varepsilon \in \mathbb{R}_+^*, \rho \in \mathbb{R}_+^*, \gamma \in (0, \bar{\gamma}_\varepsilon]$, we have*

$$W_2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}, \Pi_{\boldsymbol{\rho}}) \leq \varepsilon/3,$$

where $\boldsymbol{\gamma} = \gamma\mathbf{1}_n, \boldsymbol{\rho} = \rho\mathbf{1}_n$.

Proof For any $\varepsilon \in \mathbb{R}_+^*, \rho \in \mathbb{R}_+^*, \gamma \in (0, \bar{\gamma}_\varepsilon]$ applying Proposition 2.41, we get

$$W_2^2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}, \Pi_{\boldsymbol{\rho}}) \leq \frac{4(1 + \|\bar{\mathbf{B}}_0^{-1}\mathbf{B}_0^\top \tilde{\mathbf{D}}_0^{1/2}\|^2)}{m}\mathcal{R}^\star(\boldsymbol{\gamma}),$$

where

$$\mathcal{R}^\star(\boldsymbol{\gamma}) = \sum_{i=1}^n \left\{ d_i\gamma^2\tilde{M}^2 + \frac{d_i\gamma^2\mathfrak{f}}{\tilde{M}}\left(d_iL_i^2 + \frac{\tilde{M}^4}{\tilde{m}}\right) + d_i\gamma\tilde{M}\mathfrak{f}^3(1 + \mathfrak{f} + \mathfrak{f}^2)\right\}.$$

Since $\gamma \leq \bar{\gamma}_\varepsilon$, we have $\mathcal{R}^\star(\boldsymbol{\gamma}) \leq \mathcal{R}^\star(\bar{\boldsymbol{\gamma}}_\varepsilon)$ where we denoted $\bar{\boldsymbol{\gamma}}_\varepsilon = (\bar{\gamma}_\varepsilon, \ldots, \bar{\gamma}_\varepsilon)$. Thus, we get $W_2(\Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}, \Pi_{\boldsymbol{\rho}}) \leq \varepsilon/3$. ■

2.E.4 Discussion

Let $\boldsymbol{\rho}_\varepsilon = \rho_\varepsilon\mathbf{1}_n$ such that $W_2(\pi_{\boldsymbol{\rho}_\varepsilon}, \pi(\cdot|\mathcal{D})) \leq \varepsilon/3$. From Lemma 2.49, we can take $\bar{\rho}_\varepsilon = \mathrm{O}(\varepsilon/d)$ when $\varepsilon \to 0$ and $d \to \infty$. Similarly, under Assumption 2.1-Assumption 2.3, in the asymptotic regime $\varepsilon \to \infty$ and $d \to \infty$, we obtain by Lemma 2.50 that $\bar{\gamma}_\varepsilon = \mathrm{O}(\varepsilon^4/d^3)$, $N_\varepsilon = \mathrm{O}(d/\varepsilon^2)$ is enough to ensure that $W_2(\Pi_{\bar{\boldsymbol{\rho}}_\varepsilon}, \Pi_{\bar{\boldsymbol{\rho}}_\varepsilon,\bar{\gamma}_\varepsilon}) \leq \varepsilon/3$. On the other hand, when Assumption 2.7 is additionally assumed, we only need to suppose $\bar{\gamma}_\varepsilon = \mathrm{O}(\varepsilon^2/d^2)$ and $\bar{N}_\varepsilon = \mathrm{O}(1)$. For these step-sizes choices, Theorem 2.48 shows the number of iterations $\ell_\varepsilon = \mathrm{O}(d^2\log(d/\varepsilon)/\varepsilon^2)$ ensures that $W_2(\delta_{\mathbf{x}^\star}P^{\ell_\varepsilon}_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{N}}, \Pi_{\boldsymbol{\rho},\boldsymbol{\gamma}}) \leq \varepsilon/3$.

Chapter 3

FALD: Federated Averaging Langevin Dynamics

Contents

This chapter focuses on Bayesian inference in a federated learning context (FL). While several distributed MCMC algorithms have been proposed, few consider the specific limitations of FL such as communication bottlenecks and statistical heterogeneity. Recently, Federated Averaging Langevin Dynamics (FALD) was introduced, which extends the Federated Averaging algorithm to Bayesian inference. We obtain a novel tight non-asymptotic upper bound on the Wasserstein distance to the global posterior for FALD. This bound highlights the effects of statistical heterogeneity, which causes a drift in the local updates that negatively impacts convergence. We propose a new algorithm VR-FALD* that uses control variates to correct the client drift. We establish non-asymptotic bounds showing that VR-FALD* is not affected by statistical heterogeneity. Finally, we illustrate our results on several FL benchmarks for Bayesian inference.

3.1 Introduction

The paradigm of fully centralized machine learning is increasingly at odds with real-world use cases. Centralized machine learning leads to (a) data processing bottlenecks, (b) inefficient use of communication resources and (c) risks exposing individuals' private data. As storage and computational capacity increases at the agent level, it becomes increasingly attractive to decentralize computational tasks whenever possible. The term *federated* learning (FL) was recently coined to capture some aspects of this grand challenge (McMahan et al., 2017; Kairouz et al., 2021; Yang et al., 2019; Alistarh et al., 2017; Horváth et al., 2022; Wang et al., 2021).

Reducing communication costs has been identified as one of the major challenges of FL (Kairouz et al., 2021). Two main approaches have been proposed to achieve this goal. In the former, agents perform multiple local optimization steps before sending

a model update to the central node (McMahan et al., 2017). The latter consists in compressing the messages exchanged (Alistarh et al., 2017; Horváth et al., 2022). In this chapter, we focus on the first approach which is widely used in practice. However, due to statistical heterogeneity, performing multiple steps can hinder convergence, as model updates target each agent's local minimizer (Li et al., 2019; Ro et al., 2021). This results in a tradeoff between communication cost and convergence (Wang et al., 2020b) and a need for algorithms that mitigate *client drift* (Karimireddy et al., 2020).

Most of existing FL algorithms minimize a training loss. However, their results do not provide reliable uncertainty quantification, a strong requirement in safety-critical applications (Coglianese and Lehr, 2016; Fatima et al., 2017). We address this problem by considering the federated version of Bayesian inference (Welling and Teh, 2011; Yurochkin et al., 2019; Chen and Chao, 2021; Izmailov et al., 2021; Wilson et al., 2021). The objective is to compute the predictive distribution, the highest posterior density regions (HPD). To this end, it is required to sample the posterior distribution $\pi \propto \exp(-U)$ associated with the model at hand. This target posterior decomposes into the product of local posteriors $\pi = \prod_{i\in[n]} \pi^i$. It is well known that sampling according to product distributions (Neiswanger et al., 2014; Hoffman et al., 2013; Minsker et al., 2014; Wang et al., 2015; Al-Shedivat et al., 2021; Dai et al., 2021) raises serious computational challenges even when sampling from each local posterior π^i is reasonably easy. We tackle this question in our contributions which can be summarized as follows.

Contributions.

- We study a random loop version of the `FALD` algorithm proposed in Deng et al. (2021), and we establish non-asymptotic upper bounds in Wasserstein distance for strongly convex potentials U. An analysis of `FALD` was conducted in (Deng et al., 2021, Theorem 5.7). However, the proof is plagued by an error; see Section 3.B.1.
- We give matching lower bounds to show that even with full batch gradients, `FALD` can be slower than Stochastic Gradient Langevin Dynamics (`SGLD`) due to client-drift.
- We propose a new method `VR-FALD`* that circumvents the shortcomings of `FALD`. This algorithm extends the Shifted Local-SVRG of Gorbunov et al. (2021) to the Bayesian context. It combines Stochastic Variance Reduced Gradient (SVRG) Langevin Dynamics (LD) (Dubey et al., 2016) and adapts the bias reduction techniques from `SCAFFOLD` (Karimireddy et al., 2020).
- We derive theoretical guarantees for `VR-FALD`* which highlight its gradient variance reduction effect and its ability to deal with data heterogeneity.
- The results are based on a general framework developed in the supplement, that encompasses a broad family of federated Bayes algorithms based on Langevin dynamics. This is the first unifying study among existing works on federated Bayesian inference.
- Finally, in Section 3.4 we illustrate our results using classical FL benchmarks and provide a thorough comparison with existing FL Bayesian methods.

Related works. Many distributed MCMC algorithms have been proposed in the last decade, and it is difficult to credit all the references. The first significant contributions in this direction are the Consensus Monte Carlo (CMC) approach and "embarrassingly parallel" MCMC algorithms; see, e.g. Neiswanger et al. (2014); Wang and Dunson (2013); Scott et al. (2016). These methods require running separate MCMC chains on each client/computational node, with each chain targeting the local posterior π^i. In the final stage, the algorithms recombine the samples from these chains to generate samples from the desired global posterior π (Minsker et al., 2014). The local posteriors may differ significantly from each other due to statistical heterogeneity, data imbalance, and / or inaccurate approximation. The effectiveness of the final combinations is either based on stringent assumptions on the local likelihoods (Liu and Ihler, 2014; Nemeth and Sherlock, 2018; Mesquita et al., 2020; Chittoor and Simeone, 2021) or on "fusion" algorithms that are exact but scale badly with the dimension; see, e.g. Dai et al. (2021); De Souza et al. (2022).

Vono et al. (2020); Rendell et al. (2020); Plassier et al. (2021); Vono et al. (2022a) introduced hierarchical Bayesian models to simulate separate MCMC chains on each machine. Inspired by the alternating direction method of multipliers (Boyd et al., 2011), each client is assigned an auxiliary parameter that is conditionally independent given the server parameter. These authors developed MCMC schemes which alternate between sampling the clients parameters given the server parameter, and sampling the server parameter given the clients parameters. However, these approaches require tuning an additional hyperparameter to control the dispersion of the "local parameters". This parameter characterizes the tradeoff between computational tractability and closeness to the original target distribution.

A competing approach to Federated Averaging, the quantized-SGD scheme, has been proposed in (Alistarh et al., 2017) for non-Bayesian FL. In this framework, the agents do not adapt parameters locally, but a random subset of the agents compute at each iteration a new gradient estimator and transmit a compressed form—see Haddadpour et al. (2021), among many others, (Bernstein et al., 2018; Tang et al., 2021) for scalar quantization or (Shlezinger et al., 2020), for vector quantization. These approaches have been extended to the Bayesian inference context in Lee et al. (2020); Zhang et al. (2022); Vono et al. (2022b). Performance analysis is given in Vono et al. (2022b); Sun et al. (2022).

The Federated Gradient Stochastic Langevin Dynamics (FSGLD) algorithm introduced by El Mekkaoui et al. (2021) extends the distributed-SGLD (DSGLD) (Ahn et al., 2014) to the FL setting. Specifically, FSGLD operates passing a Markov chain between computing nodes and using only local data to estimate gradients at each step.

Methods with multiple local steps have been considered by several authors. Deng et al. (2021) designed FALD as a Bayesian version of FedAvg. Al-Shedivat et al. (2021) proposed FedPa as a generalization of FedAvg. This method performs several local steps to infer Gaussian approximations of the clients local parameters. These local parameters are then reweighed using the estimated local means and covariance matrices before being aggregated on the central server.

Notation and Convention. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$, and we set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^\star$, we refer to $\{1, \ldots, n\}$ with the notation $[n]$. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures on $\mathbb{R}^d$ with finite 2-moment. For any random

variable ξ with values in $\mathbb{R}^d$, we define $\mathrm{Var}(\xi) = \mathbb{E}[\|\xi - \mathbb{E}\xi\|^2]$. Let μ, ν be in $\mathcal{P}_2(\mathbb{R}^d)$, we define the Wasserstein distance of order 2 by $W_2(\mu,\nu) = (\inf_{\zeta\in\mathbf{\Pi}(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - x'\|^2 \mathrm{d}\zeta(x,x'))^{1/2}$, where $\mathbf{\Pi}(\mu,\nu)$ is the set of transference plans of μ and ν.

3.2 Algorithm derivation

We aim to sample a target probability density function π defined for $x \in \mathbb{R}^d$ by

$$\pi(x) \propto \prod_{i=1}^n \pi^i(x), \qquad \pi^i(x) \propto \exp(-U^i(x)), \tag{3.1}$$

where n is the number of clients and the potential U^i is a finite sum expressed by

$$U^i(x) = \varpi^i U^0(x) + \sum_{j=1}^{N_i} U^{i,j}(x),$$

with $\{\varpi^i\}_{i\in[n]} \in [0,1]^n$ and $\sum_{i\in[n]} \varpi^i = 1$. This setting encompasses the Bayesian federated learning as a particular case, in which π stands for the global posterior distribution and $\{\pi^i\}_{i\in[n]}$ are referred to as local posteriors (Wu and Robert, 2017; Dai et al., 2021). In this case U^0 is the global negative log-prior, N_i denotes the number of observations of client i, $U^{i,j}$ is the negative log-likelihood of the j-th data of client i, and $\varpi^i U^0$ is the fraction of the negative log-prior allocated to this client (Rendell et al., 2020).

Federated Averaging Langevin Dynamics (FALD). FALD, proposed in Deng et al. (2021), is an extension to the Bayesian setting of FedAvg (McMahan et al., 2017). The updates performed on the ith client define a sequence of local parameters $(X_k^i)_{k\in\mathbb{N}}$ which are transmitted according to some preset schedule (which is deterministic in Deng et al. (2021) and is random in this work) to a central server. The central server averages the local parameters to update the global parameter. This global parameter is finally transmitted back to each client, and is used as a starting point of a new round of local iterations. Hence, each iteration $k \geq 0$ of FALD can be decomposed into two steps:

(1) **Local iteration on each client.** Each client i performs one step of the Langevin Monte Carlo algorithm (Grenander and Miller, 1994; Roberts and Tweedie, 1996) with a stochastic gradient associated with its local potential:

$$\begin{aligned} G_{k+1}^i &= \hat{\nabla} U_{k+1}^i(X_k^i), \\ \tilde{X}_{k+1}^i &= X_k^i - \gamma G_{k+1}^i + \sqrt{2\gamma}\, Z_{k+1}^i, \end{aligned} \tag{3.2}$$

where $\gamma > 0$ and for $x \in \mathbb{R}^d$, $\hat{\nabla} U_{k+1}^i(x)$ is an unbiased estimator of $\nabla U^i(x)$ given by (see Welling and Teh (2011) – general updates are considered in the supplement)

$$\hat{\nabla} U_{k+1}^i = \varpi^i \nabla U^0 + (N_i/b_i) \sum_{j\in S_{k+1}^i} \nabla U^{i,j}, \tag{3.3}$$

where $(S_k^i)_{k\in\mathbb{N}^\star}$ is a sequence of i.i.d. uniform random subsets of $[N_i]$ of cardinal number b_i. Moreover, $(Z_k^i)_{k\in\mathbb{N}^*}$, $i \in [n]$ are sequence of i.i.d Gaussian random variables which might be correlated across the agents and the central server. More precisely, given independent sequences, $(\tilde{Z}_k^i)_{k\in\mathbb{N}^*}$, $i \in [n]$ and $(\tilde{Z}_k)_{k\in\mathbb{N}^*}$ of i.i.d. d-dimensional standard Gaussian random variables, for $\tau \in [0,1]$ we set

$$Z_k^i = \sqrt{\tau}\, \tilde{Z}_k + \sqrt{1-\tau}\, \tilde{Z}_k^i. \tag{3.4}$$

(2) **A local update.** With probability $p_c \in (0, 1]$, the ith client communicates its parameter $\tilde{X}^i_{k+1}$, resulting from the first step, to the central server which in turns broadcasts the average $X_{k+1} = n^{-1} \sum_{i\in[n]} \tilde{X}^i_{k+1}$. Finally, each client updates its parameter as $X^i_{k+1} = X_{k+1}$. When no communication is performed, each client updates its parameter as $X^i_{k+1} = \tilde{X}^i_{k+1}$.

The local recursions defined by FALD can be written for $i \in [n]$ and $k \geq 0$ as

$$X^i_{k+1} = (1 - B_{k+1})\tilde{X}^i_{k+1} + (B_{k+1}/n) \sum_{j\in[n]} \tilde{X}^j_{k+1}, \tag{3.5}$$

where $(B_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. Bernoulli random variables with parameter p_c.

For $k \geq 1$, denote by $\mu^{(\gamma)}_k$ the distribution of the average parameter

$$X_k = (1/n) \sum_{i\in[n]} X^i_k. \tag{3.6}$$

Non-asymptotic Wasserstein bounds between $\mu^{(\gamma)}_k$ and the target distribution π are established in Theorem 3.1 under the following assumptions.

A 1. *For any $i \in [n]$, U^i is continuously differentiable. In addition, there exist $m, L > 0$ such that for any $i \in [n]$, the function U^i is L-smooth and m-strongly convex, i.e., for any $x, x' \in \mathbb{R}^d$,*

$$(m/2)\|x' - x\|^2 \leq U^i(x') - U^i(x) - \langle \nabla U^i(x), x' - x\rangle \leq (L/2)\|x' - x\|^2.$$

A 2. *For any $i \in [n]$, $(\{\hat{\nabla} U^i_k\}_{i\in[n]})_{k\in\mathbb{N}}$ are i.i.d. unbiased estimates of $\{\nabla U^i\}_{i\in[n]}$. In addition, there exists $\hat{L} \geq 0$ such that for any $x, x' \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|\hat{\nabla} U^i_k(x') - \hat{\nabla} U^i_k(x)\|^2\right] \leq \hat{L}^2\|x' - x\|^2.$$

In the minibatch scenario (3.3), **A**2 is satisfied if for $i \in [n]$, $j \in [N_i]$ there exists $L^i_j \geq 0$ such that for any $x, x' \in \mathbb{R}^d$, $\|\nabla U^{i,j}(x') - \nabla U^{i,j}(x)\| \leq L^i_j\|x' - x\|$.

Finally, we also consider the following optional smoothness condition on the potentials $\{U^i\}_{i\in[n]}$. This additional assumption, often satisfied in applications have been considered *e.g.* in Durmus and Moulines (2019); Dalalyan and Karagulyan (2019).

HX 1. *There exists $\tilde{L} \geq 0$, such that for any $i \in [n]$, the function U^i is three times continuously differentiable and for any $x, x' \in \mathbb{R}^d$, $\|\nabla^2 U^i(x) - \nabla^2 U^i(x')\| \leq \tilde{L}\|x - x'\|$.*

We introduce some key quantities appearing in the theoretical derivations below. Denote by $x_\star$ the minimizer of $\sum_{i\in[n]} U^i$ which exists and is unique under **A**1. We define

$$\begin{aligned} \mathsf{V}_\pi &= \textstyle\int_{\mathbb{R}^d} \mathrm{Var}\{n^{-1}\sum_{i\in[n]} \hat{\nabla} U^i_1(x)\}\pi(\mathrm{d}x), \\ \mathsf{V}_\star &= \mathrm{Var}\{n^{-1}\textstyle\sum_{i\in[n]} \hat{\nabla} U^i_1(x_\star)\}, \end{aligned} \tag{3.7}$$

the average of the stochastic gradient variance under the stationary distribution π and at the minimum $x_\star$, respectively. Finally, the statistical heterogeneity between the clients is quantified by (see, e.g. Stich et al. (2018))

$$\mathsf{H} = n^{-1} \textstyle\sum_{i\in[n]} \|\nabla U^i(x_\star)\|^2.$$

For ease of presentation, for two sequences $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ we write $a_k \lesssim b_k$ if there exists $C > 0$ only depending on the constants introduced in **A**1, **A**2 and **HX**1 such that $a_k \leq C b_k$, for any $k \in \mathbb{N}$.

Theorem 3.1 (Simplified). *Assume **A**1, **A**2 and suppose for any $i \in [n]$, $X_0^i = X_0$. Then, there exist $\bar{\gamma} > 0$, such that for any $\gamma \in \left(0, \bar{\gamma}\right]$, $k \in \mathbb{N}$, $X_0 \sim \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$W_2^2(\mu_k^{(\gamma)}, \pi(\cdot|\mathcal{D})) \lesssim (1 - \gamma m/8)^k \, \mathsf{I}(\mu_0) + \frac{\gamma^e}{n}\mathsf{J} + \gamma \mathsf{V}_\pi + \frac{\gamma^2(1-p_c)}{p_c^2}\left\{\mathsf{H} + p_c \mathsf{V}_\star + \frac{d}{n}\right\} + \frac{\gamma(1-\tau)(1-n^{-1})d}{p_c},$$

*where $\mathsf{J} = d$, $e = 1$ and $\mathsf{I}(\mu_0) < \infty$ is a function of the initial condition μ_0. If **HX**1 holds, then $e = 2$ and $\mathsf{J} = d(1 + d/n)$.*

Elements of proof are provided in Section 3.3; a precise statement is given in Theorem 3.22 with detailed proofs. Note the step-size upper bound $\bar{\gamma}$ is proportional to p_c. In the single user case ($n = p_c = \tau = 1$), we recover up to numerical constants the results stated in Durmus and Moulines (2019); Dalalyan and Karagulyan (2019). Note that, under **HX**1 the leading term in the step-size γ is proportional to the stochastic gradient variance V_π, in accordance with the bounds obtained for SGLD by *e.g.* , Dalalyan and Karagulyan (2019). More discussions on these bounds are postponed after the statement of Theorem 3.3.

Lower bounding the effect of heterogeneity. Similar to FedAvg, the convergence of FALD is impaired by data heterogeneity. Multiple local SGLD steps described in (3.2) cause X_k^i to target the local posteriors $\pi^i \propto \exp(U^i)$. We now provide lower bound on the Wasserstein distance between the distribution of the samples generated by FALD and the target distribution π which is proportional to the heterogeneity $\gamma^2\mathsf{H}$.

Proposition 3.2. *There exist $\bar{\gamma} > 0$, potentials $\{U^i\}_{i=1}^2$ on $\mathbb{R}$ satisfying **A**1, **HX**1 and an instance of FALD satisfying **A**2 such that for any $\gamma \in \left(0, \bar{\gamma}\right]$, we have*

$$\liminf_{k \to \infty} W_2^2(\mu_k^{(\gamma)}, \pi(\cdot|\mathcal{D})) \gtrsim \gamma^2 \mathsf{H}.$$

This proposition extends Karimireddy et al. (2020, Theorem II) to the Bayesian context and underlines the same limitation as FedAvg. To circumvent this, various bias reduction techniques have been suggested in the stochastic optimization literature (Horváth et al., 2022; Gorbunov et al., 2021). In the next section, we adapt similar mechanisms to derive an alternative to FALD satisfying better finite bounds.

FALD with control variates and bias reduction. To mitigate the impact of local stochastic gradients, we adapt variance-reduction techniques (Wang et al., 2013; Kovalev et al., 2020) and bias-reduction techniques (Horváth et al., 2022; Gorbunov et al., 2021). This new approach introduces a different recursion rule in step (1) of FALD, while keeping step (2) unchanged. The local update rule is based on a reference point $Y_k \in \mathbb{R}^d$ common to all clients. This common point is updated with probability $q_c \in \left(0, 1\right]$ and allows the inclusion of a local shift C_k to recenter the local gradients. This mechanism eliminates the "infamous non-stationarity of the local methods" (paraphrasing Gorbunov et al. (2021)) and therefore avoids extra bias. At each iteration k, the first step of the VR-FALD* algorithm is divided into two parts:

(1.1) **Update of the reference parameter and control variate.** The variance reduced gradient requires a sporadic computation of the full local gradient. Let $(B_k^Y)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. Bernoulli random variables with parameter $q_c \in (0,1]$. If $B_{k+1}^Y = 1$, then the client reference point Y_k is updated: the clients transmit their local parameter $\{X_k^i\}_{i\in[n]}$ to the central server which computes their average $Y_{k+1} = n^{-1}\sum_{i\in[n]} X_k^i$; which is sent back to the clients. The clients then compute the full gradients $\{\nabla U^i(Y_{k+1})\}_{i\in[n]}$ and transmit them to the central server which updates the shift $C_{k+1} = n^{-1}\sum_{i\in[n]} \nabla U^i(Y_{k+1})$. To summarize, the reference point and the shift are updated according to

$$\begin{aligned} Y_{k+1} &= (1 - B_{k+1}^Y)Y_k + (B_{k+1}^Y/n)\textstyle\sum_{i\in[n]} X_k^i, \\ C_{k+1} &= (1 - B_{k+1}^Y)C_k + (B_{k+1}^Y/n)\textstyle\sum_{i\in[n]} \nabla U^i(Y_{k+1}). \end{aligned} \tag{3.8}$$

(1.2) **Local iteration on each client.** This step is similar to FALD, upon replacing the local updates (2) by the variance-reduced version

$$G_{k+1}^i = \hat{\nabla} U_{k+1}^i(X_k^i) - \hat{\nabla} U_{k+1}^i(Y_k) + C_k, \tag{3.9}$$

$$\tilde{X}_{k+1}^i = X_k^i - \gamma G_{k+1}^i + \sqrt{2\gamma} Z_{k+1}^i. \tag{3.10}$$

The VR-FALD* analysis relies on the following additional assumption.

A 3. *There exists $\omega \geq 0$ such that for any $i \in [n]$, $k \in \mathbb{N}^\star$ and $x, y \in \mathbb{R}^d$, the following inequality holds*

$$\mathbb{E}\left[\left\|\hat{\nabla} U_k^i(x) - \hat{\nabla} U_k^i(y) - \nabla U^i(x) + \nabla U^i(y)\right\|^2\right] \leq \omega\|x - y\|^2.$$

Under **A**1 and **A**2, **A**3 is satisfied with $\omega = 2L^2 + 2\hat{L}^2$. However, using this result leads to some discrepancy in previous existing analysis, since $\omega = 0$ in the non-stochastic gradient case while $2L^2 + 2\hat{L}^2 \neq 0$ in general. Finally, in the minibatch scenario (3.3), if $\{\nabla U^{i,j}\}_{j\in[N_i]}$ are L_i-Lipschitz, then **A**3 holds with $\omega = \max_{i\in[n]}\{N_i L_i^2/b_i\}$; see Remark 3.17.

For $k \geq 0$, denote by $\mu_k^{(\mathrm{VR}\star,\gamma)}$ the distribution of the average $X_k = n^{-1}\sum_{i\in[n]} X_k^i$ where X_k^i is defined as in (3.5) with $\tilde{X}_k^i$ given in (3.10). With these notations, we obtain the following theoretical guarantee on VR-FALD*.

Theorem 3.3 (Simplified)**.** *Assume **A**1, **A**2, **A**3 and suppose for $i \in [n]$, $X_0^i = Y_0 = X_0$. Then, there exist $\bar{\gamma}^{\mathrm{VR}\star} > 0$, such that for any $q_c \leq p_c$, $\gamma \in (0, \bar{\gamma}^{\mathrm{VR}\star}]$, $k \in \mathbb{N}$, $X_0 \sim \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$\begin{aligned} W_2^2(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi) \lesssim (1 - \gamma m/8)^k\, \mathsf{I}^{\mathrm{Vr}\star}(\mu_0) + \frac{\gamma^e}{n}\mathsf{J} + \frac{\gamma^2 d}{n q_c}\omega \\ + \frac{\gamma(1-\tau)(1-n^{-1})d}{p_c} + \frac{\gamma^2(1-p_c)}{p_c^2}\left\{\gamma \mathsf{V}_\star + \frac{d}{n}\right\}, \end{aligned}$$

*where $\mathsf{J} = d$, $e = 1$, $\mathsf{V}_\star$ is defined in (3.7), $\mathsf{I}^{\mathrm{Vr}\star}(\mu_0) < \infty$ is a function of the initial condition μ_0. If **HX**1 holds, then $e = 2$ and $\mathsf{J} = d(1 + d/n)$.*

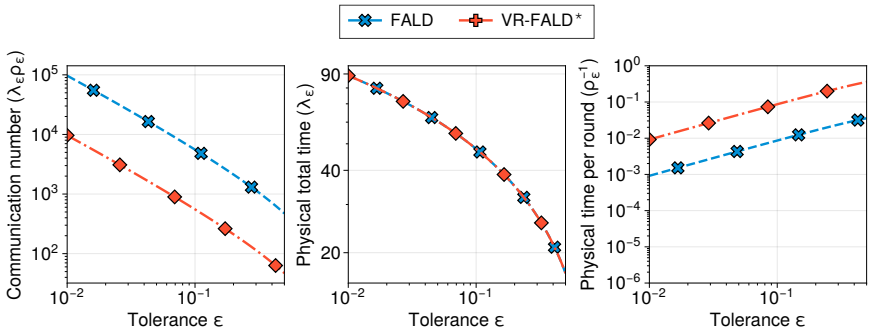
The proof is postponed to Section 3.B.2. Compared to Theorem 3.1, the client-drift term does no longer appear, highlighting the advantage of VR-FALD* in dealing with data heterogeneity between agents.

Further, the variance of the stochastic gradients of VR-FALD* only appear in the factor $\gamma^2\omega$. This result agrees with Chatterji et al. (2018) for SVRG-LD, which might be seen as a particular instance of VR-FALD* with $n = 1$, $p_c = 1$. Nevertheless, a close inspection of the proof in Chatterji et al. (2018) reveals a gap—see Remark 3.33, which is corrected in the proof of Theorem 3.32.

Complexity and Communication costs. We now discuss the complexity and communication costs of FALD and VR-FALD*. We study two extreme cases: (A) the local computation cost is negligible and only the communication cost matters, which is typical in cross-device applications. (B) the communication cost is negligible and only the local computation cost (complexity) matters. More general scenarios are discussed in the supplement Section 3.D. In this discussion, it is assumed that **HX**1 is satisfied and $\tau = 1$. In both cases, for a target precision $\epsilon > 0$, we optimize the hyperparameters (number of iterations K_ϵ, learning rate γ_ϵ, probability of communication $p_{c,\epsilon}$) to ensure $W_2(\mu_{K_\epsilon}^{(\gamma)}, \pi) \leq \epsilon$ (FALD) or $W_2(\mu_{K_\epsilon}^{(\mathrm{VR}\star,\gamma)}, \pi) \leq \epsilon$ (VR-FALD*). The values of the parameters d, m, ω, H, J, V_π and $\mathsf{V}_\star$ are reported in Table 3.5.
(Scenario A) The objective is to minimize the number of communications $p_{c,\epsilon}K_\epsilon$. As γ can be arbitrarily small, we set $K_\epsilon = \gamma^{-1}\lambda_\epsilon$, $p_{c,\epsilon} = \rho_\epsilon\gamma$, where $\lambda_\epsilon, \rho_\epsilon > 0$. Hence, the optimization problem becomes $\min\{\lambda_\epsilon\rho_\epsilon\}$ subject to $\mathsf{I}(\mu_0)\exp(-\lambda_\epsilon m/8) + \rho_\epsilon^{-2}(\mathsf{H} + d/b) \leq \epsilon^2$. As $\epsilon \downarrow 0^+$, the minimum number of communications $p_{c,\epsilon}K_\epsilon$ scales as $\tilde{O}(\epsilon^{-1}\sqrt{\mathsf{H} + n^{-1}d})$ for FALD and $\tilde{O}(\epsilon^{-1}\sqrt{n^{-1}d})$ for VR-FALD*.
(Scenario B) We take $p_{c,\epsilon} = 1$ and seek to minimize the total number of iterations K_ϵ. As $\epsilon \downarrow 0^+$, K_ϵ scales as $\tilde{O}(\epsilon^{-2}(\mathsf{V}_\pi + \epsilon\sqrt{n^{-1}\mathsf{J}}))$ for FALD and $\tilde{O}(\epsilon^{-1}\sqrt{n^{-1}\mathsf{J} + n^{-1}\omega d})$ for VR-FALD*.
In Figures 3.1a-3.1b, we display the optimal number of communications $p_{c,\epsilon}K_\epsilon$ as a function of ϵ (left panels Figures 3.1a-3.1b). We also exhibit the *physical time* which corresponds to the time of the Langevin diffusion. The total physical times – λ_ϵ for (A) and $\gamma_\epsilon K_\epsilon$ for (B) – are displayed in the middle panels Figures 3.1a-3.1b. Finally, the right panels Figures 3.1a-3.1b represent the average physical time between two consecutive communications — ρ_ϵ^{-1} for (A) and $\gamma/p_{c,\epsilon}$ for (B). Note that, the total physical time is (almost) the same for FALD, VR-FALD*, in scenarios (A) and (B). VR-FALD* significantly reduces the number of communications $p_{c,\epsilon}K_\epsilon$ in scenario (A) (top panel) and number of rounds K_ϵ (B) (bottom panel) *w.r.t.* FALD.

Figures 3.1a-3.1b also illustrate that the "embarrassingly parallel" approach of (Neiswanger et al., 2014) is far from optimal. Indeed, our results show the importance of making multiple interactions (rather than a single consensus step) and using correlated noises between clients. In scenario (A), the optimal number of communications scales inversely proportional to $1/\epsilon$ which improve the bounds $\tilde{O}(1/\epsilon^2)$ derived in Deng et al. (2021, Section 5.3.1). For scenario (B), FALD has the same complexity as QLSD Vono et al. (2022b) under similar assumptions; see also Sun et al. (2022). VR-FALD* has the lowest complexity ($\tilde{O}(1/\epsilon)$) among the Bayesian Federated algorithms reported earlier. This bound matches the one obtained by Chatterji et al. (2018) for the fully centralized SVRG-LD (corresponding to $n = 1$).

(Scenario A) Numerical results optimizing $p_{c,\epsilon}K_\epsilon$.



(Scenario B) Numerical results optimizing K_ϵ.

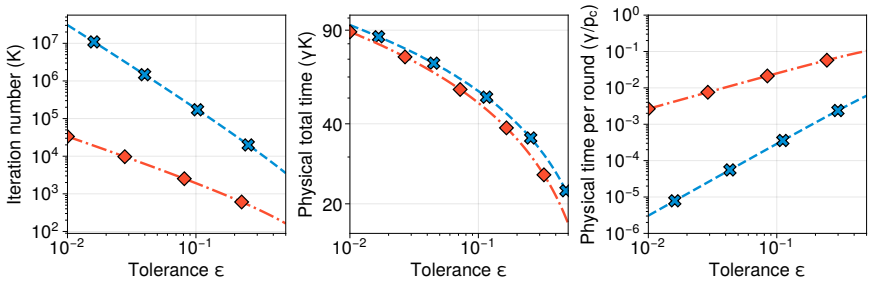


Figure 3.1 – Complexity and Communication costs.

3.3 Proofs outline

We briefly outline the main steps of the proof of Theorems 3.1 and 3.3. Details of the proofs can be found in the supplementary chapter, where we analyze the two algorithms under a common unifying framework. For both algorithms, the local parameters $(X_k^i)_{i\in[n]}$, $k \geq 0$, are given by (3.5), where $(\tilde{X}_k^i)_{i\in[n]}$ stands for local iterations, which are given in (3.2) for `FALD` and (3.9) for `VR-FALD`*. Then, we bound the Wasserstein distance between the target distribution π and the distribution of $X_k = n^{-1}\sum_{i\in[n]} X_k^i$ which is denoted by $(\mu_k^{(\gamma)})_{k\in\mathbb{N}}$. The Wasserstein distance is defined as the infimum over the coupling. We use below the synchronous coupling construction used in (Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019) for the analysis of Stochastic Gradient Langevin algorithms.

Synchronous coupling. We first construct a Brownian motion $(\mathsf{W}_t)_{t\geq 0}$ by $\mathsf{W}_t = \sqrt{\tau}\,\tilde{\mathsf{W}}_t + \sqrt{(1-\tau)/n}\,\sum_{i\in[n]}\tilde{\mathsf{W}}_t^i$, starting from $b+1$ independent d-dimensional standard Brownian motions $(\tilde{\mathsf{W}}_t^i)_{t\geq 0}$, $i \in [n]$, and $(\tilde{\mathsf{W}}_t)_{t\geq 0}$. Second, we define the following standard Gaussian random variables $\tilde{Z}_{k+1}^i = \gamma^{-1/2}(\tilde{\mathsf{W}}^i_{(k+1)\gamma} - \tilde{\mathsf{W}}^i_{k\gamma})$, $\tilde{Z}_{k+1} = \gamma^{-1/2}(\tilde{\mathsf{W}}_{(k+1)\gamma} - \tilde{\mathsf{W}}_{k\gamma})$, and we set Z_k^i as in (3.4). For $k \in \mathbb{N}$, it holds that $\sqrt{\gamma}\sum_{i\in[n]} Z_{k+1}^i = \sqrt{n}(\mathsf{W}_{(k+1)\gamma} - \mathsf{W}_{k\gamma})$. Finally, we consider $(\mathsf{X}_t)_{t\geq 0}$ the strong solution of the Langevin diffusion associated with π and starting from $\mathsf{X}_0 \sim \pi$ (see (3.1)) and driven by $(\mathsf{W}_t)_{t\geq 0}$:

$$\mathrm{d}\mathsf{X}_t = -(1/n)\sum_{i\in[n]} \nabla U^i(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2/n}\,\mathrm{d}\mathsf{W}_t. \tag{3.11}$$

Under **A**1 and **A**2, π is the unique stationary distribution for the Langevin diffusion, hence the distribution of X_t is π for all $t \geq 0$; see *e.g.* Roberts and Tweedie (1996). Hence, $(X_k, \mathsf{X}_{k\gamma})$ defines a coupling between $\mu_k^{(\gamma)}$ and π, thus for any $k \in \mathbb{N}$ we get

$$W_2^2(\mu_k^{(\gamma)}, \pi) \leq \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right].$$

The rest of the proof then consists in bounding the right-hand side. It is worth noting that in contrast to most analysis on Langevin dynamics, we consider a Langevin diffusion (3.11) we scale the gradient term by n^{-1} and the Brownian motion by $n^{-1/2}$. This scaling is adapted to the averaging procedure defining $(X_k)_{k\in\mathbb{N}}$.

Decomposition of $\mathbb{E}[\|X_k - \mathsf{X}_{k\gamma}\|^2]$. Denote by $\mathcal{F}_k$ the filtration generated by $\mathsf{X}_0, (\mathsf{W}_t)_{t\leq k\gamma}$ and $(\{X_l^i\}_{i=1}^n)_{l\leq k}$. Using the definition (3.6) of $(X_k)_{k\in\mathbb{N}}$ combined with **A**1, we show in Proposition 3.5 that for any $\gamma \lesssim 1$

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \lesssim (1 - \gamma m/2)\|\mathsf{X}_{k\gamma} - X_k\|^2 + E_k + \gamma^2 S_k + V_k, \tag{3.12}$$

where $V_k = n^{-1}\sum_{i\in[n]} \|X_k^i - X_k\|^2$ and

$$S_k = \mathrm{Var}^{\mathcal{F}_k}(n^{-1}\textstyle\sum_{i\in[n]} G_k^i),$$
$$E_k = \gamma^{-1}\|\mathbb{E}^{\mathcal{F}_k}[I_k]\|^2 + \mathbb{E}^{\mathcal{F}_k}\left[\|I_k\|^2\right],$$

with $I_k = n^{-1}\sum_{i\in[n]} \int_{k\gamma}^{(k+1)\gamma} (\nabla U^i(\mathsf{X}_s) - \nabla U^i(\mathsf{X}_{k\gamma}))\mathrm{d}s$.

Bounding E_k. The term E_k accounts for the difference between the diffusion and its discretization; the bound is the same for FALD and VR-FALD*. By adapting Durmus and Moulines (2019, Lemma 21), we establish in Lemma 3.8 that

$$\mathbb{E}\left[E_k\right] \lesssim \gamma^2 d/n. \tag{3.13}$$

Under **HX**1, for $\gamma \lesssim 1$ the bound can be sharpened in

$$\mathbb{E}\left[E_k\right] \lesssim (\gamma^3 d/n)(1 + d/n). \tag{3.14}$$

The right-hand side of (3.13) has a higher order with respect to the step-size γ in comparison to (3.14). This step is the reason why we consider the more restrictive assumption **HX**1, which leads to different guarantees depending on whether this condition is met or not.

Bounding S_k. S_k is the conditional variance of the stochastic gradient. This is the main difference between the two algorithms. For FALD, we show in Lemma 3.21 that

$$\mathbb{E}\left[S_k\right] \lesssim \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + \mathbb{E}\left[V_k\right] + \mathsf{V}_\pi. \tag{3.15}$$

On the other hand, under **A**3, we establish in Lemma 3.29 that for VR-FALD*, it holds that

$$\mathbb{E}\left[S_k\right] \lesssim \omega\mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + \omega\mathbb{E}\left[V_k\right] + \frac{\gamma\omega d}{n q_c} + \omega q_c \sum_{l=0}^{k-1}(1 - q_c)^{k-l-1}\mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right].$$

Compared to the inequality (3.15), which holds for FALD, the variance term V_π for VR-FALD* is replaced by $\gamma\omega d/nq_c$, which can be made arbitrarily small with $\gamma \to 0$. Note that this term is inversely proportional to the update probability q_c of the control variate. Interestingly, the term S_k vanishes when $\omega = 0$, i.e., when each client uses its full local gradient at each iteration.

Bounding V_k. We show in Lemma 3.20 (FALD) and Lemma 3.28 (VR-FALD*), there exist $a_0, a_1 \geq 0$ satisfying

$$\mathbb{E}[V_k] \leq (1 - \gamma m/8)^k a_0 + a_1.$$

To establish this result, we consider the sequence $(f_k)_{k\in\mathbb{N}}$ with general term given by

$$f_k = V_k + \alpha_d d_k^2 + \alpha_\sigma \sigma_k^2,$$

where $\alpha_d, \alpha_\sigma \geq 0$ are given in (3.101); $d_k = \|X_k - x_\star\|$ denotes the distance between the average parameter X_k and the minimizer $x_\star$ of the global potential U; $\sigma_k = 0$ for FALD and $\sigma_k^2 = n^{-1}\sum_{i\in[n]} \mathbb{E}^{\mathcal{F}_k}[\|\hat{\nabla} U_k^i(Y_k) - \hat{\nabla} U_k^i(x_\star)\|^2]$ for VR-FALD* with Y_k defined in (3.8). The weights α_d, α_σ are tailored to prove a contraction; more precisely, we show the existence of $a_2 > 0$ whose expression is given in Lemma 3.14, such that

$$f_{k+1} \leq (1 - \gamma m/4) f_k + \gamma^2 a_2 + 2\gamma d\left(1 - \tau\right)(1 - n^{-1}). \tag{3.16}$$

An immediate induction combines with $V_k \leq f_k$ yields a first bound for $\mathbb{E}[V_k]$ of the form (3.3) with a_1 of order γ. In a final step Lemma 3.12, we refine this bound to obtain a term a_1 of order γ^2.

Gathering all the bounds. The proof is concluded by plugging the upper bounds derived for E_k, S_k, V_k into (3.12).

3.4 Numerical experiments

To illustrate our findings, we perform three numerical experiments on both synthetic toy-examples and real datasets. We compare FALD, VR-FALD* with Bayesian federated learning benchmarks: DG-LMC (Plassier et al., 2021), the Federated Stochastic Langevin Dynamics FSGLD (El Mekkaoui et al., 2021), the Quantized Langevin Stochastic Dynamic (QLSD) and its variance-reduced version QLSD^{++} (Vono et al., 2022b). We also include in our benchmark state of the art (centralized MCMC) algorithms: HMC (Brooks et al., 2011), the Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) and the preconditioned SGLD (pSGLD) (Li et al., 2016).

Gaussian posterior. We consider $n = 100$ clients associated to local Gaussian potentials with mean $\{\mu_i\}_{i\in[n]}$ and covariance $\{\Sigma_i\}_{i\in[n]}$, i.e., $U^i(x) = (1/2)(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)$. For different values of the hyperparameters (p_c, γ, τ), we run 100 chains with $k_1 = 10^7$ iterations: $(X_k)_{k=1}^{k_1}$ and discard 10% of the samples (more details are reported in Section 3.E.1). For each chain, we estimate the posterior variance $\sigma_\star^2 = \int \|x - x_\star\|^2 \mathrm{d}\pi(x|\mathcal{D})$ using FALD and VR-FALD*, where $\pi(\cdot|\mathcal{D}) \propto \exp(-\sum_{i\in[n]} U^i)$ and $x_\star = \arg\max_{x\in\mathbb{R}^d} \pi(x|\mathcal{D})$. We compute a Monte-Carlo estimates (over 10^2 independent

Probability p_c	$p_c = 1/5$			$p_c = 1/10$			$p_c = 1/20$		
Stepsize γ	$\frac{1}{2}p_c\bar{\gamma}$	$\frac{1}{5}p_c\bar{\gamma}$	$\frac{1}{10}p_c\bar{\gamma}$	$\frac{1}{2}p_c\bar{\gamma}$	$\frac{1}{5}p_c\bar{\gamma}$	$\frac{1}{10}p_c\bar{\gamma}$	$\frac{1}{2}p_c\bar{\gamma}$	$\frac{1}{5}p_c\bar{\gamma}$	$\frac{1}{10}p_c\bar{\gamma}$
FALD ($\tau = 0$)	2.5E+01	9.5E-01	3.9E-02	3.6E+01	1.1E+00	8.2E-02	4.2E+01	2.0E+00	1.1E-01
VR-FALD* ($\tau = 0$)	4.8E-02	2.6E-02	1.4E-02	5.0E-02	4.9E-02	3.7E-02	9.8E-02	5.3E-02	3.9E-02
VR-FALD* ($\tau = 1$)	2.8E-02	2.0E-02	1.3E-02	4.1E-02	3.7E-02	1.4E-02	8.6E-02	4.3E-02	2.1E-02

Table 3.1 – Asymptotic bias in function of τ, p_c and γ.

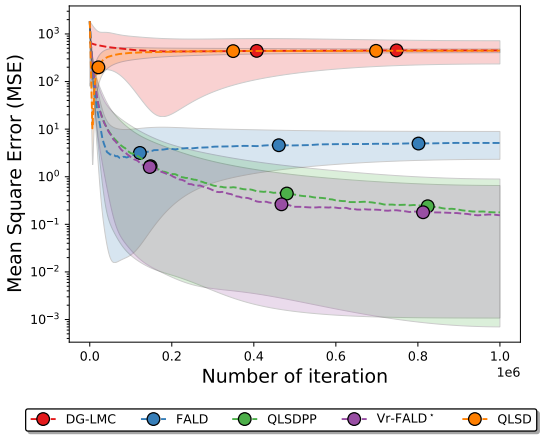


Figure 3.2 – MSE comparison with $p_c = 1/5$ and $\gamma = \bar{\gamma}/3$.

replications) of the Mean Squared Error (MSE) given by $\{(k_1 - k_0)^{-1} \sum_{k=k_0+1}^{k_1} \|X_k - x_\star\|^2 - \sigma_\star^2\}^2$ where k_1 is the total number of samples and k_0 is the burn-in period. The values of the hyperparameters are reported in Section 3.E.1. From Table 3.1, VR-FALD* always outperforms FALD for any choices of p_c, γ. This illustrates the impact of the heterogeneity and supports the theoretical findings given in Theorems 3.1 and 3.3. Furthermore, the asymptotic bias for VR-FALD* improves when $\tau = 1$ as derived in the theoretical analysis.

Bayesian Logistic Regression. We assess the performance of FALD and VR-FALD* using calibration metrics—the expected calibration error (ECE), the Brier score (BS), and the negative log likelihood (nNLL); see Guo et al. (2017)—and predictive accuracy. We consider Bayesian logistic regression applied to the Titanic dataset, which consists of $p = 2$ classes with $N = 2201$ samples in dimension $d = 4$. This dataset is allocated between $n = 10$ clients in a very heterogeneous manner, as displayed in Figure 3.3. We use an isotropic Gaussian prior with a mean of zero and variance 1. We also report the total variation distance between the predictive distribution obtained for FALD and VR-FALD* to the predictive distribution approximated by 100 long runs of Langevin Stochastic Dynamics (LSD). These metrics are evaluated on a test data sets of 441 samples, and the mean and standard deviation are reported in Table 3.2. Moreover, we illustrate the quality improvement of VR-FALD* over FALD in Figure 3.4. We compared the Wasserstein distance using POT (Flamary et al., 2021) between the empirical distributions generated by FALD, VR-FALD* to the estimated target distribution. Based on the same samples, we compute the relative highest posterior density (HPD) error; see Section 3.E.2 for details.

METHOD	Accuracy	Agreement	$10^4\times$ TV	10×ECE	10×BS	10×nNLL
LSD	72.4 ± 0.1	99.9 ± 0.1	5.53 ± 2.00	1.20 ± 0.01	3.44 ± 0.00	5.30 ± 0.00
FALD	77.0 ± 0.8	91.3 ± 0.9	533.32 ± 8.13	1.05 ± 0.09	3.37 ± 0.01	5.19 ± 0.00
VR-FALD⋆	74.9 ± 0.1	93.6 ± 0.1	287.81 ± 2.04	1.00 ± 0.05	3.51 ± 0.00	5.35 ± 0.00

Table 3.2 – Bayesian Logistic Regression on Titanic.

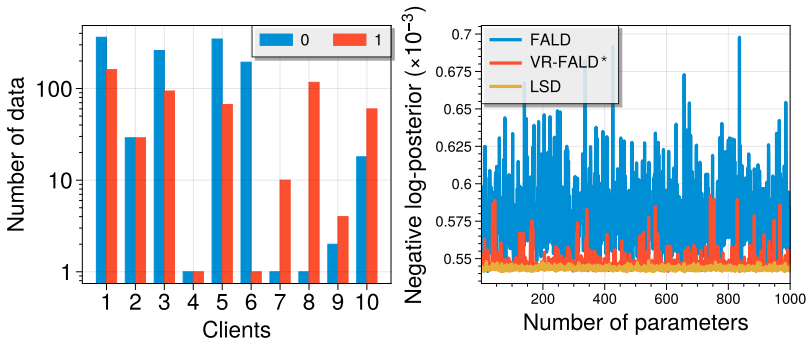


Figure 3.3 – Logistic regression – dataset distribution (Log Scale) and negative log-posterior (right).

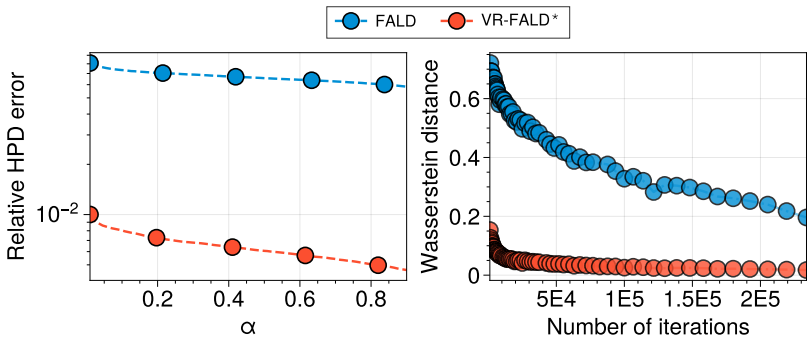


Figure 3.4 – Logistic regression – HPD relative error (left) and Wasserstein distance (right).

Bayesian Neural Network: MNIST. To illustrate the behavior of FALD and VR-FALD⋆ in a non-convex setting, we perform Bayesian Neural Network (BNN) inference on the MNIST dataset (Deng, 2012). To this end, we distribute the dataset to $n = 20$ clients as follows: 80% of the data labeled $\mathrm{y} \in \{0, \dots, 9\}$ are equally allocated to clients $i = \mathrm{y} + 1$ and $i = \mathrm{y} + 10$; the remaining data are evenly distributed among the n clients. The likelihood of the observations is computed using LeNet5 neural network (LeCun et al., 1998) with an isotropic Gaussian prior. Finally, we implement FALD and its variants with $p_c = 1/n$ and $q_c = N_n/N_d$, where N_n is the batch size used in the experiments and N_d is the total number of data. All standard deviations and the values of the other parameters are reported in Section 3.E.3.

In Table 3.3 we can observe that the best results are obtained by VR-FALD⋆: it achieves similar performance to the (fully centralized) SGLD and pSGLD. Alleviating client drift using control variates is still effective even in the highly non-convex BNN setting.

METHOD	SGLD	pSGLD	FALD	VR-FALD⋆	FSGLD
Accuracy	99.1	99.2	99.1	99.2	98.5
$10^3\times$ECE	6.88	21.6	4.07	4.34	6.34
$10^2\times$BS	1.66	1.45	1.47	1.39	2.39
$10^2\times$nNLL	3.53	4.24	3.06	3.43	4.87

Table 3.3 – Performance of Bayesian FL algorithms on MNIST.

Bayesian Neural Network: CIFAR10. We consider the CIFAR10 dataset (Krizhevsky et al., 2009) and the ResNet-20 model (He et al., 2016). We split the data across 20 clients, similar to the previous example. Denote by $\mathsf{Y} = \{\mathrm{y}_1, \ldots, \mathrm{y}_{10}\}$ the set of labels. Then 80% of the data associated with a label $\mathrm{y}_j \in \mathsf{Y}$, $j \in [10]$, is distributed among clients j and $j+10$, while the rest of the data is evenly distributed among clients. We assess the performance of FALD and VR-FALD⋆ against HMC, Deep Ensemble, and SGLD. We follow Izmailov et al. (2021) by computing the *accuracy*, *agreement*, and total deviation distance between the predictive distribution. All of these quantities are defined in the Appendix; see Section 3.E.4. We also report the calibration results and all resulting scores in Table 3.4; the results for HMC and SGLD are from Izmailov et al. (2021, Table 6). Details on the implementation and choice of hyperparameters can be found in Section 3.E.4. We can see that VR-FALD⋆ gives very similar results to SGLD and performs favorably in terms of agreement. Finally, FALD and VR-FALD⋆ outperform Deep Ensembles.

METHOD	HMC	SGD	DEEP ENS.	SGLD	FALD	VR-FALD⋆
Accuracy	89.6	91.57	91.68	89.96	**92.54**	92.03
Agreement	94.0	90.99	91.03	**92.43**	91.53	91.12
10× TV	0.74	1.45	1.49	**1.03**	1.42	1.39
$10^2\times$ECE	5.9	4.71	5.44	4.41	3.79	**3.26**
10×BS	1.4	1.69	1.45	1.53	**1.16**	1.20
10×nNLL	3.07	3.35	3.81	3.15	2.75	**2.63**

Table 3.4 – Performance of Bayesian FL algo. on CIFAR10.

3.5 Conclusion

In this work, we propose VR-FALD⋆ which extends the FALD Deng et al. (2021) algorithm by introducing control variates to mitigate client drift and reducing stochastic gradient variance. We develop a unifying framework for Bayesian FL combining ideas from Langevin Monte Carlo and Federated Averaging schemes. The theory covers a wide range of local stochastic gradient algorithms; connections can even be made with the global consensus Monte Carlo method (Rendell et al., 2020; Vono et al., 2022a). Using this theoretical framework, we develop non-asymptotic bounds for the algorithms FALD and VR-FALD⋆, and discuss the choice of hyperparameters (learning rate, communication probability, control variate update probability) to obtain optimal tradeoffs. Our analysis allows to correct some errors in the results obtained previously for FALD. The results we obtain on both toy examples and applications to BNNs clearly show the importance of variance reduction and heterogeneity, even when the potential is non-convex.

Theoretical road map. The derivations leading to Theorem 3.1 and Theorem 3.3 are split on two sections:

- Section 3.A consists of general results under mild assumptions. In this section, we derive an upper bound on V_k – see Section 3.A.3, and provide a Wasserstein upper bound holding for numerous federated averaging Langevin schemes in Theorem 3.10.
- Section 3.B is subdivided between the results on FALD (Section 3.B.1) and VR-FALD* (Section 3.B.2). In both subsections, we prove intermediate results showing that results of Section 3.A.3 hold, and finally we apply Theorem 3.10 to provide the final theoretical guarantees on FALD and VR-FALD*.

3.A General scheme and technical results

Problem statement. We consider a general recursion that includes both FALD and VR-FALD*. This general scheme is based on i.i.d. random variables $\{\xi_k\}_{k\in\mathbb{N}}$ taking values in a measurable space $(\mathsf{E},\mathcal{E})$ and whose joint distribution is denoted by ν_ξ. Moreover, we introduce a family of measurable functions $\{\mathcal{G}^i : \mathbb{R}^d \times \mathsf{Y}^2 \times \mathsf{C}^2 \times \mathsf{E} \to \mathbb{R}^d\,,\ \mathcal{Y}^i : \mathbb{R}^d \times \mathsf{Y}^2 \times \mathsf{E} \to \mathsf{Y}\,,\ \mathcal{C}^i : \mathbb{R}^d \times \mathsf{Y} \times \mathsf{C}^2 \times \mathsf{E} \to \mathsf{C}\}_{i=1}^n$, where $(\mathsf{Y},\mathcal{Y})$ and $(\mathsf{C},\mathcal{C})$ are measurable spaces. For each $i \in [n]$, the functions $(\mathcal{G}^i, \mathcal{Y}^i, \mathcal{C}^i)$ correspond to the update of the local parameter and control variate by the ith agent. To define the global control variate update, we consider the function $\mathcal{D} : \mathsf{Y} \times \mathsf{C}^{n+1} \times (\mathbb{R}^d)^{n+1} \times \mathsf{E} \to \mathsf{Y} \times \mathsf{C}$. Starting from $\{G_0^i\}_{i=1}^n, \{X_0^i\}_{i=1}^n \in (\mathbb{R}^d)^n$, $(C_0, \{C_0^i\}_{i=1}^n) \in \mathsf{C}^{n+1}$, $(Y_0, \{Y_0^i\}_{i=1}^n) \in \mathsf{Y}^{n+1}$ and set $X_0 = n^{-1}\sum_{i=1}^n X_0^i$. For each $k \in \mathbb{N}$ the random variables are updated according to

$$G_{k+1}^i = \mathcal{G}^i\left(X_k^i, Y_k^i, Y_k, C_k^i, C_k, \xi_{k+1}\right), \tag{3.17}$$

$$\tilde{X}_{k+1}^i = X_k^i - \gamma G_{k+1}^i + \sqrt{2\gamma}\left(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,Z_{k+1}^i\right),$$

$$Y_{k+1}^i = \mathcal{Y}^i\left(X_k^i, Y_k^i, Y_k, \xi_{k+1}\right), \tag{3.18}$$

$$C_{k+1}^i = \mathcal{C}^i\left(X_k^i, Y_k^i, C_k^i, C_k, \xi_{k+1}\right), \tag{3.19}$$

$$X_{k+1}^i = B_{k+1}\sum_{j=1}^n \tilde{X}_{k+1}^j + (1-B_{k+1})\tilde{X}_{k+1}^i, \tag{3.20}$$

$$(Y_{k+1}, C_{k+1}) = \mathcal{D}(Y_k, C_k, \{C_k^i\}_{i=1}^n, \{X_k^i\}_{i=1}^n, \xi_{k+1}), \tag{3.21}$$

where $\tau \in [0,1]$; $\gamma \in (0, \bar{\gamma}]$ is the step-size; $\{(B_k, \xi_k, \tilde{Z}_k, Z_k^1, \ldots, Z_k^n) : k \in \mathbb{N}^\star\}$ is a set of independent sequences of i.i.d. random variables such that for any $k \in \mathbb{N}^*$ B_k, is a Bernoulli random variable with parameter $p_c \in (0,1]$; and $(\tilde{Z}_k, Z_k^1, \ldots, Z_k^n)$ are d-dimensional standard Gaussian random variables. Recall that $(\xi_k)_{k\geq 1}$ is a set of i.i.d. random variables distributed according to ν_ξ such that Assumption 3.4 holds to ensure that the combination of functions $\{\mathcal{G}^i\}_{i\in[n]}$ provides an unbiased estimate of ∇U.

In iteration $k \geq 0$, the local parameter of the ith client is denoted by X_k^i, and G_k^i stands for its local gradient. If $B_k = 1$ (communication round), the local parameter X_k^i is set to the value of the global server parameter X_k. If $B_k = 0$, X_k^i is set to the local update $\tilde{X}_k^i$. Moreover, we write Y_k^i the reference point used to compute the control variate C_k^i. The first step (3.17) corresponds to the computation of a stochastic estimate of ∇U^i

by the ith client. Then, the client updates the reference point Y_k^i (3.18) at which the local control variate is computed. The client also update its own local control variate C_k^i in (3.19). If $B_{k+1} = 1$, then the server averages the parameter of each client, and broadcasts this average. If $B_{k+1} = 0$, then each client keeps $\tilde{X}_{k+1}^i$ as its new local parameter. Finally, the server updates the reference point Y_k and the global control variate C_k according to (3.21). Denote the filtration $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ defined for any $k \geq 0$, by

Algorithm 3.2 Stochastic Averaging Langevin Dynamics - FALD and its variants

Input: initial vectors $(X_0^i)_{i\in[n]}$, noise parameter $\tau \in [0,1]$, number of communication rounds K, probability $p_c \in (0,1]$ of communication, probability $q_c \in [0,1]$ to update the control variates, and step-size γ

Initialize: $Y_0 = (1/n)\sum_{i=1}^n X_0^i$ and $C_0 = (1/n)\nabla U(Y_0)$

for $k = 0$ **to** $K-1$ **do**

  Draw $B_{k+1} \sim \mathcal{B}(p_c), \tilde{Z}_{k+1} \sim \mathcal{N}(0_d, \mathrm{I}_d)$ // On every client

  for $i = 1$ **to** n **do** // In parallel on the n clients

    Draw $\xi_{k+1}^i \sim \nu_\xi^i$, $\tilde{Z}_{k+1}^i \sim \mathcal{N}(0_d, \mathrm{I}_d)$

    Compute G_k^i following (3.17)

    Set $\tilde{X}_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\,(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i)$

    if $B_{k+1} = 1$ **then**

      Broadcast $\tilde{X}_{k+1}^i$ to the server // Communication round

    else

      Update $X_{k+1}^i \leftarrow \tilde{X}_{k+1}^i$ // Local step

    if $\tilde{B}_{k+1} = 1$ **then** // Control variate update round

      Broadcast the necessary information to the server in order to update (Y_k^i, C_k^i, Y_k, C_k)

    else

      Set $(Y_{k+1}^i, C_{k+1}^i, Y_{k+1}, C_{k+1}) \leftarrow (Y_k^i, C_k^i, Y_k, C_k)$ // No update

  if $B_{k+1} = 1$ **then** // During communication round

    Update then broadcast $X_{k+1} \leftarrow (1/n)\sum_{i=1}^n \tilde{X}_{k+1}^i$ // On the central server

    Update the local parameter $X_{k+1}^i \leftarrow X_{k+1}$ // On every client

  if $\tilde{B}_{k+1} = 1$ **then** // During control variate update round

    If needed, update then broadcast $Y_{k+1} \leftarrow (1/n)\sum_{i=1}^n X_k^i$ // On the central server

    Update (Y_k^i, C_k^i) using the parameters $(X_k^i, Y_k^i, Y_k, Y_{k+1}, C_k)$ // On every client

    Update then broadcast $C_{k+1} \leftarrow (1/n)\sum_{i=1}^n C_{k+1}^i$ // On the central server

Output: samples $\{X_\ell\}_{\{\ell\in[K]\,:\,B_\ell=1\}}$.

$$\mathcal{F}_k = \sigma\left(\mathsf{X}_0, \left(B_l, C_l, Y_l, \tilde{Z}_l, \xi_l, \left(C_l^i, G_l^i, X_l^i, \tilde{X}_l^i, Y_l^i, Z_l^i\right)_{i=1,\dots,n}\right)_{0\leq l\leq k}\right) \tag{3.22}$$

and consider the conditional expectation and variance denoted by $\mathbb{E}^{\mathcal{F}_k}$, $\mathrm{Var}^{\mathcal{F}_k}(\cdot) = \mathbb{E}^{\mathcal{F}_k}[\|\cdot - \mathbb{E}^{\mathcal{F}_k}[\cdot]\|^2]$ respectively. For $k \in \mathbb{N}$, we introduce X_k the average of the local parameters given by

$$X_k = \frac{1}{n}\sum_{i=1}^{n} X_k^i, \tag{3.23}$$

and we set

$$V_k = \frac{1}{n}\sum_{i=1}^{n} \|X_k^i - X_k\|^2. \tag{3.24}$$

Finally, to control the distance between the average parameter X_k and the minimizer $x_\star = \arg\min U$, we consider the parameter d_k, which for $k \geq 0$ is given by

$$d_k = \|X_k - x_\star\|. \tag{3.25}$$

For each $k \in \mathbb{N}$ and $\gamma \in \left(0, \bar{\gamma}\right]$, we denote by $\mu_k^{(\gamma)}$ the distribution of X_k defined by (3.23). To ensure the quality of the samples generated by Algorithm 3.2, we control the Wasserstein distance $W_2(\pi(\cdot|\mathcal{D}), \mu_k^{(\gamma)})$. Recall that the Wasserstein distance is the infimum of $\mathbb{E}[\|\mathsf{X}_{k\gamma} - X_k\|^2]$ over all couplings $(\mathsf{X}_{k\gamma}, X_k)$ such that $\mathsf{X}_{k\gamma}$ is distributed according to $\pi(\cdot|\mathcal{D})$. Thus, to study the convergence of $(\mu_k^{(\gamma)})_{k\in\mathbb{N}}$, we introduce a synchronous coupling $(\mathsf{X}_{k\gamma}, X_k)_{k\geq 0}$ with values in $(\mathbb{R}^d)^2$ between $\pi(\cdot|\mathcal{D})$ and $\mu_k^{(\gamma)}$, starting from the couple (X_0, X_0) distributed according to $\zeta \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, *i.e.*, $\zeta(\mathbb{R}^d, \cdot) = \mu_0^{(\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$ and $\zeta(\cdot, \mathbb{R}^d) = \pi(\cdot|\mathcal{D})$. Since $\log \pi(\cdot|\mathcal{D})$ is supposed m-strongly concave by **A**1, note that $\pi(\cdot|\mathcal{D})$ belongs in $\mathcal{P}_2(\mathbb{R}^d)$. Based on independent d-dimensional standard Brownian motions $(\{\tilde{\mathsf{W}}_t, \{\tilde{\mathsf{W}}_t^i\}_{i=1}^n\})_{t\geq 0}$, we define $\mathsf{W}_t = \sqrt{\tau}\tilde{\mathsf{W}}_t + \sqrt{(1-\tau)/n}\sum_{i=1}^n \tilde{\mathsf{W}}_t^i$. For $k \in \mathbb{N}^\star$, we introduce $\tilde{Z}_k = \gamma^{-1/2}(\tilde{\mathsf{W}}_{k\gamma} - \tilde{\mathsf{W}}_{(k-1)\gamma})$, and for $i \in [n]$, we consider $\tilde{Z}_k^i = \gamma^{-1/2}(\tilde{\mathsf{W}}_{k\gamma}^i - \tilde{\mathsf{W}}_{(k-1)\gamma}^i)$. Therefore, for all $k \in \mathbb{N}^\star$ we can verify that $\mathsf{W}_{k\gamma} - \mathsf{W}_{(k-1)\gamma} = \sqrt{\gamma\tau}\tilde{Z}_k + \sqrt{\gamma(1-\tau)/n}\sum_{i=1}^n \tilde{Z}_k^i$. Moreover, consider $(\mathsf{X}_t)_{t\geq 0}$ the strong solution of the Langevin stochastic differential equation (SDE) given by

$$\mathrm{d}\mathsf{X}_t = -\frac{1}{n}\nabla U(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{\frac{2}{n}}\,\mathrm{d}\mathsf{W}_t. \tag{3.26}$$

The Langevin diffusion defines a Markov semigroup $(\tilde{P}_t)_{t\geq 0}$ satisfying $\pi(\cdot|\mathcal{D})\tilde{P}_t = \pi(\cdot|\mathcal{D})$ for any $t \geq 0$, see for example Roberts and Tweedie (1996, Theorem 2.1). Note that X_t and X_k are distributed according to $\pi(\cdot|\mathcal{D})$ and $\mu_k^{(\gamma)}$, respectively. From the definition of the Wasserstein distance of order 2 it follows that

$$W_2(\pi(\cdot|\mathcal{D}), \mu_k^{(\gamma)}) \leq \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right]^{1/2}.$$

So the proof consists mainly of upper bounding the squared norm $\|\mathsf{X}_{k\gamma} - X_k\|$, from which we derive an explicit bound on the Wasserstein distance by the previous inequality.

First upper bound on $\mathbb{E}^{\mathcal{F}_k}[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2]$. Under mild assumptions, we derive a first bound in Proposition 3.5 to control $\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2$ based on $\|\mathsf{X}_{k\gamma} - X_k\|^2$, $(1/n)\sum_{i=1}^n G_k^i$ and V_k. This decomposition highlights the different approximations brought by the discretization of the Langevin diffusion (3.26) between the averaged

parameter $(X_k)_{k\in\mathbb{N}}$ defined in (3.23) and $\{\mathsf{X}_{k\gamma}\}_{k\in\mathbb{N}}$. Recall that $x_\star = \arg\min U$ and for all $k \in \mathbb{N}$, consider I_k the approximation error defined by

$$I_k = \int_{k\gamma}^{(k+1)\gamma} \left(\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right) \mathrm{d}s. \tag{3.27}$$

For $\bar{\gamma} > 0$ small enough and $k \in \mathbb{N}$, for all $\gamma \in \left(0, \bar{\gamma}\right]$ and under the following assumption Assumption 3.4 we control the distance between the target distribution $\pi(\cdot|\mathcal{D})$ and $\mu_k^{(\gamma)}$.

Assumption 3.4. *For any $\{(x^i, y^i, c^i)\}_{i=1}^n \in \mathbb{R}^{3d}$, we have*

$$\sum_{i=1}^n \int_{\mathsf{E}} \mathcal{G}^i\left(\left\{(x^j, y^j, c^j)\right\}_{j=1}^n, \xi^i\right) \mathrm{d}\nu_\xi(\xi^i) = \sum_{i=1}^n \nabla U^i(x^i).$$

Proposition 3.5. *Assume **A**1, Assumption 3.4 hold and let $\gamma \leq 2(3L)^{-1}$. Then, for any $k \in \mathbb{N}$, we have*

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq \left[1 - \gamma m\left(1 - 3\gamma L\right)\right]\|\mathsf{X}_{k\gamma} - X_k\|^2 + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2\right)V_k \\ + \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right]\right) + \gamma^2\,\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right),\end{aligned}$$

where $V_k, \mathcal{F}_k, d_k$ are defined in (3.24), (3.22) and (3.25).

Proof Let k be in $\mathbb{N}$ and γ in $\left(0, 2(3L)^{-1}\right]$. Recall the stochastic processes $X_{k+1}, \mathsf{X}_{(k+1)\gamma}$ are defined in (3.23) and (3.26) by

$$\begin{cases}\mathsf{X}_{(k+1)\gamma} = \mathsf{X}_{k\gamma} - \gamma\nabla\bar{U}(\mathsf{X}_{k\gamma}) - I_k + \sqrt{2/n}\left(\mathsf{W}_{(k+1)\gamma} - \mathsf{W}_{k\gamma}\right), \\ X_{k+1} = \frac{1}{n}\sum_{i=1}^n\left[X_k^i - \gamma G_k^i + \sqrt{2\gamma}\left(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i\right)\right],\end{cases}$$

with I_k defined in (3.27). Substracting the two above equations gives

$$\mathsf{X}_{(k+1)\gamma} - X_{k+1} = (\mathsf{X}_{k\gamma} - X_k) - \left(\int_{k\gamma}^{(k+1)\gamma} \nabla\bar{U}(\mathsf{X}_s)\mathrm{d}s - \frac{\gamma}{n}\sum_{i=1}^n G_k^i\right).$$

Taking the conditional expectation of the above equation and developing the squared norm, we obtain

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] = \mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] - 2\gamma\left\langle\mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle \\ - 2\left\langle\mathsf{X}_{k\gamma} - X_k, \mathbb{E}^{\mathcal{F}_k}\left[I_k\right] + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{n}\sum_{i=1}^n\mathbb{E}^{\mathcal{F}_k}\left[G_k^i\right]\right\rangle \\ + \mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k + \gamma\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \frac{\gamma}{n}\sum_{i=1}^n G_k^i\right\|^2\right].\end{aligned} \tag{3.28}$$

Using that for all $\alpha > 0, (a,b) \in (\mathbb{R}^d)^2$, $2\langle a,b\rangle \leq \alpha\|a\|^2 + (1/\alpha)\|b\|^2$ combined with Assumption 3.4, for any $\epsilon > 0$ we have

$$-2\left\langle \mathsf{X}_{k\gamma} - X_k, \mathbb{E}^{\mathcal{F}_k}\left[I_k\right] + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{n}\sum_{i=1}^n \mathbb{E}^{\mathcal{F}_k}\left[G_k^i\right]\right\rangle \leq \epsilon\|\mathsf{X}_{k\gamma} - X_k\|^2 + \frac{2}{\epsilon}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + \frac{2\gamma^2}{\epsilon}\left\|\nabla\bar{U}(X_k) - \frac{1}{n}\sum_{i=1}^n \nabla U^i(X_k^i)\right\|^2 . \quad (3.29)$$

In addition, the unbiased property Assumption 3.4 implies that

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k + \gamma\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \frac{\gamma}{n}\sum_{i=1}^n G_k^i\right\|^2\right] = \gamma^2\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right) + \mathbb{E}^{\mathcal{F}_k}\left[\left\|\gamma\left(\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right) + I_k + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{n}\sum_{i=1}^n \nabla U^i(X_k^i)\right\|^2\right] . \quad (3.30)$$

The Young inequality shows that

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|\gamma\left(\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right) + I_k + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{n}\sum_{i=1}^n \nabla U^i(X_k^i)\right\|^2\right] \quad (3.31)$$

$$\leq 3\gamma^2\left\|\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right] + 3\gamma^2\left\|\nabla\bar{U}(X_k) - \frac{1}{n}\sum_{i=1}^n \nabla U^i(X_k^i)\right\|^2 .$$

By **A**1 we know that $\bar{U}$ is L-smooth and convex which imply the co-coercivity of $\bar{U}$ (Nesterov, 2003, Theorem 2.1.5), that is for all $x, y \in \mathbb{R}^d$, $\left\|\nabla\bar{U}(y) - \nabla\bar{U}(x)\right\|^2 \leq L\left\langle \nabla\bar{U}(y) - \nabla\bar{U}(x), y - x\right\rangle$. Hence, we deduce that

$$\left\|\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\|^2 \leq L\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle . \quad (3.32)$$

Setting $\epsilon = \gamma m$, we have $0 < \epsilon \leq 1$ and $1 + 1/\epsilon \leq 2(\gamma m)^{-1}$. Therefore, (3.29), (3.30) and (3.32) associated with (3.28) show that

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq (1+\gamma m)\|\mathsf{X}_{k\gamma} - X_k\|^2 + \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right]\right) - \gamma\left(2 - 3\gamma L\right)\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle + \gamma^2\left(3 + \frac{2}{\gamma m}\right)\left\|\nabla\bar{U}(X_k) - \frac{1}{n}\sum_{i=1}^n \nabla U^i(X_k^i)\right\|^2 + \gamma^2\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right) . \quad (3.33)$$

For any $i \in [n]$, by **A**1, the m-convexity of $\bar{U}$ gives that

$$\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla \bar{U}(\mathsf{X}_{k\gamma}) - \nabla \bar{U}(X_k) \right\rangle \geq m \|\mathsf{X}_{k\gamma} - X_k\|^2 \tag{3.34}$$

In addition, under **A**1 the Jensen inequality implies

$$\left\| \nabla \bar{U}(X_k) - \frac{1}{n} \sum_{i=1}^{n} \nabla U^i(X_k^i) \right\|^2 \leq L^2 V_k, \tag{3.35}$$

where V_k is defined in (3.24). Therefore, using the assumption on γ and plugging (3.34) and (3.35) in (3.33) yields the expected inequality. ■

3.A.1 General supporting lemmas

In this subsection, we consider the stochastic processes $(X_k)_{k\in\mathbb{N}}$, $(\mathsf{X}_{k\gamma})_{k\in\mathbb{N}}$ defined in (3.23) and (3.26). We derive several lemmas which allow us to derive a recursion on $\mathbb{E}[\|\mathsf{X}_{k\gamma} - X_k\|^2]$.

Lemma 3.6. *Assume **A**1 holds. Then, for any $k \in \mathbb{N}$ and $\gamma > 0$ we have*

$$\mathbb{E}\left[\left\| I_k \right\|^2 \right] \leq \frac{d\gamma^3 L^2}{n} \left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12} \right).$$

Proof Let k be in $\mathbb{N}$. Using the Jensen inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left\| I_k \right\|^2 \right] &= \mathbb{E}\left[\left\| \int_{k\gamma}^{(k+1)\gamma} \left(\nabla \bar{U}(\mathsf{X}_s) - \nabla \bar{U}(\mathsf{X}_{k\gamma}) \right) \mathrm{d}s \right\|^2 \right] \\
&\leq \gamma \int_{k\gamma}^{(k+1)\gamma} \mathbb{E}\left[\left\| \nabla \bar{U}(\mathsf{X}_s) - \nabla \bar{U}(\mathsf{X}_{k\gamma}) \right\|^2 \right] \mathrm{d}s \\
&\leq L^2 \gamma \int_{k\gamma}^{(k+1)\gamma} \mathbb{E}\left[\|\mathsf{X}_s - \mathsf{X}_{k\gamma}\|^2 \right] \mathrm{d}s.
\end{aligned} \tag{3.36}$$

Further, for any $s \in \mathbb{R}_+$, using Durmus and Moulines (2019, Lemma 21) applied to $(\mathsf{X}_{nt})_{t\in\mathbb{R}_+}$ we obtain

$$\mathbb{E}^{\mathcal{F}_{k\gamma}}\left[\|\mathsf{X}_s - \mathsf{X}_{k\gamma}\|^2 \right] \leq \frac{d(s-k\gamma)}{n} \left(2 + (s-k\gamma)^2 \frac{L^2}{3} \right) + \frac{3}{2}(s-k\gamma)^2 L^2 \|\mathsf{X}_{k\gamma} - x_\star\|^2.$$

Integrating the previous inequality on $[k\gamma, (k+1)\gamma]$, it implies

$$\int_{k\gamma}^{(k+1)\gamma} \mathbb{E}\left[\|\mathsf{X}_s - \mathsf{X}_{k\gamma}\|^2 \right] \mathrm{d}s \leq \frac{\gamma^2}{n} \left(d + \frac{nL^2\gamma}{2} \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - x_\star\|^2 \right] + \frac{dL^2\gamma^2}{12} \right). \tag{3.37}$$

Plugging (3.37) in (3.36) gives

$$\mathbb{E}\left[\left\|I_k\right\|^2\right] \leq \frac{L^2\gamma^3}{n}\left(d + \frac{nL^2\gamma}{2}\mathbb{E}\left[\left\|\mathsf{X}_{k\gamma} - x_\star\right\|^2\right] + \frac{dL^2\gamma^2}{12}\right). \tag{3.38}$$

Applying Durmus and Moulines (2019, Proposition 1) to $(\mathsf{X}_{nt})_{t\in\mathbb{R}_+}$, we get

$$\mathbb{E}\left[\left\|\mathsf{X}_{k\gamma} - x_\star\right\|^2\right] \leq \frac{d}{nclientsm}. \tag{3.39}$$

Thus, combining (3.38) with (3.39) completes the proof. ■

Lemma 3.7. *Assume **A**1 and **HX**1 hold. Then, for any $k \in \mathbb{N}$ and $\gamma > 0$ we have*

$$\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] \leq \frac{2\gamma^4 d}{3n}\left(L^3 + \frac{d\tilde{L}^2}{n}\right),$$

where I_k is defined in (3.27).

Proof Denote Δ the Laplacian defined, for all $x \in \mathbb{R}^d$, by $\Delta U(x) = \{\sum_{l=1}^d (\partial^2 U_j)(x)/\partial x_l^2\}_{j=1}^d$, moreover let $k \in \mathbb{N}$ be a fixed integer and $\gamma > 0$. Using the Itô formula, we have for $s \in \left[k\gamma, (k+1)\gamma\right]$

$$\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma}) = \int_{k\gamma}^s \frac{1}{n}\Delta(\nabla\bar{U})(\mathsf{X}_u) - \nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\mathrm{d}u + \sqrt{\frac{2}{n}}\int_{k\gamma}^s \nabla^2\bar{U}(\mathsf{X}_u)\mathrm{d}B_u. \tag{3.40}$$

We will upper bound separately the three terms of the previous equality. First, the L-Lipschitz property of $\nabla\bar{U}$ given by **A**1 implies for any $u \in \mathbb{R}_+$ that

$$\left\|\nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\right\| \leq L\left\|\nabla\bar{U}(\mathsf{X}_u) - \nabla\bar{U}(x_\star)\right\|. \tag{3.41}$$

In addition, since for $u \in \mathbb{R}_+$, the random variable X_u is distributed according to the stationary distribution $\pi(\cdot|\mathcal{D}) \propto \exp(-U)$, we know from Dalalyan (2017a, Lemma 2) that

$$\mathbb{E}\left[\left\|\nabla\bar{U}(\mathsf{X}_u) - \nabla\bar{U}(x_\star)\right\|^2\right] \leq \frac{dL}{n}. \tag{3.42}$$

Therefore, we deduce from (3.41) and (3.42) the following bound

$$\mathbb{E}\left[\left\|\nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\right\|^2\right] \leq \frac{dL^3}{n}. \tag{3.43}$$

Denote $(e_i)_{i=1}^d$ the canonical basis of $\mathbb{R}^d$; using that U is three times continuously differentiable we can apply the Schwarz's theorem which combined with **HX**1, immediately yield that

$$\left\|\Delta(\nabla\bar{U})(x)\right\|^2 = \sum_{i=1}^d \left|\sum_{j=1}^d \partial_j^2\partial_i\bar{U}(x)\right|^2 \leq d\sum_{i=1}^d\sum_{j=1}^d \left|\partial_i\partial_j^2\bar{U}(x)\right|^2$$

$$= d\sum_{i=1}^{d}\lim_{\epsilon\to 0}\left\{\epsilon^{-2}\sum_{j=1}^{d}\left|\partial_j^2\bar{U}(x+\epsilon\cdot e_i)-\partial_j^2\bar{U}(x)\right|^2\right\}$$
$$\leq d\sum_{i=1}^{d}\lim_{\epsilon\to 0}\left\{\epsilon^{-2}\left(\tilde{L}\|(x+\epsilon\cdot e_i)-x\|^{-1}\right)^2\right\}\leq\left(d\tilde{L}\right)^2. \tag{3.44}$$

Lastly, we upper bound the third term derived in (3.40). Since the potentials $\{U^i\}_{i\in[n]}$ are supposed L-smooth and $\bar{U}$ twice continuously differentiable, for $s\in\left[k\gamma,(k+1)\gamma\right]$ we know that $\int_{k\gamma}^{s}\nabla^2\bar{U}(\mathsf{X}_u)\mathrm{d}B_u$ is a $\mathcal{F}_s$-martingale. Thus, for $k\geq 0$ we deduce that

$$\mathbb{E}^{\mathcal{F}_k}\left[\int_{k\gamma}^{(k+1)\gamma}\nabla^2\bar{U}(\mathsf{X}_u)\,\mathrm{d}u\right]=0. \tag{3.45}$$

Eventually, combining (3.40), (3.43), (3.44) and (3.45) with the Jensen and Young inequalities give

$$\frac{1}{\gamma}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right]=\frac{1}{\gamma}\mathbb{E}\left[\left\|\int_{k\gamma}^{(k+1)\gamma}\mathbb{E}^{\mathcal{F}_k}\left[\nabla\bar{U}(\mathsf{X}_s)-\nabla\bar{U}(\mathsf{X}_{k\gamma})\right]\mathrm{d}s\right\|^2\right]$$
$$\leq\int_{k\gamma}^{(k+1)\gamma}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[\nabla\bar{U}(\mathsf{X}_s)-\nabla\bar{U}(\mathsf{X}_{k\gamma})\right]\right\|^2\right]\mathrm{d}s$$
$$=\int_{k\gamma}^{(k+1)\gamma}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[\int_{k\gamma}^{s}\frac{1}{n}\Delta(\nabla\bar{U})(\mathsf{X}_u)-\nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\mathrm{d}u\right]\right\|^2\right]\mathrm{d}s$$
$$\leq 2\int_{k\gamma}^{(k+1)\gamma}(s-k\gamma)\int_{k\gamma}^{s}\mathbb{E}\left[\frac{1}{n^2}\left\|\int_{k\gamma}^{s}\Delta(\nabla\bar{U})(\mathsf{X}_u)\mathrm{d}u\right\|^2+\left\|\nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\mathrm{d}u\right\|^2\right]\mathrm{d}s$$
$$\leq 2\int_{k\gamma}^{(k+1)\gamma}(s-k\gamma)^2\left(\frac{dL^3}{n}+\frac{(d\tilde{L})^2}{n^2}\right)\mathrm{d}s=\frac{2\gamma^3 d}{3n}\left(L^3+\frac{d\tilde{L}^2}{n}\right).$$

Multiplying this last inequality by $\gamma>0$ proves the expected result. ■

Lemma 3.8. *Assume **A**1 hold. Then, for any $k\in\mathbb{N}$ and $\gamma\in\left(0,(3m)^{-1}\right]$ we have*

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right]+3\mathbb{E}\left[\left\|I_k\right\|^2\right]\leq\begin{cases}\frac{3\gamma^2 dL^2}{nm}\left(1+\frac{19\gamma L^2}{36m}\right)\\ \frac{\gamma^3 d}{nm}\left(5L^3+\frac{4d\tilde{L}^2}{3n}\right)\end{cases}\quad\text{if } \boldsymbol{HX}1 \text{ holds and } \gamma\leq L^{-1}.$$

Proof Let k be in $\mathbb{N}$ and $\gamma\in\left(0,(3m)^{-1}\right]$, using Lemma 3.6 we have

$$\mathbb{E}\left[\left\|I_k\right\|^2\right]\leq\frac{\gamma^3 dL^2}{n}\left(1+\frac{\gamma L^2}{2m}+\frac{\gamma^2 L^2}{12}\right).$$

Therefore, we deduce

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right] \leq \frac{3\gamma^2 dL^2}{nm}\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right).$$

Moreover, if we additionally suppose the regularity of the Hessian of the potentials $(U^i)_{i=1}^n$ as stated in **HX**1, we sharpen the upper bound on $\mathbb{E}[\|\mathbb{E}^{\mathcal{F}_k}[I_k]\|^2]$. Indeed, we show in Lemma 3.7 that

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] \leq \frac{4\gamma^3 d}{3nm}\left(L^3 + \frac{d\tilde{L}^2}{n}\right).$$

Hence, we deduce that

$$\begin{aligned}\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right] &\leq \frac{3\gamma^3 dL^2}{n}\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right) + \frac{4\gamma^3 d}{3nm}\left(L^3 + \frac{d\tilde{L}^2}{n}\right) \\ &\leq \frac{\gamma^3 dL^3}{nm}\left(3 + \frac{4}{3} + \frac{19\gamma L}{36}\right) + \frac{4\gamma^3 d^2\tilde{L}^2}{3n^2 m}.\end{aligned}$$

■

3.A.2 Derivation of the central theorem

Assumption 3.9. *There exist* $\alpha_v \in \big(0,1\big)$ *and* $(v_1, v_2) \in (\mathbb{R}_+)^2$ *such that for any* $k \in \mathbb{N}$, V_k *satisfies*

$$\mathbb{E}\left[V_k\right] \leq v_1\alpha_v^k + v_2,$$

where V_k *is defined in* (3.24).

HX 2. *There exist* $q_c \in (0,1)$ *and* α_0, α_1, α_2, α_3, $\alpha_4 \in \mathbb{R}_+$ *satisfying*

$$(1-q_c)(1+\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1}) < 2$$

such that for $k \geq 0$ *the following inequality holds*

$$\begin{aligned}(1-q_c)^{-1}\mathbb{E}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq \alpha_0\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] + \alpha_1\sum_{l=0}^{k-1}(1-q_c)^{k-l}\mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] \\ + \alpha_2\mathbb{E}\left[V_k\right] + \alpha_3\sum_{l=0}^{k-1}(1-q_c)^{k-l}\mathbb{E}\left[V_l\right] + \alpha_4.\end{aligned}$$

With the notation introduced in **HX**2, consider

$$\delta = \frac{-1-\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1}}{2}. \tag{3.46}$$

At iteration $k \geq 0$, recall that $\mu_k^{(\gamma)}$ denotes the distribution of the average parameter X_k (3.23). The next result controls the Wasserstein distance between $\mu_k^{(\gamma)}$ and the posterior distribution π.

Theorem 3.10. *Assume **HX**2 and Assumption 3.9 hold. Then, for any probability measure* $\mu_0^{(\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, *we have*

$$W_2^2\left(\mu_k^{(\gamma)}, \pi\right) \leq \left(1+\alpha_0+\delta\right)^k \left(1-q_c\right)^k W_2^2\left(\mu_0^{(\gamma)}, \pi\right) + (1-q_c) v_1 \left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right) \frac{\alpha_v^k - \left(1+\alpha_0+\delta\right)^k \left(1-q_c\right)^k}{\alpha_v - \left(1+\alpha_0+\delta\right)\left(1-q_c\right)} + \frac{1-q_c}{q_c-(1-q_c)(\alpha_0+\delta)} \left[\left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right) v_2 + \alpha_4\right].$$

Proof For any $k \in \mathbb{N}$, define

$$u_k = \left(1-q_c\right)^{-k} \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right], \qquad S_k = \sum_{l=0}^{k} u_l,$$
$$v_k = \left(1-q_c\right)^{-k} \left(\alpha_2 \mathbb{E}\left[V_k\right] + \alpha_4\right) + \alpha_3 \sum_{l=0}^{k-1} (1-q_c)^{-l} \mathbb{E}\left[V_l\right]. \tag{3.47}$$

With the above notations, **HX**2 becomes

$$u_{k+1} \leq \alpha_0 u_k + \alpha_1 \sum_{l=0}^{k-1} u_l + v_k,$$

which can be rewritten as

$$S_{k+1} - S_k \leq \alpha_0 \left(S_k - S_{k-1}\right) + \alpha_1 S_{k-1} + v_k. \tag{3.48}$$

Since δ is solution of $\delta(1+\alpha_0+\delta)+\alpha_0-\alpha_1=0$, adding $(1+\delta)S_k$ in (3.48) gives that

$$\begin{aligned} S_{k+1} + \delta S_k &\leq \left(1+\alpha_0+\delta\right)\left(S_k - \frac{\alpha_0-\alpha_1}{1+\alpha_0+\delta} S_{k-1}\right) + v_k \\ &= \left(1+\alpha_0+\delta\right)\left(S_k + \delta S_{k-1}\right) + v_k. \end{aligned}$$

Using the fact that $\alpha_0 \leq 1 + \sqrt{(\alpha_0-1)^2+4\alpha_1}$, we obtain $2(1+\delta) = 1-\alpha_0 + \sqrt{(\alpha_0-1)^2+4\alpha_1} \geq 0$. Hence $1+\delta > 0$, which leads to the following upper bound

$$u_{k+1} \leq u_{k+1} + (1+\delta) \sum_{l=0}^{k} u_l = S_{k+1} + \delta S_k.$$

Thus, we obtain that

$$u_k \leq S_k + \delta S_{k-1} \leq \left(1+\alpha_0+\delta\right)^{k-1} \left(u_1 + (1+\delta) u_k\right) + \sum_{l=1}^{k-1} \left(1+\alpha_0+\delta\right)^{k-l-1} v_l.$$

Plugging the definition (3.47) of u_k and v_l inside the previous inequality, we get

$$\left(1-q_c\right)^{-k} \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right]$$

$$\leq \left(1+\alpha_0+\delta\right)^{k-1}\left(\left(1-q_{\mathrm{c}}\right)^{-1}\mathbb{E}\left[\|\mathsf{X}_\gamma - X_1\|^2\right] + (1+\delta)\mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right]\right)$$
$$+\sum_{l=1}^{k-1}\left(1+\alpha_0+\delta\right)^{k-l-1}\left[\left(1-q_{\mathrm{c}}\right)^{-l}\left(\alpha_2\mathbb{E}\left[V_l\right]+\alpha_4\right) + \alpha_3\sum_{j=0}^{l-1}(1-q_{\mathrm{c}})^{-j}\mathbb{E}\left[V_j\right]\right]. \tag{3.49}$$

Moreover, using **HX**2 we obtain that

$$\mathbb{E}\left[\|\mathsf{X}_\gamma - X_1\|^2\right] \leq (1-q_{\mathrm{c}})\alpha_0\mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right] + (1-q_{\mathrm{c}})\alpha_2\mathbb{E}\left[V_0\right] + \alpha_4, \tag{3.50}$$

combining (3.49) with (3.50) yield

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] \leq{}& \left(1+\alpha_0+\delta\right)^k\left(1-q_{\mathrm{c}}\right)^k\mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right]\\ &+\alpha_2\sum_{l=0}^{k-1}\left(1+\alpha_0+\delta\right)^{k-l-1}\left(1-q_{\mathrm{c}}\right)^{k-l}\mathbb{E}\left[V_l\right]\\ &+\alpha_3\sum_{j=0}^{k-2}(1-q_{\mathrm{c}})^{k-j}\mathbb{E}\left[V_j\right]\sum_{l=j+1}^{k-1}\left(1+\alpha_0+\delta\right)^{k-l-1}\\ &+(1-q_{\mathrm{c}})\alpha_4\sum_{l=0}^{k-1}\left(1+\alpha_0+\delta\right)^l\left(1-q_{\mathrm{c}}\right)^l. \end{aligned}\tag{3.51}$$

Consider the function $f : a \in \mathbb{R} \to \mathbb{R}$ defined by $f(a) = a(1+\alpha_0+a)+\alpha_0-\alpha_1$. Using the definition (3.46) of δ combined with the increasing property of f, we deduce from $f(\delta) = 0 > f(-\alpha_0) = -\alpha_1$ that $\delta > -\alpha_0$, and thus we get $1+\alpha_0+\delta > 1$ which implies that

$$\sum_{l=j+1}^{k-1}\left(1+\alpha_0+\delta\right)^{k-l-1} \leq \sum_{l=0}^{k-j-2}\left(1+\alpha_0+\delta\right)^{k-j-l-2} \leq \frac{(1+\alpha_0+\delta)^{k-j-1}}{\alpha_0+\delta}. \tag{3.52}$$

Therefore, plugging (3.52) in (3.51) gives

$$\sum_{j=0}^{k-2}(1-q_{\mathrm{c}})^{k-j}\mathbb{E}\left[V_j\right]\sum_{l=j+1}^{k-1}\left(1+\alpha_0+\delta\right)^{k-l-1} \leq \sum_{l=0}^{k-2}\frac{\left(1-q_{\mathrm{c}}\right)^{k-l}(1+\alpha_0+\delta)^{k-l-1}}{\alpha_0+\delta}\mathbb{E}\left[V_l\right]. \tag{3.53}$$

In addition, since **HX**2 ensures that $(1-q_{\mathrm{c}})(1+\alpha_0+\delta) < 1$, we have

$$\sum_{l=0}^{k-1}\left(1+\alpha_0+\delta\right)^l\left(1-q_{\mathrm{c}}\right)^l \leq \frac{1}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}. \tag{3.54}$$

The last inequality combined with (3.51) and (3.53) show that

$$\begin{aligned}&\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] \leq \left(1+\alpha_0+\delta\right)^k\left(1-q_{\mathrm{c}}\right)^k\mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right]\\ &+\left(\alpha_2+\frac{\alpha_3}{\alpha_0+\delta}\right)\sum_{l=0}^{k-1}\left(1+\alpha_0+\delta\right)^{k-l-1}\left(1-q_{\mathrm{c}}\right)^{k-l}\mathbb{E}\left[V_l\right] + \frac{(1-q_{\mathrm{c}})\alpha_4}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}. \end{aligned}\tag{3.55}$$

Further, since we assume Assumption 3.9, we have

$$\sum_{l=0}^{k-1} \left(1+\alpha_0+\delta\right)^{k-l-1} \left(1-q_{\mathrm{c}}\right)^{k-l} \mathbb{E}\left[V_l\right] \leq v_1 \sum_{l=0}^{k-1} \left(1+\alpha_0+\delta\right)^{k-l-1} \left(1-q_{\mathrm{c}}\right)^{k-l} \alpha_v^l + v_2 \sum_{l=0}^{k-1} \left(1+\alpha_0+\delta\right)^{k-l-1} \left(1-q_{\mathrm{c}}\right)^{k-l}. \quad (3.56)$$

A calculation gives that

$$\sum_{l=0}^{k-1} \left(1+\alpha_0+\delta\right)^{k-l-1} \left(1-q_{\mathrm{c}}\right)^{k-l} \alpha_v^l \leq (1-q_{\mathrm{c}}) \frac{\alpha_v^k - \left(1+\alpha_0+\delta\right)^k \left(1-q_{\mathrm{c}}\right)^k}{\alpha_v - \left(1+\alpha_0+\delta\right)\left(1-q_{\mathrm{c}}\right)} \quad (3.57)$$

and combining (3.54), (3.56) with (3.57), we find that

$$\sum_{l=0}^{k-1} \left(1+\alpha_0+\delta\right)^{k-l-1} \left(1-q_{\mathrm{c}}\right)^{k-l} \mathbb{E}\left[V_l\right] \leq (1-q_{\mathrm{c}}) v_1 \frac{\alpha_v^k - \left(1+\alpha_0+\delta\right)^k \left(1-q_{\mathrm{c}}\right)^k}{\alpha_v - \left(1+\alpha_0+\delta\right)\left(1-q_{\mathrm{c}}\right)} + \frac{(1-q_{\mathrm{c}})v_2}{q_{\mathrm{c}} - (1-q_{\mathrm{c}})(\alpha_0+\delta)}. \quad (3.58)$$

Therefore, plugging (3.58) inside (3.55) shows that

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] &\leq \left(1+\alpha_0+\delta\right)^k \left(1-q_{\mathrm{c}}\right)^k \mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right] \\ &\quad + (1-q_{\mathrm{c}}) v_1 \left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right) \frac{\alpha_v^k - \left(1+\alpha_0+\delta\right)^k \left(1-q_{\mathrm{c}}\right)^k}{\alpha_v - \left(1+\alpha_0+\delta\right)\left(1-q_{\mathrm{c}}\right)} \\ &\quad + \frac{1-q_{\mathrm{c}}}{q_{\mathrm{c}} - (1-q_{\mathrm{c}})(\alpha_0+\delta)} \left[\left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right) v_2 + \alpha_4\right]. \end{aligned} \quad (3.59)$$

Eventually, since the Wasserstein distance $W_2(\pi, \mu_k^{(\gamma)})$ is the infimum over all couplings, we obtain that $W_2^2(\pi, \mu_k^{(\gamma)}) \leq \mathbb{E}[\|\mathsf{X}_{k\gamma} - X_k\|^2]$. Moreover, it follows from the strongly convex assumption **A**1 that $\pi \in \mathcal{P}_2(\mathbb{R}^d)$. Thus, we can apply Villani (2008, Theorem 4.1) to prove the existence of an optimal coupling ζ such that taking (X_0, X_0) distributed according to ζ implies that $\mathbb{E}[\|\mathsf{X}_0 - X_0\|^2]^{1/2} = W_2(\pi, \mu_0^{(\gamma)})$. Substituting these results into (3.59) completes the proof. ∎

3.A.3 Upper bound on V_k

The goal of this subsection is to prove the upper bound derived in Lemma 3.13 for $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ to ensure that Assumption 3.9 holds. Recall that for $k \geq 0$, V_k is defined in (3.24), d_k in (3.25), G_k^i in (3.17) and we introduce $\bar{G}_k^i = \mathbb{E}^{\mathcal{F}_k}[G_k^i]$. To prove the central lemma of this subsection, we also consider the assumptions **HX**3 and **HX**4 given below.

HX 3. *There exist* $A_d, A_\sigma \in \left(0,1\right), B_d, B_\sigma, C_d, C_\sigma, D_d, D_\sigma \in \mathbb{R}_+$, *such that for any* $k \in \mathbb{N}$, *we have*

$$\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1-A_d\right) \mathbb{E}\left[d_k^2\right] + B_d \mathbb{E}\left[\sigma_k^2\right] + C_d \mathbb{E}\left[V_k\right] + D_d,$$

$$\mathbb{E}\left[\sigma_{k+1}^2\right] \leq (1 - A_\sigma)\,\mathbb{E}\left[\sigma_k^2\right] + B_\sigma \mathbb{E}\left[d_k^2\right] + C_\sigma \mathbb{E}\left[V_k\right] + D_\sigma.$$

HX4. *There exist $A, \bar{A}, B, \bar{B}, C, \bar{C}, D, \bar{D} \geq 0$ such that for any $i \in [n], k \in \mathbb{N}$, we have*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\bar{G}_k^i\right\|^2\right] \leq \bar{A}\mathbb{E}\left[V_k\right] + \bar{B}\mathbb{E}\left[d_k^2\right] + \bar{C}\mathbb{E}\left[\sigma_k^2\right] + \bar{D},$$
$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] \leq A\mathbb{E}\left[V_k\right] + B\mathbb{E}\left[d_k^2\right] + C\mathbb{E}\left[\sigma_k^2\right] + D.$$

With the notation considered in **HX**3 and **HX**4, for any $\gamma > 0$ we also introduce the following quantities:

$$\begin{aligned}
\mathrm{C}^\gamma &= \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c} - 4A_d}\left[B + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right)\right], \\
\mathrm{C}_r^\gamma &= \frac{9\gamma^2\left(1-p_\mathrm{c}\right)C_\sigma}{p_\mathrm{c} - 4A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right), \\
\mathrm{C}_\sigma^\gamma &= \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c} - 4A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + \mathrm{C}^\gamma B_d\left(2 + \frac{3}{A_\sigma - A_d}\right), \\
\mathrm{C}_{d_0}^\gamma &= 7\mathrm{C}^\gamma, \quad \mathrm{C}_V^\gamma = 1 + 2\mathrm{C}^\gamma C_d, \\
\mathrm{C}_\delta^\gamma &= \frac{4(1-p_\mathrm{c})\gamma^2 D_\sigma}{A_\sigma(p_\mathrm{c} - 4A_d)}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left(D + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{D}\right) \\
&\quad + \frac{\mathrm{C}^\gamma}{A_d}\left(1 + \frac{2B_d B_\sigma}{A_d(A_\sigma - A_d)}\right)\left(D_d + \frac{B_d D_\sigma}{A_\sigma}\right) + \frac{8\left(1-\tau\right)(n-1)\gamma d}{n p_\mathrm{c}}.
\end{aligned} \tag{3.60}$$

If $A_d \leq A_\sigma/2$ and $A_d A_\sigma \geq 8B_d B_\sigma$, we also introduce a convergence rate (proved later in Lemma 3.12) defined by

$$\alpha = A_d - \frac{2(A_\sigma - A_d)^{-1}B_d B_\sigma}{1 + \sqrt{1 + 4(1-A_d)^{-1}(A_\sigma - A_d)^{-1}B_d B_\sigma}}. \tag{3.61}$$

Lemma 3.11. *Assume **HX**3 and also that $A_d \leq A_\sigma/2$, $A_d A_\sigma \geq 8B_d B_\sigma$ hold. Then, we have*

$$A_d/2 < \alpha \leq A_d.$$

Proof First, introduce $\delta_\alpha \in \mathbb{R}_+$ the unique non-negative solution of

$$\delta_\alpha^2 + \delta_\alpha = \frac{B_d B_\sigma}{(1-A_d)(A_\sigma - A_d)}.$$

Since we suppose $A_d \leq A_\sigma/2$, thus we have $A_d \leq 1/2$ which implies that $(1-A_d)(A_d^2/4 + A_d/2) \geq A_d/4$. In addition, using $A_d A_\sigma \geq 8B_d B_\sigma$, we get that

$$\left(1 - A_d\right)\left(\frac{A_d^2}{4} + \frac{A_d}{2}\right) \geq \frac{A_d}{4} \geq \frac{2B_d B_\sigma}{A_\sigma} \geq \left(1 - A_d\right)\left(\delta_\alpha^2 + \delta_\alpha\right).$$

Hence, the increasing property of the function $x \in \mathbb{R}_+ \mapsto x^2 + x$ combined with the fact that $\delta_\alpha \geq 0$ prove that $A_d \geq 2\delta_\alpha$. Moreover, a calculation shows that α satisfies $\alpha = 1 - (1 - A_d)(1 + \delta_\alpha)$. Thus, using $0 \leq 2\delta_\alpha \leq A_d$ implies that $\alpha \in \left(A_d/2, A_d\right]$. ■

The random variable V_k given in (3.24) measures the averaged distance between the global parameter X_k and the local ones $(X_k^i)_{i\in[n]}$. The first lines of the proof of the next lemma are based on Gorbunov et al. (2021, Lemma E.3), however their purpose was to upper bound $\sum_l w_l \mathbb{E}V_l$ for some weights $w_l > 0$, while we prefer to control $\mathbb{E}V_k$ to combine this bound with that of Proposition 3.5. Moreover, the assumptions considered in this work are different, so the proof requires the development of other techniques

Lemma 3.12. *Assume **HX**3, **HX**4 hold with* $A_d < \min(A_\sigma/2, p_c/4), A_d A_\sigma \geq 8 B_d B_\sigma$ *and consider* $\gamma \leq p_c^{1/2}(2 - 2p_c)^{-1/2}[A + (1 + 2/p_c)\bar{A}]^{-1/2}$. *Then, for any* $k \in \mathbb{N}$, *we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \alpha\right)^k \left(\mathrm{C}_V^\gamma \mathbb{E}\left[V_0\right] + \mathrm{C}_{d_0}^\gamma \mathbb{E}\left[d_0^2\right] + \mathrm{C}_\sigma^\gamma \mathbb{E}\left[\sigma_0^2\right] + 2D_d\right) + \mathrm{C}_r^\gamma \sum_{i=0}^{k-2} \left(1 - \alpha\right)^{k-i-1} \mathbb{E}\left[V_i\right] + \mathrm{C}_\delta^\gamma,$$

where V_k *is defined in* (3.24).

Proof Let $k \in \mathbb{N}^\star$, using for $i \in [n]$ the definitions (3.20), (3.23) of X_k^i and X_k

$$X_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\left(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i\right),$$
$$X_{k+1} = X_k - \frac{\gamma}{n}\sum_{j=1}^n G_k^i + \sqrt{\frac{2\gamma\tau}{n}}\tilde{Z}_{k+1} + \frac{\sqrt{2(1-\tau)\gamma}}{n}\sum_{i=1}^n Z_{k+1}^i.$$

First upper bound on $\mathbb{E}\left[V_k\right]$. Substracting the two above equations combined with the Jensen inequality give

$$\mathbb{E}\left[V_{k+1}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|X_{k+1}^i - X_{k+1}\right\|^2\right]$$
$$= \frac{1-p_c}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(G_k^i - G^k) + \sqrt{2(1-\right.\right.$$
$$= \frac{1-p_c}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right]$$
$$+ \frac{(1-p_c)\gamma^2}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|(G_k^i - \bar{G}_k^i) - (G^k - \bar{G}^k)\right\|^2\right] + 2(1-\tau)\gamma\mathbb{E}\left[\left\|Z_{k+1}^i - \frac{1}{n}\sum_{j=1}^n Z_{k+1}^j\right\|^2\right],$$

where the inner product is eliminated using that $\mathbb{E}^{\mathcal{F}_k}[G_k^i - G^k] = \bar{G}_k^i - \bar{G}^k$ (with $\mathcal{F}_k$ defined in (3.22)). Recall that, for any $u_0, u_1 \in \mathbb{R}^d$ and $\epsilon > 0$, it holds $\|u_0 + u_1\|^2 \leq (1+\epsilon)\|u_0\|^2 + (1+\epsilon^{-1})\|u_1\|^2$. In addition, denoting $I \sim \mathrm{Unif}([n])$ and using the Jensen

inequality, we obtain that

$$n^{-1}\sum_{i=1}^{n}\mathbb{E}[\|(G_k^i-\bar{G}_k^i)-(G^k-\bar{G}^k)\|^2]=\mathbb{E}[\|(G_k^I-\bar{G}_k^I)-(G^k-\bar{G}^k)\|^2] \\ \leq \mathbb{E}[\|G_k^I-\bar{G}_k^I\|^2].$$

Setting $u_0=X_k^i-X_k$, $u_1=\gamma(\bar{G}_k^i-\bar{G}^k)$ and $\epsilon=2/p_{\mathrm{c}}$, we get

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\right] &\leq \frac{1-p_{\mathrm{c}}}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(X_k^i-X_k)-\gamma(\bar{G}_k^i-\bar{G}^k)\right\|^2\right] \\ &\quad+\frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|G_k^i-\bar{G}_k^i\right\|^2\right]+2\left(1-\tau\right)(1-1/n)\gamma d \\ &\leq \frac{(1-p_{\mathrm{c}})(1+p_{\mathrm{c}}/2)}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|X_k^i-X_k\right\|^2\right]+\frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|G_k^i-\bar{G}_k^i\right\|^2\right] \\ &\quad+\frac{(1-p_{\mathrm{c}})(1+2/p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\bar{G}_k^i-\bar{G}^k\right\|^2\right]+2\left(1-\tau\right)(1-1/n)\gamma d.\end{aligned}$$

Using $(1-p_{\mathrm{c}})(1+p_{\mathrm{c}}/2)\leq 1-p_{\mathrm{c}}/2$ and $n^{-1}\sum_{i=1}^{n}\mathbb{E}^{\mathcal{F}_k}[\|\bar{G}_k^i-\bar{G}^k\|^2]=\mathbb{E}^{\mathcal{F}_k}[\|\bar{G}_k^I-\bar{G}^k\|^2]\leq \mathbb{E}^{\mathcal{F}_k}[\|\bar{G}_k^I\|^2]$, we finally obtain that

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\right] &\leq \left(1-p_{\mathrm{c}}/2\right)\mathbb{E}\left[V_k\right]+\frac{(1-p_{\mathrm{c}})(2+p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\bar{G}_k^i\right\|^2\right] \\ &\quad+\frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|G_k^i-\bar{G}_k^i\right\|^2\right]+2\left(1-\tau\right)(1-1/n)\gamma d.\end{aligned}$$

Combining the last inequality with **HX**4, it shows

$$\mathbb{E}\left[V_{k+1}\right]\leq\left(1-\frac{p_{\mathrm{c}}}{2}+(1-p_{\mathrm{c}})\gamma^2\left[A+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{A}\right]\right)\mathbb{E}\left[V_k\right]+(1-p_{\mathrm{c}})\gamma^2\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)$$
$$+(1-p_{\mathrm{c}})\gamma^2\left(B+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\mathbb{E}\left[d_k^2\right]+(1-p_{\mathrm{c}})\gamma^2\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_k^2\right]+2\left(1-\tau\right)(1-1/n)\gamma d.$$

Since $\gamma\leq\frac{p_{\mathrm{c}}^{1/2}}{2(1-p_{\mathrm{c}})^{1/2}\left[A+(1+2/p_{\mathrm{c}})\bar{A}\right]^{1/2}}$, the above inequality implies that

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\right] &\leq \left(1-\frac{p_{\mathrm{c}}}{4}\right)\mathbb{E}\left[V_k\right]+(1-p_{\mathrm{c}})\gamma^2\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)+2\left(1-\tau\right)(1-1/n)\gamma d \\ &\quad+(1-p_{\mathrm{c}})\gamma^2\left(B+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\mathbb{E}\left[d_k^2\right]+(1-p_{\mathrm{c}})\gamma^2\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_k^2\right].\end{aligned}$$

Using by convention that $\sum_{l=0}^{-1}=0$, an induction shows that

$$\mathbb{E}\left[V_k\right]\leq\left(1-\frac{p_{\mathrm{c}}}{4}\right)^k\mathbb{E}\left[V_0\right]+\frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)+\frac{8\left(1-\tau\right)(n-1)\gamma d}{np_{\mathrm{c}}}$$

$$+ (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \sum_{l=0}^{k-1} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} \mathbb{E}\left[d_l^2\right]$$
$$+ (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} \mathbb{E}\left[\sigma_l^2\right]. \quad (3.62)$$

Moreover, for any $l \in \mathbb{N}^\star$ the assumption **HX**3 implies that

$$\mathbb{E}\left[d_l^2\right] \leq (1 - A_d)\, \mathbb{E}\left[d_{l-1}^2\right] + B_d \mathbb{E}\left[\sigma_{l-1}^2\right] + C_d \mathbb{E}\left[V_{l-1}\right] + D_d,$$

and unrolling the recursion gives that

$$\mathbb{E}\left[d_l^2\right] \leq (1 - A_d)^l\, \mathbb{E}\left[d_0^2\right] + \sum_{j=1}^{l} (1 - A_d)^{l-j} \left(B_d \mathbb{E}\left[\sigma_{j-1}^2\right] + C_d \mathbb{E}\left[V_{j-1}\right]\right) + \frac{D_d}{A_d}. \quad (3.63)$$

Similarly, we also have

$$\mathbb{E}\left[\sigma_l^2\right] \leq (1 - A_\sigma)^l\, \mathbb{E}\left[\sigma_0^2\right] + \sum_{j=1}^{l} (1 - A_\sigma)^{l-j} \left(B_\sigma \mathbb{E}\left[d_{j-1}^2\right] + C_\sigma \mathbb{E}\left[V_{j-1}\right]\right) + \frac{D_\sigma}{A_\sigma}. \quad (3.64)$$

Hence, by plugging (3.64) in (3.62) we obtain that

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^k \mathbb{E}\left[V_0\right] + \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}} \left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{8\,(1-\tau)\,(n-1)\gamma d}{n p_{\mathrm{c}}}$$
$$+ (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \sum_{l=0}^{k-1} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} \mathbb{E}\left[d_l^2\right]$$
$$+ (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1 - A_\sigma)^l\, \mathbb{E}\left[\sigma_0^2\right]$$
$$+ B_\sigma (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1} \sum_{j=1}^{l} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1 - A_\sigma)^{l-j}\, \mathbb{E}\left[d_{j-1}^2\right]$$
$$+ C_\sigma (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1} \sum_{j=1}^{l} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1 - A_\sigma)^{l-j}\, \mathbb{E}\left[V_{j-1}\right]$$
$$+ \frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}} - 4A_d)} \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right). \quad (3.65)$$

In addition, interchanging the summations gives

$$\sum_{l=0}^{k-1} \sum_{j=1}^{l} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1 - A_\sigma)^{l-j}\, \mathbb{E}\left[V_{j-1}^2\right]$$
$$= \sum_{i=0}^{k-2} \left[\sum_{l=0}^{k-i-2} \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^{k-i-2-l} (1 - A_\sigma)^l\right] \mathbb{E}\left[V_i\right].$$

Thus, using that $\sum_{l=0}^{k-i-2} \left(1 - p_{\mathrm{c}}/4\right)^{k-i-2-l} (1 - A_\sigma)^l \leq 4\,(1 - A_d)^{k-i-1}\,(p_{\mathrm{c}} - 4A_d)^{-1}$, we can simplify the upper bound of $\mathbb{E}\left[V_k\right]$ derived in (3.65). Indeed, we can write

$$\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_{\mathrm{c}}}{4}\right)^k \mathbb{E}\left[V_0\right] + \frac{4(1-p_{\mathrm{c}})\gamma^2\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right]$$
$$+\frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)+\frac{8\left(1-\tau\right)(n-1)\gamma d}{np_{\mathrm{c}}}+\frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)$$
$$+(1-p_{\mathrm{c}})\gamma^2\left(B+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right]$$
$$+B_\sigma(1-p_{\mathrm{c}})\gamma^2\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{j=0}^{l-1}\left(1-A_\sigma\right)^{l-j-1}\mathbb{E}\left[d_j^2\right]$$
$$+\frac{4(1-p_{\mathrm{c}})\gamma^2 C_\sigma}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\sum_{l=0}^{k-2}\left(1-A_d\right)^{k-l-1}\mathbb{E}\left[V_l\right]. \quad (3.66)$$

Upper bound on $\mathbb{E}\left[d_k^2\right]$. For $l \geq 1$, plugging (3.64) into (3.63) yields the following upper bound

$$\mathbb{E}\left[d_l^2\right] \leq \left(1-A_d\right)^l \mathbb{E}\left[d_0^2\right] + C_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\mathbb{E}\left[V_{j-1}\right]+\frac{D_d}{A_d}$$
$$+B_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\Bigg[\left(1-A_\sigma\right)^{j-1}\mathbb{E}\left[\sigma_0^2\right]$$
$$+\sum_{i=1}^{j-1}\left(1-A_\sigma\right)^{j-i-1}\left(B_\sigma\mathbb{E}\left[d_{i-1}^2\right]+C_\sigma\mathbb{E}\left[V_{i-1}\right]\right)+\frac{D_\sigma}{A_\sigma}\Bigg].$$

The above inequality leads to the next inequality

$$\mathbb{E}\left[d_l^2\right] \leq \left(1-A_d\right)^l \mathbb{E}\left[d_0^2\right] + B_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-1}\mathbb{E}\left[\sigma_0^2\right]$$
$$+C_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\mathbb{E}\left[V_{j-1}\right]+B_dC_\sigma\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_\sigma\right)^{j-i-1}\left(1-A_d\right)^{l-j}\mathbb{E}\left[V_{i-1}\right]$$
$$+B_dB_\sigma\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\mathbb{E}\left[d_{i-1}^2\right]+\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}. \quad (3.67)$$

By interchanging the double summations in (3.67), we obtain

$$\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\mathbb{E}\left[d_{i-1}^2\right]$$
$$=\sum_{i=1}^{l-1}\left[\sum_{j=i+1}^{l}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\right]\mathbb{E}\left[d_{i-1}^2\right]$$
$$=\sum_{i=0}^{l-2}\left[\sum_{j=0}^{l-i-2}\left(1-A_d\right)^{l-i-2-j}\left(1-A_\sigma\right)^{j}\right]\mathbb{E}\left[d_i^2\right] \leq \frac{1}{A_\sigma-A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[d_i^2\right]. \quad (3.68)$$

Similarly, we can also get that

$$\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\mathbb{E}\left[V_{i-1}\right] \leq \frac{1}{A_\sigma - A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right]. \tag{3.69}$$

Plugging back (3.68) and (3.69) in (3.67) shows

$$\begin{aligned}\mathbb{E}\left[d_l^2\right] \leq \left(1-A_d\right)^l\mathbb{E}\left[d_0^2\right] + \frac{B_d\left(1-A_d\right)^l}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right] + \frac{B_dB_\sigma}{A_\sigma - A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[d_i^2\right] \\ + C_d\sum_{i=0}^{l-1}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right] + \frac{B_dC_\sigma}{A_\sigma - A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right] + \frac{D_d}{A_d} + \frac{B_dD_\sigma}{A_dA_\sigma}.\end{aligned} \tag{3.70}$$

Now, we want to control $\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[d_i^2\right]$. For this, for any $l \in \mathbb{N}$ define

$$\begin{aligned}U_l = \mathbb{E}\left[d_0^2\right] + \frac{B_d}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right] + \frac{D_d\left(1-A_d\right)^{-l}}{A_d} + \frac{B_dD_\sigma\left(1-A_d\right)^{-l}}{A_dA_\sigma} \\ + C_d\sum_{i=0}^{l-1}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right] + \frac{B_dC_\sigma}{A_\sigma - A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right]\end{aligned} \tag{3.71}$$

and consider

$$S_l = \sum_{i=0}^{l}\left(1-A_d\right)^{-i}\mathbb{E}\left[d_i^2\right].$$

With the above notation, (3.70) can be rewritten as

$$S_l - S_{l-1} \leq \frac{B_dB_\sigma}{\left(1-A_d\right)\left(A_\sigma - A_d\right)}S_{l-2} + U_l. \tag{3.72}$$

For $l \geq 2$, using the upper bound derived in (3.72) gives

$$\mathbb{E}\left[d_l^2\right] = \left(1-A_d\right)^l\left(S_l - S_{l-1}\right) \leq \frac{B_dB_\sigma\left(1-A_d\right)^{l-1}S_{l-2}}{\left(A_\sigma - A_d\right)} + \left(1-A_d\right)^l U_l. \tag{3.73}$$

Finally, we define

$$\delta_\alpha = \frac{-1+\sqrt{1+4\left(1-A_d\right)^{-1}\left(A_\sigma - A_d\right)^{-1}B_dB_\sigma}}{2} \tag{3.74}$$

such that δ_α is solution of the equation

$$\delta_\alpha^2 + \delta_\alpha = \frac{B_dB_\sigma}{\left(1-A_d\right)\left(A_\sigma - A_d\right)} \tag{3.75}$$

Thus for $l \geq 2$, the definition of δ_α combined with (3.72) show

$$S_l + \delta_\alpha S_{l-1} \leq \left(1+\delta_\alpha\right)\left(S_{l-1} + \delta_\alpha S_{l-2}\right) + U_l.$$

Unrolling this recursion gives

$$S_k + \delta_\alpha S_{k-1} \leq \left(1+\delta_\alpha\right)^{k-1}\left(S_1 + \delta_\alpha S_0\right) + \sum_{l=2}^{k}\left(1+\delta_\alpha\right)^{k-l}U_l. \tag{3.76}$$

Upper bound on $\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{l-j-1}\mathbb{E}[d_j^2]$. Let consider a fixed $\tilde{\alpha}\in\{p_{\mathrm{c}}/4, A_\sigma\}$, by assumption we have $A_d<\tilde{\alpha}<1$. Since we want to control $\sum_{l=0}^{k-1}(1-p_{\mathrm{c}}/4)^{k-l-1}\mathbb{E}[d_l^2]$ and $\sum_{l=0}^{k-1}(1-p_{\mathrm{c}}/4)^{k-l-1}\sum_{j=0}^{l-1}\left(1-A_\sigma\right)^{l-j-1}\mathbb{E}[d_j^2]$ involved in the inequality (3.66), we first study $\sum_{l=0}^{k-1}(1-\tilde{\alpha})^{k-l-1}\mathbb{E}[d_l^2]$. From (3.73), we deduce that

$$\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right]\leq\frac{B_dB_\sigma}{(1-A_d)(A_\sigma-A_d)}\sum_{l=0}^{k-1}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}S_{l-2}+\sum_{l=0}^{k-1}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}U_l. \tag{3.77}$$

Since we suppose **HX**3 and $A_d\leq A_\sigma/2$, $A_dA_\sigma\geq 8B_dB_\sigma$ we can apply Lemma 3.11 which shows that $1-\alpha=(1-A_d)(1+\delta_\alpha)\in\left(0,1-\tilde{\alpha}\right)$ and leads to

$$\sum_{l=0}^{k-1}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-3}\leq\left(1+\delta_\alpha\right)^{-3}\sum_{l=0}^{k-1}\left(1-\alpha\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\leq\frac{\left(1-\alpha\right)^{k}}{(\tilde{\alpha}-\alpha)(1+\delta_\alpha)^3}. \tag{3.78}$$

Moreover, for $l\geq 2$ applying the result given by (3.76), we have

$$S_{l-2}\leq\left(1+\delta_\alpha\right)^{l-3}\left(S_1+\delta_\alpha S_0\right)+\sum_{j=2}^{l-2}\left(1+\delta_\alpha\right)^{l-j-2}U_j. \tag{3.79}$$

Using the definition of U_l given by (3.71), we can write the following equality

$$\begin{aligned}
&\sum_{l=0}^{k-1}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\sum_{j=2}^{l-2}\left(1+\delta_\alpha\right)^{l-j-2}U_j\\
&\quad=\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}\\
&\qquad+\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l-j}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}\\
&+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}\sum_{i=0}^{j-1}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right]
\end{aligned} \tag{3.80}$$

We now upper bound each quantity separately. Regarding the first double sum, since $(1-A_d)(1+\delta_\alpha)=1-\alpha$ we get

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}\\
&=\sum_{j=2}^{k-3}\left(1-A_d\right)^{j+2}\sum_{l=j+2}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1-\alpha\right)^{l-j-2}
\end{aligned}$$

$$\leq \frac{1}{\tilde{\alpha}-\alpha}\sum_{j=4}^{k-1}\left(1-A_d\right)^j\left(1-\alpha\right)^{k-j} \leq \frac{\left(1-A_d\right)^4\left(1-\alpha\right)^{k-3}}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}. \tag{3.81}$$

Using $(1-A_d)(1+\delta_\alpha) = 1-\alpha$ combined with $\sum_{l=j+2}^{k-1}(1-\alpha)^{l-j-2}(1-\tilde{\alpha})^{k-l-1} \leq (\tilde{\alpha}-\alpha)^{-1}(1-\alpha)^{k-j-2}$ give

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l-j}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2} \\
&= \left(1-A_d\right)^2\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-\alpha\right)^{l-j-2}\left(1-\tilde{\alpha}\right)^{k-l-1} \\
&= \left(1-A_d\right)^2\sum_{j=2}^{k-3}\sum_{l=j+2}^{k-1}\left(1-\alpha\right)^{l-j-2}\left(1-\tilde{\alpha}\right)^{k-l-1} \\
&\leq \frac{\left(1-A_d\right)^2}{\tilde{\alpha}-\alpha}\sum_{j=2}^{k-3}\left(1-\alpha\right)^{k-j-2} \leq \frac{(1-\alpha)\left(1-A_d\right)^2}{\alpha(\tilde{\alpha}-\alpha)}.
\end{aligned} \tag{3.82}$$

The same arguments show that

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}\sum_{i=0}^{j-1}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right] \\
&\leq \sum_{i=0}^{k-4}\sum_{j=i+1}^{k-3}\sum_{l=j+2}^{k-1}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right] \\
&\leq \sum_{i=0}^{k-4}\mathbb{E}\left[V_i\right]\sum_{j=i+1}^{k-3}\left(1-A_d\right)^{j-i+1}\sum_{l=j+2}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1-\alpha\right)^{l-j-2} \\
&\leq \frac{1}{\tilde{\alpha}-\alpha}\sum_{i=0}^{k-4}\mathbb{E}\left[V_i\right]\sum_{j=i+1}^{k-3}\left(1-A_d\right)^{j-i+1}\left(1-\alpha\right)^{k-j-2} \\
&= \frac{(1-\alpha)\left(1-A_d\right)^2}{\tilde{\alpha}-\alpha}\sum_{i=0}^{k-4}\mathbb{E}\left[V_i\right]\sum_{j=i+1}^{k-3}\left(1-A_d\right)^{j-i-1}\left(1-\alpha\right)^{k-j-3} \\
&\leq \frac{\left(1-\alpha\right)^{-1}\left(1-A_d\right)^2}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}\sum_{i=0}^{k-4}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right].
\end{aligned} \tag{3.83}$$

Therefore, plugging (3.81), (3.82), (3.83) inside (3.80) implies

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^{l}\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}U_j \\
&\leq \frac{(1-\alpha)\left(1-A_d\right)^2}{\alpha(\tilde{\alpha}-\alpha)}\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)+\frac{\left(1-A_d\right)^4\left(1-\alpha\right)^{k-3}}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \\
&\quad+\frac{\left(1-\alpha\right)^{-1}\left(1-A_d\right)^2}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-4}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right].
\end{aligned} \tag{3.84}$$

In addition, by definition of U_l provides in (3.71) we have

$$\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}U_l=\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}$$
$$+\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}$$
$$+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\sum_{i=0}^{l-1}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right].$$

Thus, a calculation yields that

$$\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}U_l\leq\frac{\left(1-A_d\right)^k}{\tilde{\alpha}-A_d}\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)$$
$$+\frac{1}{\tilde{\alpha}}\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)+\frac{1}{\tilde{\alpha}-A_d}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(1-A_d\right)^{k-i-1}\mathbb{E}\left[V_i\right]. \quad (3.85)$$

Plugging (3.79) in (3.77) shows

$$\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right]\leq\frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)}{(1-A_d)(A_\sigma-A_d)}\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-3}$$
$$+\frac{B_dB_\sigma}{(1-A_d)(A_\sigma-A_d)}\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}U_j$$
$$+\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}U_l. \quad (3.86)$$

Hence, by combining (3.78), (3.84), (3.85) and (3.86) we obtain for $A_d>\alpha$, that

$$\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right]\leq\frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(1-A_d)(A_\sigma-A_d)(\tilde{\alpha}-\alpha)(1+\delta_\alpha)^3}$$
$$+\left(\frac{\left(1-A_d\right)^k}{\tilde{\alpha}-A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^k}{(A_\sigma-A_d)(A_d-\alpha)(\tilde{\alpha}-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)$$
$$+\left(\frac{1}{\tilde{\alpha}}+\frac{B_dB_\sigma}{\alpha(\tilde{\alpha}-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)$$
$$+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(\frac{\left(1-A_d\right)^{k-i-1}}{\tilde{\alpha}-A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^{k-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(\tilde{\alpha}-\alpha)}\right)\mathbb{E}\left[V_i\right]. \quad (3.87)$$

In addition, the above bound holds even if $A_d=\alpha$ by considering that $(A_d-\alpha)^{-1}B_dB_\sigma=0$.

Upper bound on $\sum_{l=0}^{k-1}(1-p_c/4)^{k-l-1}\mathbb{E}\left[d_l^2\right]$. Applying (3.87) with $\tilde{\alpha}=p_c/4$ gives

$$\begin{aligned}\sum_{l=0}^{k-1}\left(1-\frac{p_c}{4}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right] &\leq \frac{4B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(1-A_d)(A_\sigma-A_d)(p_c-4\alpha)(1+\delta_\alpha)^3}\\ &+\left(\frac{4\left(1-A_d\right)^k}{p_c-4A_d}+\frac{4B_dB_\sigma\left(1-\alpha\right)^k}{(A_\sigma-A_d)(A_d-\alpha)(p_c-4\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\ &+\left(\frac{4}{p_c}+\frac{4B_dB_\sigma}{\alpha(p_c-4\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\ &+4\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(\frac{\left(1-A_d\right)^{k-i-1}}{p_c-4A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^{k-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(p_c-4\alpha)}\right)\mathbb{E}\left[V_i\right].\end{aligned} \tag{3.88}$$

Upper bound on $\sum_{l=0}^{k-1}(1-p_c/4)^{k-l-1}\sum_{j=0}^{l-1}\left(1-A_\sigma\right)^{l-j-1}\mathbb{E}[d_j^2]$. Recall that we consider that $(A_d-\alpha)^{-1}B_dB_\sigma=0$ in the specific case where $A_d=\alpha$. This time, setting $\tilde{\alpha}=A_\sigma$ in (3.87) shows that

$$\begin{aligned}\sum_{j=0}^{l-1}\left(1-A_\sigma\right)^{l-j-1}\mathbb{E}\left[d_l^2\right] &\leq \frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^l}{(1-A_d)(A_\sigma-A_d)(A_\sigma-\alpha)(1+\delta_\alpha)^3}\\ &+\left(\frac{\left(1-A_d\right)^l}{A_\sigma-A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^l}{\left(A_\sigma-A_d\right)\left(A_d-\alpha\right)\left(A_\sigma-\alpha\right)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\ &+\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\ &+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{l-2}\left(\frac{\left(1-A_d\right)^{l-i-1}}{A_\sigma-A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^{l-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(A_\sigma-\alpha)}\right)\mathbb{E}\left[V_i\right].\end{aligned} \tag{3.89}$$

Moreover, we have the two following bounds

$$\begin{aligned}\sum_{l=0}^{k-1}\left(1-\frac{p_c}{4}\right)^{k-l-1}\left(1-A_d\right)^l &\leq \frac{4\left(1-A_d\right)^k}{p_c-4A_d},\\ \sum_{l=0}^{k-1}\left(1-\frac{p_c}{4}\right)^{k-l-1}\left(1-\alpha\right)^l &\leq \frac{4\left(1-\alpha\right)^k}{p_c-4\alpha}.\end{aligned} \tag{3.90}$$

Therefore, permuting the summations implies

$$\begin{aligned}\sum_{l=0}^{k-1}\left(1-\frac{p_c}{4}\right)^{k-l-1}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right] &\leq \sum_{i=0}^{k-3}\mathbb{E}\left[V_i\right]\sum_{l=i+2}^{k-1}\left(1-\frac{p_c}{4}\right)^{k-l-1}\left(1-A_d\right)^{l-i-1}\\ &\leq \frac{4}{p_c-4A_d}\sum_{i=0}^{k-3}\left(1-A_d\right)^{k-i-1}\mathbb{E}\left[V_i\right].\end{aligned} \tag{3.91}$$

In a similar way, we obtain

$$\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{i=0}^{l-2}\left(1-\alpha\right)^{l-i-1}\mathbb{E}\left[V_i\right]\leq\frac{4}{p_{\mathrm{c}}-4\alpha}\sum_{i=0}^{k-3}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right]. \quad (3.92)$$

Hence, the combination of (3.89) with (3.90), (3.91), (3.92) yields

$$\begin{aligned}
&\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{j=0}^{l-1}\left(1-A_\sigma\right)^{l-j-1}\mathbb{E}\left[d_l^2\right]\\
&\qquad\leq\frac{4B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(p_{\mathrm{c}}-4\alpha)(1-A_d)(A_\sigma-A_d)(A_\sigma-\alpha)(1+\delta_\alpha)^3}\\
&+\frac{4}{A_\sigma-A_d}\left(\frac{\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^k}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&\qquad+\frac{4}{p_{\mathrm{c}}}\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+\frac{4}{A_\sigma-A_d}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-3}\left(\frac{\left(1-A_d\right)^{k-i-1}}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^{k-i-1}}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\mathbb{E}\left[V_i\right].
\end{aligned} \quad (3.93)$$

Upper bound on $\mathbb{E}\left[V_k\right]$. Plugging (3.88) and (3.93) in (3.66), we obtain

$$\begin{aligned}
\mathbb{E}\left[V_k\right]&\leq\left(1-\frac{p_{\mathrm{c}}}{4}\right)^k\mathbb{E}\left[V_0\right]+\frac{4(1-p_{\mathrm{c}})\gamma^2\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right]\\
&+\frac{4(1-p_{\mathrm{c}})\gamma^2D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)+\frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)+\frac{8\left(1-\tau\right)(n-1)\gamma d}{np_{\mathrm{c}}}\\
&\qquad+(1-p_{\mathrm{c}})\gamma^2\left(B+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\left[\frac{4B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(1-A_d)(A_\sigma-A_d)(p_{\mathrm{c}}-4\alpha)(1+\delta_\alpha)^3}\right.\\
&\quad+\left(\frac{4\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}+\frac{4B_dB_\sigma\left(1-\alpha\right)^k}{(A_\sigma-A_d)(A_d-\alpha)(p_{\mathrm{c}}-4\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&\qquad+\left(\frac{4}{p_{\mathrm{c}}}+\frac{4B_dB_\sigma}{\alpha(p_{\mathrm{c}}-4\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&\left.+4\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(\frac{\left(1-A_d\right)^{k-i-1}}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^{k-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(p_{\mathrm{c}}-4\alpha)}\right)\mathbb{E}\left[V_i\right]\right]\\
&+4(1-p_{\mathrm{c}})\gamma^2B_\sigma\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\left[\frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(p_{\mathrm{c}}-4\alpha)(1-A_d)(A_\sigma-A_d)(A_\sigma-\alpha)(1+\delta_\alpha)^3}\right.\\
&+\frac{1}{A_\sigma-A_d}\left(\frac{\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^k}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)
\end{aligned}$$

$$+\frac{1}{p_c}\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)$$
$$+\frac{1}{A_\sigma-A_d}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-3}\left(\frac{\left(1-A_d\right)^{k-i-1}}{p_c-4A_d}+\frac{B_dB_\sigma\left(1-\alpha\right)^{k-i-1}}{(p_c-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\mathbb{E}\left[V_i\right]\Bigg]$$
$$+\frac{4(1-p_c)\gamma^2C_\sigma}{p_c-4A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\sum_{l=0}^{k-2}\left(1-A_d\right)^{k-l-1}\mathbb{E}\left[V_l\right]. \quad (3.94)$$

For any negative number $j<0$, using the convention that $\sum_{l=0}^{j}=0$ and simplifying the calculations provided by (3.94), we find that

$$\mathbb{E}\left[V_k\right]\leq\left(1-\frac{p_c}{4}\right)^k\mathbb{E}\left[V_0\right]+\frac{4(1-p_c)\gamma^2\left(1-A_d\right)^k}{p_c-4A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right]$$
$$+\frac{4(1-p_c)\gamma^2B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(p_c-4\alpha)(1-A_d)(A_\sigma-A_d)(1+\delta_\alpha)^3}\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-\alpha}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]$$
$$+\frac{4(1-p_c)\gamma^2D_\sigma}{A_\sigma(p_c-4A_d)}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+\frac{4(1-p_c)\gamma^2}{p_c}\left(D+\frac{2+p_c}{p_c}\bar{D}\right)+\frac{8\left(1-\tau\right)(n-1)\gamma d}{np_c}$$
$$+4(1-p_c)\gamma^2\Bigg[\left(\frac{1}{p_c}+\frac{B_dB_\sigma}{\alpha(p_c-4\alpha)(A_\sigma-A_d)}\right)\left(B+\frac{2+p_c}{p_c}\bar{B}\right)$$
$$+\frac{B_\sigma}{p_c}\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\Bigg]\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)$$
$$+\frac{4\gamma^2\left(1-p_c\right)\left(1-A_d\right)^k}{p_c-4A_d}\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)$$
$$+\frac{4\gamma^2\left(1-p_c\right)B_dB_\sigma\left(1-\alpha\right)^k}{(p_c-4\alpha)\left(A_d-\alpha\right)(A_\sigma-A_d)}\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-\alpha}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)$$
$$+\frac{4\gamma^2\left(1-p_c\right)}{p_c-4A_d}\left[C_\sigma\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\left(B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right)\right]$$
$$\times\sum_{i=0}^{k-2}\left(1-A_d\right)^{k-i-1}\mathbb{E}\left[V_i\right]$$
$$+\frac{4\gamma^2\left(1-p_c\right)B_dB_\sigma}{(p_c-4\alpha)\left(A_d-\alpha\right)(A_\sigma-A_d)}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-\alpha}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]$$
$$\times\sum_{i=0}^{k-3}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right]. \quad (3.95)$$

As explained in (3.75), recall that

$$\delta_\alpha^2+\delta_\alpha=\frac{B_dB_\sigma}{\left(1-A_d\right)\left(A_\sigma-A_d\right)},\qquad\qquad\alpha=A_d-\delta_\alpha(1-A_d).$$

Thus, when $B_dB_\sigma\neq0$ then $\delta_\alpha\neq0$, which implies that $A_d\neq\alpha$ and gives

$$\frac{B_dB_\sigma}{\left(A_d-\alpha\right)\left(A_\sigma-A_d\right)}=1+\delta_\alpha.$$

In addition, in the proof of Lemma 3.11 we saw that $2\delta_\alpha\leq A_d\leq1/2$ and also that $A_d/2\leq\alpha\leq A_d$. Therefore, we can regroup several terms in (3.95) and write

$$\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_c}{4}\right)^k \mathbb{E}\left[V_0\right] + \frac{4(1-p_c)\gamma^2\left(1-A_d\right)^k}{p_c-4A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right]$$
$$+\frac{4(1-p_c)\gamma^2\delta_\alpha\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(p_c-4A_d)(1+\delta_\alpha)^2}\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]$$
$$+\frac{9\gamma^2\left(1-p_c\right)\left(1-\alpha\right)^k}{p_c-4A_d}\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)$$
$$+\frac{4(1-p_c)\gamma^2 D_\sigma}{A_\sigma(p_c-4A_d)}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+\frac{4(1-p_c)\gamma^2}{p_c}\left(D+\frac{2+p_c}{p_c}\bar{D}\right)+\frac{8\left(1-\tau\right)(n-1)\gamma d}{np_c}$$
$$+4(1-p_c)\gamma^2\left[\left(\frac{1}{p_c}+\frac{2B_dB_\sigma}{A_d(p_c-4A_d)(A_\sigma-A_d)}\right)\left(B+\frac{2+p_c}{p_c}\bar{B}\right)\right.$$
$$\left.+\frac{B_\sigma}{p_c}\left(\frac{1}{A_\sigma}+\frac{2B_dB_\sigma}{A_d(A_\sigma-A_d)^2}\right)\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right]\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)$$
$$+\frac{9\gamma^2\left(1-p_c\right)}{p_c-4A_d}\left[C_\sigma\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\left(B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right)\right]$$
$$\times\sum_{i=0}^{k-2}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right]. \quad (3.96)$$

Recall that we defined C^γ in (3.60) by

$$\mathrm{C}^\gamma = \frac{4(1-p_c)\gamma^2}{p_c-4A_d}\left[B+\frac{2+p_c}{p_c}\bar{B}+\frac{B_\sigma}{A_\sigma-A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)\right].$$

Hence, using (3.96) we get that

$$\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_c}{4}\right)^k \mathbb{E}\left[V_0\right]+\frac{4(1-p_c)\gamma^2 D_\sigma}{A_\sigma(p_c-4A_d)}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+\frac{4(1-p_c)\gamma^2}{p_c}\left(D+\frac{2+p_c}{p_c}\bar{D}\right)$$
$$+\frac{\mathrm{C}^\gamma}{A_d}\left(1+\frac{2B_dB_\sigma}{A_d(A_\sigma-A_d)}\right)\left(D_d+\frac{B_dD_\sigma}{A_\sigma}\right)+\frac{8\left(1-\tau\right)(n-1)\gamma d}{np_c}$$
$$+\left(\frac{4(1-p_c)\gamma^2\left(1-A_d\right)^k}{p_c-4A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+\frac{9\mathrm{C}^\gamma B_d\left(1-\alpha\right)^k}{4\left(A_\sigma-A_d\right)}\right)\mathbb{E}\left[\sigma_0^2\right]$$
$$+\frac{9}{4}\mathrm{C}^\gamma\left(1-\alpha\right)^k\mathbb{E}\left[d_0^2\right]+\mathrm{C}^\gamma\left(1-\alpha\right)^{k-2}\left(A_d-\alpha\right)\left(1-A_d\right)\left(S_1+\frac{A_d-\alpha}{1-A_d}S_0\right)$$
$$+\left[\frac{9\gamma^2\left(1-p_c\right)C_\sigma}{p_c-4A_d}\left(C+\frac{2+p_c}{p_c}\bar{C}\right)+3\mathrm{C}^\gamma\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\right]\sum_{i=0}^{k-2}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right].$$

Finally, we conclude the proof remarking that

$$\mathrm{C}^\gamma\left(1-\alpha\right)^{k-2}\left(A_d-\alpha\right)\left[(1-A_d)S_1+(A_d-\alpha)S_0\right]$$
$$\leq \mathrm{C}^\gamma\left(1-\alpha\right)^{k-2}\left(A_d-\alpha\right)\left[(2-A_d-\alpha)\mathbb{E}\left[d_0^2\right]+B_d\mathbb{E}\left[\sigma_0^2\right]+C_d\mathbb{E}\left[V_0\right]+D_d\right]$$
$$\leq \mathrm{C}^\gamma\left(1-\alpha\right)^k\left(4\mathbb{E}\left[d_0^2\right]+2B_d\mathbb{E}\left[\sigma_0^2\right]+2C_d\mathbb{E}\left[V_0\right]+2D_d\right). \quad (3.97)$$

■

In order to ease notation, with the definitions used in **HX**4 and (3.60), consider for any $\gamma \in \mathbb{R}_+$ the variable $\mathrm{C}_\epsilon^\gamma \in \mathbb{R}_+$ defined by

$$\mathrm{C}_\epsilon^\gamma = \mathrm{C}_V^\gamma \mathbb{E}\left[V_0\right] + \mathrm{C}_{d_0}^\gamma \mathbb{E}\left[d_0^2\right] + \mathrm{C}_\sigma^\gamma \mathbb{E}\left[\sigma_0^2\right] + 2D_d \tag{3.98}$$

In addition, with the previous notations consider

$$\delta = \frac{2\left(1 - A_d/2\right)^{-1} \mathrm{C}_r^\gamma}{1 + \sqrt{1 + 4\left(1 - A_d/2\right)^{-1} \mathrm{C}_r^\gamma}}$$

and define

$$\gamma_V = \frac{p_{\mathrm{c}}^{1/2}}{(2 - 2p_{\mathrm{c}})^{1/2}\left[A + (1 + 2/p_{\mathrm{c}})\bar{A}\right]^{1/2}}.$$

Lemma 3.13. *Assume **HX**3, **HX**4 hold with* $4\mathrm{C}_r^\gamma \leq A_d < \min(A_\sigma/2, p_{\mathrm{c}}/4), A_d A_\sigma \geq 8B_d B_\sigma$ *and let* $\gamma \in \left(0, \gamma_V\right]$. *Then, for any* $k \geq 1$, *we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{4}\right)^k \left(2\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma}{A_d}\right) + \mathrm{C}_\delta^\gamma,$$

where V_k *is defined in* (3.24), $\mathrm{C}_\epsilon^\gamma, \mathrm{C}_r^\gamma, \mathrm{C}_\delta^\gamma$ *in* (3.60) *and* (3.98).

Proof Let k in $\mathbb{N}$ be fixed. Since the assumptions of Lemma 3.12 are satisfied, we know that

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \alpha\right)^k \mathrm{C}_\epsilon^\gamma + \mathrm{C}_r^\gamma \sum_{l=0}^{k-2} \left(1 - \alpha\right)^{k-l-1} \mathbb{E}\left[V_l\right] + \mathrm{C}_\delta^\gamma, \tag{3.99}$$

where α is defined in (3.61). In addition, Lemma 3.11 shows that $A_d/2 \leq \alpha$. Hence, multiplying the last inequality by the weight ω_k defined for any $l \in \mathbb{N}$, by

$$\omega_l = \left(1 - A_d/2\right)^{-l},$$

we obtain the following inequality

$$\omega_k \mathbb{E}\left[V_k\right] \leq \mathrm{C}_\epsilon^\gamma + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2} \sum_{l=0}^{k-2} \omega_l \mathbb{E}\left[V_l\right] + \mathrm{C}_\delta^\gamma \omega_k.$$

Applying the sharp Grönwall inequality (Holte, 2009), we get

$$\omega_k \mathbb{E}\left[V_k\right] \leq \mathrm{C}_\epsilon^\gamma + \omega_k \mathrm{C}_\delta^\gamma + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2} \sum_{l=0}^{k-1} \left(\mathrm{C}_\epsilon^\gamma + \omega_l \mathrm{C}_\delta^\gamma\right) \left(1 + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2}\right)^{k-l-1}.$$

Therefore, a calculation shows that

$$\omega_k \mathbb{E}\left[V_k\right] \leq \mathrm{C}_\epsilon^\gamma + \omega_k \mathrm{C}_\delta^\gamma + \mathrm{C}_\epsilon^\gamma \left(1 + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2}\right)^k + \frac{\mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma}{1 - A_d/2} \sum_{l=0}^{k-1} \omega_l \left(1 + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2}\right)^{k-l-1},$$

and simplifying the previous inequality gives the following upper bound:

$$\mathbb{E}\left[V_k\right] \leq \mathrm{C}_\delta^\gamma + \omega_k^{-1}\mathrm{C}_\epsilon^\gamma + \mathrm{C}_\epsilon^\gamma \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^k + \mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma \sum_{l=0}^{k-1} \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^{k-l-1}. \quad (3.100)$$

In addition, using $4\mathrm{C}_r^\gamma < A_d < p_\mathrm{c}/4$ implies $0 < 1 - A_d/2 + \mathrm{C}_r^\gamma < 1$ which combined with (3.100) gives

$$\mathbb{E}\left[V_k\right] \leq \mathrm{C}_\delta^\gamma + \omega_k^{-1}\mathrm{C}_\epsilon^\gamma + \mathrm{C}_\epsilon^\gamma \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^k + \frac{\mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma}{A_d/2 - \mathrm{C}_r^\gamma} \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^k.$$

Eventually, combining the last inequality with the assumption $4\mathrm{C}_r^\gamma < A_d$ completes the proof. ■

With the notation of the assumptions **HX**3 and **HX**4, we define

$$\alpha_d = \frac{4\gamma^2}{p_\mathrm{c} A_d} \max\left\{p_\mathrm{c} B + 3\bar{B}, \frac{4B_\sigma}{A_\sigma}\left(p_\mathrm{c} C + 3\bar{C}\right)\right\}, \quad \alpha_\sigma = \frac{4\gamma^2\left(p_\mathrm{c} C + 3\bar{C}\right)}{p_\mathrm{c} A_\sigma}. \quad (3.101)$$

The following lemma is used in the convergence proof of `VR-FALD⋆` (see Lemma 3.28).

Lemma 3.14. *Assume **HX**3, **HX**4 hold with*

$$A_d \leq \min\left(A_\sigma, \frac{p_\mathrm{c}}{4}\right), \quad \alpha_d C_d + \alpha_\sigma C_\sigma \leq \frac{p_\mathrm{c}}{8}, \quad \alpha_d B_d + \gamma^2\left(C + \frac{3}{p_\mathrm{c}}\bar{C}\right) \leq \frac{\alpha_\sigma A_\sigma}{2},$$

and consider $\gamma \leq p_\mathrm{c}^{1/2}(2 - 2p_\mathrm{c})^{-1/2}[A + (1 + 2/p_\mathrm{c})\bar{A}]^{-1/2}$. *Then, for any* $k \in \mathbb{N}$, *we have*

$$\mathbb{E}\left[V_k\right] + \alpha_d \mathbb{E}\left[d_k^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_k^2\right] \leq \left(1 - \frac{A_d}{2}\right)^k \left(\mathbb{E}\left[V_0\right] + \alpha_d \mathbb{E}\left[d_0^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_0^2\right]\right) + \frac{2(1 - p_\mathrm{c})\gamma^2}{A_d}\left(D + \frac{2 + p_\mathrm{c}}{p_\mathrm{c}}\bar{D}\right) + \frac{2\alpha_d D_d + 2\alpha_\sigma D_\sigma}{A_d} + \frac{4\left(1 - \tau\right)(n - 1)\gamma d}{n A_d},$$

where V_k *is defined in* (3.24).

Proof Let $k \in \mathbb{N}^\star$, using for $i \in [n]$ the definitions (3.20), (3.23) of X_k^i and X_k

$$X_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\left(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1 - \tau}\,\tilde{Z}_{k+1}^i\right),$$
$$X_{k+1} = X_k - \frac{\gamma}{n}\sum_{j=1}^n G_k^i + \sqrt{\frac{2\gamma\tau}{n}}\tilde{Z}_{k+1} + \frac{\sqrt{2(1 - \tau)\gamma}}{n}\sum_{i=1}^n Z_{k+1}^i.$$

Substracting the two above equations combined with the Jensen inequality give

$$\mathbb{E}\left[V_{k+1}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|X_{k+1}^i - X_{k+1}\right\|^2\right]$$

$$= \frac{1-p_{\mathrm{c}}}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(G_k^i - G^k) + \sqrt{2(1-\tau)\gamma}Z_{k+1}^i - \frac{\sqrt{2(1-\tau)\gamma}}{n}\sum_{j=1}^{n}Z_{k+1}^j\right\|^2\right]$$

$$= \frac{1-p_{\mathrm{c}}}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right] + 2(1-\tau)\gamma\mathbb{E}\left[\left\|Z_{k+1}^i - \frac{1}{n}\sum_{j=1}^{n}Z_{k+1}^j\right\|^2\right]$$
$$+ \frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(G_k^i - \bar{G}_k^i) - (G^k - \bar{G}^k)\right\|^2\right].$$

Hence, we get

$$\mathbb{E}\left[V_{k+1}\right] \leq \frac{1-p_{\mathrm{c}}}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right]$$
$$+ \frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] + 2\left(1-\tau\right)(1-1/n)\gamma d$$
$$\leq \frac{(1-p_{\mathrm{c}})(1+p_{\mathrm{c}}/2)}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|X_k^i - X_k\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right]$$
$$+ \frac{(1-p_{\mathrm{c}})(1+2/p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\bar{G}_k^i - \bar{G}^k\right\|^2\right] + 2\left(1-\tau\right)(1-1/n)\gamma d.$$

We finally obtain

$$\mathbb{E}\left[V_{k+1}\right] \leq \left(1 - p_{\mathrm{c}}/2\right)\mathbb{E}\left[V_k\right] + \frac{(1-p_{\mathrm{c}})(2+p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\bar{G}_k^i\right\|^2\right]$$
$$+ \frac{(1-p_{\mathrm{c}})\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] + 2\left(1-\tau\right)\left(1-\frac{1}{n}\right)\gamma d.$$

Combining the last inequality with **HX**4 shows

$$\mathbb{E}\left[V_{k+1}\right] \leq \left(1 - \frac{p_{\mathrm{c}}}{2} + (1-p_{\mathrm{c}})\gamma^2\left[A + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{A}\right]\right)\mathbb{E}\left[V_k\right]$$
$$+ (1-p_{\mathrm{c}})\gamma^2\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + (1-p_{\mathrm{c}})\gamma^2\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\mathbb{E}\left[d_k^2\right]$$
$$+ (1-p_{\mathrm{c}})\gamma^2\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_k^2\right] + 2\left(1-\tau\right)\left(1-\frac{1}{n}\right)\gamma d.$$

Since $\gamma \leq \frac{p_{\mathrm{c}}^{1/2}}{2(1-p_{\mathrm{c}})^{1/2}\left[A+(1+2/p_{\mathrm{c}})\bar{A}\right]^{1/2}}$, the above inequality implies that

$$\mathbb{E}\left[V_{k+1}\right] \leq \left(1 - \frac{p_{\mathrm{c}}}{4}\right)\mathbb{E}\left[V_k\right] + (1-p_{\mathrm{c}})\gamma^2\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + 2\left(1-\tau\right)(1-1/n)\gamma d$$

$$+ (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \mathbb{E}\left[d_k^2\right] + (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \mathbb{E}\left[\sigma_k^2\right].$$

The previous bound combined with **HX**3 gives that

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] + \alpha_d \mathbb{E}\left[d_{k+1}^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_{k+1}^2\right] &\leq \left[\left(1 - \frac{p_{\mathrm{c}}}{4}\right) + \alpha_d C_d + \alpha_\sigma C_\sigma\right] \mathbb{E}\left[V_k\right] \\
&+ \left[\alpha_d(1-A_d) + \alpha_\sigma B_\sigma + (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\right] \mathbb{E}\left[d_k^2\right] \\
&+ \left[\alpha_\sigma(1-A_\sigma) + \alpha_d B_d + (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right] \mathbb{E}\left[\sigma_k^2\right] \\
&+ (1-p_{\mathrm{c}})\gamma^2 \left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + 2\left(1-\tau\right)\frac{(n-1)}{n}\gamma d + \alpha_d D_d + \alpha_\sigma D_\sigma. \quad (3.102)
\end{aligned}$$

By assumption, we have

$$\begin{aligned}
\alpha_d C_d + \alpha_\sigma C_\sigma &\leq \frac{p_{\mathrm{c}}}{8}, \\
\alpha_d B_d + \gamma^2 \left(C + \frac{3}{p_{\mathrm{c}}}\bar{C}\right) &\leq \frac{\alpha_\sigma A_\sigma}{2},
\end{aligned} \quad (3.103)$$

and by definition of α_d, α_σ given in (3.101), we know that $\alpha_\sigma B_\sigma + \gamma^2(B + 3\bar{B}/p_{\mathrm{c}}) \leq \alpha_d A_d/2$. In addition, since we suppose that $A_d \leq \min(p_{\mathrm{c}}/4, A_\sigma)$, the last inequalities combined with (3.103) imply

$$\begin{aligned}
1 - \frac{p_{\mathrm{c}}}{4} + \alpha_d C_d + \alpha_\sigma C_\sigma &\leq 1 - \frac{A_d}{2} \\
1 - A_d + \frac{\alpha_\sigma}{\alpha_d} B_\sigma + \frac{(1-p_{\mathrm{c}})\gamma^2}{\alpha_d} \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) &\leq 1 - \frac{A_d}{2} \\
1 - A_\sigma + \frac{\alpha_d}{\alpha_\sigma} B_d + \frac{(1-p_{\mathrm{c}})\gamma^2}{\alpha_\sigma} \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) &\leq 1 - \frac{A_d}{2}.
\end{aligned} \quad (3.104)$$

Thus, by taking up (3.102) and using (3.104), we get

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] + \alpha_d \mathbb{E}\left[d_{k+1}^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_{k+1}^2\right] &\leq \left(1 - \frac{A_d}{2}\right) \left(\mathbb{E}\left[V_k\right] + \alpha_d \mathbb{E}\left[d_k^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_k^2\right]\right) \\
&+ (1-p_{\mathrm{c}})\gamma^2 \left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + 2\left(1-\tau\right)\left(1 - \frac{1}{n}\right)\gamma d + \alpha_d D_d + \alpha_\sigma D_\sigma.
\end{aligned}$$

Finally, the stated result follows by induction. ■

3.B Main results

Section 3.B is divided into four subsections in which we prove theoretical results for the FALD and VR-FALD* algorithms. These analyses are presented in Theorem 3.22 and Theorem 3.30. The proofs are based on Lemma 3.13 proved in Section 3.A.3 to ensure

that the local parameters $\{X_k^i\}_{i\in[n]}$ do not deviate too much from X_k, then we apply the general result given in Section 3.A to obtain explicit upper bounds for $W_2(\pi, \mu_k^{(\gamma)})$.

Until the end of this chapter, we consider a family of independent random variables $(\xi^i)_{i=1}^n$ distributed according to $\nu_\xi^{\otimes n}$, and we denote $(H^i)_{i=1}^n$ a family of functions defined on $\mathbb{R}^d \times \mathsf{E} \to \mathbb{R}^d$ such that for each $i \in [n], x \in \mathbb{R}^d$, $H^i(x, \xi^i(\cdot))$ is measurable on $(\mathsf{E}, \mathcal{E})$ and satisfies the following condition:

A 4. *Assume there exists $\hat{L} \geq 0$, such that for any $i \in [n], x, y \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[H^i(x, \xi^i)\right] = \nabla U^i(x),$$
$$\mathbb{E}\left[\left\|H^i(y, \xi^i) - H^i(x, \xi^i)\right\|^2\right] \leq \hat{L}^2 \left\|y - x\right\|^2.$$

The assumption **A**4 is equivalent to **A**2 written in the main chapter, though for clarity we prefer to replace the stochastic gradient $\hat{\nabla} U_k^i$ by $H^i(\cdot, \xi^i)$. To simplify the notation, in what follows we consider the random variable $\xi = (\xi^1, \dots, \xi^n)$, and we denote

$$H : \begin{cases} \mathbb{R}^d \times \mathsf{E}^n \to \mathbb{R}^d \\ (x, z) \mapsto \sum_{i=1}^n H^i(x, z^i) \end{cases}.$$

Thus, for each $x \in \mathbb{R}^d$, with this notation we have $H(x, \xi) = \sum_{i=1}^n H^i(x, \xi^i)$. We also introduce the averaged versions $\bar{U}, \bar{H}$ of the local potentials $\{U^i\}_{i\in[n]}$ and the stochastic gradients $\{H^i\}_{i\in[n]}$ defined by

$$\bar{U}(x) = \frac{1}{n}\sum_{i=1}^n U^i(x), \qquad \bar{H}(x, z) = \frac{1}{n}\sum_{i=1}^n H^i(x, z^i).$$

Remark 3.15. *In the minibatch scenario without replacement, the ith client draws a minibatch $J_i \subset [N_i]$ of size $b_i = |J_i| \in [N_i]$ among N_i data and computes its stochastic gradient, which for $x \in \mathbb{R}^d$ is given by $H^i(x, \xi^i) = \sum_{j\in J_i} \nabla U^{i,j}(x)$. Using the result provided in Vono et al. (2022b, Lemma S4), we know that*

$$\mathbb{E}\left[\left\|H^i(y, \xi^i) - H^i(x, \xi^i)\right\|^2\right] = \left\|\nabla U^i(y) - \nabla U^i(x)\right\|^2 + \mathrm{Var}\left(H^i(y, \xi^i) - H^i(x, \xi^i)\right)$$
$$\leq \left(1 + \frac{b_i(N_i - b_i)\max_{j=1}^{N_i} L_j^i}{N_i(N_i - 1)L}\right) L^2 \left\|y - x\right\|^2.$$

*Therefore, **A**4 is satisfied for a choice of $\hat{L} > 0$ such that*

$$\hat{L} \leq L\sqrt{1 + \max_{i=1}^n \left\{b_i(N_i - b_i)[N_i(N_i - 1)]^{-1}(\max_{j=1}^{N_i} L_j^i)L^{-1}\right\}}.$$

A 5. *For $i \in [n]$, $j \in [N_i]$, assume that $U^{i,j}$ is continuously differentiable, convex and there exists $L_j^i > 0$ such that for any $x, y \in \mathbb{R}^d$,*

$$U^{i,j}(y) \leq U^{i,j}(x) + \left\langle \nabla U^{i,j}(x), y - x \right\rangle + \frac{L_j^i}{2}\left\|y - x\right\|^2.$$

A 6. *Assume there exists $\tilde{\omega} > 0$ such that for any $x \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\left\|H(x,\xi) - H(x_\star,\xi) - \nabla U(x)\right\|^2\right] \leq \tilde{\omega} n^2 \left\|x - x_\star\right\|^2.$$

A1 combined with **A**4 implies **A**6 with $\tilde{\omega} = 2L^2 + 2\hat{L}^2$. However, this new assumption **A**6 is interesting because without stochastic gradient we obtain $\tilde{\omega} = 0$, which allows us to recover the classical Langevin bounds.

Remark 3.16. *Consider the same scenario as detailed in Remark 3.15 and define*

$$\tilde{\omega} = \left(\sum_{i=1}^n \frac{b_i(N_i - b_i)\max_{j=1}^{N_i} L_j^i}{n^2 N_i(N_i-1)}\right) L.$$

Applying Vono et al. (2022b, Lemma S4) we have the following lines

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{H}(x,\xi) - \bar{H}(x_\star,\xi) - \nabla\bar{U}(x)\right\|^2\right] &= \mathrm{Var}\left(\bar{H}(x,\xi) - \bar{H}(x_\star,\xi)\right) \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left(H^i(x,\xi^i) - H^i(x_\star,\xi^i)\right) \leq \tilde{\omega}\left\|x - x_\star\right\|^2.\end{aligned}$$

*Therefore, **A**6 is satisfied and in the deterministic case where all data are used to calculate the gradient, we have $\tilde{\omega} = 0$.*

To deal with variance reduction based algorithms, we consider the following assumption **A**7, which is also implied by **A**1-**A**4, however the constant ω vanishes with exact gradient computation.

A 7. *Assume there exists $\omega \geq 0$ such that for any $i \in [n]$ and $x, y \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\left\|H^i(x,\xi^i) - H^i(y,\xi^i) - \nabla U^i(x) + \nabla U^i(y)\right\|^2\right] \leq \omega\left\|x - y\right\|^2.$$

Remark 3.17. *In the minibatch scenario without replacement detailed in Remark 3.15, the use of Vono et al. (2022b, Lemma S4) implies that*

$$\begin{aligned}\mathbb{E}\left[\left\|H^i(x,\xi^i) - H^i(y,\xi^i) - \nabla U^i(x) + \nabla U^i(y)\right\|^2\right] &= \mathrm{Var}\left(H^i(x,\xi^i) - H^i(y,\xi^i)\right) \\ &\leq \frac{b_i(N_i - b_i)}{N_i(N_i-1)} L \max_{j=1}^{N_i} L_j^i \left\|x - y\right\|^2.\end{aligned}$$

*Thus, **A**7 is satisfied by setting*

$$\omega = \max_{i=1}^n \left\{\frac{b_i(N_i - b_i)}{N_i(N_i-1)} \max_{j=1}^{N_i} L_j^i\right\} L.$$

In the deterministic case, we obtain $\omega = 0$. Similarly, in the minibatch scenario with replacement it is sufficient to set

$$\omega = \frac{N_i - b_i}{b_i}\sum_{j=1}^{N_i}(L_j^i)^2$$

*to ensure that **A**7 holds.*

3.B.1 Study of FALD

Remark on the theoretical analysis of Deng et al. (2021)

FALD has been proposed in Deng et al. (2021), the authors develop an MCMC algorithm targeting the distribution proportional to $\exp(-n^{-1}\sum_{i=1}^n U^i)$ and also establish non-asymptotic bounds. They introduce (Deng et al., 2021, Lemma B.2) the stochastic processes $\{(\bar{\theta}_t^i)_{t\geq 0}\}_{i\in[n]}$ satisfying the Langevin stochastic differential equations for $t \geq 0$, $\mathrm{d}\bar{\theta}_t^i = -\nabla U^i(\bar{\theta}_t^i) + \sqrt{2b}\,\mathrm{dW}_t^i$ where $\{(\mathsf{W}_t^i)_{t\geq 0}\}_{i\in[n]}$ are independent d-dimensional standard Brownian motion and define $\bar{\theta}_t = n^{-1}\sum_{i=1}^n \bar{\theta}_t^i$. Then, it is asserted (Deng et al., 2021, Lemma B.5) that $(\bar{\theta}_t)$ is solution of the Langevin stochastic differential equation $\mathrm{d}\bar{\theta}_t = -n^{-1}\sum_{i=1}^n \nabla U^i(\bar{\theta}_t) + \sqrt{2}\,\mathrm{dW}_t$, where $\mathsf{W}_t = n^{-1/2}\sum_{i=1}^n \mathsf{W}_t^i$. However, this statement cannot hold in all generalities, and we give a counter-example. For instance, consider the Gaussian potentials $\{U^i : x \in \mathbb{R}^d \mapsto \Sigma_i^{-1}(x - \mathrm{m}^i)\}_{i\in[n]}$ where $\{(\mathrm{m}^i, \Sigma_i)\}_{i\in[n]}$ are the mean and the covariance parameters; if for $i \in [n]$, $\bar{\theta}_0^i$ is distributed according to $\exp(-U^i)$, then $n^{-1}\sum_{i=1}^n \bar{\theta}_t^i$ follows $\mathcal{N}(n^{-1}\sum_{i=1}^n \mathrm{m}^i, n^{-2}\sum_{i=1}^n \Sigma_i)$ whereas $\exp(-n^{-1}\sum_{i=1}^n U^i)$ corresponds to the density of the Gaussian $\mathcal{N}(\sum_{i=1}^n (\bar{\Sigma}\Sigma_i^{-1})\mathrm{m}^i, b\bar{\Sigma})$ where $\bar{\Sigma} = (\sum_{i=1}^n \Sigma_i^{-1})^{-1}$. Therefore, for any $t \geq 0$, in this case $\bar{\theta}_t$ is distributed according to $\mathcal{N}(n^{-1}\sum_{i=1}^n \mathrm{m}^i, n^{-2}\sum_{i=1}^n \Sigma_i)$ and thus cannot be distributed according to $\exp(-n^{-1}\sum_{i=1}^n U^i)$ as crucially used in the proof of Deng et al. (2021, Lemma B.5).

Theoretical analysis

In this section, we prove the first theoretical guarantee on FALD stated in Theorem 3.22. Similar to McMahan et al. (2017), the clients update their local parameters $\{X_k^i\}_{i\in[n]}$ several times before transmitting them to the server with probability $p_c \in (0, 1]$. Then, the server aggregates the local parameters to update its own parameter X_k as in (3.23). For all $i \in [n], k \in \mathbb{N}$, consider the stochastic gradients defined by

$$G_k^i = H^i(X_k^i, \xi_{k+1}^i), \tag{3.105}$$

$$\bar{G}_k^i = \nabla U^i(X_k^i). \tag{3.106}$$

Lemma 3.18. *Assume **A**1, **A**4 and **A**6 hold. Then for any $k \in \mathbb{N}$, we have*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\bar{G}_k^i\|^2\right] \leq 3L^2\mathbb{E}\left[V_k\right] + 3L^2\mathbb{E}\left[d_k^2\right] + \frac{3}{n}\sum_{i=1}^n \|\nabla U^i(x_\star)\|^2,$$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|G_k^i - \bar{G}_k^i\|^2\right] \leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\tilde{\omega}\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right].$$

For any $i \in [n], k \in \mathbb{N}$, recall the stochastic gradients $G_k^i, \bar{G}_k^i$ are defined in (3.105) and (3.106), respectively

Proof Using the Young inequality combined with the Lipschitz property **A**1 of the gradients $(U^i)_i^n$, for $k \geq 0$ we get

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\bar{G}_k^i\|^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla U^i(X_k^i) - \nabla U^i(X_k) + \nabla U^i(X_k) - \nabla U^i(x_\star) + \nabla U^i(x_\star)\|^2\right]$$

Algorithm 3.3 Stochastic Averaging Langevin Dynamics - FALD

Input: initial vectors $(X_0^i)_{i\in[n]}$, noise parameter $\tau \in [0,1]$, number of communication rounds K, probability p_c of communication, step-size γ.
for $k = 0$ **to** $K-1$ **do**
  // On each client
  Draw $B_{k+1} \sim \mathcal{B}(p_c), \tilde{Z}_{k+1} \sim \mathcal{N}(0_d, \mathrm{I}_d)$
  // In parallel on the n clients
  for $i = 1$ **to** n **do**
    Draw $\xi_{k+1}^i \sim \nu_\xi$ and $\tilde{Z}_{k+1}^i \sim \mathcal{N}(0_d, \mathrm{I}_d)$
    Compute $G_k^i = H^i(X_k^i, \xi_{k+1}^i)$
    Set $\tilde{X}_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\,(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i)$
    if $B_{k+1} = 1$ **then**
      Broadcast $\tilde{X}_{k+1}^i$ to the server
    else
      Update $X_{k+1}^i \leftarrow \tilde{X}_{k+1}^i$
  if $B_{k+1} = 1$ **then**
    // On the central server
    Update then broadcast the global parameter $X_{k+1} = \frac{1}{n}\sum_{i=1}^n \tilde{X}_{k+1}^i$
    // On each client
    Update the local parameter $X_{k+1}^i \leftarrow X_{k+1}$

Output: samples $\{X_\ell\}_{\{\ell\in[K]\,:\,B_\ell=1\}}$.

$$\leq 3L^2\mathbb{E}\left[V_k\right] + 3L^2\mathbb{E}\left[d_k^2\right] + \frac{3}{n}\sum_{i=1}^n \|\nabla U^i(x_\star)\|^2.$$

In addition, since the random variables $(G_k^i - \bar{G}_k^i)_{i=1}^n$ are centered and independent, the Young and the Jensen inequality imply that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(G_k^i - \bar{G}_k^i\right)\right\|^2\right]$$

$$= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i) - \bar{H}(X_k, \xi_{k+1}) + \bar{H}(X_k, \xi_{k+1}) - \bar{H}(x_\star, \xi_{k+1})\right.\right.$$
$$\left.\left. + \bar{H}(x_\star, \xi_{k+1}) - \nabla\bar{U}(X_k) + \nabla\bar{U}(X_k) - \frac{1}{n}\sum_{i=1}^n \nabla U^i(X_k^i)\right\|^2\right]$$

$$\leq 3\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i) - \frac{1}{n}\sum_{i=1}^n \nabla U^i(X_k^i) - \bar{H}(X_k, \xi_{k+1}) + \nabla\bar{U}(X_k)\right\|^2\right]$$
$$+ 3\mathbb{E}\left[\left\|\bar{H}(X_k, \xi_{k+1}) - \nabla\bar{U}(X_k) - \bar{H}(x_\star, \xi_{k+1})\right\|^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right]$$

$$\leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\tilde{\omega}\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right].$$

■

Lemma 3.19. *Assume **A**1 and **A**4 hold. Then, for any $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, we have*

$$\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1 - \frac{\gamma m}{2}\right)\mathbb{E}\left[d_k^2\right] + \frac{2\gamma L^2}{m}\mathbb{E}\left[V_k\right] + 3\gamma^2\mathbb{E}\left[\|\bar{H}(x_\star, \xi)\|^2\right] + \frac{2\gamma d}{n},$$

where V_k, d_k are defined in (3.24) and (3.25).

Proof Let k be in $\mathbb{N}$. Rewriting the expression of X_{k+1} defined in (3.23), we obtain

$$\begin{aligned}
\mathbb{E}\left[d_{k+1}^2\right] &= \mathbb{E}\left[\left\|X_{k+1} - x_\star\right\|^2\right] \\
&= \mathbb{E}\left[\left\|X_k - x_\star - \frac{\gamma}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i) + \sqrt{2\gamma}\left(\sqrt{\frac{\tau}{n}}\,\tilde{Z}_{k+1} + \frac{\sqrt{1-\tau}}{n}\sum_{i=1}^n Z_{k+1}^i\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|X_k - x_\star\right\|^2\right] - 2\gamma\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i)\right\rangle\right] \\
&\quad + \gamma^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i)\right\|^2\right] + \frac{2\gamma d}{n}. \qquad (3.107)
\end{aligned}$$

Further, the Young inequality combined with **A**4 give

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i)\right\|^2\right] &\leq \frac{3}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|H^i(X_k^i, \xi_{k+1}^i) - H^i(X_k, \xi_{k+1}^i)\right\|^2\right] \\
&\quad + 3\mathbb{E}\left[\|\bar{H}(x_\star, \xi)\|^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(X_k, \xi_{k+1}) - \bar{H}(x_\star, \xi)\right\|^2\right] \\
&\leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\hat{L}^2\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star, \xi)\|^2\right]. \qquad (3.108)
\end{aligned}$$

In addition, using the fact that for any vectors $a, b \in \mathbb{R}^d$, $2|\langle a, b\rangle| \leq m\|a\|^2 + \|b\|^2/m$ we can upper bound the inner product derived in (3.107) as follows

$$\begin{aligned}
&-\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i)\right\rangle\right] = -\mathbb{E}\left[\left\langle X_k - x_\star, \nabla\bar{U}(X_k)\right\rangle\right] \\
&+\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{n}\sum_{i=1}^n \left[H^i(X_k, \xi_{k+1}^i) - H^i(X_k^i, \xi_{k+1}^i)\right]\right\rangle\right] \\
&\leq -\mathbb{E}\left[\left\langle X_k - x_\star, \nabla\bar{U}(X_k)\right\rangle\right] + m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m) \\
&\leq -m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m). \qquad (3.109)
\end{aligned}$$

Therefore, plugging (3.108) and (3.109) in (3.107) shows

$$\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1-\gamma\left[m-3\gamma\hat{L}^2\right]\right)\mathbb{E}\left[d_k^2\right] + \gamma\left(3\gamma\hat{L}^2+\frac{L^2}{m}\right)\mathbb{E}\left[V_k\right] + 3\gamma^2\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + \frac{2\gamma d}{n}.$$

Eventually, the assumption $\gamma \leq m(6\hat{L}^2)^{-1}$ completes the proof. ■

For any $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, under **A**1, **A**4 and **A**6 using Lemma 3.18 and Lemma 3.19 we have shown that **HX**3 and **HX**4 hold with the following quantities

$$\begin{aligned}
&A = 3\hat{L}^2, \quad B = 3\tilde{\omega}, \quad C = 0, \quad D = 3\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right],\\
&\bar{A} = 3L^2, \quad \bar{B} = 3L^2, \quad \bar{C} = 0, \quad \bar{D} = (3/n)\textstyle\sum_{i=1}^n \|\nabla U^i(x_\star)\|^2,\\
&A_d = \gamma m/2, \quad B_d = 0, \quad C_d = 2\gamma L^2/m, \quad D_d = 3\gamma^2\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + 2\gamma d/n,\\
&A_\sigma = 1, \quad B_\sigma = 0, \quad C_\sigma = 0, \quad D_\sigma = 0.
\end{aligned} \tag{3.110}$$

For any $\gamma > 0$, consider the following variables

$$\begin{aligned}
&\mathrm{C}^\gamma = \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}-4A_d}\left(B+\frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{B}\right), \quad \mathrm{C}_r^\gamma = 3\mathrm{C}^\gamma C_d, \quad \mathrm{C}_V^\gamma = 1+2C_d\mathrm{C}^\gamma,\\
&\mathrm{C}_\epsilon^\gamma = \mathrm{C}_V^\gamma\mathbb{E}\left[V_0\right] + 7\mathrm{C}^\gamma\mathbb{E}\left[d_0^2\right] + 2D_d,\\
&\mathrm{C}_\delta^\gamma = \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left(D+\frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{D}\right) + \frac{\mathrm{C}^\gamma D_d}{A_d} + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}}.
\end{aligned} \tag{3.111}$$

We also introduce γ_1 and I_γ, which are defined for any $\gamma > 0$ by

$$\begin{aligned}
\gamma_1 &= \frac{p_\mathrm{c}^{1/2}}{(2-2p_\mathrm{c})^{1/2}\left[A+(1+2/p_\mathrm{c})\bar{A}\right]^{1/2}} \wedge \frac{m}{6\hat{L}^2} \wedge \frac{p_\mathrm{c}}{2m} \wedge \frac{q_\mathrm{c}}{m},\\
I_\gamma &= \left\{\gamma \in (0,\gamma_1) \,:\, \gamma m \geq 8\mathrm{C}_r^\gamma\right\}.
\end{aligned} \tag{3.112}$$

Based on Lemma 3.13, we derive the following result.

Lemma 3.20. *Assume **A**1, **A**4 and **A**6 hold. Then, for any $\gamma \in I_\gamma$ and $k \geq 1$, we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1-\frac{A_d}{4}\right)^k\left(2\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_\delta^\gamma\mathrm{C}_r^\gamma}{A_d}\right) + \mathrm{C}_\delta^\gamma, \tag{3.113}$$

where V_k is defined in (3.24) and $\mathrm{C}_\epsilon^\gamma, \mathrm{C}_r^\gamma, \mathrm{C}_\delta^\gamma$ in (3.111).

Proof For any $\gamma \in I_\gamma$, we have $4\mathrm{C}_r^\gamma \leq A_d$ and moreover it is easy to check that $A_d < \min(A_\sigma/2, p_\mathrm{c}/4)$, $A_dA_\sigma \geq 8B_dB_\sigma = 0$. In addition, since **A**1, **A**4 and **A**6 are satisfied we can apply Lemma 3.18 and Lemma 3.19 which show that **HX**3, **HX**4 hold with the variables introduced in (3.110). Therefore, we can use Lemma 3.13 to complete

the proof. ■

Based on the results presented in this section, we can rewrite the upper bound on $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ given in Lemma 3.20 into the format of Assumption 3.9. We consider for $\gamma > 0$,

$$v_1 = 2\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_\delta^\gamma \mathrm{C}_r^\gamma}{A_d}, \qquad v_2 = \mathrm{C}_\delta^\gamma. \tag{3.114}$$

Lemma 3.21. *Assume* ***A***1, *Assumption 3.4,* ***A***4 *hold and let* $\gamma \le 2(3L)^{-1}$. *Then for any* $k \in \mathbb{N}$, *we have*

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] &\le \left[1 - \gamma m\left(1 - 3\gamma L\right) + 3\gamma^2\hat{L}^2\right]\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] \\ &+ \gamma\left(\frac{2L^2}{m} + 3\gamma(L^2 + \hat{L}^2)\right)\mathbb{E}\left[V_k\right] + \left(\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right]\right) \\ &+ \frac{3\gamma^2}{n^2}\int_{\mathbb{R}^d}\mathrm{Var}^{\mathcal{F}_0}\left(H(x,\xi)\right)\pi(\mathrm{d}x).\end{aligned}$$

Proof For any $k \in \mathbb{N}$, recall that $\mathcal{F}_k$ is defined in (3.22) and using Proposition 3.5 we obtain

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] &\le \left[1 - \gamma m\left(1 - 3\gamma L\right)\right]\|\mathsf{X}_{k\gamma} - X_k\|^2 + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2\right)V_k \\ &+ \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right]\right) + \gamma^2\,\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right). \end{aligned} \tag{3.115}$$

Since the stochastic gradients $(H^i(\cdot, \xi_{k+1}^i))_{i=1}^n$ are unbiased, **A**4 with the Young inequality imply that

$$\begin{aligned}&\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right) = \mathbb{E}^{\mathcal{F}_k}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left[H^i(X_k^i, \xi_{k+1}^i) - \nabla U^i(X_k^i)\right]\right\|^2\right] \\ &= \mathbb{E}^{\mathcal{F}_k}\left[\left\|\frac{1}{n}\sum_{i=1}^n H^i(X_k^i, \xi_{k+1}^i) - \bar{H}(X_k, \xi_{k+1}) - \frac{1}{n}\sum_{i=1}^n \nabla U^i(X_k^i) + \nabla\bar{U}(X_k)\right.\right. \\ &\left.\left. + \bar{H}(X_k, \xi_{k+1}) - \bar{H}(\mathsf{X}_{k\gamma}, \xi_{k+1}) - \nabla\bar{U}(X_k) + \nabla\bar{U}(\mathsf{X}_{k\gamma}) + \bar{H}(\mathsf{X}_{k\gamma}^i, \xi_{k+1}) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right\|^2\right] \\ &\le 3\hat{L}^2 V_k + 3\hat{L}^2\|X_k - \mathsf{X}_{k\gamma}\|^2 + 3\,\mathrm{Var}^{\mathcal{F}_k}\left(\bar{H}(\mathsf{X}_{k\gamma}, \xi_{k+1})\right).\end{aligned}$$

Taking the expectation and using that $\mathsf{X}_{k\gamma}$ has distribution π combined with (3.115) complete the proof. ■

For notational convenience, we also introduce the time step-size γ_2 defined by

$$\gamma_2 = \frac{p_\mathrm{c}}{4m} \wedge \frac{1}{6(L + \hat{L}^2/m)} \wedge \frac{p_\mathrm{c} m}{38(1-p_\mathrm{c})^{1/2}\left(p_\mathrm{c}\tilde{\omega} + 3L^2\right)^{1/2} L}.$$

Theorem 3.22. *Assume **A**1, **A**4 and **A**6 hold and let $\gamma \in (0, \gamma_1 \wedge \gamma_2)$. Then, for any initial probability measure $\mu_0^{(\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, we have*

$$W_2^2\left(\mu_k^{(\gamma)}, \pi\right) \leq \left(1 - \frac{\gamma m}{2}\right)^k W_2^2\left(\mu_0^{(\gamma)}, \pi\right) + \frac{8L^2}{m^2} v_1 \left(1 - \frac{\gamma m}{8}\right)^k + \frac{6L^2}{m^2} v_2 + \frac{6\gamma d}{nm^2}\kappa_I + \frac{6\gamma}{n^2 m} \int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}(H(x, \xi_1))\pi(\mathrm{d}x),$$

*where v_1, v_2 are defined in (3.114) and $\kappa_I = L^2(1 + \gamma L^2/m)$. If in addition we suppose **HX**1, set $\kappa_I = 2\gamma(L^3 + d\tilde{L}^2/n)$.*

Proof We know that Assumption 3.4 is satisfied since for any $i \in [n], x \in \mathbb{R}^d$ the stochastic gradient $H^i(x, \xi_1^i)$ is unbiased. The constraint $\gamma \leq \gamma_1$ combined with Lemma 3.19 implies **HX**3 and plugging the expression of $A_d, A_\sigma, B_d, C, \bar{C}, C_d, C_\sigma$ provided in (3.110) into C_r^γ defined in (3.111) gives that

$$\mathrm{C}_r^\gamma = \frac{72\gamma^3(1 - p_\mathrm{c})L^2(\tilde{\omega} + (1 + 2/p_\mathrm{c})L^2)}{(p_\mathrm{c} - 2\gamma m)m}.$$

For any $\gamma \in \left(0, \gamma_2\right]$, we have $(p_\mathrm{c} - 2\gamma m)m^2 \geq 576(1 - p_\mathrm{c})\gamma^2 L^2(\tilde{\omega} + (1 + 2/p_\mathrm{c})L^2)$ which shows that $\gamma \in I_\gamma$. Thus, we can apply Lemma 3.20 which proves that Assumption 3.9 holds with $q_\mathrm{c} = \gamma m$ and $\alpha_v = 1 - A_d/4$ and v_1, v_2 defined in (3.114). Since the assumptions of Lemma 3.21 are satisfied, **HX**2 holds, and therefore we can apply Theorem 3.10 with

$$(1 - q_\mathrm{c})\alpha_0 = 1 - \gamma m\left(1 - 3\gamma L\right) + 3\gamma^2 \hat{L}^2, \quad \alpha_1 = 0,$$
$$(1 - q_\mathrm{c})\alpha_2 = \gamma\left(\frac{2L^2}{m} + 3\gamma(L^2 + \hat{L}^2)\right), \quad \alpha_3 = 0,$$
$$(1 - q_\mathrm{c})\alpha_4 = \left(\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right]\right) + \frac{3\gamma^2}{n^2}\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}\left(H(x, \xi_1)\right)\pi(\mathrm{d}x).$$

Furthermore, using Lemma 3.8 we have

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right] \leq \frac{3\gamma^2 dL^2}{nm}\left(1 + \frac{19\gamma L^2}{36m}\right). \tag{3.116}$$

Moreover, if we suppose **HX**1, we obtain

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right] \leq \frac{\gamma^3 d}{nm}\left(5L^3 + \frac{4d\tilde{L}^2}{3n}\right). \tag{3.117}$$

Finally, with the notation of Theorem 3.10 we obtain $1 + \delta = 0$, and using $\gamma \leq (6(L + m^{-1}\hat{L}^2))^{-1}$ combined with (3.116) or (3.117) if we suppose **HX**1 give the expected result. ■

Now, consider the time step-sizes γ_3 and $\gamma_\star$ defined by

$$\gamma_3 = \frac{p_\mathrm{c} m}{3L^2 + p_\mathrm{c}\tilde{\omega}}, \qquad \gamma_\star = \gamma_1 \wedge \gamma_2 \wedge \gamma_3. \tag{3.118}$$

From the previous result, the next corollary controls the asymptotic bias obtained by Algorithm 3.3.

Corollary 3.23. *Assume* ***A***1, ***A***4 *and* ***A***6 *hold and let* $\gamma \in (0, \gamma_\star)$, $\tau = 1$. *Then, for any initial probability measure* $\mu_0^{(\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, *we have*

$$\frac{6^{-4}n}{\gamma d} \limsup_{k\to\infty} W_2^2\left(\mu_k^{(\gamma)}, \pi\right) \leq \frac{\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}(H\left(x, \xi_1\right))\pi(\mathrm{d}x)}{ndm} + \frac{\tilde{\kappa}_I}{m^2} + \frac{(1-p_\mathrm{c})\gamma L^2}{p_\mathrm{c}^2 m^2}\left(\frac{1}{d}\sum_{i=1}^n \left\|\nabla U^i(x_\star)\right\|^2 + \frac{p_\mathrm{c}}{nd}\mathbb{E}\left[\left\|H(x_\star, \xi)\right\|^2\right] + \frac{L^2 + p_\mathrm{c}\tilde{\omega}}{m}\right),$$

where $\tilde{\kappa}_I = L^2$ *and if we suppose* ***HX***1, $\tilde{\kappa}_I = \gamma(L^3 + d\tilde{L}^2/n)$.

Proof Using Theorem 3.22 combined with $\gamma \leq \gamma_1 \wedge \gamma_2$ gives that

$$\limsup_{k\to\infty} W_2^2\left(\mu_k^{(\gamma)}, \pi\right) \leq \frac{6\gamma}{n^2 m}\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}(H\left(x, \xi_1\right))\pi(\mathrm{d}x) + \frac{6\gamma d}{nm^2}\kappa_I + \frac{6L^2}{m^2}v_2. \quad (3.119)$$

Further, recall that $A_d, B, \bar{B}, D, \bar{D}, D_d$ are provided in (3.110) and C_δ^γ is defined in (3.111) by

$$\begin{aligned}
\mathrm{C}_\delta^\gamma &= \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left(D + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{D}\right) + \frac{\mathrm{C}^\gamma D_d}{A_d} + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}} \\
&\leq \frac{12(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left[1 + \frac{12\gamma}{m}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right)\right]\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}} \\
&\quad + \frac{36(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}^2 n}\sum_{i=1}^n \left\|\nabla U^i(x_\star)\right\|^2 + \frac{96(1-p_\mathrm{c})\gamma^2 d}{p_\mathrm{c} nm}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right) \\
&\leq \frac{156(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{36(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}^2 n}\sum_{i=1}^n \left\|\nabla U^i(x_\star)\right\|^2 \\
&\quad + \frac{96(1-p_\mathrm{c})\gamma^2 d}{p_\mathrm{c} nm}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right) + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}}. \quad (3.120)
\end{aligned}$$

Finally, setting $\tau = 1$ combined with (3.119) and (3.120) show that

$$\limsup_{k\to\infty} W_2^2\left(\mu_k^{(\gamma)}, \pi\right) \leq \frac{6\gamma}{n^2 m}\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}(H\left(x, \xi_1\right))\pi(\mathrm{d}x) + \frac{6\gamma d}{nm^2}\kappa_I + \frac{8(1-p_\mathrm{c})\gamma^2 L^2}{np_\mathrm{c}m^2}\left[\frac{156}{n}\mathbb{E}\left[\left\|H(x_\star, \xi)\right\|^2\right] + \frac{36}{p_\mathrm{c}}\sum_{i=1}^n \left\|\nabla U^i(x_\star)\right\|^2 + \frac{96d}{m}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right)\right].$$

■

3.B.2 Study of VR-FALD⋆

In this alternative of FALD derived in Section 3.B.1, we introduce control variates to cope with both heterogeneity and variance in local gradients. Instead of using $H^i(X_k^i)$ to update the local parameter X_k^i, this time the ith client uses the proxy $H^i(X_k^i, \xi_{k+1}^i) -$

$H^i(Y_k, \xi^i_{k+1}) + \nabla U^i(Y_k)$ based on an analog of the SVRG algorithm (Johnson and Zhang, 2013; Karimireddy et al., 2020) and where Y_k is a global reference point updated with probability $q_c \in (0, 1]$. We derive an explicit upper bound on the Wasserstein distance between the distribution of the server parameter $\mathsf{X}_{k\gamma}$ and the target distribution π. We also show how this new global control variate mitigates the effect of heterogeneity in the convergence rate. To do so, we consider the stochastic gradients defined for any $i \in [n], k \in \mathbb{N}$, by

$$G^i_k = H^i(X^i_k, \xi^i_{k+1}) - H^i(Y_k, \xi^i_{k+1}) + C_k, \tag{3.121}$$

$$\bar{G}^i_k = \nabla U^i(X^i_k) - \nabla U^i(Y_k) + C_k \tag{3.122}$$

and denote

$$\sigma_k = \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}^{\mathcal{F}_k} \left[\left\| H^i(Y_k, \xi^i_{k+1}) - H^i(x_\star, \xi^i_{k+1}) \right\|^2 \right] \right)^{1/2} . \tag{3.123}$$

Lemma 3.24. *Assume **A**1, **A**4 and **A**6 hold. Then for any $k \in \mathbb{N}$, we have*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\bar{G}^i_k\|^2 \right] \le 3L^2 \mathbb{E} \left[V_k \right] + 3L^2 \mathbb{E} \left[d_k^2 \right] + 3\mathbb{E} \left[\sigma_k^2 \right],$$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|G^i_k - \bar{G}^i_k\|^2 \right] \le 3\hat{L}^2 \mathbb{E} \left[V_k \right] + 3\tilde{\omega} \mathbb{E} \left[d_k^2 \right] + 3\mathbb{E} \left[\sigma_k^2 \right].$$

For any $i \in [n], k \in \mathbb{N}$, recall the stochastic gradients $G^i_k, \bar{G}^i_k$ are defined in (3.121) and (3.122), respectively

Proof For $k \ge 0$, Lipschitz property of $\{\nabla U^i\}_{i \in [n]}$ supposed in **A**1 gives that

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\bar{G}^i_k\|^2 \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla U^i(X^i_k) - \nabla U^i(Y_k) + \nabla \bar{U}(Y_k)\|^2 \right] \\
&\le \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla U^i(X^i_k) - \nabla U^i(X_k)\|^2 \right] + \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla U^i(Y_k) - \nabla U^i(x_\star)\|^2 \right] \\
&\qquad + \frac{3}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla U^i(X_k) - \nabla U^i(x_\star)\|^2 \right] \\
&\le 3L^2 \mathbb{E} \left[V_k \right] + 3L^2 \mathbb{E} \left[d_k^2 \right] + 3\mathbb{E} \left[\sigma_k^2 \right]
\end{aligned}$$

and the proof is concluded by noting that **A**4 gives

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \mathbb{E} \|G^i_k - \bar{G}^i_k\|^2 = \mathbb{E} \left[\mathrm{Var}^{\mathcal{F}_k} \left(\frac{1}{n} \sum_{i=1}^n G^i_k \right) \right] \\
&\le \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n H^i(X^i_k, \xi^i_{k+1}) - \bar{H}(X_k, \xi_{k+1}) \right\|^2 \right]
\end{aligned}$$

Algorithm 3.4 VR-FALD*

Input: initial vectors $(X_0^i)_{i\in[n]}$, noise parameter $\tau \in \big[0,1\big]$, number of communication rounds K, probability p_{c} of communication, probability q_{c} to update the control variates, step-size γ and batch size r.
Initialize $Y_0 = (1/n)\sum_{i=1}^n X_0^i$ and $C_0 = (1/n)\nabla U(Y_0)$
for $k = 0$ **to** $K-1$ **do**
  // On each client
  Draw $B_{k+1} \sim \mathcal{B}(p_{\mathrm{c}}), \tilde{Z}_{k+1} \sim \mathcal{N}(0_d, \mathrm{I}_d)$
  // In parallel on the n clients
  for $i = 1$ **to** n **do**
    Draw $\xi_{k+1}^i \sim \nu_\xi$, $\tilde{Z}_{k+1}^i \sim \mathcal{N}(0_d, \mathrm{I}_d)$
    Compute $G_k^i = H^i(X_k^i, \xi_{k+1}^i) - H^i(Y_k, \xi_{k+1}^i) + C_k$
    Set $\tilde{X}_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\,(\sqrt{\tau/n}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i)$
    if $B_{k+1} = 1$ **then**
      Broadcast $\tilde{X}_{k+1}^i$ to the server
    else
      Update $X_{k+1}^i \leftarrow \tilde{X}_{k+1}^i$
    if $\tilde{B}_{k+1} = 1$ **then**
      Broadcast X_k^i to the server
    else
      Update $Y_{k+1} \leftarrow Y_k$ and $C_{k+1} \leftarrow C_k$
  if $B_{k+1} = 1$ **then**
    // On the central server
    Update then broadcast the global parameter $X_{k+1} \leftarrow (1/n)\sum_{i=1}^n \tilde{X}_{k+1}^i$
    // On each client
    Update the local parameter $X_{k+1}^i \leftarrow X_{k+1}$
  if $\tilde{B}_{k+1} = 1$ **then**
    // On the central server
    Update then broadcast $Y_{k+1} \leftarrow (1/n)\sum_{i=1}^n X_k^i$
    // On each client
    Compute and broadcast $\nabla U^i(Y_{k+1})$
    // On the central server
    Update then broadcast $C_{k+1} \leftarrow (1/n)\nabla U(Y_{k+1})$

Output: samples $\{X_\ell\}_{\{\ell\in[K]\,:\,B_\ell=1\}}$.

$$\leq 3\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} H^i(X_k^i,\xi_{k+1}^i) - \bar{H}(X_k,\xi_{k+1})\right\|^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1}) - \bar{H}(x_\star,\xi_{k+1})\right\|^2\right]$$
$$+ 3\mathbb{E}\left[\left\|\bar{H}(X_k,\xi_{k+1}) - \bar{H}(x_\star,\xi_{k+1}) - \nabla\bar{U}(X_k)\right\|^2\right].$$

■

Lemma 3.25. *Assume **A**1 and **A**4 hold. Then, for any $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, we have*

$$\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1 - \frac{\gamma m}{2}\right)\mathbb{E}\left[d_k^2\right] + \frac{2\gamma L^2}{m}\mathbb{E}\left[V_k\right] + 4\gamma^2\mathbb{E}\left[\sigma_k^2\right] + 10\gamma^2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] + \frac{2\gamma d}{n},$$

where V_k, d_k, σ_k are defined in (3.24), (3.25) and (3.123).

Proof Let k be in $\mathbb{N}$. Writing the expression of X_{k+1} defined in (3.23) and developing the expectation of the squared norm give

$$\mathbb{E}\left[d_{k+1}^2\right] = \mathbb{E}\left[\left\|X_{k+1} - x_\star\right\|^2\right]$$
$$= \mathbb{E}\left[\left\|X_k - x_\star - \frac{\gamma}{n}\sum_{i=1}^{n} H^i(X_k^i,\xi_{k+1}^i) + \gamma\bar{H}(Y_k,\xi_{k+1}) - \gamma\nabla\bar{U}(Y_k)\right.\right.$$
$$\left.\left. + \sqrt{2\gamma}\left(\sqrt{\frac{\tau}{n}}\,\tilde{Z}_{k+1} + \frac{\sqrt{1-\tau}}{n}\sum_{i=1}^{n} Z_{k+1}^i\right)\right\|^2\right]$$
$$= \mathbb{E}\left[\left\|X_k - x_\star\right\|^2\right] - 2\gamma\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{n}\sum_{i=1}^{n} H^i(X_k^i,\xi_{k+1}^i)\right\rangle\right]$$
$$+ \gamma^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} H^i(X_k^i,\xi_{k+1}^i)\right\|^2\right] + \gamma^2\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1}) - \nabla\bar{U}(Y_k)\right\|^2\right]$$
$$- 2\gamma^2\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^{n} H^i(X_k^i,\xi_{k+1}^i), \bar{H}(Y_k,\xi_{k+1}) - \gamma\nabla\bar{U}(Y_k)\right\rangle\right] + \frac{2\gamma d}{n}$$
$$= \mathbb{E}\left[d_k^2\right] - 2\gamma\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{n}\sum_{i=1}^{n}\nabla U^i(X_k^i)\right\rangle\right] + 2\gamma^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} H^i(X_k^i,\xi_{k+1}^i)\right\|^2\right]$$
$$+ 2\gamma^2\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1}) - \nabla\bar{U}(Y_k)\right\|^2\right] + \frac{2\gamma d}{n}. \tag{3.124}$$

Using the Young inequality combined with **A**4 show

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}H^i(X_k^i,\xi_{k+1}^i)\right\|^2\right] \leq \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|H^i(X_k^i,\xi_{k+1}^i)-H^i(X_k,\xi_{k+1}^i)\right\|^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(X_k,\xi_{k+1})-\bar{H}(x_\star,\xi)\right\|^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\hat{L}^2\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right]. \quad (3.125)$$

We also have that

$$\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1})-\nabla\bar{U}(Y_k)\right\|^2\right] \leq 2\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1})-\bar{H}(x_\star,\xi_{k+1})\right\|^2\right] + 2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \leq 2\mathbb{E}\left[\sigma_k^2\right] + 2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right]. \quad (3.126)$$

In addition, using the fact that for any vectors $a,b\in\mathbb{R}^d$, $2\left|\langle a,b\rangle\right| \leq m\left\|a\right\|^2 + \left\|n\right\|^2/m$, we can upper bound the inner product derived in (3.124) as follows

$$\begin{aligned}
&-\mathbb{E}\left[\left\langle X_k-x_\star, \frac{1}{n}\sum_{i=1}^{n}\nabla U^i(X_k^i)\right\rangle\right] = -\mathbb{E}\left[\left\langle X_k-x_\star,\nabla\bar{U}(X_k)\right\rangle\right]\\
&+\mathbb{E}\left[\left\langle X_k-x_\star, \frac{1}{n}\sum_{i=1}^{n}\left[H^i(X_k,\xi_{k+1}^i)-H^i(X_k^i,\xi_{k+1}^i)\right]\right\rangle\right]\\
&\leq -\mathbb{E}\left[\left\langle X_k-x_\star,\nabla\bar{U}(X_k)\right\rangle\right] + m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m)\\
&\leq -m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m). \qquad (3.127)
\end{aligned}$$

Hence, combining (3.124), (3.125), (3.126) and (3.127) implies that

$$\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1-\gamma m+6\gamma^2\hat{L}^2\right)\mathbb{E}\left[d_k^2\right] + \left(\frac{\gamma L^2}{m}+6\gamma^2\hat{L}^2\right)\mathbb{E}\left[V_k\right] + 4\gamma^2\mathbb{E}\left[\sigma_k^2\right] + 10\gamma^2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] + \frac{2\gamma d}{n}.$$

Using the assumption on γ completes the proof. ■

Lemma 3.26. *Assume the L-smoothness of the potentials $\{U^i\}_{i\in[n]}$ and **A**4 hold. Then, for any $k\in\mathbb{N}$, we have*

$$\mathbb{E}\left[\sigma_{k+1}^2\right] \leq (1-q_c)\mathbb{E}\left[\sigma_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[d_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[V_k\right],$$

where V_k, d_k, σ_k are defined in (3.24), (3.25) and (3.123).

Proof Let's consider $k \geq 0$, using **A**4 implies that

$$\begin{aligned}
\mathbb{E}\left[\sigma_{k+1}^2\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|H^i(Y_{k+1}^i, \xi_{k+1}^i) - H^i(x_\star, \xi_{k+1}^i)\right\|^2\right] \\
&= \frac{1-q_{\mathrm{c}}}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|H^i(Y_k^i, \xi_{k+1}^i) - H^i(x_\star, \xi_{k+1}^i)\right\|^2\right] \\
&\quad + \frac{q_{\mathrm{c}}}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|H^i(X_k^i, \xi_{k+1}^i) - H^i(x_\star, \xi_{k+1}^i)\right\|^2\right] \\
&= (1-q_{\mathrm{c}})\mathbb{E}\left[\sigma_k^2\right] + \frac{2q}{n}\sum_{i=1}^n \mathbb{E}\Big[\left\|H^i(X_k^i, \xi_{k+1}^i) - H^i(X_k, \xi_{k+1}^i)\right\|^2 \\
&\quad + \left\|H^i(X_k, \xi_{k+1}^i) - H^i(x_\star, \xi_{k+1}^i)\right\|^2\Big] \\
&\leq (1-q_{\mathrm{c}})\mathbb{E}\left[\sigma_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[d_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[V_k\right].
\end{aligned}$$

Which shows the expected result. ■

For any $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, under **A**1, **A**4 and **A**6 we have shown that Lemma 3.24 and Lemma 3.25 imply **HX**3 and **HX**4 with

$$\begin{aligned}
&A = c_V = 3\hat{L}^2, \quad B = c_d = 3\tilde{\omega}, \quad C = c_\sigma = 3, \quad D = c = 0, \\
&\bar{A} = 3L^2, \quad \bar{B} = 3L^2, \quad \bar{C} = 3, \quad \bar{D} = 0, \\
&A_d = \gamma m/2, \quad B_d = 4\gamma^2, \quad C_d = 2\gamma L^2/m, \quad D_d = (10\gamma^2)\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + 2\gamma d/n, \\
&A_\sigma = q, \quad B_\sigma = 2q\hat{L}^2, \quad C_\sigma = 2q\hat{L}^2, \quad D_\sigma = 0.
\end{aligned} \tag{3.128}$$

For any $\gamma > 0$, consider the following variables

$$\alpha_d = \frac{4\gamma^2}{p_{\mathrm{c}} A_d}\max\left\{p_{\mathrm{c}}B + 3\bar{B}, \frac{4B_\sigma}{A_\sigma}\left(p_{\mathrm{c}}C + 3\bar{C}\right)\right\}, \quad \alpha_\sigma = \frac{4\gamma^2\left(p_{\mathrm{c}}C + 3\bar{C}\right)}{p_{\mathrm{c}}A_\sigma}. \tag{3.129}$$

Lemma 3.27. *Assume **A**1, **A**4 and **A**6 hold with*

$$A_d \leq \min\left(A_\sigma, \frac{p_{\mathrm{c}}}{4}\right), \quad \alpha_d C_d + \alpha_\sigma C_\sigma \leq \frac{p_{\mathrm{c}}}{8}, \quad \alpha_d B_d + \gamma^2\left(C + \frac{3}{p_{\mathrm{c}}}\bar{C}\right) \leq \frac{\alpha_\sigma A_\sigma}{2},$$

and consider $\gamma \leq m(6\hat{L}^2)^{-1} \wedge p_{\mathrm{c}}^{1/2}(2-2p_{\mathrm{c}})^{-1/2}[A + (1+2/p_{\mathrm{c}})\bar{A}]^{-1/2}$. Then, for any $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{2}\right)^k\left(\mathbb{E}\left[V_0\right] + \alpha_d\mathbb{E}\left[d_0^2\right] + \alpha_\sigma\mathbb{E}\left[\sigma_0^2\right]\right) + \frac{2\alpha_d D_d}{A_d} + \frac{4\left(1-\tau\right)(n-1)\gamma d}{nA_d},$$

where V_k is defined in (3.24).

Proof Applying Lemma 3.14 with the variables provided in (3.128) gives the result. ■

Let's introduce $\gamma_1 > 0$ such that

$$\gamma_1 \leq \frac{m}{128\hat{L}^2} \wedge \frac{m}{8\max\left(3L^2 + p_c\tilde{\omega}, 24\hat{L}^2\right)} \wedge \frac{2q}{m} \wedge \frac{p_c}{2m} \wedge \frac{p_c}{\left[2(1-p_c)(p_c A + 3\bar{A})\right]^{1/2}} \wedge \frac{p_c}{8\left[6\left(\frac{L^2}{m^2}\max\left(3L^2 + p_c\tilde{\omega}, 24\hat{L}^2\right)\right) + \frac{2}{q_c}\right]^{1/2}}.$$

Under **A**1, **A**4 and **A**6, for all $\gamma \in \left(0, \gamma_1\right]$ the assumptions of Lemma 3.27 are satisfied. The upper bound on $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ derived in Lemma 3.27 can be rewritten into the format of Assumption 3.9 by considering

$$\tilde{v}_1 = \mathbb{E}\left[V_0\right] + \alpha_d \mathbb{E}\left[d_0^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_0^2\right], \quad \tilde{v}_2 = \frac{2\alpha_d D_d}{A_d} + \frac{4\left(1-\tau\right)(n-1)\gamma d}{nA_d}. \tag{3.130}$$

In addition, for any $\gamma > 0$, consider the following variables

$$\begin{aligned}
\mathrm{C}^\gamma &= \frac{4(1-p_c)\gamma^2}{p_c - 4A_d}\left[B + \frac{2+p_c}{p_c}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_c}{p_c}\bar{C}\right)\right], \\
\mathrm{C}_r^\gamma &= \frac{9\gamma^2\left(1-p_c\right)C_\sigma}{p_c - 4A_d}\left(C + \frac{2+p_c}{p_c}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right), \\
\mathrm{C}_\sigma^\gamma &= \frac{4(1-p_c)\gamma^2}{p_c - 4A_d}\left(C + \frac{2+p_c}{p_c}\bar{C}\right) + \mathrm{C}^\gamma B_d\left(2 + \frac{3}{A_\sigma - A_d}\right), \\
\mathrm{C}_{d_0}^\gamma &= 7\mathrm{C}^\gamma, \qquad \mathrm{C}_V^\gamma = 1 + 2\mathrm{C}^\gamma C_d, \\
\mathrm{C}_\delta^\gamma &= \frac{\mathrm{C}^\gamma D_d}{A_d}\left(1 + \frac{2B_d B_\sigma}{A_d(A_\sigma - A_d)}\right) + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_c}, \\
\mathrm{C}_\epsilon^\gamma &= \mathrm{C}_V^\gamma \mathbb{E}\left[V_0\right] + \mathrm{C}_{d_0}^\gamma \mathbb{E}\left[d_0^2\right] + \mathrm{C}_\sigma^\gamma \mathbb{E}\left[\sigma_0^2\right] + 2D_d.
\end{aligned} \tag{3.131}$$

Based on Lemma 3.12, we derive the following result.

Lemma 3.28. *Assume **A**1, **A**4 and **A**6 hold and consider $\gamma \in \left(0, \gamma_1\right]$. Then, for any $k \in \mathbb{N}$, we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{4}\right)^k \left(\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma \tilde{v}_1}{A_d}\right) + \frac{2\mathrm{C}_r^\gamma \tilde{v}_2}{A_d} + \mathrm{C}_\delta^\gamma,$$

where V_k is defined in (3.24) and $\mathrm{C}_\epsilon^\gamma, \mathrm{C}_r^\gamma, \mathrm{C}_\delta^\gamma$ in (3.131).

Proof Since we suppose **A**1, **A**4 and **A**6 hold with $\gamma \leq \gamma_1$, the assumptions of Lemma 3.27 are satisfied. Therefore, for any $l \in \mathbb{N}$, we obtain

$$\mathbb{E}\left[V_l\right] \leq \left(1 - \frac{A_d}{2}\right)^l \tilde{v}_1 + \tilde{v}_2. \tag{3.132}$$

Moreover, the condition $\gamma \leq m/128\hat{L}^2$ ensures that $A_d A_\sigma = q\gamma m/2 \geq 8B_d B_\sigma = 64q\gamma^2\hat{L}^2$, hence we can apply Lemma 3.12. Then, plugging (3.132) in the bound derived in Lemma 3.12 gives

$$\mathbb{E}\left[V_k\right] \leq \left(1-\alpha\right)^k \mathrm{C}_\epsilon^\gamma + \mathrm{C}_r^\gamma \sum_{i=0}^{k-2}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right] + \mathrm{C}_\delta^\gamma, \tag{3.133}$$

where α is defined in (3.61) by

$$\alpha = A_d - \frac{2(A_\sigma - A_d)^{-1} B_d B_\sigma}{1 + \sqrt{1 + 4(1 - A_d)^{-1}(A_\sigma - A_d)^{-1} B_d B_\sigma}}. \tag{3.134}$$

Using Lemma 3.11, we know that $A_d/2 < \alpha \leq A_d$ and combining this bound with (3.132) and (3.133) leads to

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{4}\right)^k \left(\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma \tilde{v}_1}{A_d}\right) + \frac{2\mathrm{C}_r^\gamma \tilde{v}_2}{A_d} + \mathrm{C}_\delta^\gamma.$$

■

In order to rewrite the upper bound on $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ given in Lemma 3.28 in the format of Assumption 3.9, we consider for $\gamma > 0$

$$v_1 = \mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma \tilde{v}_1}{A_d}, \qquad v_2 = \frac{2\mathrm{C}_r^\gamma \tilde{v}_2}{A_d} + \mathrm{C}_\delta^\gamma. \tag{3.135}$$

Lemma 3.29. *Assume* ***A****1,* ***A****7, Assumption 3.4 and hold and let* $\gamma \leq (6L)^{-1}$. *Using the convention that* $\sum_0^{-1} = 0$, *then for any* $k \in \mathbb{N}$, *we have*

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] &\leq \left[1 - \gamma m + \gamma^2\left(3mL + 4\omega\right)\right]\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] \\ &+ 4\gamma^2\omega q_{\mathrm{c}} \sum_{l=0}^{k-1}(1 - q_{\mathrm{c}})^{k-l-1}\mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2 + 4\gamma\omega\right)\mathbb{E}\left[V_k\right] \\ &+ \left(\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right]\right) + \frac{16\gamma^3\omega d}{n q_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right).\end{aligned}$$

Proof For $k \in \mathbb{N}$, using the independence of $(\xi_{k+1}^i)_{i\in[n]}$ combined with Assumption 3.4 and **A**7, we obtain

$$\begin{aligned}\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right) &= \mathbb{E}^{\mathcal{F}_k}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left[\nabla U^i(X_k^i) - \nabla U^i(Y_k) - H^i(X_k^i, \xi_{k+1}^i) + H^i(Y_k, \xi_{k+1}^i)\right]\right\|^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U^i(X_k^i) - \nabla U^i(Y_k) - H^i(X_k^i, \xi_{k+1}^i) + H^i(Y_k, \xi_{k+1}^i)\right\|^2\right] \\ &\leq \frac{\omega}{n}\sum_{i=1}^n \left\|X_k^i - Y_k\right\|^2. \end{aligned} \tag{3.136}$$

Denote $t_k \in \mathbb{N}$ the time when the reference point of the control variate is updated, therefore we have

$$t_k = \begin{cases} 0, & \text{if } k = 0 \\ \max\left\{l \in \{0, \ldots, k-1\} \,:\, Y_k = n^{-1}\sum_{i=1}^n X_k^i\right\}, & \text{if } k \geq 1 \end{cases}. \tag{3.137}$$

Hence, for any $i \in [n], k \geq 0$, we have

$$X_k^i - Y_k = (X_k^i - X_k) + (X_k - \mathsf{X}_{k\gamma}) + (\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}) + (\mathsf{X}_{t_k\gamma} - Y_k).$$

Thus, for $k \geq 0$, combining the previous line with Young's inequality, it yields that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left\|X_k^i - Y_k\right\|^2\right] \leq 4\mathbb{E}\left[V_k\right] + 4\mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + 4\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}\|^2\right] + 4\mathbb{E}\left[\|\mathsf{X}_{t_k\gamma} - Y_k\|^2\right]. \tag{3.138}$$

For $k \geq 1$, by definition of t_k, we have

$$\mathbb{E}\left[V_{t_k}\right] = \sum_{l=0}^{k-1} \mathbb{P}\left(t_k = l\right) \mathbb{E}\left[V_l\right] = q\sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[V_l\right].$$

Moreover, for $k \geq 1$ we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}\right\|^2\right] &= \sum_{l=0}^{k-1} \mathbb{P}\left(t_k = l\right) \mathbb{E}\left[\left\|\mathsf{X}_{k\gamma} - \mathsf{X}_{l\gamma}\right\|^2\right] \\
&= q\sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\left\|-\int_{l\gamma}^{k\gamma} \nabla \bar{U}(\mathsf{X}_s)\mathrm{d}s + \sqrt{\frac{2}{n}}\left(\mathsf{W}_{k\gamma} - \mathsf{W}_{l\gamma}\right)\right\|^2\right] \\
&\leq 2\gamma q\sum_{l=0}^{k-1} (k-l)(1-q_{\mathrm{c}})^{k-l-1} \left(\int_{l\gamma}^{k\gamma} \mathbb{E}\left[\left\|\nabla \bar{U}(\mathsf{X}_s)\right\|^2\right] \mathrm{d}s + \frac{2d}{n}\right).
\end{aligned} \tag{3.139}$$

Using Dalalyan (2017a, Lemma 2) with $s \in \mathbb{R}_+$, we obtain

$$\mathbb{E}\left[\|\nabla \bar{U}(\mathsf{X}_s)\|^2\right] \leq dL/n.$$

Using by convention that $\sum_{l=1}^{0} = 0$, for any $k \in \mathbb{N}$ and $x \neq 1$ we have

$$\sum_{l=1}^{k} l^2 x^{l-1} = (1-x)^{-3}\left(1 + x - x^k\left[2x + kx(1-x) + (k+1)(1+k(1-x))(1-x)\right]\right).$$

Thus, setting $x = 1 - q$ inside the last shows that

$$\sum_{l=1}^{k} l^2 (1-q_{\mathrm{c}})^{l-1} \leq 2/q_{\mathrm{c}}^3.$$

Hence, the above line combined with $\sum_{l=1}^{k} l(1-q_{\mathrm{c}})^{l-1} = q^{-2}\left[1 - (1+kq)(1-q_{\mathrm{c}})^k\right]$ and (3.139) yield the following upper bound

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}\right\|^2\right] &\leq \frac{2\gamma dq}{n}\sum_{l=0}^{k-1}\left[(k-l)(1-q_{\mathrm{c}})^{k-l-1}\left(2 + (k-l)\gamma L\right)\right] \\
&\leq \frac{2\gamma dq}{n}\sum_{l=0}^{k-1}\left[(k-l)(1-q_{\mathrm{c}})^{k-l-1}\left(2 + (k-l)\gamma L\right)\right]
\end{aligned}$$

$$\leq \frac{4\gamma d}{n q_{\mathrm{c}}} \left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right). \tag{3.140}$$

In addition, by definition (3.137) of t_k, we immediately get for any $k \geq 1$, that

$$\begin{aligned}\mathbb{E}\left[\left\|\mathsf{X}_{t_k\gamma} - X_{t_k}\right\|^2\right] &= \sum_{l=0}^{k-1} \mathbb{P}\left(t_k = l\right) \mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] \\ &= q \sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right].\end{aligned}$$

Combining (3.136), (3.138) with (3.140), for any $k \geq 1$ we obtain

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right)\right] &\leq 4\omega \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] \\ &+ 4\omega q_{\mathrm{c}} \sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] + 4\omega \mathbb{E}\left[V_k\right] + \frac{16\gamma\omega d}{n q_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right). \end{aligned} \tag{3.141}$$

Since $Y_0 = n^{-1}\sum_{i=1} X_0^i$, we have $\mathrm{Var}^{\mathcal{F}_k}(n^{-1}\sum_{i=1}^n G_k^i) \leq \omega V_k$ and therefore the above inequality also holds for $k = 0$. Lastly, using Proposition 3.5 gives

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] &\leq \left[1 - \gamma m\left(1 - 3\gamma L\right)\right] \|\mathsf{X}_{k\gamma} - X_k\|^2 + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2\right) V_k \\ &+ \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right]\right) + \gamma^2 \mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{n}\sum_{i=1}^n G_k^i\right).\end{aligned}$$

Hence, plugging (3.141) in the above inequality yields the expected result. ■

Based on Lemma 3.29, for any $\gamma > 0$ introduce the following notations

$$\alpha_0 = \left(1 - q_{\mathrm{c}}\right)^{-1}\left[1 - \gamma m + \gamma^2\left(3mL + 4\omega\right)\right], \qquad \alpha_1 = \frac{4\gamma^2\omega q}{\left(1 - q_{\mathrm{c}}\right)^2}, \tag{3.142}$$

$$\alpha_2 = \frac{\gamma}{1 - q_{\mathrm{c}}}\left(\frac{2L^2}{m} + 3\gamma L^2 + 4\gamma\omega\right), \qquad \alpha_3 = 0,$$

$$\alpha_4 = (1-q_{\mathrm{c}})^{-1}\left(\frac{2\sup_{l\in\mathbb{N}}}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_l}\left[I_l\right]\right\|^2\right] + 3\sup_{l\in\mathbb{N}}\mathbb{E}\left[\left\|I_l\right\|^2\right] + \frac{16\gamma^3\omega d}{n q_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right)\right).$$

For ease of reading, we also introduce the time step-size γ_2 defined by

$$\gamma_2 \leq \frac{q_{\mathrm{c}}}{L} \wedge \frac{q_{\mathrm{c}}}{2m} \wedge \frac{1}{6(L + 4m^{-1}\omega)}. \tag{3.143}$$

Theorem 3.30. *Assume* ***A***1, ***A***4, ***A***6, ***A***7 *and let* $\gamma \in (0, \gamma_1 \wedge \gamma_2)$. *Then, for any initial probability measure* $\mu_0^{(\mathrm{VR}\star,\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, *we have*

$$W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)},\pi\right) \leq \left(1-\frac{\gamma m}{2}\right)^k W_2^2\left(\mu_0^{(\mathrm{VR}\star,\gamma)},\pi\right) + \left(1-\frac{\gamma m}{8}\right)^k \frac{3L^2}{m^2} v_1 + \frac{6L^2}{m^2} v_2 + \frac{6\gamma d}{nm^2}\kappa_I + \frac{32\gamma^2\omega d}{nmq},$$

*where v_1, v_2 are defined in (3.135) and $\kappa_I = L^2(1+\gamma L^2/m)$. If in addition we suppose **HX**1, set $\kappa_I = 2\gamma(L^3 + d\tilde{L}^2/n)$.*

Proof We know that Assumption 3.4 is satisfied since for any $i \in [n], x \in \mathbb{R}^d$ the stochastic gradient $H^i(x,\xi^i)$ is unbiased. Lemma 3.28 proves that Assumption 3.9 holds with $\alpha_v = 1 - A_d/4$ and v_1, v_2 defined in (3.135). Lemma 3.29 implies that **HX**2 holds with the choice of $(\alpha_i)_{i=0}^4$ detailed in (3.142). Finally, since **HX**2 and Assumption 3.9 hold, we can apply Theorem 3.10 to show that

$$\begin{aligned} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)},\pi\right) &\leq \left(1+\alpha_0+\delta\right)^k \left(1-q_\mathrm{c}\right)^k W_2^2\left(\mu_0^{(\mathrm{VR}\star,\gamma)},\pi\right) \\ &\quad + (1-q_\mathrm{c}) v_1 \left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right) \frac{\alpha_v^k - \left(1+\alpha_0+\delta\right)^k \left(1-q_\mathrm{c}\right)^k}{\alpha_v - \left(1+\alpha_0+\delta\right)\left(1-q_\mathrm{c}\right)} \\ &\quad + \frac{1-q_\mathrm{c}}{q_\mathrm{c} - (1-q_\mathrm{c})(\alpha_0+\delta)} \left[\left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right) v_2 + \alpha_4\right], \end{aligned} \tag{3.144}$$

where $\delta = 2^{-1}(\sqrt{(\alpha_0-1)^2+4\alpha_1} - 1 - \alpha_0)$ is defined in (3.46). Using for any $a > 0, b \geq 0$, that $\sqrt{a+b} \leq \sqrt{a} + b/(2\sqrt{a})$, we obtain

$$\begin{aligned} \alpha_0 + \sqrt{(\alpha_0-1)^2+4\alpha_1} &= 1 + \left(\alpha_0 - 1\right)\left(1 + \sqrt{1 + \frac{4\alpha_1}{\left(\alpha_0-1\right)^2}}\right) \\ &\leq 1 + 2\left(\alpha_0-1\right)\left(1 + \frac{\alpha_1}{\left(\alpha_0-1\right)^2}\right) = 2\alpha_0 - 1 + \frac{2\alpha_1}{\alpha_0-1}. \end{aligned}$$

Since $\gamma \leq \gamma_2 \leq q(2m)^{-1} \wedge \{6(L+4m^{-1}\omega)\}^{-1}$, the previous line implies that

$$\begin{aligned} 2\left(1-q_\mathrm{c}\right)\left(1+\alpha_0+\delta\right) &= (1-q_\mathrm{c})\left(1+\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1}\right) \\ &\leq 2(1-q_\mathrm{c})\left(\alpha_0 + \frac{\alpha_1}{\alpha_0-1}\right) \\ &= 2\left(1-\gamma m + \gamma^2\left(3mL + 4\omega + \frac{4q\omega}{q_\mathrm{c} - \gamma m + \gamma^2\left(3mL+4\omega\right)}\right)\right) \\ &\leq 2(1-\gamma m/2). \end{aligned} \tag{3.145}$$

This upper bound gives that

$$(1-q_\mathrm{c})(\alpha_0+\delta) = (1-q_\mathrm{c})\left(1+\alpha_0+\delta\right) + q - 1 \leq q - \gamma m/2.$$

Thus, we deduce that

$$\frac{1}{q_\mathrm{c} - (1-q_\mathrm{c})(\alpha_0+\delta)} \leq \frac{2}{\gamma m}. \tag{3.146}$$

Further, using $\gamma \leq \gamma_2$ combined with the definitions of $\alpha_0, \alpha_2, \alpha_3, \alpha_v$ and δ show that

$$\begin{aligned} &\frac{\alpha_v^k - \left(1+\alpha_0+\delta\right)^k \left(1-q_{\mathrm{c}}\right)^k}{\alpha_v - \left(1+\alpha_0+\delta\right)\left(1-q_{\mathrm{c}}\right)} \leq \frac{8}{3\gamma m}\left(1-\frac{\gamma m}{8}\right)^k, \\ &\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta} = \frac{\gamma}{1-q_{\mathrm{c}}}\left(\frac{2L^2}{m} + 3\gamma L^2 + 4\gamma\omega\right) \leq \frac{3\gamma L^2}{(1-q_{\mathrm{c}})m}. \end{aligned} \tag{3.147}$$

Lastly, plugging (3.145), (3.146) and (3.147) in (3.144) yields

$$\begin{aligned} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi\right) \leq \left(1-\frac{\gamma m}{2}\right)^k W_2^2\left(\mu_0^{(\mathrm{VR}\star,\gamma)}, \pi\right) + \left(1-\frac{\gamma m}{8}\right)^k \frac{3L^2}{m^2} v_1 \\ + \frac{6L^2}{m^2} v_2 + \frac{2(1-q_{\mathrm{c}})\alpha_4}{\gamma m}. \end{aligned} \tag{3.148}$$

In addition, following the lines provided in the proof of Theorem 3.22, we deduce

$$\frac{2(1-q_{\mathrm{c}})\alpha_4}{\gamma m} \leq \frac{6\gamma d L^2}{nm^2}\left(1 + \frac{19\gamma L^2}{36m}\right) + \frac{32\gamma^2\omega d}{nmq}. \tag{3.149}$$

If in addition we suppose **HX**1, then we obtain

$$\frac{2(1-q_{\mathrm{c}})\alpha_4}{\gamma m} \leq \gamma m L^2\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right) + \frac{4\gamma}{9}\left(L^3 + \frac{d\tilde{L}^2}{n}\right) + \frac{32\gamma^2\omega d}{nmq}. \tag{3.150}$$

Finally, plugging (3.149) or (3.150) if **HX**1 holds inside (3.148) combined with $\gamma \leq qL^{-1}$ lead to the expected result. ■

Now, consider the time step-sizes γ_3 and $\gamma_\star$ defined by

$$\gamma_3 = \frac{p_{\mathrm{c}} m}{3L^2 + 16\hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}}, \qquad \gamma_\star = \gamma_1 \wedge \gamma_2 \wedge \gamma_3.$$

From the previous result, the next corollary controls the asymptotic bias obtained by Algorithm 3.4.

Corollary 3.31. *Assume **A**1, **A**4, **A**6, **A**7 and let $\gamma \in (0, \gamma_\star)$ with $\tau = 1$. Then, for any initial probability measure $\mu_0^{(\mathrm{VR}\star,\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, we have*

$$\begin{aligned} \frac{9^{-9}b}{\gamma d} \limsup_{k\to\infty} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi\right) \leq \frac{\kappa_I}{m^2} + \frac{\gamma\omega}{mq} \\ + \frac{(1-p_{\mathrm{c}})\gamma L^2}{p_{\mathrm{c}}^2 m^5}\left(L^2 + \hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right)\left(1 + \frac{\gamma}{nd}\mathbb{E}\left[\left\|H(x_\star, \xi)\right\|^2\right]\right)\left(L^2 + \frac{q_{\mathrm{c}}}{p_{\mathrm{c}}}\hat{L}^2\right), \end{aligned}$$

*where $\tilde{\kappa}_I = L^2(1 + \gamma L^2 m^{-1})$ and if we suppose **HX**1, $\tilde{\kappa}_I = \gamma(L^3 + d\tilde{L}^2 n^{-1})$.*

Proof Applying Theorem 3.30 with $\gamma \in (0, \gamma_1 \wedge \gamma_2)$ shows that

$$\begin{aligned} \limsup_{k\to\infty} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi\right) &\leq \frac{6L^2}{m^2} v_2 + \frac{6\gamma d}{nm^2}\kappa_I + \frac{32\gamma^2\omega d}{nmq} \\ &\leq \frac{6L^2 \mathrm{C}_\delta^\gamma}{m^2} + \frac{12L^2 \mathrm{C}_r^\gamma \tilde{v}_2}{A_d m^2} + \frac{6\gamma d}{nm^2}\kappa_I + \frac{32\gamma^2\omega d}{nmq}. \end{aligned} \tag{3.151}$$

Plugging the definitions of $\tilde{v}_1, \tilde{v}_2$ provided in (3.130) combined with the previous inequality, we obtain

$$\limsup_{k\to\infty} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi\right) \leq \frac{6L^2\mathrm{C}_\delta^\gamma}{m^2} + \frac{24L^2\mathrm{C}_r^\gamma \alpha_d D_d}{A_d^2 m^2} + \frac{48L^2\mathrm{C}_r^\gamma\left(1-\tau\right)(n-1)\gamma d}{nA_d^2m^2} + \frac{6\gamma d}{nm^2}\kappa_I + \frac{32\gamma^2\omega d}{nmq}.$$

Further, recall that $A_d, B, \bar{B}, D, \bar{D}, D_d$ are provided in (3.128) and α_d is defined in (3.129) by

$$\begin{aligned}\alpha_d &= \frac{4\gamma^2}{p_\mathrm{c}A_d}\max\left\{p_\mathrm{c}B + 3\bar{B}, \frac{4B_\sigma}{A_\sigma}\left(p_\mathrm{c}C + 3\bar{C}\right)\right\} \\ &= \frac{24\gamma}{p_\mathrm{c}m}\max\left\{3L^2 + p_\mathrm{c}\tilde{\omega}, 8(p_\mathrm{c}+3)\hat{L}^2\right\} \leq \frac{768\gamma}{p_\mathrm{c}m}\left(L^2 + \hat{L}^2 + p_\mathrm{c}\tilde{\omega}\right).\end{aligned}$$

Moreover, $\mathrm{C}_\delta^\gamma, \mathrm{C}_r^\gamma$ are defined in (3.131) by

$$\begin{aligned}\mathrm{C}_\delta^\gamma &= \frac{\mathrm{C}^\gamma D_d}{A_d}\left(1 + \frac{2B_dB_\sigma}{A_d(A_\sigma - A_d)}\right) + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}} \\ &= \frac{10\mathrm{C}^\gamma}{m}\left(1 + \frac{64\gamma q\hat{L}^2}{(2q-\gamma m)m}\right)\left(5\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + \frac{d}{n}\right) + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}} \\ &\leq \frac{360(1-p_\mathrm{c})\gamma^2}{mp_\mathrm{c}^2}\left(3L^2 + 11\hat{L}^2 + p_\mathrm{c}\tilde{\omega}\right)\left(5\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + \frac{d}{n}\right) \\ &\quad + \frac{8\left(1-\tau\right)(n-1)\gamma d}{np_\mathrm{c}}, \qquad (3.152) \\ \mathrm{C}_r^\gamma &= \frac{9\gamma^2\left(1-p_\mathrm{c}\right)C_\sigma}{p_\mathrm{c}-4A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_dC_\sigma}{A_\sigma - A_d}\right) \\ &\leq \frac{144\gamma^2(1-p_\mathrm{c})}{p_\mathrm{c}^2}\left[3q\hat{L}^2 + \gamma\left(\frac{L^2}{m} + 8\gamma\hat{L}^2\right)\left(p_\mathrm{c}\tilde{\omega} + 3L^2 + 16\hat{L}^2\right)\right] \\ &\leq \frac{432\gamma^2(1-p_\mathrm{c})}{p_\mathrm{c}^2}\left(p_\mathrm{c}L^2 + q\hat{L}^2\right)\end{aligned}$$

Eventually, for the specific choice $\tau = 1$ combined with (3.151) and (3.152), it yields that

$$\begin{aligned}\limsup_{k\to\infty} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi\right) &\leq \frac{6\gamma d}{nm^2}\kappa_I + \frac{32\gamma^2\omega d}{nmq} + \frac{18432\gamma\mathrm{C}_r^\gamma D_dL^2}{A_d^2m^3p_\mathrm{c}}\left(L^2 + \hat{L}^2 + p_\mathrm{c}\tilde{\omega}\right) \\ &\quad + \frac{2160(1-p_\mathrm{c})\gamma^2L^2}{p_\mathrm{c}^2m^3}\left(3L^2 + 11\hat{L}^2 + p_\mathrm{c}\tilde{\omega}\right)\left(5\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + \frac{d}{n}\right). \qquad (3.153)\end{aligned}$$

Therefore, using (3.152) and (3.153) we can finally conclude that

$$9^9\limsup_{k\to\infty} W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)}, \pi\right) \leq \frac{\gamma d}{nm^2}\kappa_I + \frac{\gamma^2\omega d}{nmq}$$

$$+\frac{(1-p_{\mathrm{c}})\gamma^2 L^2}{p_{\mathrm{c}}^2 m^5}\left(L^2+\hat{L}^2+p_{\mathrm{c}}\tilde{\omega}\right)\left(\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right]+\frac{d}{n}\right)\left(L^2+\frac{q_{\mathrm{c}}}{p_{\mathrm{c}}}\hat{L}^2\right).$$

■

The single client case corresponds to $n = p_{\mathrm{c}} = 1$ and leads for $k \geq 0$ to $V_k = 0$. Moreover, the assumption Assumption 3.9 holds with $v_1 = v_2 = 0$. Thus, we obtain a convergence bound for SVRG-LD from Theorem 3.30.

Theorem 3.32. *Assume **A**1, **A**4, **A**6, **A**7 and let $\gamma \in (0, \gamma_1 \wedge \gamma_2)$. Then, for any initial probability measure $\mu_0^{(\mathrm{VR}\star,\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, we have*

$$W_2^2\left(\mu_k^{(\mathrm{VR}\star,\gamma)},\pi\right) \leq \left(1-\frac{\gamma m}{2}\right)^k W_2^2\left(\mu_0^{(\mathrm{VR}\star,\gamma)},\pi\right)+\frac{6\gamma d}{nm^2}\kappa_I+\frac{32\gamma^2\omega d}{mq},$$

*where $\kappa_I = L^2(1+\gamma L^2/2m+\gamma^2 L^2/12)$. If in addition we suppose **HX**1, set $\kappa_I = 3\gamma(L^3+d\tilde{L}^2/n)$.*

Remark 3.33.

- *The constants obtained in this result can be refined by directly using that $\mathbb{E}[V_k] = 0$ in the proof of Lemma 3.29 and by simplifying the calculations detailed in Theorem 3.30.*
- *The proof given in Chatterji et al. (2018, Theorem 4.2-Option 2) on the convergence of SVRG-LD seems to have some gaps since the authors use Grönwall's inequality (Clark, 1987) as if $\spadesuit = \tau^2(8\delta d + 4M\delta^2 d + 4\delta^2 M\Omega_1)$ were constant, which is not the case because $\Omega_1 = \left\langle \nabla f(y_k) - \nabla f(x_k), y_k - x_k \right\rangle$ depends on the iteration k. If we denote $\spadesuit_k$ instead of $\spadesuit$ and adopt their other notation (we also correct a typography in the right-hand term), we obtain*

$$\mathbb{E}\left[\left\|x_k-\tilde{x}\right\|_2^2\right] \leq \spadesuit_k + \sum_{j=\tau s}^{k-1}\mathbb{E}\left[\left\|x_j-\tilde{x}\right\|_2^2\right]. \tag{3.154}$$

  Then, it is claimed in the proof of Chatterji et al. (2018, Theorem 4.2-Option 2) that (3.154) implies $\mathbb{E}[\|x_k - \mathsf{X}_k\|^2] \leq \spadesuit_k \exp(\tau\rho)$. But this inequality cannot hold in all generalities, for example if we consider : $\tau s = 0$, for $j < k$, $\spadesuit_j = 1$, $x_j = \tilde{x} + \sqrt{2^j/d}\cdot\mathbf{1}$ and $\spadesuit_k = 0$, $x_k = \tilde{x} + \mathbf{1}/\sqrt{d}$, then (3.154) holds for $j \in [k]$ but $\mathbb{E}[\|x_k - \mathsf{X}_k\|^2] = 1$ whereas $\spadesuit_k \exp(\tau\rho) = 0$.

3.C Lower bound on the heterogeneity in a Gaussian case

In this section, we want to illustrate the heterogeneity problem by lower bounding the Wasserstein distance W_2 in a simple case. To simplify the calculations, we assume that each client performs 2 local iterations following the FALD update before communicating its local parameter to the central server. More specifically, take $(\mu_1, \mu_2, \sigma_1, \sigma_2) \in \mathbb{R}^2 \times (\mathbb{R}_+^*)^2$ and define the potentials $U^1 : x \in \mathbb{R}^d \mapsto \sigma_1^{-2}\left(x-\mu_1\right)^2$, $U^2 : x \in \mathbb{R}^d \mapsto$

$\sigma_2^{-2}(x-\mu_2)^2$. Thus, the global posterior distribution π is Gaussian with mean $\bar{\mathrm{m}}$ and variance $\bar{\sigma}^2$ given by

$$\bar{\mathrm{m}} = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2} \qquad \bar{\sigma} = \left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)^{-1/2}. \tag{3.155}$$

The objective is to illustrate the problem of heterogeneity in the basic version of FALD. To do so, we first show that this algorithm generates samples targeting a distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ where the distance $W_2(\pi,\pi_\gamma)$ is lower bounded by a heterogeneity term. To this end, we introduce the Markov kernel, which for each $\gamma>0, \mathsf{B}\in\mathcal{B}(\mathbb{R}^d)$ is given by

$$P_\gamma(x,\mathsf{B}) = \int_{\mathsf{B}} \exp\left(-\frac{\left\|x' - \left(1-\frac{\gamma}{\bar{\sigma}^2}+\frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)\right)x - \frac{\gamma\bar{\mathrm{m}}}{\bar{\sigma}^2}+\frac{\gamma^2}{2}\left(\frac{\mu_1}{\sigma_1^4}+\frac{\mu_2}{\sigma_2^4}\right)\right\|^2}{2\gamma\left(1+\left(1-\frac{\gamma}{2\bar{\sigma}^2}\right)^2\right)}\right)\frac{\mathrm{d}x'}{(2\pi)^{d/2}},$$

and we define the stochastic processes $(A_k,\tilde{A}_k)_{k\geq 0}$ on $\mathbb{R}^d\times\mathbb{R}^d$ starting from $(X_0,X_0)=(x,\tilde{x})$ and following the recursion for $k\geq 0$,

$$\begin{aligned}
A_{k+1} &= A_k - \frac{\gamma}{\bar{\sigma}^2}(A_k-\bar{\mathrm{m}}) + \frac{\gamma^2}{2}\left(\frac{A_k-\mu_1}{\sigma_1^4}+\frac{A_k-\mu_2}{\sigma_2^4}\right) + \sqrt{\gamma}\left[\left(1-\frac{\gamma}{2\bar{\sigma}^2}\right)Z_{k+1}+Z_{k+2}\right],\\
\tilde{A}_{k+1} &= \tilde{A}_k - \frac{\gamma}{\bar{\sigma}^2}(\tilde{A}_k-\bar{\mathrm{m}}) + \frac{\gamma^2}{2}\left(\frac{\tilde{A}_k-\mu_1}{\sigma_1^4}+\frac{\tilde{A}_k-\mu_2}{\sigma_2^4}\right) + \sqrt{\gamma}\left[\left(1-\frac{\gamma}{2\bar{\sigma}^2}\right)Z_{k+1}+Z_{k+2}\right].
\end{aligned} \tag{3.156}$$

It is possible to verify that $(A_k,\tilde{A}_k)$ is distributed according to $(\delta_x P_\gamma^k, \delta_{\tilde{x}}P_\gamma^k)$.

Lemma 3.34. *Let $\gamma\in\left(0, 2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4+\sigma_2^4)]^{-1}\right)$. Then, there exists $\pi_\gamma\in\mathcal{P}_2(\mathbb{R}^d)$ such that for any distribution $\pi^0\in\mathcal{P}_2(\mathbb{R}^d)$, the sequence $(\pi^0P_\gamma^k)_{k\in\mathbb{N}}$ converges to π_γ in $\mathcal{P}_2(\mathbb{R}^d)$.*

Proof Let $k\in\mathbb{N}$ and consider the stochastic processes $(A_l,\tilde{A}_l)_{l\in\mathbb{N}}$ defined in (3.156), subtracting the two recursions we obtain

$$A_{k+1}-\tilde{A}_{k+1} = \left(1-\frac{\gamma}{\bar{\sigma}^2}+\frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)\right)\left(A_k-\tilde{A}_k\right).$$

Since $0<\gamma<2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4+\sigma_2^4)]^{-1}$, taking the norm in the previous inequality implies that

$$\left\|A_{k+1}-\tilde{A}_{k+1}\right\| = \left(1-\frac{\gamma}{\bar{\sigma}^2}+\frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)\right)\left\|A_k-\tilde{A}_k\right\|. \tag{3.157}$$

Finally, combining (3.157) with Douc et al. (2018, Lemma 20.3.2), we deduce that the c-Dobrushin coefficient of P_γ is upper bounded by $1-\gamma/\bar{\sigma}^2+\gamma^2/2\left(1/\sigma_1^4+1/\sigma_2^4\right)$. Hence, applying Douc et al. (2018, Theorem 20.3.4) we deduce the existence and uniqueness of

a stationary distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ for the Markov Kernel P_γ such that

$$W_2(\pi^0 P_\gamma^k, \pi) \leq \left(1 - \gamma/\bar{\sigma}^2 + \gamma^2/2\left(1/\sigma_1^4 + 1/\sigma_2^4\right)\right)^k W_2(\pi^0, \pi_\gamma).$$

■

Lemma 3.34 shows the existence of an invariant distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ for P_γ and the next lemma specifies this distribution of π_γ.

Lemma 3.35. *Assume* $\gamma \in \left(0, 2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4 + \sigma_2^4)]^{-1}\right)$. *Then, the stationarity distribution* π_γ *is Gaussian with parameters given by*

$$\mathrm{m}_{(\gamma)} = \frac{\bar{\mathrm{m}} - \frac{\gamma\bar{\sigma}^2}{2}\left(\frac{\mu_1}{\sigma_1^4} + \frac{\mu_2}{\sigma_2^4}\right)}{1 - \frac{\gamma\bar{\sigma}^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)}, \quad \sigma^2_{(\gamma)} = \frac{\bar{\sigma}^2 - \frac{\gamma}{2} + \frac{\gamma^2}{8\sigma^2}}{1 - \frac{\gamma}{2}\left(\frac{\bar{\sigma}^2}{\sigma_1^4} + \frac{\bar{\sigma}^2}{\sigma_2^4}\right) - \frac{\gamma}{2}\left(\frac{1}{\bar{\sigma}} - \frac{\gamma}{2}\left(\frac{\bar{\sigma}}{\sigma_1^4} + \frac{\bar{\sigma}}{\sigma_2^4}\right)\right)^2}.$$

Proof First, let $k \in \mathbb{N}$ be fixed and introduce

$$\alpha = 1 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right), \qquad \beta = \frac{\gamma\bar{\mathrm{m}}}{\bar{\sigma}^2} - \frac{\gamma^2}{2}\left(\frac{\mu_1}{\sigma_1^4} + \frac{\mu_2}{\sigma_2^4}\right),$$
$$\tilde{Z}_k = \left(1 - \frac{\gamma}{2\bar{\sigma}^2}\right) Z_{2k-1} + Z_{2k}.$$

Moreover, consider $(A_l)_{l\in\mathbb{N}}$ the stochastic process following (3.156) and initialized at π_γ. By induction, we know that

$$A_k = \alpha^k A_0 + \beta \sum_{l=0}^{k-1} \alpha^l + \sqrt{\gamma} \sum_{l=0}^{k-1} \alpha^{k-l-1} \tilde{Z}_l. \tag{3.158}$$

Since A_k is distributed according to $\pi_\gamma P_\gamma^k$, we have that A_k follows π_γ. Denote ν_γ^k the distribution of $\sqrt{\gamma}\sum_{l=0}^{k-1} \alpha^{k-l-1}\tilde{Z}_l - \beta\sum_{l=0}^{k-1}\alpha^l$, combining (3.158) with the definition of the Wasserstein, we have

$$W_2^2\left(\pi_\gamma, \nu_\gamma^k\right) \leq \mathbb{E}\left[\left\| A_k - \sqrt{\gamma}\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l - \beta\sum_{l=0}^{k-1}\alpha^l \right\|^2\right] = \alpha^{2k}\mathbb{E}\left[\left\|A_0\right\|^2\right]. \tag{3.159}$$

Since A_0 is distributed according to π_γ belonging to $\mathcal{P}_2(\mathbb{R}^d)$, we deduce that $\mathbb{E}[\|A_0\|^2] < \infty$. Consequently, (3.159) implies that $(\nu_\gamma^k)_{k\in\mathbb{N}}$ converges to π_γ, but using the fact that $(\nu_\gamma^k)_{k\in\mathbb{N}}$ converges to a Gaussian distribution, we obtain by uniqueness of the limit in metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ that π_γ is a Gaussian distribution. Recalling that $\mathrm{m}_{(\gamma)}$ denotes the expectation of the random variable distributed according to π_γ, using (3.156) at stationarity yields

$$\mathrm{m}_{(\gamma)} = \mathrm{m}_{(\gamma)} - \frac{\gamma}{\bar{\sigma}^2}\left(\mathrm{m}_{(\gamma)} - \bar{\mathrm{m}}\right) + \frac{\gamma^2}{2}\left(\frac{\mathrm{m}_{(\gamma)} - \mu_1}{\sigma_1^4} - \frac{\mathrm{m}_{(\gamma)} - \mu_2}{\sigma_2^4}\right)$$

Thus, we deduce that

$$\mathrm{m}_{(\gamma)} = \frac{\bar{\mathrm{m}} - (\gamma\bar{\sigma}^2/2)\left(\mu_1/\sigma_1^4 + \mu_2/\sigma_2^4\right)}{1 - (\gamma\bar{\sigma}^2/2)\left(1/\sigma_1^4 + 1/\sigma_2^4\right)}.$$

In addition, we can obtain the standard deviation $\sigma_{(\gamma)}$ of π_γ since we have

$$\begin{aligned}
\mathrm{Var}\left(\beta\sum_{l=0}^{k-1}\alpha^l + \sqrt{\gamma}\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l\right) &= \gamma\,\mathrm{Var}\left(\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l\right) = \frac{\gamma(1-\alpha^{2k})}{1-\alpha^2}\,\mathrm{Var}(\tilde{Z}_0) \\
&\xrightarrow[k\to\infty]{} \frac{\gamma\,\mathrm{Var}(\tilde{Z}_0)}{1-\alpha^2} \\
&= \frac{\gamma\left(2 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{4\bar{\sigma}^4}\right)}{1 - \left(1 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)\right)^2} \\
&= \frac{1 - \frac{\gamma}{2\bar{\sigma}^2} + \frac{\gamma^2}{8\bar{\sigma}^4}}{\frac{1}{\bar{\sigma}^2} - \frac{\gamma}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right) - \frac{\gamma}{2}\left(\frac{1}{\bar{\sigma}^2} - \frac{\gamma}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)\right)^2}.
\end{aligned}$$

■

Theorem 3.36. *Assume* $\gamma \in \left(0, 2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4 + \sigma_2^4)]^{-1}\right)$. *Then, the Wasserstein distance between the stationnary distribution* π_γ *and the target* π *of* FALD *is lower bounded as*

$$W_2\left(\pi_\gamma, \pi\right) \geq \frac{\gamma}{2}\left|\mu_1 - \mu_2\right|\left|\frac{\bar{\sigma}^2}{\sigma_1^2} - \frac{\bar{\sigma}^2}{\sigma_2^2}\right|.$$

Proof Based on Lemma 3.35, we know that π_γ is Gaussian with parameters $(\mathrm{m}_{(\gamma)}, \sigma^2_{(\gamma)})$ and using that π is Gaussian too with parameters $(\bar{\mathrm{m}}, \bar{\sigma}^2)$ given in (3.155), we have that

$$\begin{aligned}
W_2^2\left(\pi_\gamma, \pi\right) = \left(\mathrm{m}_{(\gamma)} - \bar{\mathrm{m}}\right)^2 + \left(\sigma_{(\gamma)} - \bar{\sigma}\right)^2 &\geq \frac{\gamma^2\bar{\sigma}^4}{4}\left|\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)\bar{\mathrm{m}} - \frac{\mu_1}{\sigma_1^4} - \frac{\mu_2}{\sigma_2^4}\right|^2 \\
&= \frac{\gamma^2\bar{\sigma}^4\left(\mu_1 - \mu_2\right)^2}{4}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)^2.
\end{aligned}$$

■

3.D Analysis of the complexity and communication cost

In this section, we study the optimal choices of k, γ when p_c is fixed. For $c_0, c_1, c_2 \geq 0$ fixed, we consider the following optimization problem:

$$\begin{cases} \min_{k \in \mathbb{N}^\star, \gamma > 0} \{k\} \\ \text{Subject to } \left\{ c_0 \exp\left(-8k\gamma/m\right) + c_1\gamma + c_2\gamma^2 \leq \epsilon^2 \right\} \end{cases}.$$

Using that the constraint must be saturated at the optimum (which can be proved), we can write k as a function of γ. Hence, the problem becomes

$$\begin{cases} \min_{k,\gamma} \left\{ \frac{8}{\gamma m} \log\left(\frac{c_0}{\epsilon^2 - c_1\gamma - c_2\gamma^2} \right) \right\} \\ \text{Subject to } 0 < \gamma \text{ and } \epsilon^2 - c_1\gamma - c_2\gamma^2 > 0 \end{cases}. \tag{3.160}$$

Let us introduce $x \in \mathbb{R}_+^*$, defined by $x = \epsilon^{-2}\gamma$ and let $\tilde{c}_2 = \epsilon^2 c_2$. We can rewrite (3.160) as

$$\begin{cases} \min_{k,x} \left\{ \frac{8}{\epsilon^2 m x} \log\left(\frac{c_0}{\epsilon^2(1 - c_1 x - \tilde{c}_2 x^2)} \right) \right\} \\ \text{Subject to } 0 < x \text{ and } 1 - c_1 x - \tilde{c}_2 x^2 > 0 \end{cases}. \tag{3.161}$$

Consider $\mu = -c_1/(2\tilde{c}_2)$, $\sigma = \sqrt{c_1^2/(4\tilde{c}_2^2) + 1/\tilde{c}_2}$, and denote $z = (x - \mu)/\sigma$. Since $x = \mu + z\sigma$, we can verify that $1 - c_1 x - \tilde{c}_2 x^2 = \tilde{c}_2 \sigma^2 (1 - z^2)$. Hence, (3.161) is equivalent to

$$\begin{cases} \min_{k,\gamma} \left\{ \frac{8}{\epsilon^2 m(\mu + z\sigma)} \log\left(\frac{c_0}{\epsilon^2 \tilde{c}_2 \sigma^2 (1 - z^2)} \right) \right\} \\ \text{Subject to } -\mu/\sigma < z < 1 \end{cases}.$$

According to the intermediate value theorem, we have the existence of z_ϵ (not necessarily unique, but we can consider one of the solutions) such that

$$z_\epsilon = \underset{-\mu/\sigma < z < 1}{\arg\max} \left\{ \frac{\log(1 - z^2)}{\mu + z\sigma} \right\}.$$

Thus, the solution is

$$\begin{aligned} \gamma_\epsilon &= \epsilon^2 \times \frac{z_\epsilon^2 + (4\epsilon^2 c_2)^{-1}(z_\epsilon^2 - 1)c_1^2}{c_1/2 + z_\epsilon \sqrt{4^{-1}\ c_1^2 + \epsilon^2 c_2}}, \\ K_\epsilon &= \frac{8\left(c_1/2 + z_\epsilon \sqrt{4^{-1}\ c_1^2 + \epsilon^2 c_2}\right)}{\epsilon^2 m \left(z_\epsilon^2 + (4\epsilon^2 c_2)^{-1}(z_\epsilon^2 - 1)c_1^2\right)} \log\left(\frac{c_0}{\epsilon^2 (c_1^2/4 + \epsilon^2 c_2)^{1/2}(1 - z_\epsilon^2)} \right). \end{aligned}$$

FALD. According to the Theorem 3.1, we have

$$\begin{cases} c_0 = \mathsf{I}(\mu_0) \\ c_1 = \mathsf{V}_\pi + \left(1 - 1_{\mathbf{HX}1}\right) \mathsf{J}/n + (1 - \tau)(1 - n^{-1})d/p_c \\ c_2 = 1_{\mathbf{HX}1}\mathsf{J}/n + (1 - p_c)\left\{\mathsf{H} + p_c \mathsf{V}_\epsilon + d/n\right\}/p_c^2 \end{cases}.$$

If $c_1 > 0$, define $w = \epsilon^2 c_2 / c_1^2$. For $\epsilon \in (0, c_1/\sqrt{2c_2}]$, we have $0 < w \leq 1/2$. Consider $z = 1 - w$, we get that

$$\left(\frac{\mu}{\sigma}\right)^2 = \frac{1}{1 + 4\epsilon^2 c_2 / c_1^2} < \frac{1}{1+2w} \leq 1 - w \leq 1 - 2w + w^2 = z^2 < 1.$$

Hence, the previous inequalities show that $-\mu/\sigma < z < 1$, and for this choice

$$\frac{c_1/2 + z\sqrt{4^{-1}c_1^2 + \epsilon^2 c_2}}{z^2 + (4\epsilon^2 c_2)^{-1}(z^2-1)c_1^2} \leq \frac{c_1 + \epsilon(1-w)\sqrt{c_2}}{7/8 + (w - 2 + 1/64)w}.$$

Thus, for any $\epsilon \in (0, c_1(2\sqrt{c_2})^{-1}]$, we deduce that $w < 1/4$. Therefore, we have shown that $K_\epsilon = \tilde{\mathrm{O}}((\epsilon^2 m)^{-1}(c_1 + \epsilon\sqrt{c_2}))$. Moreover, this result is immediately valid when $c_1 = 0$ since $z_\epsilon = \arg\max_{0<z<1}\{z^{-1}\log(1-z^2)\}$. Furthermore, when $p_{\mathrm{c},\epsilon} \downarrow 0^+$, $p_{\mathrm{c},\epsilon} K_\epsilon = \tilde{\mathrm{O}}((\epsilon m)^{-1}\sqrt{n^{-1}\mathsf{J}})$ as stressed in the main chapter.

VR-FALD⋆. Using Theorem 3.3, we obtain

$$\begin{cases} c_0 = \mathsf{I}^{\mathrm{Vr}\star}(\mu_0) \\ c_1 = \left(1 - 1_{\mathbf{HX}1}\right)\mathsf{J}/n + (1-\tau)(1-n^{-1})d/p_{\mathrm{c}} \\ c_2 = 1_{\mathbf{HX}1}\mathsf{J}/n + (1-p_{\mathrm{c}})\left\{p_{\mathrm{c}}\mathsf{V}_\epsilon + d/n\right\}/p_{\mathrm{c}}^2 \end{cases}.$$

When assuming **HX**1 and $\tau = 1$, we have $c_1 = 0$. Hence, $z_\epsilon = \arg\max_{0<z<1}\{z^{-1}\log(1-z^2)\}$ and therefore

$$K_\epsilon = \frac{8\sqrt{c_2}}{\epsilon m z_\epsilon} \log\left(\frac{c_0}{\epsilon^3\sqrt{c_2}(1-z_\epsilon^2)}\right).$$

When $p_{\mathrm{c},\epsilon} \downarrow 0^+$, the minimum number of communications becomes $p_{\mathrm{c},\epsilon} K_\epsilon = \tilde{O}(\epsilon^{-1}\sqrt{n^{-1}d})$. Finally, setting $p_{\mathrm{c},\epsilon} = 1$ gives $K_\epsilon = \tilde{O}(\epsilon^{-1}\sqrt{n^{-1}\mathsf{J} + n^{-1}\omega d})$.

Table 3.5 – Complexity and communication settings of Figure 3.1.

Parameter	d	m	ω	H	J	V_π	$\mathsf{V}_\star$
Value	10	1	10	100	20	10	30

3.E Numerical experiments

3.E.1 Gaussian example

In this first experiment, we consider $n = 100$ clients associated with potentials: $\forall i \in [n]$, $U^i : x \in \mathbb{R}^d \mapsto (1/2)(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)$ in dimension $d = 20$. In this particular case, we know, that the posterior distribution $\pi \propto \exp(-\sum_{i=1}^n U^i)$ is Gaussian with mean $x_\star = \sum_{i=1}^n (\Sigma_\star \Sigma_i^{-1}\mu_i)$ and covariance $\Sigma_\star = (\sum_{i=1}^n \Sigma_i^{-1})^{-1}$. Also, we have a close formula to calculate $\int \|x - x_\star\|^2 \mathrm{d}\pi(x)$, since this quantity is equal to $\mathrm{Trace}(\Sigma_\star)$. To speed up the calculations, we initialize all chains at $x_\star$, we discard the first 10% of the samples and keep all others. Moreover, we consider the step-size $\bar{\gamma} = 2[\lambda_{\min}(\Sigma_\star^{-1}) + \lambda_{\max}(\Sigma_\star^{-1})]^{-1}$ for Langevin Monte Carlo (Dalalyan and Karagulyan, 2019; Durmus and Moulines,

2019), and we run the algorithms for the step-sizes $\gamma \in \{\frac{p_c\bar{\gamma}}{2}, \frac{p_c\bar{\gamma}}{5}, \frac{p_c\bar{\gamma}}{10}\}$ associated with $p_c \in \{\frac{1}{5}, \frac{1}{10}, \frac{1}{20}\}$. We set the probability of updating the control variates $q_c = p_c$ so as not to increase the communication cost too much. We also consider the two extreme values of the parameter $\tau \in \{0, 1\}$ to determine whether it is preferable to have independent Gaussian noise on each client or if it is better to have a common one.

3.E.2 Bayesian Logistic Regression

The second experiment is performed on the Titanic dataset, which is in the public domain and licensed under the Commons Public Domain Dedication License (PDDL-1.0). We distribute this dataset heterogeneously across $n = 10$ clients by drawing a Dirichlet random variable for each label on the standard $n - 1$ simplex. Since the sum of the coordinates of these random variables equals 1, each coordinate indicates the fraction of labels to be distributed to each client. To have access to ground truth, we also implement Langevin Stochastic Dynamics (LSD). We compute $K = 250000$ iterations, each time considering a burn-in period of length 10% initialized with a warm start provided by SGD. The ith client uses its local dataset $\{(z_{ij}, o_{ij}) \in \mathbb{R}^4 \times \{0, 1\} : j \in [N_i]\}$ to calculate the local potential $U^i(x) = \sum_{j=1}^{N_i} [o_{ij} \log(1 + \exp(-z_{ij}^T x)) + (1 - o_{ij}) \log(1 + \exp(z_{ij}^T x))] + \lambda \|x\|^2$, where $\lambda = 1$ is associated with the Gaussian prior. Denote $\mathsf{Z}_{\text{train}}$ the matrix whose lines are the covariates z_{ij}^T, and write $\Sigma = \mathsf{Z}_{\text{train}}^T \mathsf{Z}_{\text{train}}$. We run the algorithms with minibatches of size $b_i = 1$; a step-size $\gamma = 2[\lambda_{\min}(\Sigma) + \lambda_{\max}(\Sigma)]^{-1}$ for FALD, VR-FALD* and equal to γ/n for LSD with thinning inversely proportional to the step-size. Moreover, we consider a communication probability of $p_c = 1/20$ and clients update their control variates with probability $q_c = p_c$. Finally, to evaluate the obtained results, we consider the accuracy, agreement, and total variation, as well as the calibration results such as ECE, BS, and NLL, which are described below.

Accuracy. Based on samples from the approximate posterior distribution, we compute the minimum mean squared estimator (*i.e.*, which corresponds to the posterior mean) and use it to make predictions for the test dataset. The *Accuracy* metric corresponds to the percentage of well-predicted labels.

Agreement. Let p_{ref} and p denote the predictive densities associated with HMC and an approximate simulation-based algorithm, respectively. Similar to Izmailov et al. (2021), we define the agreement between p_{ref} and p as the proportion of test data points for which the top-1 predictions of p_{ref} and p, *i.e.*

$$\text{agreement}(p_{\text{ref}}, p) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \mathbf{1}\left[\arg\max_{y'} p_{\text{ref}}(y' \mid x) = \arg\max_{y'} p(y' \mid x)\right].$$

Total variation (TV). By denoting $\mathcal{Y}$ as the set of possible labels, we consider the total variation metric between p_{ref} and p, *i.e.*

$$\text{TV}(p_{\text{ref}}, p) = \frac{1}{2|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \sum_{y' \in \mathcal{Y}} \left| p_{\text{ref}}(y' \mid x) - p(y' \mid x) \right|.$$

Expected Calibration Error (ECE). To measure the difference between the accuracy and confidence of the predictions, we group the data into $M \geq 1$ buckets defined for each $m \in [M]$ by $\mathrm{B_m} = \{(x, y) \in \mathcal{D}_{\text{test}} : p(y_{\text{pred}}(x)|x) \in \Big](m-1)/M, m/M\Big]\}$. As in the previous work of Ovadia et al. (2019), we denote the model accuracy on $\mathrm{B_m}$ by

$$\mathrm{acc}\left(\mathrm{B_m}\right) = \frac{1}{\left|\mathrm{B_m}\right|} \sum_{(x,y)\in \mathrm{B_m}} \mathbf{1}_{y_{\text{pred}}(x)=y}$$

and define the confidence on $\mathrm{B_m}$ by

$$\mathrm{conf}\left(\mathrm{B_m}\right) = \frac{1}{\left|\mathrm{B_m}\right|} \sum_{(x,y)\in \mathrm{B_m}} p(y_{\text{pred}}(x)|x).$$

As emphasized in Guo et al. (2017), for any $m \in [M]$ the accuracy $\mathrm{acc}\left(\mathrm{B_m}\right)$ is an unbiased and consistent estimator of $\mathbb{P}\left(y_{\text{pred}}(x) = y \mid (m-1)/M < p(y_{\text{pred}}(x)|x) \leq m/M\right)$. Therefore, the ECE is defined by

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{\left|\mathrm{B_m}\right|}{\left|\mathcal{D}_{\text{test}}\right|} \left|\mathrm{acc}\left(\mathrm{B_m}\right) - \mathrm{conf}\left(\mathrm{B_m}\right)\right|$$

and is an estimator of

$$\mathbb{E}_{(x,y)}\left[\left| PP\left(y_{\text{pred}}(x) = y \mid p(y_{\text{pred}}(x)|x)\right) - p(y_{\text{pred}}(x)|x)\right|\right].$$

Thus, the ECE measures the absolute difference between the confidence level of a prediction and its accuracy.

Brier Score (BS). The BS is a proper scoring rule (see for example Dawid and Musio (2014)) that can only evaluate random variables taking a finite number of values. Denote by $\mathcal{Y}$ the finite set of possible labels, the BS measures the confidence of the model in its predictions and is defined by

$$\mathrm{BS} = \frac{1}{\left|\mathcal{D}_{\text{test}}\right|} \sum_{(x,y)\in \mathcal{D}_{\text{test}}} \sum_{c\in\mathcal{Y}} (p(y=c|x) - \mathbf{1}_{y=c})^2.$$

Normalized Negative Log Likelihood (nNLL). This classical score defined by

$$\mathrm{nNLL} = -\frac{1}{\left|\mathcal{D}_{\text{test}}\right|} \sum_{(x,y)\in \mathcal{D}_{\text{test}}} \log p(y|x)$$

measures the ability of the model to predict good labels with high probability.

Highest posterior density (HPD). Under the Bayesian paradigm, we are interested in quantifying uncertainty by estimating the regions of high probability. For all $\alpha \in (0, 1)$, we run each algorithm to estimate $\eta_\alpha^{\text{algo}} > 0$ such that $\int_{R_\alpha} \pi(x)\mathrm{d}x = 1 - \alpha$, where $\mathcal{R}_\alpha = \{x \in \mathbb{R}^d : \pi(x) \geq \exp(-\eta_\alpha^{\text{algo}})\}$. Then we define the relative HPD error as $|\eta_\alpha^{\text{algo}}/\eta_\alpha^{\text{LSD}} - 1|$, where η_α^{LSD} is estimated based on the samples drawn with the Langevin Stochastic Dynamics method.

METHOD	SGLD	pSGLD	FALD	VR-FALD⋆	FSGLD
Accuracy	99.1 ± 0.1	99.2 ± 0.1	99.1 ± 0.1	99.2 ± 0.1	98.5 ± 0.2
$10^3 \times$ECE	6.88 ± 27.07	21.6 ± 11.1	4.07 ± 0.80	4.34 ± 1.26	6.34 ± 1.90
$10^2 \times$BS	1.66 ± 1.76	1.45 ± 0.12	1.47 ± 0.45	1.39 ± 0.07	2.39 ± 1.72
$10^2 \times$nNLL	3.53 ± 5.08	4.24 ± 1.14	3.06 ± 0.43	3.43 ± 0.37	4.87 ± 0.51
Weight Decay	5	5	5	5	5
Batch Size	64	64	8	8	64
Learning rate	1e-07	1e-08	1e-07	1e-07	1e-08
Local steps	N/A	N/A	20	20	20
Burn-in	100epch.	100epch.	1e04	1e04	1e04
Thinning	1	1	1e03	1e03	1e03
Training	1e03epch.	1e03epch.	1e05it.	1e05it.	1e05it.

Table 3.6 – Performance of Bayesian FL algorithms on MNIST.

3.E.3 Bayesian Neural Network: MNIST

To investigate the behavior of the proposed algorithms in a highly non-convex setting, we perform a first Deep Learning experiment on the MNIST dataset (Deng, 2012), which can be publicly downloaded using the torchvision package and is available under the Creative Commons Attribution-Share Alike 3.0 license. To this end, we distribute the entire dataset across $n = 20$ clients in a highly heterogeneous manner to train the LeNet5 neural network (LeCun et al., 1998). The MNIST real-world dataset consists of 70000 grayscale images of size 28×28 associated with the 10 digits. This dataset is divided into two subsets: the training set, which contains 60000 images, and the test set, which consists of the remaining 10000 images. We report the median of the scores with their associated hyperparameters in Table 3.6. The burn-in corresponds to the number of steps performed before we start storing the samples, and the thinning is the frequency with which we keep the samples. We also consider a Gaussian prior which corresponds to a squared norm regularizer with weight decay. We initialized FSGLD (El Mekkaoui et al., 2021) with a global SGD warm start combined with local SWAG (Maddox et al., 2019) to learn Gaussian conducive gradients.

3.E.4 Bayesian Neural Network: CIFAR10

In this last experiment, we consider the more challenging dataset CIFAR10 (Krizhevsky et al., 2009), which is available under license MIT and contains images of size $(3, 32, 32)$. We used different approaches to sample the weights for the ResNet-20 model (He et al., 2016), which is publicly available in the pytorchcv library. We initialized the algorithms with 10 different parameters using SGD (400 epochs) trained with a OneCycleLR scheduler (Smith and Topin, 2019), and we also use data augmentation with a minibatch of size 128 and a learning rate of 2e-7. Based on these initializations, we ran 10 chains in parallel for SGLD, FALD, and VR-FALD⋆ with step-sizes of 1e-7, 2e-8, 1e-8. We considered 1e4 iterations with only one stored sample every 1e3 iterations (we did not keep the initial weights obtained by SGD to make the predictions). For each chain, we can see that Bayesian model averaging increases the accuracy. To compare the behavior of the mentioned algorithms, we compute the accuracy, the agreement, i.e., the percentage of time the top-1 prediction of an algorithm matches that given by the HMC, and the

METHOD	HMC	SGD	DEEP ENS.	SGLD	FALD	VR-FALD*
Accuracy	89.6 ± 0.25	91.57 ± 0.34	91.68 ± 0.17	89.96 ± 0.72	**92.54** ± 0.04	92.03 ± 0.09
Agreement	94.0 ± 0.25	90.99 ± 0.35	91.03 ± 0.43	**92.43** ± 0.03	91.53 ± 0.39	91.12 ± 0.39
$10\times$ TV	0.74 ± 0.03	1.45 ± 0.05	1.49 ± 0.05	**1.03** ± 0.03	1.42 ± 0.01	1.39 ± 0.01
$10^2\times$ECE	5.9±NA	4.71 ± 1.35	5.44 ± 0.67	4.41 ± 0.37	3.79 ± 0.11	**3.26** ± 0.09
$10\times$BS	1.4±NA	1.69 ± 0.11	1.45 ± 0.10	1.53 ± 0.10	**1.16** ± 0.03	1.20 ± 0.03
$10\times$nNLL	3.07±NA	3.35 ± 0.70	3.81 ± 0.51	3.15 ± 0.21	2.75 ± 0.04	**2.63** ± 0.04

Table 3.7 – Performance of Bayesian FL algorithms on CIFAR10.

total variation (TV) between the predictive distribution given by an algorithm with the one associated with the HMC sampler. We also give some classical calibration scores (Guo et al., 2017), such as the expected calibration error (ECE), the Brier score (BS), and the negative log-likelihood (nNLL).

Part III

Federated Uncertainty Quantification via Bayesian & Frequentist approaches

"Uncertainty is the fuel of curiosity and the foundation of exploration."

Chapter 4

QLSD: Quantized Langevin Stochastic Dynamics

Contents

The objective of Federated Learning (FL) is to perform statistical inference for data which are decentralized and stored locally on networked clients. FL raises many constraints which include privacy and data ownership, communication overhead, statistical heterogeneity, and partial client participation. In this chapter, we address these problems in the framework of the Bayesian paradigm. To this end, we propose a novel federated Markov Chain Monte Carlo algorithm, referred to as Quantized Langevin Stochastic Dynamics which may be seen as an extension to the FL setting of Stochastic Gradient Langevin Dynamics, which handles the communication bottleneck using gradient compression. To improve performance, we then introduce variance reduction techniques, which lead to two improved versions coined QLSD* and QLSD^{++}. We give both non-asymptotic and asymptotic convergence guarantees for the proposed algorithms. We illustrate their performances using various Bayesian Federated Learning benchmarks.

4.1 Introduction

A paradigm shift has occurred with *Federated Learning* (FL) (McMahan et al., 2017; Kairouz et al., 2021). In FL, multiple entities (called clients) which own locally stored data collaborate in learning a "global" model which can then be "adapted" to each client. In the canonical FL, this task is coordinated by a central server. The initial focus of FL was on mobile and edge device applications, but recently there has been a surge of interest in applying the FL framework to other scenarios; in particular, those involving a small number of trusted clients (*e.g.* multiple organisations, enterprises, or other stakeholders).

Table 4.1 – Overview of the main existing distributed/federated approximate Bayesian approaches. Column *Comm. overhead* gives the scheme employed to address the communication bottleneck. Column *Heterogeneity* means that the proposed approach tackles the impact of data heterogeneity on convergence while column *Bounds* highlights available non-asymptotic convergence guarantees.

Method	Comm. overhead	Heterogeneity	Partial participation	Bounds
Hasenclever et al. (2017)	local steps	✗	✗	✗
Nemeth and Sherlock (2018)	one-shot	✗	✗	✗
Bui et al. (2018)	local steps	✗	✓	✗
Jordan et al. (2019)	one-shot	✗	✗	✓
Corinzia et al. (2019)	local steps	✗	✓	✗
Kassab and Simeone (2022)	local steps	✗	✓	✗
El Mekkaoui et al. (2021)	local steps	✗	✗	✓
Plassier et al. (2021)	local steps	✗	✗	✓
Chen and Chao (2021)	local steps	✓	✓	✗
Liu and Simeone (2021a)	one-shot	✗	✗	✗
This work	compression	✓	✓	✓

FL has become one of the most active areas of artificial intelligence research over the past 5 years. FL differs significantly from the classical (distributed) ML setup (McMahan et al., 2017): the storage, computational, and communication capacities of each client vary amongst each other. This poses considerable challenges to successfully deal with many constraints raised by (i) partial client participation (*e.g.* in mobile applications, a client is not always active); (ii) communication bottleneck (clients are communication-constrained with limited bandwidth usage); (iii) model update synchronization and merging.

Many methods derived from stochastic gradient descent techniques have been proposed in the literature to meet the specific FL constraints (McMahan et al., 2017; Alistarh et al., 2017; Horváth et al., 2022; Karimireddy et al., 2020; Li et al., 2020b; Philippenko and Dieuleveut, 2020), see Wang et al. (2021) for a recent comprehensive overview. Whilst these approaches have successfully solved important issues associated to FL, they are unfortunately unable to capture and quantify epistemic predictive uncertainty which is essential in many applications such as autonomous driving or precision medicine (Hunter, 2016; Franchi et al., 2020). Indeed, these methods only provide a point estimate being a minimizer of a target empirical risk function. In contrast, the Bayesian paradigm (Robert, 2001) stands for a natural candidate to quantify uncertainty by providing a full description of the posterior distribution of the parameter of interest, and as such has become ubiquitous in the machine learning community (Andrieu et al., 2003; Hoffman et al., 2013; Izmailov et al., 2020, 2021).

In the last decade, many research efforts have been made to adapt serial workhorses of Bayesian computational methods such as variational inference, expectation-propagation, and Markov chain Monte Carlo (MCMC) algorithms to massively distributed architectures (Wang and Dunson, 2013; Ahn et al., 2014; Wang et al., 2015; Hasenclever et al., 2017; Bui et al., 2018; Jordan et al., 2019; Rendell et al., 2020; Vono et al., 2022a). Since the main bottleneck in distributed computing is the communication overhead, these approaches mainly focus on deriving efficient algorithms specifically designed to meet such a constraint, requiring only periodic or few rounds of communication between

a central server and clients; see Plassier et al. (2021, Section 4) for a recent overview. As highlighted in Table 4.1, most current Bayesian FL methods adapt these approaches and focus almost exclusively on Federated Averaging type updates (McMahan et al., 2017), performing multiple local steps on each client. This is in contrast with predictive FL algorithms (which are **not** estimating predictive uncertainty), for which a variety of schemes have been explored, *e.g.* via gradient compression or client subsampling (Wang et al., 2021, Section 3.1.2). Moreover, very few Bayesian FL works have attempted to address the challenges raised by partial device participation or the impact of statistical heterogeneity; see Liu and Simeone (2021b); Chen and Chao (2021). Convergence results in Bayesian FL lag far behind "canonical" FL.

In this chapter, we attempt to fill this gap, by proposing novel MCMC methods that extend Stochastic Langevin Dynamics to the FL context. It is assumed that the clients' data are independent and that the global posterior density is therefore the product of the *non-identical* local posterior densities of each client. To meet the specificity of Bayesian FL, each iteration of the proposed approaches only requires that a subset of active clients compute a stochastic gradient oracle for their associated negative log posterior density and send a lossy compression of these stochastic gradient oracles to the central server. The first scheme we derive, referred to as *Quantized Langevin Stochastic Dynamics* (`QLSD`), can interestingly be seen as the MCMC counterpart of the `QSGD` approach in FL (Alistarh et al., 2017), just as the Stochastic Gradient Langevin Dynamics (`SGLD`) (Welling and Teh, 2011) extends the Stochastic Gradient Descent (`SGD`). However, `QLSD` has the same drawbacks as `SGLD`: in particular, the invariant distribution of `QLSD` may deviate from the target distribution and become similar to the invariant measure of `SGD` when the number of observations is large (Brosse et al., 2018). We overcome this problem by deriving two variance-reduced versions $\texttt{QLSD}^\star$ and $\texttt{QLSD}^{++}$ that both include control variates.

Contributions. (1) We propose a general MCMC algorithm called `QLSD` specifically designed for Bayesian inference under the FL paradigm and two variance-reduced alternatives, especially tackling *heterogeneity*, *communication overhead* and *partial participation*. (2) We provide a non-asymptotic convergence analysis of the proposed algorithms. The theoretical analysis highlights the impact of statistical heterogeneity measured by the discrepancy between local posterior distributions. (3) We propose efficient mechanisms to mitigate the impact of statistical heterogeneity on convergence, either by using biased stochastic gradients or by introducing a *memory* mechanism that extends Horváth et al. (2022) to the Bayesian setting. In particular, we find that variance reduction indeed allows the proposed MCMC algorithm to converge towards the desired target posterior distribution when the number of observations becomes large. (4) We illustrate the advantages of the proposed methods using several FL benchmarks. We show that the proposed methodology performs well compared to state-of-the-art Bayesian FL methods.

Notations and Conventions. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$, and we set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^*$, we refer to $\{1, \dots, n\}$ with the notation $[n]$. For $N \in \mathbb{N}^*$, we use $\wp_N$ to denote the power set of $[N]$ and define $\wp_{N,n} = \{x \in \wp_N : \mathrm{Card}(x) = n\}$ for any $n \in [N]$. We denote by $\mathrm{N}(m, \Sigma)$ the Gaussian distribution with mean vector m and covariance matrix Σ. We define the sign function, for any $x \in \mathbb{R}$, as $\mathrm{sign}(x) = \mathbf{1}\{x \geq 0\} - \mathbf{1}\{x < 0\}$, and define the Wasserstein distance of

order 2 for any probability measures μ, ν on $\mathbb{R}^d$ with finite 2-moment by $W_2(\mu,\nu) = (\inf_{\zeta\in\mathcal{T}(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\theta-\theta'\|^2 \mathcal{D}\zeta(\theta,\theta'))^{1/2}$, where $\mathcal{T}(\mu,\nu)$ is the set of transference plans of μ and ν.

4.2 Quantized Langevin Stochastic Dynamics

In this section, we present the Bayesian FL framework and introduce the proposed methodology called QLSD along with two variance-reduced instances.

Problem Statement. We are interested in performing Bayesian inference on a parameter $\theta \in \mathbb{R}^d$ based on a training dataset $\mathcal{D}$. We assume that the posterior distribution admits a product-form density with respect to the d-dimensional Lebesgue measure, *i.e.*

$$\pi(\theta \mid \mathcal{D}) = \mathrm{Z}_\pi^{-1} \prod_{i=1}^n \mathrm{e}^{-U_i(\theta)}, \tag{4.1}$$

where $n \in \mathbb{N}^\star$ and $\mathrm{Z}_\pi = \int_{\mathbb{R}^d} \prod_{i=1}^n \mathrm{e}^{-U_i(\theta)} \,\mathrm{d}\theta$ is a normalization constant. This framework naturally encompasses the considered Bayesian FL problem. In this context, $\{\mathrm{e}^{-U_i}\}_{i\in[n]}$ stand for the unnormalized local posterior density functions associated to n clients, where each client $i \in [n]$ is assumed to own a local dataset $\mathcal{D}_i$ such that $\mathcal{D} = \sqcup_{i=1}^n \mathcal{D}_i$. The dependency of U_i on the local dataset $\mathcal{D}_i$ is omitted for brevity. A real-world illustration of the considered Bayesian problem is "multi-site fMRI classification" where each site (or client) owns a dataset coming from a local distribution because the methods of data generation and collection differ between sites. This results in different local likelihood functions, which combined with a local prior distribution, lead to heterogeneous local posteriors.

As in embarrassingly parallel MCMC approaches (Neiswanger et al., 2014), (4.1) implicitly assumes that the prior can be factorized across clients, which can always be done although the choice of this factorization is an open question. This product-form formulation can be alleviated by considering a global prior on θ and only calculating its gradient contribution to the central server during computations, see Algorithm 4.5.

A popular approach to sample from a target distribution with density π defined in (4.1) is based on Langevin dynamics with stochastic gradient which, starting from an initial point θ_0, defines a Markov chain $(\theta_k)_{k\in\mathbb{N}}$ by recursion:

$$\forall k \in \mathbb{N}, \qquad \theta_{k+1} = \theta_k - \gamma H_{k+1}(\theta_k) + \sqrt{2\gamma} Z_{k+1}, \tag{4.2}$$

where $\gamma \in \left(0, \bar{\gamma}\right]$, for some $\bar{\gamma} > 0$, is a discretization time step, $(Z_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. standard Gaussian random variables and $(H_k)_{k\in\mathbb{N}^\star}$ stand for unbiased estimators of ∇U with $U = \sum_{i=1}^n U_i$ (Parisi, 1981; Grenander and Miller, 1994; Roberts and Tweedie, 1996). In a serial setting involving a single client which owns a dataset of size $N \in \mathbb{N}^\star$, the potential U writes $U = U_1 = \sum_{j=1}^N U_{1,j}$ for some functions $U_{1,j} : \mathbb{R}^d \to \mathbb{R}$, and a popular instance of this framework is SGLD (Welling and Teh, 2011). This algorithm consists in the recursion (4.2) with the specific choice $H_{k+1}(\theta) = (N/n) \sum_{j\in\mathcal{S}_{k+1}} \nabla U_{1,j}(\theta)$, where $(\mathcal{S}_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. uniform random subsets of $[N]$ of cardinal n.

In the FL framework, we assume that at each iteration k, the i-th client has access to an oracle $H_{k+1}^{(i)}$ based on its local negative log posterior density U_i, depending only on $\mathcal{D}_i$, so that $H_{k+1} = \sum_{i=1}^n H_{k+1}^{(i)}$ is a stochastic gradient oracle of U. Note that we do

not assume that $H_{k+1}^{(i)}$ is an unbiased estimator of ∇U_i, but only assume that H_{k+1} is unbiased. This allows us to consider biased local stochastic gradient oracles with better convergence guarantees, see Section 4.3 for more details. A simple adaptation of SGLD to the FL framework under consideration is given by recursion:

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^n H_{k+1}^{(i)}(\theta_k) + \sqrt{2\gamma} Z_{k+1}, \qquad k \in \mathbb{N}. \tag{4.3}$$

If for any $i \in [n]$, every potential function U_i also admits a finite-sum expression *i.e.* $U_i = \sum_{j=1}^{N_i} U_{i,j}$, similar to SGLD, we can for example use the local stochastic gradient oracles $H_{k+1}^{(i)}(\theta) = (N_i/b_i) \sum_{j \in \mathcal{S}_{k+1}^{(i)}} \nabla U_{i,j}(\theta)$, where $(\mathcal{S}_{k+1}^{(i)})_{k \in \mathbb{N}^*, i \in [n]}$ stand for i.i.d. uniform random subsets of $[N_i]$ of cardinal b_i. However, considering the MCMC algorithm associated with the recursion (4.3) is not adapted to the FL context. Indeed, this algorithm would assume that each client is reliable and suffers from the same issues as SGD in a risk-based minimization context, especially a prohibitive communication overhead (Girgis et al., 2020).

Proposed Methodology. To address this problem, we propose to both account for the *partial participation of clients* and *reduce the number of bits transmitted* during the upload period by performing a lossy compression of a subset of $\{H_{k+1}^{(i)}\}_{i \in [n], k \in \mathbb{N}^\star}$. This method has been used extensively in the "canonical" FL literature (Alistarh et al., 2017; Lin et al., 2018; Haddadpour et al., 2021; Sattler et al., 2020), but interestingly has never been considered in Bayesian FL; see Table 4.1.

To this end, we introduce a compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ that is unbiased, *i.e.* for any $v \in \mathbb{R}^d$, $\mathbb{E}[\mathcal{C}(v)] = v$. In recent years, numerous compression operators have been proposed (Seide et al., 2014; Aji and Heafield, 2017; Stich et al., 2018). For example, the QSGD approach proposed in Alistarh et al. (2017) is based on stochastic quantization.

QSGD considers for $\mathcal{C}$ a component-wise quantization operator parameterized by a number of quantization levels $s \geq 1$, which for each $j \in [d]$ and $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ are given by

$$\mathcal{C}^{(s,j)}(v) = \frac{\|v\| \operatorname{sign}(v_j)}{s} \left(l_j + \mathbf{1} \left[\xi_j \leq \frac{s|v_j|}{\|v\|} - l_j \right] \right), \tag{4.4}$$

where $l_j = \left\lfloor s|v_j|/\|v\| \right\rfloor$ and $\{\xi_j\}_{j \in [d]}$ is a sequence of i.i.d. uniform random variables on $[0,1]$. In this particular case, we will denote the quantization of v via (4.4) by $\mathcal{C}^{(s)}(v) = \{\mathcal{C}^{(s,j)}(v)\}_{j \in [d]}$.

The proposed general methodology, called *Quantized Langevin Stochastic Dynamics* (QLSD) stands for a compressed and FL version of the specific instance of SGLD defined in (4.3). More precisely, QLSD is an MCMC algorithm associated with the Markov chain $(\theta_k)_{k \in \mathbb{N}}$ starting from θ_0 and defined for $k \in \mathbb{N}$ as

$$\theta_{k+1} = \theta_k - \gamma \frac{n}{|\mathcal{A}_{k+1}|} \sum_{i \in \mathcal{A}_{k+1}} \mathcal{C}_{k+1} \left[H_{k+1}^{(i)}(\theta_k) \right] + \sqrt{2\gamma} Z_{k+1},$$

where $(\mathcal{A}_k)_{k \in \mathbb{N}^\star}$ denotes the subset of active (*i.e.* available) clients at iteration k, possibly random. Note that we indexed $\mathcal{C}$ by $k+1$ to emphasize that this compression operator is a stochastic operator and hence varies across iterations, see *e.g.* (4.4). The

derivation of QLSD in the considered Bayesian FL context is described in details in Algorithm 4.5. A generalization of QLSD taking into account *heterogeneous communication constraints* between clients by considering different compression operators $\{\mathcal{C}^{(i)}\}_{i\in[n]}$ is available in Section 4.A. In the particular case of the finite-sum setting where each client owns a dataset of size N_i, *i.e.* for the choice $H_{k+1}^{(i)}(\theta) = (N_i/b_i)\sum_{j\in\mathcal{S}_{k+1}^{(i)}} \nabla U_{i,j}(\theta)$ for $\theta \in \mathbb{R}^d$, $\mathcal{S}_{k+1}^{(i)} \in \wp_{N_i,b_i}$, we denote the corresponding instance of QLSD as QLSD$^{\#}$.

In this chapter, we have decided to focus only on a non-adjusted sampling algorithm (QLSD) since the derivations of non-asymptotic results are already consequent. Moreover, up to authors'knowledge, a consensus on the choice between Metropolis-adjusted algorithms and their unadjusted counterparts has not been achieved yet.

Algorithm 4.5 Quantized Langevin Stochastic Dynamics (QLSD)

Input: number of iterations K, compression operators $\{\mathcal{C}_{k+1}\}_{k\in\mathbb{N}}$, stochastic gradients $\{H_{k+1}^{(i)}\}_{i\in[n],k\in\mathbb{N}}$, step-size $\gamma \in (0, \bar{\gamma}]$ and initial point θ_0.
for $k = 0$ **to** $K-1$ **do**
  for $i \in \mathcal{A}_{k+1}$ **do** // On active clients $\mathcal{A}_{k+1}$
    Compute $g_{i,k+1} = \mathcal{C}_{k+1}[H_{k+1}^{(i)}(\theta_k)]$.
    Send $g_{i,k+1}$ to the central server.
  // On the central server
  Compute $g_{k+1} = \frac{n}{|\mathcal{A}_{k+1}|}\sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the n clients.
Output: samples $\{\theta_k\}_{k=0}^K$.

Variance-Reduced Alternatives. Consider the finite-sum setting *i.e.* for any $i \in [n]$, $U_i = \sum_{j=1}^{N_i} U_{i,j}$ where N_i is the size of the local dataset $\mathcal{D}_i$. As highlighted in Section 4.1, SGLD-based approaches, including Algorithm 4.5, involve an invariant distribution that may deviate from the target posterior distribution when $\min_{i\in b} N_i$ goes to infinity, as stochastic gradients with large variance are used (Brosse et al., 2018; Baker et al., 2019). We deal with this problem by proposing two variance-reduced alternatives of QLSD$^{\#}$ that use control variates. The simplest variance-reduced approach, referred to as QLSD* (see Algorithm 4.7) and discussed in more details in Section 4.B, considers a fixed-point approach that uses a minimizer $\theta^{\star}$ of the potential U (Brosse et al., 2018; Baker et al., 2019) defined as

$$\theta^{\star} \in \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} U_i(\theta). \tag{4.5}$$

In this scenario, the stochastic gradient oracles write for each $i \in [n]$, $k \in \mathbb{N}^{\star}$, $\theta \in \mathbb{R}^d$ and $\mathcal{S}_{k+1}^{(i)} \in \wp_{N_i,b_i}$, $H_{k+1}^{(i)}(\theta) = (N_i/b_i)\sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^{\star})]$. Although $\mathbb{E}[H_{k+1}] = \nabla U$, note that for each $i \in [n]$, $\mathbb{E}[H_{k+1}^{(i)}] \neq \nabla U_i$ so $H_{k+1}^{(i)}$ is not an unbiased estimate of ∇U_i. We show in Section 4.3 that introducing this bias improves the convergence properties of QLSD$^{\#}$ with respect to the discrepancy between local posterior distributions. Since estimating $\theta^{\star}$ in a FL context might impose an additional computational burden on the sampling procedure, we propose another variance-

reduced alternative referred to as QLSD^{++} (see Algorithm 4.6). This method builds on the Stochastic Variance Reduced Gradient (SVRG): it uses control variates $(\zeta_k)_{k\in\mathbb{N}}$ that are updated every $l \in \mathbb{N}^\star$ iterations (Johnson and Zhang, 2013) and at each iteration $k \in \mathbb{N}$ and for any client $i \in [n]$, the stochastic gradient oracle $H_{k+1}^{(i)}$ defined by $H_{k+1}^{(i)}(\theta) = (N_i/b_i)\sum_{j\in\mathcal{S}_{k+1}^{(i)}}[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\zeta_k)] + \nabla U_i(\zeta_k)$. To reduce the impact of local posterior discrepancy on convergence, we take inspiration from the "canonical" FL literature and consider a *memory term* $(\eta_k^{(i)})_{k\in\mathbb{N}}$ on each client $i \in [n]$ (Horváth et al., 2022; Dieuleveut et al., 2020). At each iteration k, instead of directly compressing $H_{k+1}^{(i)}$, we compress the difference $H_{k+1}^{(i)} - \eta_k^{(i)}$, store it in $g_{i,k+1}$, and then compute the global stochastic gradient $g_{k+1} = \frac{n}{|\mathcal{A}_{k+1}|}\sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1} + \sum_{i=1}^n \eta_k^{(i)}$. The memory term $(\eta_k^{(i)})_{k\in\mathbb{N}}$ is then updated on each client $i \in [n]$, by the recursion $\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha \mathbf{1}_{\mathcal{A}_{k+1}}(i) g_{i,k+1}$. The benefits of using this memory mechanism will be assessed theoretically in Section 4.3 and illustrated numerically in Figure 4.5.

Algorithm 4.6 Variance-reduced Quantized Langevin Stochastic Dynamics (QLSD^{++})

Input: minibatch sizes $\{b_i\}_{i\in[n]}$, number of iterations K, compression operators $\{\mathcal{C}_{k+1}\}_{k\in\mathbb{N}^\star}$, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, initial point θ_0 and $\alpha \in (0, \bar{\alpha}]$ with $\bar{\alpha} > 0$.

// Memory mechanism initialization

Initialize $\{\eta_0^{(1)}, \ldots, \eta_0^{(n)}\}$ and $\eta_0 = \sum_{i=1}^n \eta_0^{(i)}$.

for $k = 0$ **to** $K - 1$ **do**

  // Update of the control variates

  if $k \equiv 0 \ (\mathrm{mod} l)$ **then**

    Set $\zeta_k = \theta_k$.

  else

    Set $\zeta_k = \zeta_{k-1}$

  for $i \in \mathcal{A}_{k+1}$ **do** // On active clients

    Draw $\mathcal{S}_{k+1}^{(i)} \sim \mathrm{Uniform}(\wp_{N_i,b_i})$.

    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/b_i)\sum_{j\in\mathcal{S}_{k+1}^{(i)}}[\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\zeta_k)] + \nabla U_i(\zeta_k)$.

    Compute $g_{i,k+1} = \mathcal{C}_{k+1}(H_{k+1}^{(i)}(\theta_k) - \eta_k^{(i)})$.

    Send $g_{i,k+1}$ to the central server.

    Set $\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha g_{i,k+1}$.

  // On the central server

  Compute $g_{k+1} = \eta_k + \frac{n}{|\mathcal{A}_{k+1}|}\sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.

  Set $\eta_{k+1} = \eta_k + \alpha\sum_{i\in\mathcal{A}_{k+1}}^n g_{i,k+1}$.

  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.

  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.

  Send θ_{k+1} to the n clients.

Output: samples $\{\theta_k\}_{k=0}^K$.

4.3 Theoretical analysis

This section provides a detailed theoretical analysis of the proposed methodology. In particular, we will show the *impact of using stochastic gradients*, *partial participation* and *compression* by deriving quantitative convergence bounds for QLSD, which is detailed in Algorithm 4.5. We then derive non-asymptotic convergence bounds for QLSD* and QLSD^{++}, and explicitly show that these variance-reduced algorithms indeed succeed in reducing both the variance caused by stochastic gradients and the effects of *local posterior discrepancy* in the bounds we obtain for QLSD$^\#$. We consider the following assumptions on the potential U.

Assumption 4.1. *For any $i \in [n]$, U_i is continuously differentiable. In addition, suppose that the following hold.*

(i) U is $\mathtt{m}$-strongly convex, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\langle \nabla U(\theta_1) - \nabla U(\theta_2), \theta_1 - \theta_2 \rangle \geq \mathtt{m} \left\|\theta_1 - \theta_2\right\|^2$.*
(ii) U is $\mathtt{L}$-Lipschitz, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla U\left(\theta_1\right) - \nabla U\left(\theta_2\right)\| \leq \mathtt{L}\|\theta_1 - \theta_2\|$.*

Note that Assumption 4.1-(i) implies that U admits a unique minimizer denoted by $\theta^\star \in \mathbb{R}^d$.

The compression operators $\{\mathcal{C}_{k+1}\}_{k\in\mathbb{N}}$ are assumed to satisfy the following assumption.

Assumption 4.2. *The compression operators $\{\mathcal{C}_{k+1}\}_{k\in\mathbb{N}}$ are independent and satisfy the following conditions.*

(i) For any $k \in \mathbb{N}^\star$, $v \in \mathbb{R}^d$, $\mathbb{E}[\mathcal{C}_k(v)] = v$.
(ii) There exists $\omega \geq 1$, such that for any $k \in \mathbb{N}^\star$, $v \in \mathbb{R}^d$, $\mathbb{E}[\|\mathcal{C}_k(v) - v\|^2] \leq \omega \|v\|^2$.

As an example, the assumption on the variance of the compression operator detailed in Assumption 4.2-(ii) is verified for the quantization operator $\mathcal{C}^{(s)}$ defined in (4.4) with $\omega = \min(d/s^2, \sqrt{d}/s)$ (Alistarh et al., 2017, Lemma 3.1).

Non-Asymptotic Analysis for Algorithm 4.5. We consider the following assumptions on the stochastic gradient oracles used in QLSD.

Assumption 4.3. *The random fields $\{H_{k+1}^{(i)} : \mathbb{R}^d \to \mathbb{R}^d\}_{i\in[n],k\in\mathbb{N}}$ are independent and satisfy the following conditions.*

(i) For any $\theta \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $\sum_{i=1}^n \mathbb{E}[H_{k+1}^{(i)}(\theta)] = \nabla U\left(\theta\right)$.
(ii) There exist $\{\mathtt{M}_i > 0\}_{i\in[n]}$, such that for any $i \in [n]$, $k \in \mathbb{N}$, $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|H_{k+1}^{(i)}(\theta_1) - H_{k+1}^{(i)}(\theta_2)\right\|^2\right] \leq \mathtt{M}_i \left\langle \theta_1 - \theta_2, \nabla U_i\left(\theta_1\right) - \nabla U_i\left(\theta_2\right)\right\rangle.$$

(iii) There exist $\sigma_\star, \mathtt{B}^\star \in \mathbb{R}_+$ such that for any $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}$, we have $\mathbb{E}[\|H_{k+1}^{(i)}(\theta^\star)\|]^2 \leq \mathtt{B}^\star/n$, and $\mathbb{E}[\|\sum_{i=1}^n H_{k+1}^{(i)}(\theta^\star)\|^2] \leq \sigma_\star^2$, where $\theta^\star$ is defined in (4.5).

Table 4.2 – Order of the asymptotic biases $\{B_{\bar\gamma}, B_{\circledast,\bar\gamma}, B_{\oplus,\bar\gamma}\}$, associated to the three proposed MCMC algorithms, in squared 2-Wasserstein distance for two types of asymptotic. Red dependencies prevent from (quick) convergence while green dependencies ensure convergence of associated MCMC algorithms. $\theta^\star$ is defined in (4.5).

Algo.	Bias	Dependencies of the asymptotic bias when $\bar\gamma \downarrow 0$					Dependencies of the asymptotic bias as $N_i \to \infty$
		d	$H_{k+1}^{(i)}$	$\mathtt{B}^\star$	partial particip.	ω	
QLSD	$B_{\bar\gamma}$	d	$\sigma_\star^2$	$\mathtt{B}^\star$	$(1-p)/p$	ω	$\mathrm{O}(N_i)$
QLSD$^\#$	$B_{\bar\gamma}$	d	N_i^2	$\sum_{i=1}^n \left\|\nabla U_i(\theta^\star)\right\|^2$	$(1-p)/p$	ω	$\mathrm{O}(N_i)$
QLSD*	$B_{\circledast,\bar\gamma}$	d	N_i	-	$(1-p)/p$	ω	$d\,\mathrm{O}(1)$
QLSD^{++}	$B_{\oplus,\bar\gamma}$	d	N_i	-	$(1-p)/p$	ω	$d\,\mathrm{O}(1)$

We can notice that Assumption 4.3-(ii) implies that ∇U_i is $\mathtt{M}_i$-Lipschitz continuous since by the Cauchy-Schwarz inequality, for any $i \in [n]$ and any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have that $\|\nabla U_i(\theta_1) - \nabla U_i(\theta_2)\|^2 \leq \mathtt{M}_i \langle \theta_1 - \theta_2, \nabla U_i(\theta_1) - \nabla U_i(\theta_2)\rangle$. Conversely, in the finite-sum setting, Assumption 4.3-(ii) is satisfied by QLSD$^\#$ with $\mathtt{M}_i = N_i \bar{\mathtt{M}}$ if for any $i \in [n]$ and $j \in [N_i]$, $U_{i,j}$ is convex and $\nabla U_{i,j}$ is $\bar{\mathtt{M}}$-Lipschitz continuous, for $\bar{\mathtt{M}} \geq 0$ by Nesterov (2003, Theorem 2.1.5).

In addition, it is worth mentioning that the first inequality in Assumption 4.3-(iii) is also required for our derivation in the deterministic case where $H_{k+1}^{(i)} = \nabla U_i$ due to the compression operator. In this particular case, $\mathtt{B}^\star$ stands for an upper-bound on $\sum_{i=1}^n \|\nabla U_i(\theta^\star)\|^2$ and corresponds to some discrepancy between local posterior density functions meaning that $\nabla U_i \neq \nabla U$ for $i \in [n]$. This phenomenon, referred to as *data heterogeneity* in the risk-based literature (Horváth et al., 2022; Karimireddy et al., 2020), is ubiquitous in the FL context.

Finally, we assume for simplicity that *clients' partial participation* is realized by each client having probability $p \in (0, 1]$ of being active in each communication round.

Assumption 4.4. *For any $k \in \mathbb{N}^\star$, $\mathcal{A}_k = \{i \in [n] : B_{i,k} = 1\}$ where $\{B_{i,k} : i \in [n], k \in \mathbb{N}^*\}$ is a family of i.i.d. Bernoulli random variables with success probability $p \in (0, 1]$.*

A generalization of this scheme considering different probabilities p_i per client can be found in Section 4.A.1. Under the above assumptions and by denoting Q_γ the Markov kernel associated to Algorithm 4.5, the following convergence result holds.

Theorem 4.5. *Assume Assumption 4.1, Assumption 4.2, Assumption 4.3 and Assumption 4.4. Then, there exists $\bar\gamma_\infty$ such that for $\bar\gamma < \bar\gamma_\infty$, there exist $A_{\bar\gamma}, B_{\bar\gamma} > 0$, explicitly given in Section 4.A, satisfying for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step-size $\gamma \in (0, \bar\gamma]$ and $k \in \mathbb{N}$,*

$$W_2^2\left(\mu Q_\gamma^k, \pi\right) \leq (1 - \gamma \mathtt{m}/2)^k \cdot W_2^2(\mu, \pi) + \gamma B_{\bar\gamma} + \gamma^2 A_{\bar\gamma} (1 - \mathtt{m}\gamma/2)^{k-1} k \cdot \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta),$$

where $\theta^\star$ is defined in (4.5).

Similar to ULA (Dalalyan, 2017b; Durmus and Moulines, 2019) and SGLD (Dalalyan and Karagulyan, 2019; Durmus et al., 2019), the upper bound given in Theorem 4.5 includes a contracting term that depends on the initialization and a bias term $\gamma B_{\bar{\gamma}}$ that does not vanish with $k \to \infty$ due to the use of a fixed step-size γ. In the asymptotic scenario, *i.e.* $\bar{\gamma} \downarrow 0$, Table 4.1 gives the dependencies of $B_{\bar{\gamma}}$ for QLSD and its particular instance $\texttt{QLSD}^{\#}$, in terms of key quantities associated with the setting we consider. Similar to SGLD, we can observe that the use of stochastic gradients entails a bias term of order $\sigma_\star^2\,\mathrm{O}(\gamma)$. On the other hand, the use of partial participation and compression compared to SGLD introduces an *additional bias* of order $(\omega/p)(\mathtt{m}\mathtt{B}^\star + \mathtt{L}\mathtt{M}d)\,\mathrm{O}(\gamma)$, which grows with in particular $\mathtt{B}^\star$, corresponding to the impact of the local posterior discrepancy on convergence.

Non-Asymptotic Analysis for Variance-Reduced Alternatives. We assume in the sequel that the potential functions $\{U_i\}_{i\in[n]}$ admit the finite-sum decomposition $U_i = \sum_{j=1}^{N_i} U_{i,j}$ for each $i \in [n]$ and consider the following assumptions.

Assumption 4.6. *For any $i \in [n], j \in [N_i]$, $U_{i,j}$ is continuously differentiable and the following holds.*

(i) There exists $\mathtt{M}_i > 0$ such that, for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\|\nabla U_i(\theta_2) - \nabla U_i(\theta_1)\right\|^2 \le \mathtt{M}\left\langle \theta_2 - \theta_1, \nabla U_i(\theta_2) - \nabla U_i(\theta_1)\right\rangle .$$

(ii) There exists $\bar{\mathtt{M}} \ge 0$ such that, for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\|\nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1)\right\|^2 \le \bar{\mathtt{M}}\left\langle \nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1), \theta_2 - \theta_1\right\rangle .$$

As mentioned earlier, Assumption 4.6 is satisfied if for every $i \in [n]$ and $j \in [N_i]$, $U_{i,j}$ is convex and $\nabla U_{i,j}$ is $\bar{\mathtt{M}}$-Lipschitz continuous. Under these additional conditions, the following non-asymptotic convergence results hold for the two reduced-variance MCMC algorithms described in Section 4.2. Denote by $Q_{\circledast,\gamma}$ the Markov kernel associated to $\texttt{QLSD}^\star$ with a step-size $\gamma \in \left(0, \bar{\gamma}\right]$.

Theorem 4.7. *Assume Assumption 4.1, Assumption 4.2, Assumption 4.4 and Assumption 4.6. Then, there exists $\bar{\gamma}_{\circledast,\infty}$ such that for $\bar{\gamma} < \bar{\gamma}_{\circledast,\infty}$, there exist $A_{\circledast,\bar{\gamma}}, B_{\circledast,\bar{\gamma}} > 0$, explicitly given in Section 4.B, satisfying for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step-size $\gamma \in \left(0, \bar{\gamma}\right]$ and $k \in \mathbb{N}$,*

$$W_2^2\left(\mu Q_{\circledast,\gamma}^k, \pi\right) \le (1 - \gamma\mathtt{m}/2)^k \cdot W_2^2\left(\mu, \pi\right) + \gamma B_{\circledast,\bar{\gamma}} + \gamma^2 A_{\circledast,\bar{\gamma}}(1 - \mathtt{m}\gamma/2)^{k-1} k \cdot \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta),$$

where $\theta^\star$ is defined in (4.5).

Compared to QLSD and $\texttt{QLSD}^\star$, $\texttt{QLSD}^{++}$ only defines an inhomogeneous Markov chain, see Section 4.C.3 for more details. For a step-size $\gamma \in \left(0, \bar{\gamma}\right]$ and an iteration $k \in \mathbb{N}$, we denote by $\mu Q_{\oplus,\gamma}^{(k)}$ the distribution of θ_k defined by $\texttt{QLSD}^{++}$ starting from θ_0 with distribution μ.

Theorem 4.8. *Assume Assumption 4.1, Assumption 4.2, Assumption 4.4 and Assumption 4.6, and let $l \in \mathbb{N}^\star$ and $\alpha \in (0, 1/(\omega+1)]$. Then, there exists $\bar{\gamma}_{\oplus,\infty}$ such that for $\bar{\gamma} < \bar{\gamma}_{\oplus,\infty}$, there exist $A_{\oplus,\bar{\gamma}}, B_{\oplus,\bar{\gamma}}, C_{\oplus,\bar{\gamma}} > 0$, explicitly given in Section 4.C, satisfying for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step-size $\gamma \in \left(0, \bar{\gamma}\right]$ and $k \in \mathbb{N}$,*

$$W_2^2(\mu Q_{\oplus,\gamma}^{(k)}, \pi) \leq (1-\gamma\mathfrak{m}/2)^k \cdot W_2^2\left(\mu, \pi\right) + \gamma^2 A_{\oplus,\bar{\gamma}} (1-\gamma\mathfrak{m}/2)^{\lfloor k/l \rfloor} \cdot \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta)$$
$$+ \gamma B_{\oplus,\bar{\gamma}} + \gamma C_{\oplus,\bar{\gamma}} [(1-\alpha)^k \wedge (1-\gamma\mathfrak{m}/2)^{\lfloor k/l \rfloor}] \sum_{i=1}^n \|\nabla U_i(\theta^\star)\|^2,$$

where $\theta^\star$ is defined in (4.5).

Table 4.2 provides the dependencies of the asymptotic bias terms $B_{\circledast,\bar{\gamma}}, B_{\oplus,\bar{\gamma}}$ as $\bar{\gamma} \downarrow 0$ with respect to key quantities associated to the problem we consider. For comparison, we do the same regarding the specific instance of Algorithm 4.5, QLSD$^\#$. Remarkably, thanks to biased local stochastic gradients for QLSD* and the memory mechanism for QLSD^{++}, we can notice that their associated asymptotic biases do not depend on local posterior discrepancy in contrast to QLSD$^\#$. This is in line with non-asymptotic convergence results in risk-based FL which also show that the impact of data heterogeneity can be alleviated using such a memory mechanism (Philippenko and Dieuleveut, 2020). The impact of stochastic gradients is discussed in further details in the next paragraph.

Consistency Analysis in the Big Data Regime. In Brosse et al. (2018), it was shown that ULA and SGLD define homogeneous Markov chains, each of which admits a unique stationary distribution. However, while the invariant distribution of ULA gets closer to π as N_i increases, conversely the invariant measure of SGLD never approaches π and is in fact very similar to the invariant measure of SGD. Moreover, the non-compressed counterpart of QLSD* has been shown not to suffer from this problem, and it has been theoretically proven to be a viable alternative to ULA in the Big Data environment. Since QLSD is a generalization of SGLD, the conclusions of Brosse et al. (2018) hold. On the other hand, we show that the reduced-variance alternatives to QLSD that we introduced provide more accurate estimates of π as N_i increases, see the last column in Table 4.2. Detailed calculations are deferred to Section 4.D.

4.4 Numerical experiments

This section illustrates our methodology with three numerical experiments that include both synthetic and real datasets. For all experiments, we consider the finite-sum setting and use the stochastic quantization operator $\mathcal{C}^{(s)}$ for $s \geq 1$ defined in (4.4) to perform the compression step. In this case Assumption 4.2-(ii) is verified with $\omega = \min(d/s^2, \sqrt{d}/s)$. Further experimental results are provided in Section 4.E.

Toy Gaussian Example. This first experiment aims at illustrating the general behavior of Algorithm 4.5 with respect to the use of stochastic gradients and compression scheme. To this purpose, we set $n = 20$ and $d = 50$ and consider a Gaussian posterior distribution with density defined in (4.1) where, for any $i \in [n]$ and $\theta \in \mathbb{R}^d$, $U_i(\theta) = \sum_{j=1}^{N_i} \|\theta - y_{i,j}\|^2/2$, $\{y_{i,j}\}_{i\in[n], j\in[N_i]}$ being a set of synthetic independent but

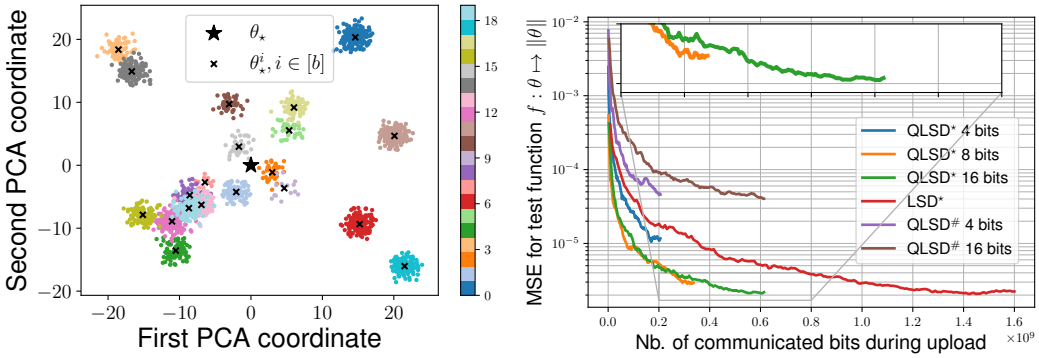


Figure 4.1 – Toy Gaussian example. (top) 2D projection of the heterogeneous synthetic dataset where each color refers to a client and each dot is an observation $y_{i,j}$. (bottom) Estimation performances of the considered Bayesian FL algorithms.

not identically distributed observations across clients and $N_i \in [10, 200]$, see Figure 4.1 (top row). Note that in this specific case, $\theta^\star$ admits a closed form expression. For all the algorithms, we choose the (optimized) step-size $\gamma = 4.9 \times 10^{-4}$ and choose a minibatch size $b_i = \lfloor N_i/10 \rfloor$. Instances of QLSD$^\#$ and QLSD* using $s = 2^p$ are referred to as p-bits instances of these MCMC algorithms. We compare these algorithms with the non-compressed counterpart of QLSD* referred to as LSD*, see Algorithm 4.8. Figure 4.1 shows the behavior of the mean squared error (MSE) associated to the test function $f : \theta \mapsto \|\theta\|$, computed using 30 independent runs of each algorithm, with respect to the number of bits transmitted. We can notice that QLSD* always outperforms QLSD$^\#$and that decreasing the value of ω does not significantly reduce the bias associated to QLSD*. This illustrates the impact of the variance of the stochastic gradients and supports our theoretical analysis summarized in Table 4.2. On the other hand, QLSD* with $s = 2^{16}$ achieves a similar MSE as LSD* while requiring roughly 2.5 times less number of bits.

Bayesian Logistic Regression. In this experiment, we compare the proposed methodology based on gradient compression with two existing FedAvg-type MCMC algorithms. Since $\theta^\star$ defined in (4.5) is not easily available, we implement QLSD^{++} detailed in Algorithm 4.6. We adopt a zero-mean Gaussian prior with covariance matrix $2 \cdot 10^{-2} \mathrm{I}_d$ and use the FEMNIST dataset (Caldas et al., 2018). We set $n = 50$, $l = 100$, $\alpha = 1/(\omega+1)$ and $\gamma = 10^{-5}$. We launch QLSD^{++} for $s \in \{2^4, 2^8, 2^{16}\}$ and compare its performances with DG-SGLD (Plassier et al., 2021) and FSGLD (El Mekkaoui et al., 2021) which use multiple local steps to address the communication bottleneck. We are interested in performing uncertainty quantification by estimating highest posterior density (HPD) regions. For any $\alpha \in (0, 1)$, we define $\mathcal{C}_\alpha = \{\theta \in \mathbb{R}^d; -\log \pi(\theta|\mathcal{D}) \leq \eta_\alpha\}$ where $\eta_\alpha \in \mathbb{R}$ is chosen such that $\int_{\mathcal{C}_\alpha} \pi(\theta|\mathcal{D}) \mathrm{d}\theta = 1 - \alpha$. We compute the relative HPD error based on the scalar summary η_α, *i.e.* $|\eta_\alpha - \eta_\alpha^{\mathrm{LSD}}|/\eta_\alpha^{\mathrm{LSD}}$ where $\eta_\alpha^{\mathrm{LSD}}$ has been estimated using the non-compressed counterpart of QLSD^{++}, referred to as LSD^{++} and standing for a serial variance-reduced SGLD, see Algorithm 4.9. Table 4.3 gives this relative HPD error for $\alpha = 0.01$ and provides the relative efficiency of QLSD^{++} and competitors corresponding to the savings in terms of transmitted bits per iteration. One can notice that the proposed approach provides similar results as its non-compressed counterpart while being 3 to 7 times more efficient. In addition, we show that QLSD^{++} provides similar performances as DG-SGLD and FSGLD which highlight that gradient compression and periodic

Table 4.3 – Bayesian Logistic Regression.

Algorithm	99% HPD error	Rel. efficiency
FSGLD	5.4e-3	6.2
DG-SGLD	5.2e-3	6.4
QLSD^{++} 4 bits	6.1e-3	7.6
QLSD^{++} 8 bits	4.3e-3	6.7
QLSD^{++} 16 bits	6.9e-4	3.1

Table 4.4 – Performances of Bayesian FL algorithms on the considered Bayesian neural networks problem.

Method	HMC	SGLD	QLSD^{++}	QLSD^{++} PP	FedBe-Dirichlet	FedBe-Gauss.	DG-SGLD	FSGLD
Accuracy	89.6	88.8	88.1	86.6	90.7	90.2	92.2	87.5
Agreement	0.94	0.91	0.90	0.90	0.90	0.89	0.91	0.91
TV	0.07	0.11	0.12	0.12	0.16	0.16	0.13	0.13

communication are competing approaches.

Bayesian Neural Networks. In our third experiment, we go beyond the scope of our theoretical analysis by performing posterior inference in Bayesian neural networks. We use the ResNet-20 model (He et al., 2016), choose a zero-mean Gaussian prior distribution with variance 1/5 and consider the classification problem associated with the CIFAR-10 dataset (Krizhevsky et al., 2009). We run QLSD^{++} with $s = 2$, $l = 20$, $\alpha = 1/(\omega + 1)$, and with either $p = 1$ (full participation) or $p = 0.25$ (partial participation). We compare the proposed methodology with a long-run Hamiltonian Monte Carlo (HMC) considered as a "ground truth" (Izmailov et al., 2021) and SGLD. For completeness, we also implement four other distributed/federated approximate sampling approaches, namely two instances of FedBe (Chen and Chao, 2021), DG-SGLD and FSGLD. Following Wilson et al. (2021), we compare the aforementioned algorithms through three metrics: classification *accuracy* on the test dataset using the minimum mean-square estimator, *agreement* between the top-1 prediction given by each algorithm and the one given by HMC and *total variation* between approximate and "true" (associated with HMC) predictive distributions. More details about algorithms' hyperparameters and considered metrics are given in Section 4.E.3. The results we obtain are gathered in Table 4.4. In terms of agreement and total variation, QLSD^{++} (even with partial participation) gives similar results as SGLD and competes favorably with other existing federated approaches. Figure 4.2 complements this empirical analysis by showing calibration curves of posterior predictive distributions.

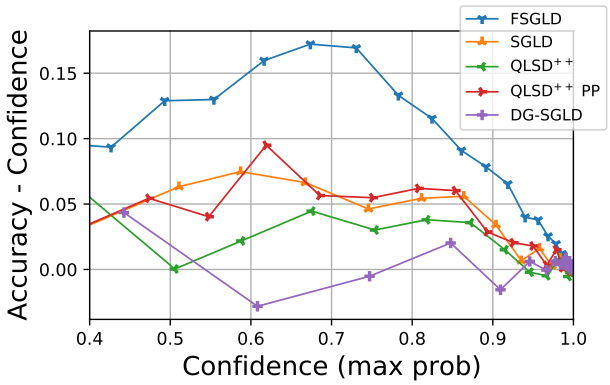


Figure 4.2 – Bayesian Neural Networks.

4.5 Conclusion

In this chapter, we presented a general methodology based on Langevin stochastic dynamics for Bayesian FL. In particular, we addressed the challenges associated with this new ML paradigm by assuming that a subset of clients sends compressed versions of its local stochastic gradient oracles to the central server. Moreover, the proposed method was found to have favorable convergence properties, as evidenced by numerical illustrations. In particular, it compares favorably to `FedAvg`-type Bayesian FL algorithms. A limitation of this work is that the proposed method does not target the initial posterior distribution due to the use of a fixed discretization time step. Therefore, this work paves the way for more advanced Bayesian FL approaches based, for example, on Metropolis-Hastings schemes to remove asymptotic biases. In addition, although the data ownership issue is implicitly tackled by the FL paradigm by not sharing data, stronger privacy guarantees can be ensured, typically by combining differential privacy, secure multi-party computation and homomorphic encryption methods. Proposing a differentially private version of our methodology is a possible extension of our work, that is left for further work. This work has no direct societal impact.

4.A Proof of Theorem 4.5

This section aims at proving Theorem 4.5 in the main chapter.

4.A.1 Generalized quantized Langevin stochastic dynamics

We show that QLSD defined in Algorithm 4.5 in the main chapter can be cast into a more general framework that we refer to as generalized quantized Langevin stochastic dynamics. Then, the guarantees for QLSD will be a simple consequence of the ones that we will establish for generalized QLSD. For ease of reading, we recall first the setting and the assumptions that we consider all along the chapter. Recall that the dataset $\mathcal{D}$ is assumed to be partitioned into n *shards* $\{\mathcal{D}_i\}_{i=1}^n$ such that $\sqcup_{i=1}^n \mathcal{D}_i = \mathcal{D}$ and the posterior distribution of interest is assumed to admit a density with respect to the d-dimensional Lebesgue measure which factorizes across clients, i.e. for any $\theta \in \mathbb{R}^d$,

$$\pi(\theta) = \exp\{-U(\theta)\} / \int_{\mathbb{R}^d} \exp\{-U(\theta)\} \, \mathrm{d}\theta, \qquad U(\theta) = \sum_{i=1}^n U_i(\theta).$$

We consider the following assumptions on the potential U.

Assumption 4.9. *For any $i \in [n]$, U_i is continuously differentiable. In addition, suppose that the following conditions hold.*

(i) U is $\mathtt{m}$-strongly convex, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$U(\theta_1) \geq U(\theta_2) + \left\langle \theta_1 - \theta_2, \nabla U(\theta_2) \right\rangle + \mathtt{m} \left\| \theta_1 - \theta_2 \right\|^2 / 2.$$

(ii) U is $\mathtt{L}$-Lipschitz, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$\left\| \nabla U(\theta_1) - \nabla U(\theta_2) \right\| \leq \mathtt{L} \left\| \theta_1 - \theta_2 \right\|.$$

Note that Assumption 4.9-(i) implies that U admits a unique minimizer denoted by $\theta^\star \in \mathbb{R}^d$. Moreover, for any $(\theta_1, \theta_1) \in \mathbb{R}^d$, Assumption 4.9-(i)-(ii) combined with Nesterov (2003, Equation 2.1.24) shows that

$$\left\langle \nabla U(\theta_2) - \nabla U(\theta_1), \theta_2 - \theta_1 \right\rangle \geq \frac{\mathtt{mL}}{\mathtt{m} + \mathtt{L}} \left\| \theta_2 - \theta_1 \right\|^2 + \frac{1}{\mathtt{m} + \mathtt{L}} \left\| \nabla U(\theta_2) - \nabla U(\theta_1) \right\|^2. \quad (4.6)$$

We consider the following assumptions on the family $\{H_i : \mathbb{R}^d \times \mathsf{X}_1 \to \mathbb{R}^d\}_{i \in [n]}$ and $\mathcal{C}$.

Assumption 4.10. *There exists a probability measure ν_2 on a measurable space $(\mathsf{X}_2, \mathcal{X}_2)$ and a family of measurable functions $\{\mathcal{C}_i : \mathbb{R}^d \times \mathsf{X}_2 \to \mathbb{R}^d\}_{i \in [n]}$ such that the following conditions hold.*

(i) For any $\theta \in \mathbb{R}^d$ and any $i \in [n]$, $\int_{\mathsf{X}_2} \mathcal{C}_i(\theta, x^{(2)}) \, \nu_2(\mathrm{d}x^{(2)}) = \theta$.
(ii) There exist $\{\omega_i \in \mathbb{R}_+\}_{i \in [n]}$, such that for any $\theta \in \mathbb{R}^d$ and any $i \in [n]$,

$$\int_{\mathsf{X}_2} \left\| \mathcal{C}_i(\theta, x^{(2)}) - \theta \right\|^2 \nu_2(\mathrm{d}x^{(2)}) \leq \omega_i \left\| \theta \right\|^2.$$

Assumption 4.11. *There exist a family of probability measures $\{\nu_1^{(i)}\}_{i \in [n]}$ defined on measurable spaces $\{(\mathsf{X}_1^{(i)}, \mathcal{X}_1^{(i)})\}_{i \in [n]}$ and a family of measurable functions $\{H_i : \mathbb{R}^d \times \mathsf{X}_1^{(i)} \to \mathbb{R}^d\}_{i \in [n]}$ such that the following conditions hold.*

(i) For any $\theta \in \mathbb{R}^d$,

$$\sum_{i=1}^{n} \int_{\mathsf{X}_1^{(i)}} H_i(\theta, x^{(1,i)}) \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = \nabla U(\theta).$$

(ii) There exist $\{\mathtt{M}_i > 0\}_{i\in[n]}$, *such that for any* $i \in [n]$, $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta_2, x^{(1,i)}) - H_i(\theta_1, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \leq \mathtt{M}_i \left\langle \theta_2 - \theta_1, \nabla U_i(\theta_2) - \nabla U_i(\theta_1) \right\rangle .$$

(iii) There exists $\sigma_\star, \mathtt{B}^\star \in \mathbb{R}_+$ *such that for any* $i \in [n]$, $\theta \in \mathbb{R}^d$, *we have*

$$\int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta^\star, x^{(1)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1)}) \leq \mathtt{B}^\star / n, \tag{4.7}$$

$$\int_{\mathsf{X}_1^{(1)} \times \cdots \times \mathsf{X}_1^{(n)}} \left\| \sum_{i=1}^{n} H_i(\theta^\star, x^{(1,i)}) \right\|^2 \otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \leq \sigma_\star^2.$$

We can notice that Assumption 4.11-(ii) implies that ∇U_i is $\mathtt{M}_i$-Lipschitz continuous since by the Cauchy Schwarz inequality, for any $i \in [n]$ and any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\| \nabla U_i(\theta_1) - \nabla U_i(\theta_2) \right\|^2 \leq \mathtt{M}_i \left\langle \theta_1 - \theta_2, \nabla U_i(\theta_1) - \nabla U_i(\theta_2) \right\rangle .$$

In addition, it is worth mentioning that the first inequality in (4.7) is also required for our derivation in the deterministic case where $H_i = \nabla U_i$ for any $i \in [n]$ due to the compression step. For $k \geq 1$, consider $(X_k^{(1,1)}, \ldots, X_k^{(1,n)})_{k\in\mathbb{N}}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,n)})_{k\in\mathbb{N}}$ two independent sequences of random variables distributed according to $\nu_1^{(1:n)} = \nu_1^{(1)} \otimes \cdots \otimes \nu_1^{(n)}$ and $\nu_2^{\otimes n}$, respectively.

In addition, we consider the partial device participation context where at each communication round $k \geq 1$, each client has a probability $p_i \in (0, 1]$ of participating, independently of other clients.

Assumption 4.12. *For any* $k \in \mathbb{N}^\star$, $\mathcal{A}_k = \{i \in [n] : B_{i,k} = 1\}$ *where for any* $i \in [n]$, $\{B_{i,k} :, k \in \mathbb{N}^*\}$ *is a family of i.i.d. Bernoulli random variables with success probability* $p_i \in \big(0, 1\big]$.

In other words, there exists a sequence $(X_k^{(3,1)}, \cdots, X_k^{(3,n)})_{k\in\mathbb{N}}$ of i.i.d. random variables distributed according $\nu_3 = \mathrm{Uniform}((0, 1])$, such that for any $k \geq 1$ and $i \in [n]$, client i is active at step k if $X_k^{(3,i)} \leq p_i$. We denote $\mathcal{A}_{k+1} = \{i \in [n]; X_{k+1}^{(3,i)} \leq p_i\}$ the set of active clients at round k. Given a step-size $\gamma \in (0, \bar{\gamma}]$ for some $\bar{\gamma} > 0$ and starting from $\theta_0 \in \mathbb{R}^d$, QLSD recursively defines $(\theta_k)_{k\in\mathbb{N}}$, for any $k \in \mathbb{N}$, as

$$\theta_{k+1} = \theta_k - \gamma \textstyle\sum_{i\in\mathcal{A}_{k+1}} (1/p_i) \mathcal{C}_i(H_i(\theta_k, X_{k+1}^{(1,i)}), X_{k+1}^{(2,i)}) + \sqrt{2\gamma} Z_{k+1}, \tag{4.8}$$

where $(Z_{k+1})_{k\in\mathbb{N}}$ is a sequence of standard Gaussian random variables. Let $\mathsf{X}_3 = [0, 1]$. For any $i \in [n]$, consider the unbiased partial participation operator $\mathcal{S}_i : \mathbb{R}^d \times \mathsf{X}_3 \to \mathbb{R}^d$ defined, for any $\theta \in \mathbb{R}^d$ and $x^{(3)} \in \mathsf{X}_3$ by

$$\mathcal{S}_i(\theta, x^{(3)}) = \frac{\mathbf{1}\{x^{(3)} \leq p_i\}}{p_i} \theta. \tag{4.9}$$

Then, (4.8) can be written of the form

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^{n} \tilde{H}_i(\theta_k, X_{k+1}^{(i)}) + \sqrt{2\gamma} Z_{k+1}, \qquad k \in \mathbb{N}, \tag{4.10}$$

where for any $i \in [n]$, we denote $X_{k+1}^{(i)} = (X_{k+1}^{(1,i)}, X_{k+1}^{(2,i)}, X_{k+1}^{(3,i)})$ and for any $\theta \in \mathbb{R}^d, x^{(1,i)} \in \mathsf{X}_1^{(i)}, x^{(2)} \in \mathsf{X}_2$ and $x^{(3)} \in \mathsf{X}_3$,

$$\tilde{H}_i\Big(\theta, (x^{(1,i)}, x^{(2)}, x^{(3)})\Big) = \mathcal{S}_i\left(\mathcal{C}_i\Big(H_i(\theta, x^{(1,i)}), x^{(2)}\Big), x^{(3)}\right). \tag{4.11}$$

With this notation and setting for any $i \in [n]$, $\tilde{\mathsf{X}}^{(i)} = \mathsf{X}_1^{(i)} \times \mathsf{X}_2 \times \mathsf{X}_3$ and $\tilde{\nu}^{(i)} = \nu_1^{(i)} \otimes \nu_2 \otimes \nu_3$, the Markov kernel associated with (4.8) is given for any $(\theta, \mathsf{A}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ by

$$Q_\gamma(\theta, \mathsf{A}) = \int_{\mathsf{A} \times \tilde{\mathsf{X}}^{(1)} \times \cdots \times \tilde{\mathsf{X}}^{(n)}} \exp\left(-\|\theta - \theta + \gamma \sum_{i=1}^{n} \tilde{H}_i(\theta, x^{(i)})\|^2/(4\gamma)\right) \frac{\mathrm{d}\tilde{\theta}\, \tilde{\nu}^{(1)}(\mathrm{d}x^{(1)}) \otimes \cdots \otimes \tilde{\nu}^{(n)}(\mathrm{d}x^{(n)})}{(4\pi\gamma)^{d/2}}. \tag{4.12}$$

The following result establishes an essential property of $\{\tilde{H}_i\}_{i \in [n]}$ under Assumption 4.10 and Assumption 4.11.

Lemma 4.13. *Assume Assumption 4.10, Assumption 4.11 and Assumption 4.12. Then, for any $\theta \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^{n} \int_{\tilde{\mathsf{X}}^{(i)}} \tilde{H}_i(\theta, x^{(i)})\, \mathrm{d}\tilde{\nu}^{(i)}(x^{(i)}) = \nabla U(\theta), \tag{4.13}$$

$$\int_{\tilde{\mathsf{X}}^{(1:n)}} \left\| \sum_{i=1}^{n} \tilde{H}_i(\theta, x^{(i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{n} \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)})$$

$$\leq 2 \max_{i \in [n]} \left\{ \frac{\mathtt{M}_i(\omega_i + 1)}{p_i} \right\} \Big\langle \theta - \theta^\star, \nabla U(\theta) \Big\rangle + 2 \left[\sigma_\star^2 + \frac{\mathtt{B}^\star}{n} \sum_{i=1}^{n} \frac{1 - p_i + \omega_i}{p_i} \right], \tag{4.14}$$

where for any $i \in [n]$, $\tilde{H}_i$ is defined in (4.11).

Proof The first identity (4.13) is straightforward using Assumption 4.11-(i) and Assumption 4.10-(i). We now show the inequality (4.14). Let $\theta \in \mathbb{R}^d$. Using Assumption 4.10-(i) or Assumption 4.11-(i), we get

$$\int_{\tilde{\mathsf{X}}^{(1:n)}} \left\| \textstyle\sum_{i=1}^{n} \tilde{H}_i(\theta, x^{(i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{n} \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)})$$

$$= \int_{\tilde{\mathsf{X}}^{(1:n)}} \left\| \sum_{i=1}^{n} \left[\tilde{H}_i(\theta, x^{(i)}) - \mathcal{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right) \right] \right\|^2 \otimes_{i=1}^{n} \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)})$$

$$+\int_{\mathsf{X}_1^{(1:n)}\times\mathsf{X}_2^n}\left\|\sum_{i=1}^n \mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)-\nabla U(\theta)\right\|^2 \nu_2^{\otimes n}(\mathrm{d}x^{(2,1:n)})\otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)}). \tag{4.15}$$

In addition, by Assumption 4.10-(i) and Assumption 4.10-(ii), we obtain

$$\begin{aligned}
&\int_{\tilde{\mathsf{X}}^{(1:n)}}\left\|\sum_{i=1}^n\left[\tilde{H}_i(\theta,x^{(i)})-\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)\right]\right\|^2\otimes_{i=1}^n\tilde{\nu}^{(i)}(\mathrm{d}x^{(i)})\\
&=\sum_{i=1}^n\int_{\tilde{\mathsf{X}}^{(i)}}\left\|\tilde{H}_i(\theta,x^{(i)})-\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)})\nu_3(\mathrm{d}x^{(3,i)})\\
&\le\sum_{i=1}^n\left(\frac{1-p_i}{p_i}\right)\int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2}\left\|\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)})\\
&=\sum_{i=1}^n\left(\frac{1-p_i}{p_i}\right)\int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2}\left\|\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)-H_i(\theta,x^{(1,i)})+H_i(\theta,x^{(1,i)})\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)})\\
&=\sum_{i=1}^n\left(\frac{1-p_i}{p_i}\right)\int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2}\left\|\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)-H_i(\theta,x^{(1,i)})\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)})\\
&+\sum_{i=1}^n\left(\frac{1-p_i}{p_i}\right)\int_{\mathsf{X}_1^{(i)}}\left\|H_i(\theta,x^{(1,i)})\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\\
&\le\sum_{i=1}^n\left[\left(\frac{1-p_i}{p_i}\right)(\omega_i+1)\right]\int_{\mathsf{X}_1^{(i)}}\left\|H_i(\theta,x^{(1,i)})\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)}).
\end{aligned}\tag{4.16}$$

Using $\|a\|^2\le 2\|a-b\|^2+2\|b\|^2$ and Assumption 4.11-(ii)-(iii), for any $i\in[n]$, we obtain

$$\begin{aligned}
\int_{\mathsf{X}_1^{(i)}}\left\|H_i(\theta,x^{(1,i)})\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)}) &\le 2\mathtt{M}_i\left\langle\theta-\theta^\star,\nabla U_i(\theta)-\nabla U_i(\theta^\star)\right\rangle\\
&\quad+2\int_{\mathsf{X}_1^{(i)}}\left\|H_i(\theta^\star,x^{(1,i)})\right\|^2\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\\
&\le 2\mathtt{M}_i\left\langle\theta-\theta^\star,\nabla U_i(\theta)-\nabla U_i(\theta^\star)\right\rangle+2\mathtt{B}^\star/n.
\end{aligned}$$

Therefore, combining this result and (4.16) gives

$$\int_{\tilde{\mathsf{X}}^{(1:n)}}\left\|\sum_{i=1}^n\left[\tilde{H}_i(\theta,x^{(i)})-\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)\right]\right\|^2\otimes_{i=1}^n\tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) \tag{4.17}$$

$$\le 2\sum_{i=1}^n\mathtt{M}_i\left(\frac{1-p_i}{p_i}\right)(\omega_i+1)\left\langle\theta-\theta^\star,\nabla U_i(\theta)-\nabla U_i(\theta^\star)\right\rangle+\frac{2\mathtt{B}^\star}{n}\sum_{i=1}^n\left(\frac{1-p_i}{p_i}\right)(\omega_i+1). \tag{4.18}$$

Similarly, by Assumption 4.10-(i) and Assumption 4.10-(ii), we have

$$\begin{aligned}
&\int_{\mathsf{X}_1^{(1:n)}\times\mathsf{X}_2^n}\left\|\sum_{i=1}^n\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)-\nabla U(\theta)\right\|^2\nu_2^{\otimes n}(\mathrm{d}x^{(2,1:n)})\otimes_{i=1}^n\nu_1^{(i)}(\mathrm{d}x^{(1,i)})\\
&=\int_{\mathsf{X}_1^{(1:n)}\times\mathsf{X}_2^n}\left\|\sum_{i=1}^n\left[\mathcal{C}_i\left(H_i(\theta,x^{(1,i)}),x^{(2,i)}\right)-H_i(\theta,x^{(1,i)})\right]\right.\\
&\left.+\sum_{i=1}^n\{H_i(\theta,x^{(1,i)})\}-\nabla U(\theta)\right\|^2\nu_2^{\otimes n}(\mathrm{d}x^{(2,1:n)})\otimes_{i=1}^n\nu_1^{(i)}(\mathrm{d}x^{(1,i)})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \int_{\mathsf{X}_1^{(i)} \times \mathsf{X}_2} \left\| \mathcal{C}_i \left(H_i(\theta, x^{(1,i)}), x^{(2)} \right) - H_i(\theta, x^{(1,i)}) \right\|^2 \nu_2(\mathrm{d}x^{(2)}) \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \int_{\mathsf{X}_1^{(1:n)}} \left\| \sum_{i=1}^{n} H_i(\theta, x^{(1,i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{n} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \sum_{i=1}^{n} \omega_i \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \int_{\mathsf{X}_1^{(1:n)}} \left\| \sum_{i=1}^{n} H_i(\theta, x^{(1,i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{n} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}).
\end{aligned}
$$

Since for any $a, b \in \mathbb{R}^d$, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have by Assumption 4.11-(i)

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(1:n)}} \left\| \sum_{i=1}^{n} H_i(\theta, x^{(1,i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{n} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \qquad (4.19) \\
&= \int_{\mathsf{X}_1^{(1:n)}} \left\| \sum_{i=1}^{n} \left[H_i(\theta, x^{(1,i)}) - \int_{\mathsf{X}_1^{(i)}} H_i(\theta, x^{(1)}) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right] \right\|^2 \otimes_{i=1}^{n} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \sum_{i=1}^{n} \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) - \int_{\mathsf{X}_1^{(i)}} H_i(\theta, x^{(1)}) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2 \sum_{i=1}^{n} \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) - H_i(\theta^\star, x^{(1,i)}) - \left[\int_{\mathsf{X}_1^{(i)}} (H_i(\theta, x^{(1)}) - H_i(\theta^\star, x^{(1)})) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ 2 \sum_{i=1}^{n} \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta_\star, x^{(1,i)}) - \int_{\mathsf{X}_1^{(i)}} H_i(\theta_\star, x^{(1)}) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2\sigma_\star^2 + 2 \sum_{i=1}^{n} \mathtt{M}_i \left\langle \nabla U_i(\theta) - \nabla U_i(\theta^\star), \theta - \theta^\star \right\rangle. \qquad (4.20)
\end{aligned}
$$

By combining (4.17), (4.19) and (4.20), we obtain

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(1:n)}} \left\| \sum_{i=1}^{n} \mathcal{C}_i \left(H_i(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes n}(\mathrm{d}x^{(2,1:n)}) \otimes_{i=1}^{n} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2 \sum_{i=1}^{n} \mathtt{M}_i(\omega_i + 1) \left\langle \nabla U_i(\theta) - \nabla U_i(\theta^\star), \theta - \theta^\star \right\rangle + 2 \left(\sigma_\star^2 + \frac{2\mathsf{B}^\star}{n} \sum_{i=1}^{n} \omega_i \right).
\end{aligned}
$$

Finally, the last inequality combined with (4.15) and (4.18) completes the proof. ■

In view of Lemma 4.13, it suffices to study the recursion specified in (4.10) under the following assumption on $(\tilde{H}_i)_{i\in[n]}$ gathered in Assumption 4.14. Indeed, Lemma 4.13 shows that Condition Assumption 4.14 below holds with $\mathsf{X}^{(i)} = \tilde{\mathsf{X}}^{(i)} = \mathsf{X}_1^{(i)} \times \mathsf{X}_2 \times \mathsf{X}_3$, $\mathcal{X}^{(i)} = \tilde{\mathcal{X}}^{(i)} = \mathcal{X}_1^{(i)} \otimes \mathcal{X}_2 \otimes \mathcal{X}_3$, $\tilde{\nu}^{(i)} = \nu_1^{(i)} \otimes \nu_2 \otimes \nu_3$, $\{\tilde{H}_i\}_{i=1}^n = \{F_i\}_{i=1}^n$,

$$
\tilde{\mathtt{M}} = 2 \max_{i\in[n]} \{\mathtt{M}_i(1+\omega_i)/p_i\},
$$

$$\tilde{\mathtt{B}}^\star = 2[\sigma_\star^2 + (\mathtt{B}^\star/n)\sum_{i=1}^n (1-p_i+\omega_i)/p_i].$$

Assumption 4.14. *There exists a family of probability measure $\{\nu^{(i)}\}_{i\in[n]}$ on a measurable space $\{(\tilde{\mathsf{X}}^{(i)}, \tilde{\mathcal{X}}^{(i)})\}_{i\in[n]}$ and a family of measurable functions $\{F_i : \mathbb{R}^d \times \mathsf{X}^{(i)} \to \mathbb{R}^d\}_{i\in[n]}$ such that the following conditions hold.*

(i) For any $\theta \in \mathbb{R}^d$, we have

$$\sum_{i=1}^n \int_{\tilde{\mathsf{X}}^{(i)}} F_i(\theta, x^{(i)})\nu^{(i)}(\mathrm{d}x^{(i)}) = \nabla U(\theta).$$

(ii) There exists $(\tilde{\mathtt{M}}, \tilde{\mathtt{B}}^\star) \in \mathbb{R}_+^2$ such that for any $\theta \in \mathbb{R}^d$, we have

$$\int_{\tilde{\mathsf{X}}^{(1:n)}} \left\| \sum_{i=1}^n F_i(\theta, x^{(i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) \le \tilde{\mathtt{M}} \left\langle \theta - \theta^\star, \nabla U(\theta) - \nabla U(\theta^\star) \right\rangle + \tilde{\mathtt{B}}^\star.$$

Then under Assumption 4.14, consider $(X_k^{(1)}, \ldots, X_k^{(n)})_{k\in\mathbb{N}^\star}$ an independent sequence distributed according to $\otimes_{i=1}^n \nu^{(i)}$. Define the general recursion

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - \gamma \sum_{i=1}^n F_i(\tilde{\theta}_k, X_{k+1}^{(i)}) + \sqrt{2\gamma} Z_{k+1}, \qquad k \in \mathbb{N}.$$

and the corresponding the Markov kernel given for any $\gamma \in \mathbb{R}_+^*$, $\theta \in \mathbb{R}^d, \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\tilde{Q}_\gamma(\theta, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}\times\tilde{\mathsf{X}}^{(1:n)}} \exp(-(4\gamma)^{-1} \left\| \bar{\theta} - \theta + \gamma \sum_{i=1}^n F_i(\theta, x^{(i)}) \right\|^2) \, \mathrm{d}\bar{\theta} \, \mathrm{d}\otimes_{i=1}^n \nu^{(i)}(x^{(i)}). \tag{4.21}$$

We refer to this Markov kernel as the generalized QLSD kernel. In our next section, we establish quantitative bounds between the iterates of this kernel and π in W_2. We then apply this result to QLSD and QLSD* as particular cases.

4.A.2 Quantitative bounds for the generalized QLSD kernel

Define

$$\bar{\gamma} = \bar{\gamma}_1 \wedge \bar{\gamma}_2 \wedge \bar{\gamma}_3, \quad \bar{\gamma}_1 = 2/[5(\mathtt{m}+\mathtt{L})], \quad \bar{\gamma}_2 = (\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}})^{-1}, \quad \bar{\gamma}_3 = (10\mathtt{m})^{-1}.$$

Theorem 4.15. *Assume Assumption 4.9 and Assumption 4.14. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step-size $\gamma \in (0, \bar{\gamma}]$, any $k \in \mathbb{N}$, we have*

$$W_2^2(\mu\tilde{Q}_\gamma^k, \pi) \le (1-\gamma\mathtt{m}/2)^k W_2^2(\mu,\pi) + \gamma\tilde{B}_{\bar{\gamma}} + \gamma^2\tilde{A}_{\bar{\gamma}}(1-\mathtt{m}\gamma/2)^{k-1}k \int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2\mu(\mathrm{d}\theta),$$

where $\tilde{Q}_\gamma$ is defined in (4.21) and

$$\tilde{B}_{\bar{\gamma}} = (2d\mathtt{L}^2/\mathtt{m})(1/\mathtt{m} + 5\bar{\gamma})\left[1 + \bar{\gamma}\mathtt{L}^2/(2\mathtt{m}) + \bar{\gamma}^2\mathtt{L}^2/12\right] + 2\tilde{\mathtt{B}}^\star/\mathtt{m} + 2\mathtt{L}\tilde{\mathtt{M}}(2d + \bar{\gamma}\tilde{\mathtt{B}}^\star)/\mathtt{m}^2$$
$$\tilde{A}_{\bar{\gamma}} = \mathtt{L}\tilde{\mathtt{M}}.$$

Let $\xi \in \mathcal{P}_2(\mathbb{R}^{2d})$ be a probability measure on $(\mathbb{R}^{2d}, \mathcal{B}(\mathbb{R}^{2d}))$ with marginals ξ_1 and ξ_2, *i.e.* $\xi(\mathsf{A} \times \mathbb{R}^d) = \xi_1(\mathsf{A})$ and $\xi(\mathsf{A} \times \mathbb{R}^d) = \xi_2(\mathsf{A})$ for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$. Note that under Assumption 4.9, the Langevin diffusion defines a Markov semigroup $(P_t)_{t\geq 0}$ satisfying $\pi P_t = \pi$ for any $t \geq 0$, see *e.g.* Roberts and Tweedie (1996, Theorem 2.1). We introduce a synchronous coupling $(\vartheta_{k\gamma}, \theta_k)$ between $\xi_1 P_{k\gamma}$ and $\xi_2 \tilde{Q}_\gamma^k$ for any $k \in \mathbb{N}$ based on a d-dimensional standard Brownian motion $(B_t)_{t\geq 0}$ and a couple of random variables (θ_0, ϑ_0) with distribution ξ independent of $(B_t)_{t\geq 0}$. Consider $(\vartheta_t)_{t\geq 0}$ the strong solution of the Langevin stochastic differential equation (SDE)

$$\mathrm{d}\vartheta_t = -\nabla U(\vartheta_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{4.22}$$

starting from ϑ_0. Note that under Assumption 4.9-(i), this SDE admits a unique strong solution (Revuz and Yor, 2013, Theorem (2.1) in Chapter IX). In addition, define $(\theta_k)_{k\in\mathbb{N}}$ starting from θ_0 and satisfying the recursion: for $k \geq 0$,

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^{n} F_i(\theta_k, x_{k+1}^{(i)}) + \sqrt{2}(B_{\gamma(k+1)} - B_{\gamma k}), \tag{4.23}$$

where $(x_j^{(1)}, \ldots, x_j^{(n)})_{j\in\mathbb{N}^\star}$ is an independent sequence of random variables with distribution $\otimes_{i=1}^n \nu^{(i)}$. Then, by definition, $(\vartheta_{k\gamma}, \theta_k)$ is a coupling between $\xi_1 P_{k\gamma}$ and $\xi_2 \tilde{Q}_\gamma^k$ for any $k \in \mathbb{N}$ and therefore

$$W_2(\xi_1 P_{k\gamma}, \xi_2 \tilde{Q}_\gamma^k) \leq \mathbb{E}\left[\|\vartheta_{\gamma k} - \theta_k\|^2\right]^{1/2}. \tag{4.24}$$

We can now give the proof of Theorem 4.15.

Proof By Villani (2008, Theorem 4.1), for any couple of probability measures on $\mathbb{R}^d$, there exists an optimal transference plan $\xi^\star$ between ν and π since $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ by the strong convexity assumption Assumption 4.9-(i). Let (ϑ_0, θ_0) be a corresponding coupling which therefore satisfies $W_2(\mu, \pi) = \mathbb{E}^{1/2}[\|\vartheta_0 - \theta_0\|^2]$. Consider then $(\vartheta_k)_{k\in\mathbb{N}}, (\theta_k)_{k\in\mathbb{N}}$ defined in (4.22)-(4.23) starting from (ϑ_0, θ_0). Note that since $\pi P_t = \pi$ by Roberts and Tweedie (1996, Theorem 2.1) for any $t \geq 0$ and θ_0 has distribution π, we get by Durmus and Moulines (2019, Proposition 1) that for any $k \in \mathbb{N}$, $\mathbb{E}[\|\vartheta_{k\gamma} - \theta^\star\|^2] \leq d/\mathtt{m}$ and then Lemma 4.17 below shows that for any $k \in \mathbb{N}$,

$$\mathbb{E}[\|\vartheta_{(k+1)\gamma} - \theta_{k+1}\|^2] \leq \kappa_\gamma \mathbb{E}[\|\vartheta_{k\gamma} - \theta_k\|^2] + \gamma^2 \mathtt{L}\tilde{\mathtt{M}}\,\mathbb{E}\left[\|\theta_0 - \theta^\star\|^2\right] \tilde{\kappa}_\gamma^k + \gamma^2 \mathrm{D}_\gamma,$$

where we have set

$$\kappa_\gamma = 1 - \gamma\mathtt{m}(1 - 5\gamma\mathtt{m}), \quad \tilde{\kappa}_\gamma = 1 - \gamma\mathtt{m}\left[2 - \gamma(\mathtt{m} + \tilde{\mathtt{M}})\right], \quad \mathrm{D}_\gamma = \mathrm{D}_{0,\gamma} + (1/\mathtt{m} + 5\gamma)(\gamma d \mathtt{L}^4/2\mathtt{m}).$$

A straightforward induction shows that

$$\mathbb{E}[\|\vartheta_{k\gamma} - \theta_k\|^2] \leq \kappa_\gamma^k W_2^2(\mu, \pi(\cdot|\mathcal{D})) + \gamma^2 \mathtt{L}\tilde{\mathtt{M}}\,\mathbb{E}\left[\|\theta_0 - \theta^\star\|^2\right] \sum_{l=0}^{k-1} \kappa_\gamma^l \tilde{\kappa}_\gamma^{k-1-l} + \frac{\gamma^2 \mathrm{D}_\gamma}{1 - \kappa_\gamma}.$$

Using $\kappa_\gamma \wedge \tilde{\kappa}_\gamma \leq 1 - \mathtt{m}\gamma/2$ since $\gamma \leq \bar{\gamma}$, (4.24) and $\pi P_t = \pi$ for any $t \geq 0$ completes the proof. ■

Supporting Lemmata

In this subsection, we derived two lemmas. Taking $(\theta_k)_{k\in\mathbb{N}}$ defined by the recursion (4.23), Lemma 4.16 aims to upper bound the squared deviation between θ_k and the minimizer of U denoted $\theta^\star$, for any $k \in \mathbb{N}$.

Lemma 4.16. *Assume Assumption 4.9 and Assumption 4.14. Let* $\gamma \in \left(0, 2/(\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}})\right]$. *Then, for any* $k \in \mathbb{N}, \theta_0 \in \mathbb{R}^d$, *we have*

$$\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \tilde{Q}_\gamma^k(\theta_0, \mathrm{d}\theta) \le (1-\gamma\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right])^k \|\theta_0-\theta^\star\|^2 + \frac{2d+\gamma\tilde{\mathtt{B}}^\star}{\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right]},$$

where $\tilde{Q}_\gamma$ *is defined in* (4.21).

Proof For any $\theta_0 \in \mathbb{R}^d$, by definition (4.21) of $\tilde{Q}_\gamma$ and using Assumption 4.14-(i), we obtain

$$\begin{aligned}\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \tilde{Q}_\gamma(\theta_0, \mathrm{d}\theta) &= \|\theta_0-\theta^\star\|^2 - 2\gamma\left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle \\ &\quad + \gamma^2 \int_{\tilde{\mathsf{X}}^{(1:n)}} \left\|\textstyle\sum_{i=1}^n F_i(\theta_0, x^{(i)})\right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) + 2\gamma d. \quad (4.25)\end{aligned}$$

Moreover, using Assumption 4.9, Assumption 4.14 and (4.6), it follows that

$$\begin{aligned}&\int_{\tilde{\mathsf{X}}^{(1:n)}} \left\|\textstyle\sum_{i=1}^n F_i(\theta_0, x^{(i)})\right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) \\ &\quad= \int_{\tilde{\mathsf{X}}^{(1:n)}} \left\|\textstyle\sum_{i=1}^n F_i(\theta_0, x^{(i)}) - \nabla U(\theta_0)\right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) + \left\|\nabla U(\theta_0)\right\|^2 \\ &\quad\le \tilde{\mathtt{M}}\left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle + \tilde{\mathtt{B}}^\star + \left\|\nabla U(\theta_0) - \nabla U(\theta^\star)\right\|^2 \\ &\quad\le [\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}}]\left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle + \tilde{\mathtt{B}}^\star - \mathtt{L}\mathtt{m}\left\|\theta_0-\theta^\star\right\|^2. \quad (4.26)\end{aligned}$$

Plugging (4.26) in (4.25) implies

$$\begin{aligned}\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \tilde{Q}_\gamma(\theta_0, \mathrm{d}\theta) &\le (1-\gamma^2\mathtt{m}\mathtt{L})\|\theta_0-\theta^\star\|^2 \\ &\quad - \gamma\{2-\gamma[\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}}]\}\left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle + \gamma^2\tilde{\mathtt{B}}^\star + 2\gamma d.\end{aligned}$$

Using Assumption 4.9-(i), we have $\langle \theta_0-\theta^\star, \nabla U(\theta_0)\rangle \ge \mathtt{m}\|\theta_0-\theta^\star\|^2$ which, combined with the condition $\gamma \le 1/(\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}})$, gives

$$\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \tilde{Q}_\gamma(\theta_0, \mathrm{d}\theta) \le (1-\gamma\mathtt{m}[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})])\|\theta_0-\theta^\star\|^2 + \gamma(2d+\gamma\tilde{\mathtt{B}}^\star).$$

Using $0 < \gamma < 2/(\mathtt{m}+\tilde{\mathtt{M}})$ and the Markov property combined with a straightforward induction completes the proof. ∎

For any $k \in \mathbb{N}$, the following lemma gives an explicit upper bound on the expected squared norm between ϑ_{k+1} and θ_{k+1} in function of ϑ_k, θ_k. The purpose of this lemma is to derive a contraction property involving a contracting term and a bias term which is easy to control.

Lemma 4.17. *Assume Assumption 4.9 and Assumption 4.14. Consider $(\vartheta_t)_{t\geq 0}$ and $(\theta_k)_{k\in\mathbb{N}}$ defined in (4.22) and (4.23), respectively, for some initial distribution $\xi \in \mathcal{P}_2(\mathbb{R}^{2d})$. For any $k \in \mathbb{N}$ and $\gamma \in \left(0, 2/[(5(\mathtt{m}+\mathtt{L}))\vee(\mathtt{m}+\tilde{\mathtt{M}}+\mathtt{L})]\right)$, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\vartheta_{\gamma(k+1)}-\theta_{k+1}\right\|^2\right] &\leq \{1-\gamma\mathtt{m}(1-5\gamma\mathtt{m})\}\mathbb{E}\left[\|\vartheta_k-\theta_k\|^2\right]+\gamma^2\mathrm{D}_{0,\gamma}\\ &\quad+\gamma^2\mathtt{L}\tilde{\mathtt{M}}(1-\gamma\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right])^k\mathbb{E}[\|\theta_0-\theta^\star\|^2]\\ &\quad+\gamma^3(1/\mathtt{m}+5\gamma)\mathtt{L}^4\mathbb{E}\left[\|\vartheta_{k\gamma}-\theta^\star\|^2\right]/2,\end{aligned}$$

where

$$\mathrm{D}_{0,\gamma}=d\mathtt{L}^2(1/\mathtt{m}+5\gamma)\left[1+\gamma^2\mathtt{L}^2/12\right]+\tilde{\mathtt{B}}^\star+\frac{\mathtt{L}\tilde{\mathtt{M}}(2d+\gamma\tilde{\mathtt{B}}^\star)}{\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right]}.$$

Proof Let $k\in\mathbb{N}$. By (4.22) and (4.23), we have

$$\begin{aligned}\vartheta_{\gamma(k+1)}-\theta_{k+1}&=\vartheta_{\gamma k}-\theta_k-\gamma\left[\nabla U(\vartheta_{\gamma k})-\nabla U(\theta_k)\right]\\ &\quad-\int_0^\gamma\left[\nabla U(\vartheta_{\gamma k+s})-\nabla U(\vartheta_{\gamma k})\right]\mathrm{d}s+\gamma\sum_{i=1}^n\left[F_i(\theta_k,X_{k+1}^{(i)})-\nabla U_i(\theta_k)\right].\end{aligned}$$

Define the filtration $(\mathcal{F}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ as $\mathcal{F}_0=\sigma(\vartheta_0,\theta_0)$ and for $\tilde{k}\in\mathbb{N}^\star$,

$$\mathcal{F}_{\tilde{k}}=\sigma(\vartheta_0,\theta_0,(X_l^{(1)},\ldots,X_l^{(n)})_{1\leq l\leq\tilde{k}},(B_t)_{0\leq t\leq\gamma\tilde{k}}).$$

Note that since $(\vartheta_t)_{t\geq 0}$ is a strong solution of (4.22), then is easy to see that $(\vartheta_{\gamma\tilde{k}},\theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ is $(\mathcal{F}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$-adapted. Taking the squared norm and the conditional expectation with respect to $\mathcal{F}_k$, we obtain using Assumption 4.14-(i) that

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\left\|\vartheta_{\gamma(k+1)}-\theta_{k+1}\right\|^2\right]&=\left\|\vartheta_{\gamma k}-\theta_k\right\|^2-2\gamma\left\langle\vartheta_{\gamma k}-\theta_k,\nabla U(\vartheta_{\gamma k})-\nabla U(\theta_k)\right\rangle\\ &\quad+2\gamma\int_0^\gamma\left\langle\nabla U(\vartheta_{\gamma k})-\nabla U(\theta_k),\mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s})-\nabla U(\vartheta_{\gamma k})\right]\right\rangle\mathrm{d}s\\ &\quad-2\int_0^\gamma\left\langle\vartheta_{\gamma k}-\theta_k,\mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s})-\nabla U(\vartheta_{\gamma k})\right]\right\rangle\mathrm{d}s+\gamma^2\left\|\nabla U(\vartheta_{\gamma k})-\nabla U(\theta_k)\right\|^2\\ &\quad+\mathbb{E}^{\mathcal{F}_k}\left[\left\|\int_0^\gamma\left[\nabla U(\vartheta_{\gamma k+s})-\nabla U(\vartheta_{\gamma k})\right]\mathrm{d}s\right\|^2\right]+\gamma^2\mathbb{E}^{\mathcal{F}_k}\left[\left\|\sum_{i=1}^n F_i(\theta_k,X_{k+1}^{(i)})-\nabla U(\theta_k)\right\|^2\right].\end{aligned}\tag{4.27}$$

First, using Jensen inequality and the fact that for any $a,b\in\mathbb{R}^d$, $|\langle a,b\rangle|\leq 2\|a\|^2+2\|b\|^2$, we get

$$\begin{aligned}&\int_0^\gamma\left\langle\nabla U(\vartheta_{\gamma k})-\nabla U(\theta_k),\mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s})-\nabla U(\vartheta_{\gamma k})\right]\right\rangle\mathrm{d}s\\ &\quad\leq 2\gamma\left\|\nabla U(\vartheta_{\gamma k})-\nabla U(\theta_k)\right\|^2+2\int_0^\gamma\mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s})-\nabla U(\vartheta_{\gamma k})\right\|^2\right]\mathrm{d}s,\end{aligned}\tag{4.28}$$

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|\int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right] \mathrm{d}s\right\|^2\right] \leq \gamma \int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s.$$

In addition, given that for any $\varepsilon > 0, a, b \in \mathbb{R}^d$, $|\langle a, b\rangle| \leq \varepsilon \|a\|^2 + (4\varepsilon)^{-1}\|b\|^2$, we get

$$\begin{aligned}\left|\int_0^\gamma \left\langle \theta_k - \vartheta_{\gamma k}, \mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}s\right| &\leq \gamma\varepsilon \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 \\ &+ (4\varepsilon)^{-1} \int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s. \quad (4.29)\end{aligned}$$

By Assumption 4.9, for $k \in \mathbb{N}$ we get by (4.6)

$$\left\|\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\|^2 \leq (\mathtt{m} + \mathtt{L}) \left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle - \mathtt{mL} \left\|\vartheta_{\gamma k} - \theta_k\right\|^2. \quad (4.30)$$

Lastly, Assumption 4.14-(ii) yields

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|\sum_{i=1}^n F_i(\theta_k, X_{k+1}^{(i)}) - \nabla U(\theta_k)\right\|^2\right] \leq \tilde{\mathtt{M}} \left\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\right\rangle + \tilde{\mathtt{B}}^\star. \quad (4.31)$$

Combining (4.28), (4.29), (4.30) and (4.31) into (4.27), for $k \in \mathbb{N}$ we get for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &\leq (1 + 2\gamma\varepsilon - 5\gamma^2\mathtt{mL}) \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 \\ &- \gamma\left[2 - 5\gamma(\mathtt{m} + \mathtt{L})\right] \left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle \\ &+ (5\gamma + (2\varepsilon)^{-1}) \int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s \\ &+ \gamma^2 \tilde{\mathtt{M}} \left\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\right\rangle + \gamma^2 \tilde{\mathtt{B}}^\star.\end{aligned}$$

Next, we use that under Assumption 4.9, $\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\rangle \geq \mathtt{m}\|\vartheta_{\gamma k} - \theta_k\|^2$ and $|\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle| \leq \mathtt{L}\|\theta_k - \theta^\star\|^2$, which implies taking $\varepsilon = \mathtt{m}/2$ and since $2 - 5\gamma(\mathtt{m} + \mathtt{L}) \geq 0$,

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &\leq (1 - \gamma\mathtt{m}(1 - 5\gamma\mathtt{m})) \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 + \gamma^2 \tilde{\mathtt{M}}\mathtt{L} \left\|\theta_k - \theta^\star\right\|^2 + \gamma^2 \tilde{\mathtt{B}}^\star \\ &+ (5\gamma + \mathtt{m}^{-1}) \int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s \\ &. \quad (4.32)\end{aligned}$$

Further, for any $s \in \mathbb{R}_+$, using Durmus and Moulines (2019, Lemma 21) we have

$$\mathtt{L}^{-2}\, \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \leq ds\left(2 + s^2\mathtt{L}^2/3\right) + 3s^2\mathtt{L}^2/2 \left\|\vartheta_{\gamma k} - \theta^\star\right\|^2.$$

Integrating the previous inequality on $[0, \gamma]$, for $k \geq 0$ we obtain

$$\mathtt{L}^{-2} \int_0^\gamma \mathbb{E}^{\mathcal{F}_k} \left[\left\| \nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k}) \right\|^2 \right] \mathrm{d}s \leq d\gamma^2 + d\gamma^4 \mathtt{L}^2/12 + \gamma^3 \mathtt{L}^2/2 \left\| \vartheta_{\gamma k} - \theta^\star \right\|^2 .$$

Plugging this bounds in (4.32) and taking the expectation combined with Lemma 4.16 conclude the proof. ■

4.A.3 Proof of Theorem 4.5

Based on Theorem 4.15, the next corollary provides an upper bound in Wasserstein distance between π and μQ_γ^k, where we consider $(\theta_k)_{k\in\mathbb{N}}$ defined in (4.8) and starting from θ following $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

Theorem 4.18. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.11 and Assumption 4.12. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step-size $\gamma \in (0, \bar{\gamma}]$ where $\bar{\gamma}$ is defined in (4.A.2), any $k \in \mathbb{N}$, we have*

$$W_2^2(\mu Q_\gamma^k, \pi) \leq (1 - \gamma \mathtt{m}/2)^k W_2^2(\mu, \pi) + \gamma B_{\bar{\gamma}} + \gamma^2 A_{\bar{\gamma}} (1 - \mathtt{m}\gamma/2)^{k-1} k \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta),$$

where Q_γ is defined in (4.12) and

$$\begin{aligned} B_{\bar{\gamma}} &= \frac{2d\mathtt{L}^2}{\mathtt{m}} \left(\frac{1}{\mathtt{m}} + 5\bar{\gamma} \right) \left[1 + \bar{\gamma}\mathtt{L}^2/(2\mathtt{m}) + \bar{\gamma}^2 \mathtt{L}^2/12 \right] + \frac{4}{\mathtt{m}} \left[\sigma_\star^2 + (\mathtt{B}^\star/n) \sum_{i=1}^n \frac{1 - p_i + \omega_i}{p_i} \right] \\ &\quad + \frac{8\mathtt{L}}{\mathtt{m}^2} \max_{i\in[n]} \left\{ \frac{\mathtt{M}_i(1+\omega_i)}{p_i} \right\} \left[d + \bar{\gamma}[\sigma_\star^2 + \frac{\mathtt{B}^\star}{n} \sum_{i=1}^n \frac{1 - p_i + \omega_i}{p_i}] \right] \\ A_{\bar{\gamma}} &= 2\mathtt{L} \max_{i\in[n]} \left\{ \frac{\mathtt{M}_i(1+\omega_i)}{p_i} \right\}. \end{aligned} \tag{4.33}$$

Proof By Lemma 4.13, the assumption Assumption 4.14 is satified for a choice of $\tilde{\mathtt{M}} = 2\max_{i\in[n]}\{\mathtt{M}_i(1+\omega_i)/p_i\})$ and $\tilde{\mathtt{B}}^\star = 2[\sigma_\star^2 + (\mathtt{B}^\star/n)\sum_{i=1}^n (1 - p_i + \omega_i)/p_i]$. Therefore, applying Theorem 4.15 completes the proof. ■

4.B Proof of Theorem 4.7

We assume here that $\{U_i\}_{i\in[n]}$ are defined, for any $i \in [n]$ and $\theta \in \mathbb{R}^d$, by

$$U_i(\theta) = \sum_{j=1}^{N_i} U_{i,j}(\theta), \quad N_i \in \mathbb{N}^\star.$$

We consider the following set of assumptions on $\{U_i\}_{i\in[n]}$ and $\{U_{i,j} : j \in [N_i]\}_{i\in[n]}$.

Assumption 4.19. *For any $i \in [n], j \in [N_i]$, $U_{i,j}$ is continuously differentiable and the following conditions hold.*

(i) There exist $\{\mathtt{M}_i > 0\}_{i\in[n]}$, such that for any $i \in [n]$, $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\|\nabla U_i(\theta_2) - \nabla U_i(\theta_2)\right\|^2 \leq \mathtt{M}_i \left\langle \theta_2 - \theta_1, \nabla U_i(\theta_2) - \nabla U_i(\theta_1)\right\rangle .$$

(ii) There exists $\bar{\mathtt{M}} \geq 0$ such that, for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\|\nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1)\right\|^2 \leq \bar{\mathtt{M}} \left\langle \nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1), \theta_2 - \theta_1\right\rangle .$$

In all this section, we assume for any $i \in [n]$ that $b_i \in \mathbb{N}^*$, $b_i \leq N_i$ is fixed. For any $i \in [n]$, recall that $\wp_{N_i}$ denotes the power set of $[N_i]$ and

$$\wp_{N_i,b_i} = \{x \in \wp_{N_i} \,:\, \mathrm{Card}(x) = b_i\}.$$

We set in this section $\nu_1^{(i)}$ as the uniform distribution on $\wp_{N_i,b_i}$. We consider the family of measurable functions $\{H_i^\star : \mathbb{R}^d \times \mathbb{R}^d \times \wp_{N_i} \to \mathbb{R}^d\}_{i\in[n]}$, defined for any $i \in [n]$, $\theta \in \mathbb{R}^d$, $x \in \wp_{N_i,b_i}$ by

$$H_i^\star(\theta, x) = \frac{N_i}{b_i} \sum_{j=1}^{N_i} \mathbf{1}_x(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right]. \tag{4.34}$$

Using this specific family of gradient estimators boils down to the QLSD* algorithm detailed in Algorithm 4.7.

Algorithm 4.7 Variance-reduced Quantised Langevin Stochastic Dynamics (QLSD*)

Input: minibatch sizes $\{b_i\}_{i\in[n]}$, number of iterations K, compression operators $\{\mathcal{C}_{k+1}\}_{k\in\mathbb{N}^\star}$, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and initial point θ_0.
for $k = 0$ **to** $K - 1$ **do**
  for $i \in \mathcal{A}_{k+1}$ // On active clients **do**
    Draw $\mathcal{S}_{k+1}^{(i)} \sim \mathrm{Uniform}(\wp_{N_i,b_i})$.
    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/b_i) \sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\theta^\star)]$.
    Compute $g_{i,k+1} = \mathcal{C}_{k+1}(H_{k+1}^{(i)}(\theta_k))$.
    Send $g_{i,k+1}$ to the central server.
  // On the central server
  Compute $g_{k+1} = \frac{n}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the n clients.
Output: samples $\{\theta_k\}_{k=0}^K$.

Let $(X_k^{(1,1)}, \ldots, X_k^{(1,n)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,n)})_{k\in\mathbb{N}^*}$ be two independent i.i.d. sequences with distribution $\otimes_{i=1}^n \nu_1^{(i)}$ and $\nu_2^{\otimes n}$. Let $(Z_k)_{k\in\mathbb{N}^*}$ be an i.i.d. sequence of d-dimensional standard Gaussian random variables independent of $(X_k^{(1,1)}, \ldots, X_k^{(1,n)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,n)})_{k\in\mathbb{N}^*}$. Similarly, as before, we consider the partial device participation context where at each communication round $k \geq 1$, each client has a probability $p_i \in (0, 1]$ of participating, independently of other clients. In other words, there exists a sequence $(X_k^{(3,1)}, \cdots, X_k^{(3,n)})_{k\in\mathbb{N}^\star}$ of i.i.d. random variables distributed according

$\nu_3 = \text{Uniform}((0,1])$, such that for any $k \geq 1$ and $i \in [n]$, client i is active at step k if $X_k^{(3,i)} \leq p_i$. We denote $\mathcal{A}_{k+1} = \{i \in [n]; X_{k+1}^{(3,i)} \leq p_i\}$ the set of active clients at round k. For ease of notation, denote for any $k \in \mathbb{N}^*$, $X_k^{(1)} = (X_k^{(1,1)}, \ldots, X_k^{(1,n)})$, $X_k^{(2)} = (X_k^{(2,1)}, \ldots, X_k^{(2,n)})$, $X_k^{(3)} = (X_k^{(3,1)}, \ldots, X_k^{(3,n)})$ and $X_k = (X_k^{(1)}, X_k^{(2)}, X_k^{(3)})$.

Note that with this notation and under Assumption 4.10, QLSD* can be cast into the framework of the generalized QLSD scheme defined in (4.8) since the recursion associated to QLSD* can be written as

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - \gamma \sum_{i=1}^n \mathcal{S}_i \left[\mathcal{C}_i \left(H_i^\star(\tilde{\theta}_k, X_{k+1}^{(1,i)}), X_{k+1}^{(2,i)} \right), X_{k+1}^{(3,i)} \right] + \sqrt{2\gamma} Z_{k+1}, \qquad k \in \mathbb{N}, \tag{4.35}$$

where, for any $i \in [n]$, $\mathcal{S}_i$ is defined in (4.9). Therefore, we only need to verify that Assumption 4.14 is satisfied with $\mathsf{X}^{(i)} = \tilde{\mathsf{X}}^{(i)} = \mathsf{X}_1^{(i)} \times \mathsf{X}_2 \times \mathsf{X}_3$, $\mathcal{X}^{(i)} = \tilde{\mathcal{X}}^{(i)} = \mathcal{X}_1^{(i)} \otimes \mathcal{X}_2 \otimes \mathcal{X}_3$, $\tilde{\nu}^{(i)} = \nu_1^{(i)} \otimes \nu_2 \otimes \nu_3$ for $i \in [n]$ and $\{F_i\}_{i=1}^n = \{F_i^\star\}_{i=1}^n = \{\mathcal{S}_i \circ \mathcal{C}_i \circ H_i^\star\}_{i=1}^n$. This is done in Section 4.B.2.

4.B.1 Proof of Theorem 4.7

The Markov kernel associated with (4.35) is given for any $(\theta, \mathsf{A}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ by

$$Q_{\circledast,\gamma}(\theta, \mathsf{A}) = \int_{\mathsf{A} \times \tilde{\mathsf{X}}^n} \exp\left(-\|\theta - \theta + \gamma \sum_{i=1}^n F_i^\star(\theta, x^{(i)})\|^2/(4\gamma) \right) \frac{\mathrm{d}\tilde{\theta} \otimes_{i=1}^n \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)})}{(4\pi\gamma)^{d/2}}. \tag{4.36}$$

Then, the following non-asymptotic convergence result holds for QLSD*.

Theorem 4.20. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step-size $\gamma \in (0, \bar{\gamma}]$ where $\bar{\gamma}$ is defined in (4.A.2), any $k \in \mathbb{N}$, we have*

$$W_2^2(\mu Q_{\circledast,\gamma}^k, \pi) \leq (1 - \gamma \mathtt{m}/2)^k W_2^2(\mu, \pi) + \gamma B_{\circledast,\bar{\gamma}} + \gamma^2 A_{\circledast,\bar{\gamma}} (1 - \mathtt{m}\gamma/2)^{k-1} k \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta),$$

where $Q_{\circledast,\gamma}$ is defined in (4.36) and

$$\begin{aligned} B_{\circledast,\bar{\gamma}} &= (2d\mathtt{L}^2/\mathtt{m}) \left(1/\mathtt{m} + 5\bar{\gamma}\right) \left[1 + \bar{\gamma}\mathtt{L}^2/(2\mathtt{m}) + \bar{\gamma}^2\mathtt{L}^2/12\right] \\ &\quad + (4\mathtt{L}d\bar{\mathtt{M}}/\mathtt{m}^2) \max_{i \in [n]} \left[\omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{b_i,N_i})\right] \\ A_{\circledast,\bar{\gamma}} &= \mathtt{L}\bar{\mathtt{M}} \max_{i \in [n]} \left[\omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{b_i,N_i})\right], \end{aligned} \tag{4.37}$$

A_{b_i,N_i} being defined in (4.38) for any $i \in [n]$.

Proof Using Lemma 4.22, Assumption 4.14 is satisfied and applying Theorem 4.15 completes the proof. ∎

4.B.2 Supporting Lemmata

In this subsection, we derive two key lemmata in order to prove Theorem 4.20.

Lemma 4.21. *For any $i \in [n]$ and any sequence $\{a_j\}_{j=1}^{N_i} \in (\mathbb{R}^d)^{\otimes N_i}$ where $N_i \geq 2$, we have*

$$\int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \left[\mathbf{1}_{x^{(1)}}(j) - \frac{b_i}{N_i} \right] a_j \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1)}) \leq \frac{b_i(N_i - b_i)}{N_i(N_i-1)} \sum_{j=1}^{N_i} \|a_j\|^2.$$

Proof Let $i \in [n]$ and $X^{(1,i)}$ distributed according to $\nu_1^{(i)}$. Since $\sum_{j=1}^{N_i} \mathbf{1}_{X^{(1,i)}}(j) = b_i$, we have

$$\sum_{l=1}^{N_i} \mathbf{1}_{X^{(1,i)}}(l) + \sum_{j \neq j'} \mathbf{1}_{X^{(1,i)}}(j) \mathbf{1}_{X^{(1,i)}}(j') = b_i^2.$$

Integrating this equality over $\mathsf{X}_1^{(i)}$ gives

$$N_i \times \frac{b_i}{N_i} + N_i(N_i-1) \times \int_{\mathsf{X}_1^{(i)}} \left[\mathbf{1}_{x^{(1,i)}}(1) \mathbf{1}_{x^{(1,i)}}(2) \right] \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = b_i^2.$$

Thus, we deduce that $\int_{\mathsf{X}_1^{(i)}} [\mathbf{1}_{x^{(1,i)}}(1) \mathbf{1}_{x^{(1,i)}}(2)] \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = b_i(b_i-1)[N_i(N_i-1)]^{-1}$. In addition, using that

$$\int_{\mathsf{X}_1^{(i)}} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{b_i}{N_i} \right) \left(\mathbf{1}_{x^{(1,i)}}(j') - \frac{b_i}{N_i} \right) \nu_1^{(i)}(\mathrm{d}x^{(1,i)})$$
$$= \int_{\mathsf{X}_1^{(i)}} [\mathbf{1}_{x^{(1,i)}}(1) \mathbf{1}_{x^{(1,i)}}(2)] \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) - \frac{b_i^2}{N_i^2},$$

we obtain

$$\int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \left[\mathbf{1}_{x^{(1,i)}}(j) - \frac{b_i}{N_i} \right] a_j \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)})$$
$$= \frac{b_i(N_i-b_i)}{N_i^2} \left[\sum_{l=1}^{N_i} \|a_l\|^2 - \sum_{j \neq j'} \frac{\langle a_j, a_{j'} \rangle}{N_i - 1} \right] = \frac{b_i(N_i-b_i)}{N_i^2(N_i-1)} \left[N_i \sum_{l=1}^{N_i} \|a_l\|^2 - \left\| \sum_{l=1}^{N_i} a_l \right\|^2 \right].$$

■

For any $i \in [n]$, denote

$$A_{b_i,N_i} = \frac{N_i(N_i-b_i)}{b_i(N_i-1)}. \tag{4.38}$$

The next lemma aims at controlling the variance of the global stochastic gradient considered in QLSD*, required to apply Theorem 4.15.

Lemma 4.22. *Assume Assumption 4.10, Assumption 4.12 and Assumption 4.19. Then, for any $\theta \in \mathbb{R}^d$, we have*

$$\int_{\mathsf{X}^{(1:n)}} \left\| \sum_{i=1}^n \mathcal{S}_i \left[\mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right), x^{(3,i)} \right] - \nabla U(\theta) \right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) \\ \leq \bar{\mathtt{M}} \max_{i \in [n]} \left[\omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{b_i, N_i}) \right] \left\langle \theta - \theta^\star, \nabla U(\theta) - \nabla U(\theta^\star) \right\rangle,$$

where $\{H_i^\star\}_{i \in [n]}$ and $\{A_{b_i,N_i}\}_{i \in [n]}$ are defined in (4.34) and (4.38), respectively. Hence, Assumption 4.14 is satisfied with $\tilde{\mathtt{B}}^\star = 0$ and

$$\tilde{\mathtt{M}} = \bar{\mathtt{M}} \max_{i \in [n]} \left[\omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{b_i, N_i}) \right].$$

Proof Let $\theta \in \mathbb{R}^d$, using Assumption 4.10 gives

$$\begin{aligned}
&\int_{\mathsf{X}^{(1:n)}} \left\| \sum_{i=1}^n \mathcal{S}_i \left[\mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right), x^{(3,i)} \right] - \nabla U(\theta) \right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) \\
&= \int_{\mathsf{X}^{(1:n)}} \left\| \sum_{i=1}^n \mathcal{S}_i \left[\mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right), x^{(3,i)} \right] - \mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) \right\|^2 \otimes_{i=1}^n \nu^{(i)}(\mathrm{d}x^{(i)}) \\
&+ \int_{\mathsf{X}_1^{(1:n)} \times \mathsf{X}_2^n} \left\| \sum_{i=1}^n \mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes n}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \sum_{i=1}^n \left(\frac{1 - p_i}{p_i} \right) (\omega_i + 1) \int_{\mathsf{X}_1^{(i)}} \left\| H_i^\star(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \int_{\mathsf{X}_1^{(1:n)} \times \mathsf{X}_2^n} \left\| \sum_{i=1}^n \mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes n}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \bar{\mathtt{M}} \sum_{i=1}^n \left(\frac{1 - p_i}{p_i} \right) (\omega_i + 1) N_i \left\langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \right\rangle \\
&+ \int_{\mathsf{X}_1^{(1:n)} \times \mathsf{X}_2^n} \left\| \sum_{i=1}^n \mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes n}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)}).
\end{aligned} \tag{4.39}$$

Again using Assumption 4.10, it follows that

$$\begin{aligned}
&\int_{\mathsf{X}_1^{(1:n)} \times \mathsf{X}_2^n} \left\| \sum_{i=1}^n \mathcal{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes n}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \int_{\mathsf{X}_1^{(1:n)} \times \mathsf{X}_2^n} \left\| \sum_{i=1}^n \mathcal{C}_i \left(\frac{N_i}{b_i} \sum_{j=1}^{N_i} \mathbf{1}_{x^{(1,i)}}(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star) \right], x^{(2,i)} \right) \right. \\
&\left. - \sum_{i=1}^n \frac{N_i}{b_i} \sum_{j=1}^{N_i} \mathbf{1}_{x^{(1,i)}}(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star) \right] \right\|^2 \\
&+ \int_{\mathsf{X}_1^{(1:n)}} \left\| \sum_{i=1}^n \frac{N_i}{b_i} \sum_{j=1}^{N_i} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{b_i}{N_i} \right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star) \right] \right\|^2 \otimes_{i=1}^n \nu_1^{(i)}(\mathrm{d}x^{(1,i)})
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \omega_i \left(\frac{N_i}{b_i}\right)^2 \int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \mathbf{1}_{x^{(1,i)}}(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \sum_{i=1}^{n} \left(\frac{N_i}{b_i}\right)^2 \int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{b_i}{N_i}\right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \sum_{i=1}^{n} \omega_i \left\|\nabla U_i(\theta) - \nabla U_i(\theta^\star)\right\|^2 \\
&+ \sum_{i=1}^{n} (\omega_i + 1) \left(\frac{N_i}{b_i}\right)^2 \int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{b_i}{N_i}\right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}).
\end{aligned} \tag{4.40}
$$

Using Lemma 4.21 combined with Assumption 4.19 yields, for any $i \in [n]$,

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(i)}} \left\| \textstyle\sum_{j=1}^{N_i} \left(\mathbf{1}_{x^{(1,i)}}(j) - b_i/N_i\right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\qquad \leq \frac{b_i(N_i - b_i)}{N_i(N_i - 1)} \bar{\mathtt{M}} \left\langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \right\rangle .
\end{aligned} \tag{4.41}
$$

In addition, Jensen inequality implies, for any $i \in [n]$, that

$$
\left\|\nabla U_i(\theta) - \nabla U_i(\theta^\star)\right\|^2 \leq N_i \sum_{j=1}^{N_i} \left\|\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right\|^2,
$$

and therefore, using Assumption 4.19, we have for any $i \in [n]$,

$$
\left\|\nabla U_i(\theta) - \nabla U_i(\theta^\star)\right\|^2 \leq \bar{\mathtt{M}} N_i \left\langle \nabla U_i(\theta) - \nabla U_i(\theta^\star), \theta - \theta^\star \right\rangle . \tag{4.42}
$$

Injecting (4.41) and (4.42) into (4.40) and using (4.39) conclude the proof. ■

4.C Proof of Theorem 4.8

4.C.1 Problem formulation.

We assume here that U is still of the form (4.1) and that there exist $\{N_i \in \mathbb{N}^\star\}_{i\in[n]}$ such that for any $i \in [n]$, there exist N_i functions $\{U_{i,j} : \theta \in \mathbb{R}^d \to \mathbb{R}\}_{j\in[N_i]}$ such that for any $\theta \in \mathbb{R}^d$,

$$
U_i(\theta) = \sum_{j=1}^{N_i} U_{i,j}(\theta).
$$

In all this section, we assume for any $i \in [n]$ that $b_i \in \mathbb{N}^*$, $b_i \leq N_i$ is fixed. Recall that $\wp_N$ denotes the power set of $[N]$ and

$$
\wp_{N,n} = \{x \in \wp_N \,:\, \mathrm{Card}(x) = n\}.
$$

In addition, we set in this section $\nu_1^{(i)}$ as the uniform distribution on $\wp_{N_i,b_i}$. We consider the family of measurable functions $\{G_i : \mathbb{R}^d \times \mathbb{R}^d \times \wp_{N_i} \to \mathbb{R}^d\}_{i\in[n]}$, defined for any $i \in [n]$, $\theta \in \mathbb{R}^d$, $\zeta \in \mathbb{R}^d$, $x \in \wp_{N_i,b_i}$ by

$$G_i(\theta, \zeta; x) = \frac{N_i}{b_i} \sum_{j=1}^{N_i} \mathbf{1}_x(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\zeta)\right] + \nabla U_i(\zeta). \tag{4.43}$$

For ease of reading, we formalise more precisely the recursion associated with QLSD^{++} under Assumption 4.10. Let $(X_k^{(1,1)}, \ldots, X_k^{(1,n)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,n)})_{k\in\mathbb{N}^*}$ be two independent i.i.d. sequences with distribution $\otimes_{i=1}^n \nu_1^{(i)}$ and $\nu_2^{\otimes n}$. Let $(Z_k)_{k\in\mathbb{N}^*}$ be an i.i.d. sequence of d-dimensional standard Gaussian random variables independent of $(X_k^{(1,1)}, \ldots, X_k^{(1,n)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,n)})_{k\in\mathbb{N}^*}$. Similarly as before, we consider the partial device participation context where at each communication round $k \geq 1$, each client has a probability $p_i \in (0, 1]$ of participating, independently of other clients. In other words, there exists a sequence $(X_k^{(3,1)}, \cdots, X_k^{(3,n)})_{k\in\mathbb{N}^\star}$ of i.i.d. random variables distributed according $\nu_3 = \text{Uniform}((0, 1])$, such that for any $k \geq 1$ and $i \in [n]$, client i is active at step k if $X_k^{(3,i)} \leq p_i$. We denote $\mathcal{A}_{k+1} = \{i \in [n]; X_{k+1}^{(3,i)} \leq p_i\}$ the set of active clients at round k. For ease of notation, denote for any $k \in \mathbb{N}^*$, $X_k^{(1)} = (X_k^{(1,1)}, \ldots, X_k^{(1,n)})$, $X_k^{(2)} = (X_k^{(2,1)}, \ldots, X_k^{(2,n)})$, $X_k^{(3)} = (X_k^{(3,1)}, \ldots, X_k^{(3,n)})$ and $X_k = (X_k^{(1)}, X_k^{(2)}, X_k^{(3)})$. Let $l \in \mathbb{N}^\star$, $\gamma \in (0, \bar{\gamma}]$ and $\alpha \in (0, \bar{\alpha}]$ for $\bar{\gamma}, \bar{\alpha} > 0$. Given $\Theta_0 = (\theta_0, \zeta_0, \{\eta_0^{(i)}\}_{i\in[n]}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times n}$, with $\zeta_0 = \theta_0$, we recursively define the sequence $(\Theta_k)_{k\in\mathbb{N}} = (\theta_k, \zeta_k, \{\eta_k^{(i)}\}_{i\in[n]})_{k\in\mathbb{N}}$, for any $k \in \mathbb{N}$ as

$$\theta_{k+1} = \theta_k - \gamma \tilde{G}(\Theta_k; X_{k+1}) + \sqrt{2\gamma} Z_{k+1}, \tag{4.44}$$

where

$$\tilde{G}(\Theta_k; X_{k+1}) = \sum_{i=1}^n \left[\mathcal{S}_i \left(\mathcal{C}_i \left\{ G_i \left(\theta_k, \zeta_k; X_{k+1}^{(1,i)} \right) - \eta_k^{(i)}; X_{k+1}^{(2,i)} \right\}, X_{k+1}^{(3,i)} \right) + \eta_k^{(i)} \right], \tag{4.45}$$

$$\zeta_{k+1} = \begin{cases} \theta_{k+1}, \text{ if } k+1 \equiv 0 \pmod{l}, \\ \zeta_k, \text{ otherwise,} \end{cases} \tag{4.46}$$

and for any $i \in [n]$,

$$\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha \mathcal{S}_i \left(\mathcal{C}_i \left\{ G_i \left(\theta_k, \zeta_k; X_{k+1}^{(1,i)} \right) - \eta_k^{(i)}; X_{k+1}^{(2,i)} \right\}, X_{k+1}^{(3,i)} \right). \tag{4.47}$$

Since QLSD^{++} involves auxiliary variables gathered with $(\theta_k)_{k\in\mathbb{N}}$ in $(\Theta_k)_{k\in\mathbb{N}}$, we cannot follow the same proof as for QLSD* by verifying Assumption 4.14 and then applying Theorem 4.15. Instead, we will adapt the proof Theorem 4.15 and in particular Lemma 4.16 and bound the variance associated to the stochastic gradient defined in (4.45). Once this variance term will be tackled, the proof of Theorem 4.8 will follow the same lines as the proof of Theorem 4.15 upon using specific moment estimates for QLSD^{++}. In the next section, we focus on these two goals: we provide uniform bounds in the number of iterations k on the variance of the sequence of stochastic gradients associated with QLSD^{++}, $(\mathbb{E}[\|\tilde{G}_i(\Theta_k, X_{k+1}) - \nabla U(\theta_k)\|^2])_{k\in\mathbb{N}}$ for any $i \in [n]$, and $(\mathbb{E}[\|\theta_k - \theta^\star\|^2])_{k\in\mathbb{N}}$, see Proposition 4.29 and Corollary 4.28. To this end, a key ingredient is the design of an appropriate Lyapunov function defined in (4.59).

4.C.2 Uniform bounds on the stochastic gradients and moment estimates for QLSD^{++}

Consider the filtration associated with $(\Theta_k)_{k\in\mathbb{N}}$ defined by $\mathcal{G}_0 = \sigma(\Theta_0)$ and for $k \in \mathbb{N}^*$,

$$\mathcal{G}_k = \sigma(\Theta_0, (X_{\tilde{k}})_{\tilde{k}\le k}, (Z_{\tilde{k}})_{\tilde{k}\le k}).$$

We denote for any $i \in [n]$, $\theta, \zeta \in \mathbb{R}^d$,

$$\Delta_i(\theta,\zeta) = \nabla U_i(\theta) - \nabla U_i(\zeta). \tag{4.48}$$

Similarly, we consider, for any $i \in [n]$, $j \in [N]$, $\theta, \zeta \in \mathbb{R}^d$,

$$\Delta_{i,j}(\theta,\zeta) = \nabla U_{i,j}(\theta) - \nabla U_{i,j}(\zeta). \tag{4.49}$$

The following lemma provides a first upper bound on the variance of the stochastic gradients used in QLSD^{++}.

Lemma 4.23. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19 and let $\gamma \in (0,\bar{\gamma}]$, $\alpha \in (0,\bar{\alpha}]$ for some $\bar{\gamma},\bar{\alpha} > 0$. Then, for any $s \in \mathbb{N}$, $r \in \{0,\dots,l-1\}$, we have*

$$\begin{aligned}
&\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_{sl+r})\right\|^2\right] \\
&\qquad \le \left[2\sum_{i=1}^n \frac{\mathtt{M}_i^2}{p_i}(\omega_i + 1 - p_i) + \left(\frac{\omega_i+1}{p_i}\right) A_{b_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right]\left\|\theta_{sl+r} - \theta^\star\right\|^2 \\
&+ \left[2\sum_{i=1}^n \frac{\omega_i + 1 - p_i}{p_i}\right]\left\|\nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)}\right\|^2 + 2\bar{\mathtt{M}}\sum_{i=1}^n\left[\left(\frac{\omega_i+1}{p_i}\right)A_{b_i,N_i}\mathtt{M}_i\right]\left\|\theta_{sl} - \theta^\star\right\|^2,
\end{aligned}$$

where $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^{(i)}\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$, $\tilde{G}$ and $A_{b,N}$ are defined in (4.44), (4.46), (4.47), (4.45) and (4.38), respectively.

Proof Let $s \in \mathbb{N}$ and $r \in \{0,\dots,l-1\}$. Using Assumption 4.10, (4.48) and (4.49), we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_{sl+r})\right\|^2\right] \\
&= \sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r}}\Bigg[\Bigg\|\mathcal{S}_i\left(\mathcal{C}_i\left\{G_i\left(\theta_{sl+r},\zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\}, X_{sl+r+1}^{(3,i)}\right) \\
&- \mathcal{C}_i\left\{G_i\left(\theta_{sl+r},\zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\}\Bigg\|^2\Bigg] \\
&+ \sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\mathcal{C}_i\left\{G_i\left(\theta_{sl+r},\zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\} + \eta_{sl+r}^{(i)} - \nabla U_i(\theta_{sl+r})\right\|^2\right] \\
&\le \sum_{i=1}^n \left(\frac{1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\mathcal{C}_i\left\{G_i\left(\theta_{sl+r},\zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\}\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{n}\omega_i\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r},\zeta_{sl+r};X_{sl+r+1}^{(1,i)}\right)-\eta_{sl+r}^{(i)}\right\|\right]\\
&+\sum_{i=1}^{n}\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r},\zeta_k;X_{sl+r+1}^{(1,i)}\right)-\nabla U_i(\theta_{sl+r})\right\|^2\right]\\
&\leq\sum_{i=1}^{n}\left(\frac{\omega_i+1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r},\zeta_{sl+r};X_{sl+r+1}^{(1,i)}\right)-\eta_{sl+r}^{(i)}\right\|^2\right]\\
&+\sum_{i=1}^{n}\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r},\zeta_{sl+r};X_{sl+r+1}^{(1,i)}\right)-\nabla U_i(\theta_{sl+r})\right\|^2\right]\\
&\leq\sum_{i=1}^{n}\left(\frac{\omega_i+1}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\frac{N_i}{b_i}\sum_{j=1}^{N_i}\left[\mathbf{1}_{X_{sl+r+1}^{(1,i)}}(j)\Delta_{i,j}(\theta_{sl+r},\zeta_{sl+r})\right]-\Delta_i(\theta_{sl+r},\zeta_{sl+r})\right\|^2\right]\\
&+\sum_{i=1}^{n}\left(\frac{\omega_i+1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\nabla U_i(\theta_{sl+r})-\eta_{sl+r}^{(i)}\right\|^2\right]\\
&\leq\sum_{i=1}^{n}\left(\frac{\omega_i+1}{p_i}\right)\frac{N_i(N_i-b_i)}{b_i(N_i-1)}\bar{\mathtt{M}}\langle\theta_{sl+r}-\zeta_{sl+r},\nabla U_i(\theta_{sl+r})-\nabla U_i(\zeta_{sl+r})\rangle\\
&+\sum_{i=1}^{n}\left(\frac{\omega_i+1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\nabla U_i(\theta_{sl+r})-\eta_{sl+r}^{(i)}\right\|^2\right],
\end{aligned}
$$

where the last line follows from Assumption 4.19 and Lemma 4.21. The proof is concluded by using the Cauchy-Schwarz inequality, Assumption 4.9 and $\zeta_{sl+r}=\theta_{sl}$. ■

The two following lemmas aim at controlling the terms that appear in Lemma 4.23.

Lemma 4.24. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19, and let $\gamma\in(0,\bar{\gamma}]$, $\alpha\in(0,\bar{\alpha}]$ for some $\bar{\gamma},\bar{\alpha}>0$. Then, for any $s\in\mathbb{N}$ and $r\in[l]$, we have*

$$
\begin{aligned}
\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\theta_{sl+r}-\theta^\star\right\|^2\right]&\leq\left(1-2\gamma m+\gamma^2B_{\mathbf{b},\mathbf{N}}\right)\left\|\theta_{sl+r-1}-\theta^\star\right\|^2\\
&\quad+\gamma^2\left[2\sum_{i=1}^{n}\frac{\omega_i+1-p_i}{p_i}\right]\left\|\nabla U_i(\theta^\star)-\eta_{sl+r-1}^{(i)}\right\|^2\\
&\quad+2\bar{\mathtt{M}}\gamma^2\sum_{i=1}^{n}\left[\left(\frac{\omega_i+1}{p_i}\right)A_{b_i,N_i}\mathtt{M}_i\right]\left\|\theta_{sl}-\theta^\star\right\|^2+2\gamma d,
\end{aligned}
$$

where

$$
B_{\mathbf{b},\mathbf{N}}=2\sum_{i=1}^{n}\left[\frac{\mathtt{M}_i^2}{p_i}(\omega_i+1-p_i)+\left(\frac{\omega_i+1}{p_i}\right)A_{b_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right]+\mathtt{L}^2, \tag{4.50}
$$

$(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}=(\theta_{\tilde{k}},\zeta_{\tilde{k}},\{\eta_{\tilde{k}}^i\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$ and $A_{b,N}$ are defined in (4.44), (4.46), (4.47) and (4.38) respectively.

Proof Let $s \in \mathbb{N}$ and $r \in [l]$. Using (4.44) and Assumption 4.10, it follows

$$\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\theta_{sl+r} - \theta^\star\right\|^2\right] = \left\|\theta_{sl+r-1} - \theta^\star\right\|^2 + 2\gamma d - 2\gamma\langle \nabla U(\theta_{sl+r-1}), \theta_{sl+r-1} - \theta^\star\rangle + \gamma^2 \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\tilde{G}(\Theta_{sl+r-1}; X_{sl+r})\right\|^2\right]. \quad (4.51)$$

Using Assumption 4.10 and (4.43)-(4.45), we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\tilde{G}(\Theta_{sl+r-1}; X_{sl+r})\right\|^2\right] \\
&= \sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\| \mathcal{S}_i\left(\mathcal{C}_i\left\{G_i\left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}; X_{sl+r}^{(2,i)}\right\}, X_{sl+r}^{(3,i)}\right)\right.\right. \\
&\left.\left. - \mathcal{C}_i\left\{G_i\left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}; X_{sl+r}^{(2,i)}\right\}\right\|^2\right] \\
&+ \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\sum_{i=1}^n \mathcal{C}_i\left\{G_i\left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}; X_{sl+r}^{(2,i)}\right\} + \eta_{sl+r-1}^{(i)}\right\|^2\right] \\
&\leq \sum_{i=1}^n \left(\frac{\omega_i + 1 - p_i}{p_i}\right) \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|G_i\left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}\right\|^2\right] \\
&+ \sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\frac{N_i}{b_i}\sum_{j=1}^{N_i}\left[\mathbf{1}_{X_{sl+r}^{(1,i)}}(j)\Delta_{i,j}(\theta_{sl+r-1}, \zeta_{sl+r-1})\right] - \Delta_i(\theta_{sl+r-1}, \zeta_{sl+r-1})\right\|^2\right] \\
&+ \left\|\nabla U(\theta_{sl+r-1})\right\|^2 \\
&= \sum_{i=1}^n \left(\frac{\omega_i + 1}{p_i}\right) \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\frac{N_i}{b_i}\sum_{j=1}^{N_i}\left[\mathbf{1}_{X_{sl+r}^{(1,i)}}(j)\Delta_{i,j}(\theta_{sl+r-1}, \zeta_{sl+r-1})\right] - \Delta_i(\theta_{sl+r-1}, \zeta_{sl+r-1})\right\|^2\right] \\
&+ \sum_{i=1}^n \left(\frac{\omega_i + 1 - p_i}{p_i}\right) \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\nabla U_i(\theta_{sl+r-1}) - \eta_{sl+r-1}^{(i)}\right\|^2\right] + \left\|\nabla U(\theta_{sl+r-1})\right\|^2 \\
&\leq \sum_{i=1}^n \left(\frac{\omega_i + 1}{p_i}\right) \frac{N_i(N_i - b_i)}{b_i(N_i - 1)} \bar{\mathtt{M}}\langle \theta_{sl+r-1} - \zeta_{sl+r-1}, \nabla U_i(\theta_{sl+r-1}) - \nabla U_i(\zeta_{sl+r-1})\rangle \\
&+ \sum_{i=1}^n \left(\frac{\omega_i + 1 - p_i}{p_i}\right) \left\|\nabla U_i(\theta_{sl+r-1}) - \eta_{sl+r-1}^{(i)}\right\|^2 + \left\|\nabla U(\theta_{sl+r-1})\right\|^2, \quad (4.52)
\end{aligned}$$

where the last line follows from Assumption 4.19 and Lemma 4.21. The proof is concluded by injecting (4.52) into (4.51), using the Cauchy-Schwarz inequality, $\nabla U(\theta^\star) = 0$, Assumption 4.9 and $\zeta_{sl+r-1} = \theta_{sl}$. ■

Lemma 4.25. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. Let $\gamma \in (0, \bar{\gamma}]$ for some $\bar{\gamma} > 0$ and $\alpha \in (0, 1/(\max_{i\in[n]} \omega_i + 1)]$. Then, for any $s \in \mathbb{N}$ and $r \in [l]$, we have*

$$\sum_{i=1}^{n} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)} \right\|^2 \right] \leq (1-\alpha) \sum_{i=1}^{n} \left\| \nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)} \right\|^2 + \alpha C_{\mathbf{b},\mathbf{N}} \left\| \theta_{sl+r-1} - \theta^\star \right\|^2 + 2\alpha \left[\sum_{i=1}^{n} A_{b_i,N_i} \bar{\mathtt{M}}\mathtt{M}_i \right] \left\| \theta_{sl} - \theta^\star \right\|^2,$$

where

$$C_{\mathbf{b},\mathbf{N}} = 2 \sum_{i=1}^{n} \left[A_{b_i,N_i} \bar{\mathtt{M}}\mathtt{M}_i + \mathtt{M}_i^2 \right], \tag{4.53}$$

$(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^i\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$ *and* $A_{b,N}$ *are defined in* (4.44), (4.46), (4.47) *and* (4.38)*, respectively.*

Proof Let $s \in \mathbb{N}$ and $r \in [l]$. Then, it follows

$$\sum_{i=1}^{n} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)} \right\|^2 \right] = \sum_{i=1}^{n} \left\| \nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)} \right\|^2 + \sum_{i=1}^{n} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right\|^2 \right] + 2 \sum_{i=1}^{n} \left\langle \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right], \eta_{sl+r-1}^{(i)} - \nabla U_i(\theta^\star) \right\rangle. \tag{4.54}$$

Using (4.47) and Assumption 4.10, we have for any $i \in [n]$,

$$\mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right\|^2 \right] \leq \alpha^2 (\omega_i + 1) \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right\|^2 \right], \tag{4.55}$$

$$\mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right] = \alpha \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right]. \tag{4.56}$$

Plugging (4.55) and (4.56) into (4.54) yields

$$\sum_{i=1}^{n} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)} \right\|^2 \right] \leq \sum_{i=1}^{n} \left\| \nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)} \right\|^2 + \alpha^2 \sum_{i=1}^{n} (\omega_i + 1) \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right\|^2 \right] + 2\alpha \sum_{i=1}^{n} \langle \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right], \eta_{sl+r-1}^{(i)} - \nabla U_i(\theta^\star) \rangle.$$

Using for any $i \in [n]$ $\alpha(1+\omega_i) \leq 1$ and the fact, for any $a, b, c \in \mathbb{R}^d$, that $\|a-c\|^2 + 2\langle a-c, c-b\rangle = \|a-b\|^2 - \|c-b\|^2$, we have

$$\sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)}\right\|^2\right] \leq (1-\alpha)\sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)}\right\|^2 + \alpha\sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|G_i\left(\theta_{sl+r-1}, \zeta_{sl+r}; X_{sl+r}^{(1,i)}\right) - \nabla U_i(\theta^\star)\right\|^2\right]. \tag{4.57}$$

Using (4.43), Assumption 4.19 and Lemma 4.21, it follows

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}^{\mathcal{G}_{sl+r-1}}&\left[\left\|G_i\left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \nabla U_i(\theta^\star)\right\|^2\right] \\ &\leq \sum_{i=1}^n \frac{N_i(N_i-b_i)}{b_i(N_i-1)}\bar{\mathtt{M}}\langle \theta_{sl+r-1} - \zeta_{sl+r-1}, \nabla U_i(\theta_{sl+r-1}) - \nabla U_i(\zeta_{sl+r-1})\rangle \\ &\quad + \sum_{i=1}^n \left\|\nabla U_i(\theta_{sl+r-1}) - \nabla U_i(\theta^\star)\right\|^2. \end{aligned} \tag{4.58}$$

The proof is concluded by plugging (4.58) into (4.57), using the Cauchy-Schwarz inequality, Assumption 4.9 and $\zeta_{sl+r-1} = \theta_{sl}$. ■

Lemma 4.24 and Lemma 4.25 involve two dependent terms which prevents us from using a straightforward induction. To cope with this issue, we consider a Lyapunov function $\psi : \mathbb{R}^d \times \mathbb{R}^{d\times n} \to \mathbb{R}$ defined, for any $\theta \in \mathbb{R}^d$ and $\eta = (\eta^{(1)}, \dots, \eta^{(n)})^\top \in \mathbb{R}^{d\times n}$ by

$$\psi(\theta, \eta) = \|\theta - \theta^\star\|^2 + \frac{3\gamma^2}{\alpha}\max_{i\in[n]}\left\{\frac{\omega_i + 1 - p_i}{p_i}\right\}\sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta^{(i)}\right\|^2. \tag{4.59}$$

The following lemma provides an upper bound on this Lyapunov function. Define for $\alpha > 0$,

$$\bar{\gamma}_{\alpha,1} = \{\mathtt{m}(B_{\mathbf{b},\mathbf{N}} + 3\omega C_{\mathbf{b},\mathbf{N}})^{-1}\} \wedge \{\alpha(3\mathtt{m})^{-1}\}, \tag{4.60}$$

where $B_{\mathbf{b},\mathbf{N}}$ and $C_{\mathbf{b},\mathbf{N}}$ are defined in (4.50) and (4.53) respectively.

Lemma 4.26. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. Let $\alpha \in (0, 1/(1+\max_{i\in[n]}\omega_i)]$, $\gamma \in (0, \bar{\gamma}_{\alpha,1}]$. Then, for any $s \in \mathbb{N}$ and $r \in [l]$, we have*

$$\begin{aligned}\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\psi(\theta_{sl+r}, \eta_{sl+r})\right] &\leq (1-\gamma\mathtt{m})\,\psi(\theta_{sl+r-1}, \eta_{sl+r-1}) \\ &\quad + 8\bar{\mathtt{M}}\gamma^2\max_{i\in[n]}\{(\omega_i+1)/p_i\}\sum_{i=1}^n A_{b_i,N_i}\mathtt{M}_i\left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d,\end{aligned}$$

where ψ is defined in (4.59) and $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^i\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$ and $A_{b,N}$ are defined in (4.44), (4.46), (4.47) and (4.38), respectively.

Proof Let $s \in \mathbb{N}$ and $r \in [l]$. Using Lemma 4.24 and Lemma 4.25, we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\psi(\theta_{sl+r}, \eta_{sl+r})\right] \\
&\qquad \leq \left(1 - 2\gamma\mathtt{m} + \gamma^2\left[B_{\mathbf{b},\mathbf{N}} + 3\omega C_{\mathbf{b},\mathbf{N}}\right]\right)\left\|\theta_{sl+r-1} - \theta^\star\right\|^2 \\
&\quad + \left[(2/3)\alpha + (1-\alpha)\right](3\gamma^2/\alpha)\max_{i\in[n]}\{(\omega_i + 1 - p_i)/p_i\}\sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta^{(i)}_{sl+r-1}\right\|^2 \\
&\qquad + 8\bar{\mathtt{M}}\gamma^2\max_{i\in[n]}\{(\omega_i+1)/p_i\}\sum_{i=1}^n A_{b_i,N_i}\mathtt{M}_i\left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d.
\end{aligned}$$

Since $\gamma \leq \bar{\gamma}_{\alpha,1}$ with $\bar{\gamma}_{\alpha,1}$ given in (4.60), it follows that

$$1 - 2\gamma\mathtt{m} + \gamma^2\left[B_{\mathbf{b},\mathbf{N}} + 3\omega C_{\mathbf{b},\mathbf{N}}\right] \leq 1 - 2\gamma\mathtt{m} + \gamma\mathtt{m} = 1 - \gamma\mathtt{m}.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\psi(\theta_{sl+r}, \eta_{sl+r})\right] &\leq \left(1 - \gamma\mathtt{m}\right)\psi(\theta_{sl+r-1}, \eta_{sl+r-1}) \\
&\quad + 8\bar{\mathtt{M}}\gamma^2\max_{i\in[n]}\{(\omega_i+1)/p_i\}\sum_{i=1}^n A_{b_i,N_i}\mathtt{M}_i\left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d.
\end{aligned}$$

■

Lemma 4.27. *Let $j \in \mathbb{N}^\star$ and fix $\gamma > 0$ such that*

$$\gamma \leq \frac{\mathtt{m}}{16j\bar{\mathtt{M}}\gamma^2\max_{i\in[n]}\{(\omega_i+1)/p_i\}\sum_{i=1}^n A_{b_i,N_i}\mathtt{M}_i} \wedge \frac{1}{\mathtt{m}}.$$

Then,

$$\left(1 - \gamma\mathtt{m}\right)^j + 8j\gamma^2\bar{\mathtt{M}}\max_{i\in[n]}\{(\omega_i+1)/p_i\}\sum_{i=1}^n A_{b_i,N_i}\mathtt{M}_i \leq 1 - \gamma\mathtt{m}/2,$$

where $A_{b,N}$ is defined in (4.38).

Proof The proof is straightforward using $(1-\gamma\mathtt{m})^j \leq 1 - \gamma\mathtt{m}$. ■

We have the following corollary regarding the Lyapunov function defined in (4.59). Denote for $\alpha > 0$,

$$\bar{\gamma}_{\alpha,2} = \bar{\gamma}_{\alpha,1} \wedge [\mathtt{m}/\{16l\bar{\mathtt{M}}\max_{i\in[n]}\{(\omega_i+1)/p_i\}\textstyle\sum_{i=1}^n A_{b_i,N_i}\mathtt{M}_i\}]^{1/3}, \tag{4.61}$$

where $\bar{\gamma}_{\alpha,1}$ is given in (4.60).

Corollary 4.28. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. Let $\alpha \in (0, 1/(1+\max_{i\in[n]}\omega_i)]$ and $\gamma \in \left(0, \bar{\gamma}_{\alpha,2}\right]$. Then, for any $s \in \mathbb{N}$ and $r \in \{0, \ldots, l-1\}$ we have*

$$\mathbb{E}^{\mathcal{G}_{sl}}\left[\psi(\theta_{(s+1)l-r}, \eta_{(s+1)l-r})\right] \leq \left(1 - \gamma\mathtt{m}/2\right)\psi(\theta_{sl}, \eta_{sl}) + 2\gamma(l-r)d,$$

where ψ is defined in (4.59) and $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta^i_{\tilde{k}}\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$ is defined in (4.44), (4.46), (4.47).

Proof The proof follows from a straightforward induction of Lemma 4.26 combined with Lemma 4.27. ■

We are now ready to control explicitly the variance of the stochastic gradient defined in (4.45).

Proposition 4.29. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. Let* $\alpha \in (0, 1/(1+\max_{i\in[n]}\omega_i)]$ *and* $\gamma \in \left(0, \bar{\gamma}_{\alpha,2}\right]$, *where* $\bar{\gamma}_{\alpha,2}$ *is defined in (4.61). Then, for any* $k = sl + r$ *with* $s \in \mathbb{N}$, $r \in \{0, \ldots, l-1\}$, $\theta_0 \in \mathbb{R}^d$ *and* $\eta_0 = (\eta_0^{(1)}, \ldots, \eta_0^{(n)})^\top \in \mathbb{R}^{d\times n}$, *we have*

$$\mathbb{E}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_k)\right\|^2\right] \leq \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^s D_{\mathbf{b},\mathbf{N}}\psi(\theta_0, \eta_0) + 4ldD_{\mathbf{b},\mathbf{N}}/\mathtt{m} + \left[2\sum_{i=1}^n \frac{\omega_i + 1 - p_i}{p_i}\right](1-\alpha)^k \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2\right],$$

where

$$D_{\mathbf{b},\mathbf{N}} = \left[2\sum_{i=1}^n \frac{\mathtt{M}_i^2}{p_i}(\omega_i + 1 - p_i) + \left(\frac{\omega_i+1}{p_i}\right) A_{b_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right] + 2\bar{\mathtt{M}}\sum_{i=1}^n \left[\left(\frac{\omega_i+1}{p_i}\right) A_{b_i,N_i}\mathtt{M}_i\right] + 4C_{\mathbf{b},\mathbf{N}}\sum_{i=1}^n (\omega_i + 1 - p_i)/p_i, \quad (4.62)$$

$A_{b,N}$ and $C_{n,N}$ are defined in (4.38) and (4.53) respectively, ψ is defined in (4.59), and $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^i\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$ is defined in (4.44), (4.46), (4.47).

Proof Let $k \in \mathbb{N}$ and write $k = sl + r$ with $s \in \mathbb{N}$, $r \in \{0, \ldots, l-1\}$ Then, using Lemma 4.23, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_k)\right\|^2\right] &\leq \left[2\sum_{i=1}^n \frac{\mathtt{M}_i^2}{p_i}(\omega_i + 1 - p_i) + \left(\frac{\omega_i+1}{p_i}\right) A_{b_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right] \mathbb{E}\left[\left\|\theta_k - \theta^\star\right\|^2\right] \\ &\quad + \left[2\sum_{i=1}^n \frac{\omega_i + 1 - p_i}{p_i}\right] \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_k^{(i)}\right\|^2\right] \\ &\quad + 2\bar{\mathtt{M}}\sum_{i=1}^n \left[\left(\frac{\omega_i+1}{p_i}\right) A_{b_i,N_i}\mathtt{M}_i\right] \mathbb{E}\left[\left\|\theta_{sl} - \theta^\star\right\|^2\right]. \quad (4.63)\end{aligned}$$

We now use our previous results to upper bound the three expectations at the right-hand side of (4.63). First, using Corollary 4.28 and a straightforward induction gives

$$\mathbb{E}\left[\left\|\theta_{sl} - \theta^\star\right\|^2\right] \leq \left(1 - \frac{\gamma\mathtt{m}}{2}\right)^s \psi(\theta_0, \eta_0) + 2\gamma ld \sum_{j=0}^{s-1}\left(1 - \frac{\gamma\mathtt{m}}{2}\right)^j$$

$$\leq \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^s \psi(\theta_0, \eta_0) + \frac{4ld}{\mathtt{m}}. \tag{4.64}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_k - \theta^\star\right\|^2\right] &\leq \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^{s+1} \psi(\theta_0, \eta_0) + 2\gamma ld \sum_{j=0}^{s} \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^j \\ &\leq \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^s \psi(\theta_0, \eta_0) + \frac{4ld}{\mathtt{m}}. \end{aligned} \tag{4.65}$$

Finally, using Lemma 4.25 combined with (4.64) and (4.65), we obtain

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_k^{(i)}\right\|^2\right] &\leq (1-\alpha) \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_{k-1}^{(i)}\right\|^2\right] \\ &\quad + 2\alpha C_{\mathbf{b},\mathbf{N}} \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^s \psi(\theta_0, \eta_0) + \frac{8ld\alpha C_{\mathbf{b},\mathbf{N}}}{\mathtt{m}}.\end{aligned}$$

Then, a straightforward induction leads to

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_k^{(i)}\right\|^2\right] &\leq (1-\alpha)^k \sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2 \\ &\quad + 2C_{\mathbf{b},\mathbf{N}} \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^s \psi(\theta_0, \eta_0) + \frac{8ldC_{\mathbf{b},\mathbf{N}}}{\mathtt{m}}. \end{aligned} \tag{4.66}$$

Combining (4.64), (4.65) and (4.66) in (4.63) concludes the proof. ■

4.C.3 Proof of Theorem 4.8

Note that $\gamma \in \left(0, \bar{\gamma}\right]$, $\alpha \in \left(0, \bar{\alpha}\right]$ and $l \in \mathbb{N}^*$, $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^{(i)}\}_{i\in[n]})_{\tilde{k}\in\mathbb{N}}$ defined in (4.44), (4.46), (4.47) is an inhomogeneous Markov chain associated with the sequence of Markov kernel $(Q_{\gamma,\alpha,l}^{(k)})_{k\in\mathbb{N}}$ defined by as follows. Define for any $(\theta, \zeta, \eta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, and $x^{(1)} \in \wp_{N_i,b_i}$, $x^{(2)} \in \mathsf{X}_2$ and $x^{(3)} \in \mathsf{X}_3$,

$$\begin{aligned}\mathcal{F}_i\left((\theta,\zeta,\eta);(x^{(1)},x^{(2)},x^{(3)})\right) &= \mathcal{S}_i\left(\mathcal{C}_i\left\{G_i\left(\theta,\zeta;x^{(1)}\right) - \eta; x^{(2)}\right\}; x^{(3)}\right) \\ \mathcal{G}_i\left((\theta,\zeta,\eta);(x^{(1)},x^{(2)},x^{(3)})\right) &= \eta + \alpha\mathcal{F}_i\left((\theta,\zeta,\eta);(x^{(1)},x^{(2)},x^{(3)})\right),\end{aligned}$$

and for $\tilde{\theta} \in \mathbb{R}^d$, $\{\eta^{(i)}\}_{i=1}^n \in \mathbb{R}^{d\times n}$, $\{x^{(1,i)}\}_{i=1}^n \in \otimes_{i=1}^n \wp_{N_i,b_i}$, $\{x^{(2,i)}\}_{i=1}^n \in \mathsf{X}_2^n$, $\{x^{(3,i)}\}_{i=1}^n \in \mathsf{X}_3^n$, setting $x^{(1:n)} = \{(x^{(1,i)}, x^{(2,i)}, x^{(3,i)})\}_{i=1}^n$,

$$\begin{aligned}&\varphi_\gamma\left((\tilde{\theta},\theta,\zeta,\{\eta^{(i)}\}_{i=1}^n);x^{(1:n)}\right) \\ &\qquad = (4\pi\gamma)^{-d/2} \exp\left(-(4\gamma)^{-1}\left\|\theta - \theta + \gamma\sum_{i=1}^n \mathcal{F}_i\left((\theta,\zeta,\eta^{(i)});x^{(i)}\right)\right\|^2\right).\end{aligned}$$

Denote $\tilde{\mathsf{X}}^{(i)} = \wp_{N_i,b_i} \times \mathsf{X}_2 \times \mathsf{X}_3$ and $\tilde{\nu}^{(i)} = \nu_1^{(i)} \times \nu_2 \times \nu_3$. Set $Q^{(0)}_{\gamma,\alpha,l} = \mathrm{Id}$ and for $k \geq 0$, $k = ls+r$, $s \in \mathbb{N}$, $r \in \{0,\dots,l-1\}$, $(\theta,\zeta,\eta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times n}$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times n})$,

$$Q^{(k+1)}_{\gamma,\alpha,l}((\theta,\zeta,\eta),\mathsf{A}) =$$
$$\begin{cases} \int_{\otimes_{i=1}^n \tilde{\mathsf{X}}^{(i)}} \mathbf{1}_{\mathsf{A}}(\tilde{\theta},\tilde{\zeta},\tilde{\eta}) \varphi_\gamma((\tilde{\theta},\theta,\zeta,\{\eta^{(i)}\}_{i=1}^n);x^{(1:n)}) \{\prod_{i=1}^n \delta_{\mathcal{G}_i((\theta,\zeta,\eta);x^{(i)})}(\mathrm{d}\tilde{\eta}^{(i)})\} \delta_\theta(\mathrm{d}\tilde{\zeta})\,\mathrm{d}\tilde{\theta} \otimes_{i=1}^n \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) & \text{if } r = 0 \\ \int_{\otimes_{i=1}^n \tilde{\mathsf{X}}^{(i)}} \mathbf{1}_{\mathsf{A}}(\tilde{\theta},\tilde{\zeta},\tilde{\eta}) \varphi_\gamma((\tilde{\theta},\theta,\zeta,\{\eta^{(i)}\}_{i=1}^n);x^{(1:n)}) \{\prod_{i=1}^n \delta_{\mathcal{G}_i((\theta,\zeta,\eta);x^{(i)})}(\mathrm{d}\tilde{\eta}^{(i)})\} \delta_\zeta(\mathrm{d}\tilde{\zeta})\,\mathrm{d}\tilde{\theta} \otimes_{i=1}^n \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) & \text{otherwise.} \end{cases}.$$

Consider then, the Markov kernel on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$,

$$R^{(k)}_{\gamma,\alpha,l,\eta_0}(\theta_0,\mathsf{A}) = Q^{(k)}_{\gamma,\alpha,l}\left((\theta_0,\theta_0,\eta_0),\mathsf{A}\times\mathbb{R}^d\times\mathbb{R}^{d\times n}\right). \tag{4.67}$$

Define

$$\bar{\gamma}_\alpha = \bar{\gamma}_{\alpha,2} \wedge \bar{\gamma}_4, \qquad \bar{\gamma}_4 = 1/(10\mathtt{m}), \tag{4.68}$$

where $\bar{\gamma}_{\alpha,2}$ is defined in (4.61). The following theorem provides a non-asymptotic convergence bound for the $\mathtt{QLSD}^{++}$ kernel.

Theorem 4.30. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $l \in \mathbb{N}^\star$, $\eta_0 \in \mathbb{R}^{d\times n}$, $\alpha \in \left(0, 1/(1+\max_{i\in[n]}\omega_i)\right]$, $\gamma \in \left(0,\bar{\gamma}_\alpha\right]$, and $k = sl+r \in \mathbb{N}$ with $s \in \mathbb{N}$, $r \in \{0,\dots,l-1\}$, we have*

$$\begin{aligned} &W_2^2\left(\mu R^{(k)}_{\gamma,\alpha,l,\eta_0}, \pi(\cdot|\mathcal{D})\right) \\ &\quad\leq \left(1-\frac{\gamma\mathtt{m}}{2}\right)^k W_2^2(\mu,\pi(\cdot|\mathcal{D})) + \frac{2\gamma D_{\mathbf{b},\mathbf{N}}}{\mathtt{m}}\left(1-\frac{\gamma\mathtt{m}}{2}\right)^s \int_{\mathbb{R}^d} \psi(\theta_0,\eta_0)\mathrm{d}\mu(\theta_0) \\ &\qquad + \frac{4\gamma(1-\alpha)^k}{\mathtt{m}}\left[\sum_{i=1}^n (\omega_i+1-p_i)/p_i\right]\sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2 + \gamma \mathrm{B}_{\oplus,\bar{\gamma}_\alpha}, \end{aligned}$$

where $R^{(k)}_{\gamma,\alpha,l,\eta_0}$ is defined in (4.67), ψ is defined in (4.59), $D_{n,N}$ in (4.62) and

$$\begin{aligned} \mathrm{B}_{\oplus,\bar{\gamma}_\alpha} = \frac{2d\mathrm{L}^2}{\mathtt{m}}(1/\mathtt{m}+5\bar{\gamma}_\alpha)\left[1+\bar{\gamma}_\alpha \mathrm{L}^2/(2\mathtt{m}) + \bar{\gamma}_\alpha^2\mathrm{L}^2/12\right] \\ + \frac{96ld}{\mathtt{m}^2}\left(\sum_{i=1}^n \frac{\mathtt{M}_i}{p_i}(\omega_i+1)(\mathtt{M}_i + \bar{\mathtt{M}}A_{b_i,N_i})\right). \end{aligned} \tag{4.69}$$

Proof Let $k \in \mathbb{N}$. The proof follows from the same lines as Theorem 4.18. By (4.22) and (4.44), we have

$$\begin{aligned} \vartheta_{\gamma(k+1)} - \theta_{k+1} = \vartheta_{\gamma k} - \theta_k - \gamma\left[\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right] \\ - \int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right]\mathrm{d}s + \gamma\left[\tilde{G}(\Theta_k;X_{k+1}) - \nabla U(\theta_k)\right]. \end{aligned}$$

Define the filtration $(\mathcal{H}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ as $\mathcal{H}_0 = \sigma(\vartheta_0,\Theta_0)$ and for $\tilde{k} \in \mathbb{N}^\star$,

$$\mathcal{H}_{\tilde{k}} = \sigma(\vartheta_0,\Theta_0,(X_l^{(1)},\dots,X_l^{(n)})_{1\leq l\leq\tilde{k}},(B_t)_{0\leq t\leq\gamma\tilde{k}}).$$

Note that since $(\vartheta_t)_{t\geq 0}$ is a strong solution of (4.22), then is easy to see that $(\vartheta_{\gamma\tilde{k}}, \Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ is $(\mathcal{H}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$-adapted. Taking the squared norm and the conditional expectation with respect to $\mathcal{H}_k$, we obtain using Assumption 4.14-(i) that

$$\begin{aligned}
\mathbb{E}^{\mathcal{H}_k}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &= \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 - 2\gamma\left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle \\
&\quad + 2\gamma\int_0^\gamma \left\langle \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k), \mathbb{E}^{\mathcal{H}_k}\left[\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}u \\
&\quad - 2\int_0^\gamma \left\langle \vartheta_{\gamma k} - \theta_k, \mathbb{E}^{\mathcal{H}_k}\left[\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}u \\
&\quad + \gamma^2\left\|\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\|^2 \\
&\quad + \mathbb{E}^{\mathcal{H}_k}\left[\left\|\int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right]\mathrm{d}u\right\|^2\right] \\
&\quad + \gamma^2\mathbb{E}^{\mathcal{H}_k}\left[\left\|\tilde{G}(\Theta_k; X_{k+1}) - \nabla U(\theta_k)\right\|^2\right]. \qquad (4.70)
\end{aligned}$$

Using Proposition 4.29, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{G}(\Theta_k; X_{k+1}) - \nabla U(\theta_k)\right\|^2\right] &\leq \left(1 - \frac{\gamma\mathtt{m}}{2}\right)^{\lfloor k/l\rfloor} D_{\mathbf{b},\mathbf{N}}\psi(\theta_0, \eta_0) + 4ldD_{\mathbf{b},\mathbf{N}}/\mathtt{m} \\
&\quad + \left[2\sum_{i=1}^n \frac{\omega_i + 1 - p_i}{p_i}\right](1-\alpha)^k \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2\right]. \qquad (4.71)
\end{aligned}$$

Then, we control the remaining terms in (4.70) using (4.28), (4.29) and (4.30). Combining these bounds and (4.71) into (4.70), for any $\varepsilon > 0$, yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &\leq (1 + 2\gamma\varepsilon - 5\gamma^2\mathtt{m}\mathtt{L})\mathbb{E}\left[\left\|\vartheta_{\gamma k} - \theta_k\right\|^2\right] \\
&\quad - \gamma\left[2 - 5\gamma(\mathtt{m} + \mathtt{L})\right]\mathbb{E}\left[\left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle\right] \\
&\quad + (5\gamma + (2\varepsilon)^{-1})\int_0^\gamma \mathbb{E}\left[\left\|\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right]\mathrm{d}u \\
&\quad + \gamma^2\left(1 - \frac{\gamma\mathtt{m}}{2}\right)^{\lfloor k/l\rfloor} D_{\mathbf{b},\mathbf{N}}\mathbb{E}\left[\psi(\theta_0, \eta_0)\right] + 4ldD_{\mathbf{b},\mathbf{N}}/\mathtt{m} \\
&\quad + 2\gamma^2\left[\sum_{i=1}^n (\omega_i + 1 - p_i)/p_i\right](1-\alpha)^k \sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2.
\end{aligned}$$

Next, we use that under Assumption 4.9, $\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\rangle \geq \mathtt{m}\|\vartheta_{\gamma k} - \theta_k\|^2$ and $|\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle| \leq \mathtt{L}\|\theta_k - \theta^\star\|^2$, which implies taking $\varepsilon = \mathtt{m}/2$ and since $2 - 5\gamma(\mathtt{m} + \mathtt{L}) \geq 0$,

$$\mathbb{E}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] \leq (1 - \gamma\mathtt{m}(1 - 5\gamma\mathtt{m}))\mathbb{E}\left[\left\|\vartheta_{\gamma k} - \theta_k\right\|^2\right]$$

$$
\begin{aligned}
&+ (5\gamma + \mathtt{m}^{-1}) \int_0^\gamma \mathbb{E}\left[\left\|\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}u \\
&+ \gamma^2 \left(1 - \frac{\gamma \mathtt{m}}{2}\right)^{\lfloor k/l \rfloor} D_{\mathbf{b},\mathbf{N}} \mathbb{E}\left[\psi(\theta_0, \eta_0)\right] + 4ldD_{\mathbf{b},\mathbf{N}}/\mathtt{m} \\
&+ 2\gamma^2 \left[\sum_{i=1}^n (\omega_i + 1 - p_i)/p_i\right] (1-\alpha)^k \sum_{i=1}^n \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2 .
\end{aligned} \tag{4.72}
$$

Further, for any $u \in \mathbb{R}_+$, using Durmus and Moulines (2019, Lemma 21) we have

$$
\mathtt{L}^{-2}\, \mathbb{E}\left[\left\|\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \leq du\left(2 + u^2\mathtt{L}^2/3\right) + 3u^2\mathtt{L}^2/2\mathbb{E}\left[\left\|\vartheta_{\gamma k} - \theta^\star\right\|^2\right].
$$

Integrating the previous inequality on $[0, \gamma]$, we obtain

$$
\mathtt{L}^{-2} \int_0^\gamma \mathbb{E}\left[\left\|\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}u \leq d\gamma^2 + d\gamma^4\mathtt{L}^2/12 + \gamma^3\mathtt{L}^2/2\mathbb{E}\left[\left\|\vartheta_{\gamma k} - \theta^\star\right\|^2\right].
$$

Plugging this bounds in (4.72) and using Durmus and Moulines (2019, Proposition 1) complete the proof. ■

4.D Consistency analysis in the big data regime

In this section, we assume that the number of observations on each client $i \in [n]$ writes $N_i = \lfloor c_i N \rfloor$ where $\{c_i > 0\}_{i\in[n]}$, $N \in \mathbb{N}^*$, and provide upper bounds on the asymptotic bias associated to each algorithm when N tends towards infinity. For simplicity, we assume for any $i \in [n]$, that $b_i = \lfloor c_i b \rfloor$ with $b \in [N]$, $\mathtt{M}_i = \mathtt{M}$ with $\mathtt{M} > 0$, $p_i = 1$ and $\omega_i = \omega$ with $\omega > 0$ but note that our conclusions also hold for the general setting considered in this chapter.

4.D.1 Asymptotic analysis for Algorithm 4.5

The following corollary is associated with QLSD defined in Algorithm 4.5 in the main chapter.

Corollary 4.31. *Assume Assumption 4.9, Assumption 4.10 Assumption 4.11 and Assumption 4.12. In addition, assume that* $\liminf_{N\to\infty} \mathtt{m}/N > 0$ *and* $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ *for* $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}, \mathtt{B}^\star, \sigma_\star\}$. *Then, we have* $\bar\gamma = \bar\eta/N$ *where* $\bar\eta > 0$ *and* $\bar\gamma$ *is defined in* (4.A.2). *In addition,*

$$
B_{\bar\gamma} = (\omega + 1)\, \mathrm{O}(N),
$$

where $B_{\bar\gamma}$ *is defined in* (4.33).

Proof Since we assume that $\liminf_{N\to\infty} \mathtt{m}/N > 0$ and $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ for $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}, \mathtt{B}^\star, \sigma_\star\}$, there exist $C_\mathtt{m}$, $C_\mathtt{L}$, $C_\mathtt{M}$, $C_{\mathtt{B}^\star}$ and $C_{\sigma_\star} > 0$ such that $\mathtt{m} \geq C_\mathtt{m} N$, $\mathtt{L} \leq C_\mathtt{L} N$, $\mathtt{M} \leq C_\mathtt{M} N$, $\mathtt{B}^\star \leq C_{\mathtt{B}^\star} N$ and $\sigma_\star \leq C_{\sigma_\star} N$. Under these assumptions, it is straightforward

from (4.A.2) to see that there exists $\bar{\eta} > 0$ such that $\bar{\gamma} = \bar{\eta}/N$. In addition, it follows from (4.33) that

$$B_{\bar{\gamma}} \leq \frac{2dC_{\mathtt{L}}^2}{C_{\mathtt{m}}}\left(\frac{1}{C_{\mathtt{m}}} + 5\bar{\eta}\right)\left[1 + \frac{\bar{\eta}C_{\mathtt{L}}^2}{2C_{\mathtt{m}}} + \frac{\bar{\eta}^2 C_{\mathtt{L}}^2}{12}\right] + \frac{4}{C_{\mathtt{m}}}(\omega C_{\mathtt{B}^\star} + C_{\sigma_\star}^2 N) + \frac{8(\omega+1)C_{\mathtt{L}}C_{\mathtt{M}}}{C_{\mathtt{m}}^2}\left[d + \bar{\eta}\left(\omega C_{\mathtt{B}^\star} + C_{\sigma_\star}^2 N\right)\right].$$

The proof is concluded by letting N tend towards infinity. ■

Regarding the specific instance $\mathtt{QLSD}^{\#}$ of Algorithm 4.5 in the main chapter, a similar result holds. Indeed, by using Lemma 4.21, we can notice that Assumption 4.11-(iii) is verified with $\sigma_\star = C_{\sigma_\star} N$ for some $C_{\sigma_\star} > 0$ and we can apply Corollary 4.31.

4.D.2 Asymptotic analysis for Algorithm 4.6

The following corollary is associated with $\mathtt{QLSD}^\star$ defined in Algorithm 4.6 in the main chapter.

Corollary 4.32. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. In addition, assume that* $\liminf_{N\to\infty} \mathtt{m}/N > 0$ *and* $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ *for* $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$. *Then, we have* $\bar{\gamma} = \bar{\eta}/N$ *where* $\bar{\eta} > 0$ *and* $\bar{\gamma}$ *is defined in* (4.A.2). *In addition,*

$$B_{\circledast,\bar{\gamma}} = d(\omega+1)\,\mathrm{O}(1),$$

where $B_{\circledast,\bar{\gamma}}$ *is defined in* (4.37).

Proof Since we assume that $\liminf_{N\to\infty} \mathtt{m}/N > 0$ and $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ for $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$, there exist $C_{\mathtt{m}}$, $C_{\mathtt{L}}$ and $C_{\mathtt{M}} > 0$ such that $\mathtt{m} \geq C_{\mathtt{m}} N$, $\mathtt{L} \leq C_{\mathtt{L}} N$ and $\mathtt{M} \leq C_{\mathtt{M}} N$. Under these assumptions, it is straightforward from (4.A.2) to see that there exists $\bar{\eta} > 0$ such that $\bar{\gamma}_\alpha = \bar{\eta}/N$. In addition, it follows from (4.27) that

$$B_{\circledast,\bar{\gamma}} \leq \frac{2dC_{\mathtt{L}}^2}{C_{\mathtt{m}}}\left(\frac{1}{C_{\mathtt{m}}} + 5\bar{\eta}\right)\left[1 + \frac{\bar{\eta}C_{\mathtt{L}}^2}{2C_{\mathtt{m}}} + \frac{\bar{\eta}^2 C_{\mathtt{L}}^2}{12}\right] + \frac{4d\bar{\mathtt{M}}C_{\mathtt{L}}}{C_{\mathtt{m}}^2}\max_{i\in[n]}\left\{c_i\omega + (\omega+1)\cdot\frac{N-b}{b(\lfloor c_i N\rfloor - 1)}\right\}.$$

The proof is concluded by letting N tend towards infinity. ■

Lastly, we have the following asymptotic convergence result regarding $\mathtt{QLSD}^{++}$ defined in Algorithm 4.6 in the main chapter.

Corollary 4.33. *Assume Assumption 4.9, Assumption 4.10, Assumption 4.12 and Assumption 4.19. In addition, assume that* $\liminf_{N\to\infty} \mathtt{m}/N > 0$ *and* $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ *for* $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$. *Then, we have* $\bar{\gamma}_\alpha = \bar{\eta}/N$ *where* $\bar{\eta} > 0$ *and* $\bar{\gamma}_\alpha$ *is defined in* (4.68). *In addition,*

$$B_{\oplus,\bar{\gamma}_\alpha} = d(\omega+1)\,\mathrm{O}(1),$$

where $B_{\oplus,\bar{\gamma}_\alpha}$ *is defined in* (4.69).

Proof Since we assume that $\liminf_{N\to\infty} \mathtt{m}/N > 0$ and $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ for $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$, there exist $C_\mathtt{m}$, $C_\mathtt{L}$ and $C_\mathtt{M} > 0$ such that $\mathtt{m} \geq C_\mathtt{m} N$, $\mathtt{L} \leq C_\mathtt{L} N$ and $\mathtt{M} \leq C_\mathtt{M} N$. Under these assumptions, it is straightforward from (4.68) to see that there exists $\bar{\eta} > 0$ such that $\bar{\gamma}_\alpha = \bar{\eta}/N$. In addition, it follows from (4.69) that

$$\mathrm{B}_{\oplus,\bar{\gamma}_\alpha} \leq \frac{2dC_\mathtt{L}^2}{C_\mathtt{m}} \left(\frac{1}{C_\mathtt{m}} + 5\bar{\eta} \right) \left[1 + \frac{\bar{\eta} C_\mathtt{L}^2}{2C_\mathtt{m}} + \frac{\bar{\eta}^2 C_\mathtt{L}^2}{12} \right] + \frac{96(\omega+1) l d n C_\mathtt{M}}{C_\mathtt{m}^2} \left(\frac{(N-b)\bar{\mathtt{M}}}{b(\min_{i\in[n]}\{\lfloor c_i N \rfloor\} - 1)} + C_\mathtt{M} \right).$$

The proof is concluded by letting N tend towards infinity. ■

4.E Experimental details

In this section, we provide additional details regarding our numerical experiments. The code, data and instructions to reproduce our experimental results can be downloaded [here].

4.E.1 Toy Gaussian example

Pseudocode of LSD⋆. For completeness, we provide in Algorithm 4.8 the pseudocode of the non-compressed counterpart of QLSD⋆, namely LSD⋆.

Algorithm 4.8 Variance-reduced Langevin Stochastic Dynamics (LSD⋆)

Input: minibatch sizes $\{b_i\}_{i\in[n]}$, number of iterations K, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and initial point θ_0.
for $k = 0$ **to** $K-1$ **do**
  for $i \in \mathcal{A}_{k+1}$ // On active clients **do**
    Draw $\mathcal{S}_{k+1}^{(i)} \sim \mathrm{Uniform}(\wp_{N_i,b_i})$.
    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/b_i) \sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\theta^\star)]$.
    Compute $g_{i,k+1} = H_{k+1}^{(i)}(\theta_k)$.
    Send $g_{i,k+1}$ to the central server.
  // On the central server
  Compute $g_{k+1} = \frac{n}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the n clients.
Output: samples $\{\theta_k\}_{k=0}^K$.

Additional experimental details. As highlighted in Section 4.4 (*Toy Gaussian example* paragraph) in the main chapter, the synthetic dataset has been generated so that each client owns a heterogeneous and unbalanced dataset. An illustration of the

unbalancedness is given in Figure 4.3. The precise procedure to generate such a dataset can be found in the aforementioned notebook.

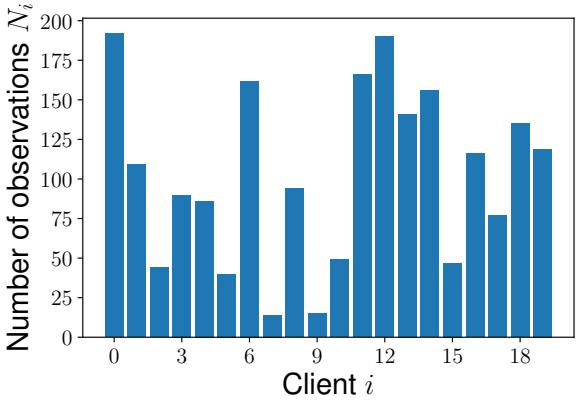


Figure 4.3 – Illustration of the unbalancedness of the synthetic dataset used in the Toy Gaussian experiment.

To obtain the figure at the bottom row of Figure 1 in the main chapter, we launched all the MCMC algorithms with $K = 500,000$ outer iterations and considered a burn-in period of $450,000$ iterations. Hence, only the last $50,000$ samples have been used to compute the MSE associated to the test function $f : \theta \mapsto \|\theta\|$. In order to compute the expected number of bits transmitted during each upload period, we considered the `Elias` encoding scheme and used the upper-bounds given in Alistarh et al. (2017, Theorem 3.2 and Lemma A.2).

- **License of the assets:** No existing asset has been used for this experiment.
- **Total amount of compute and type of resources used:** This experiment has been run on a laptop running Windows 10 and equipped with Intel(R) Core(TM) i7_8565U CPU 1.80GHz with 16Go of RAM. The total amount of compute is roughly 33 hours.
- **Training details:** All training details (here hyperparameters) are detailed in Section 4.4 in the main chapter.

discretization step-size and compression tradeoff. We complement the analysis made in the main chapter by showing on Figure 4.4 that the saving in terms of number of transmitted bits can be further improved by decreasing the value of γ. This numerical finding illustrates our theory which in particular shows that the asymptotic bias associated to $\texttt{QLSD}^\star$ is of the order $\omega\, \mathrm{O}(\gamma)$, see Table 4.1 in the main chapter.

4.E.2 Bayesian logistic regression

Pseudo-code of $\texttt{LSD}^{++}$. For completeness, we provide in Algorithm 4.9 the pseudo-code of the non-compressed counterpart of $\texttt{QLSD}^{++}$, namely $\texttt{LSD}^{++}$.

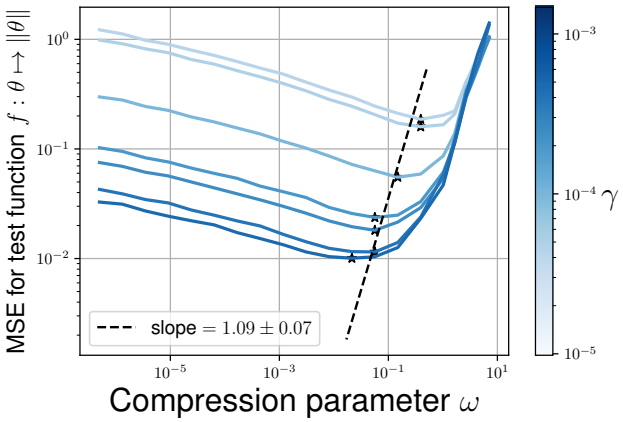


Figure 4.4 – Toy Gaussian example. tradeoff between step-size and compression parameter values.

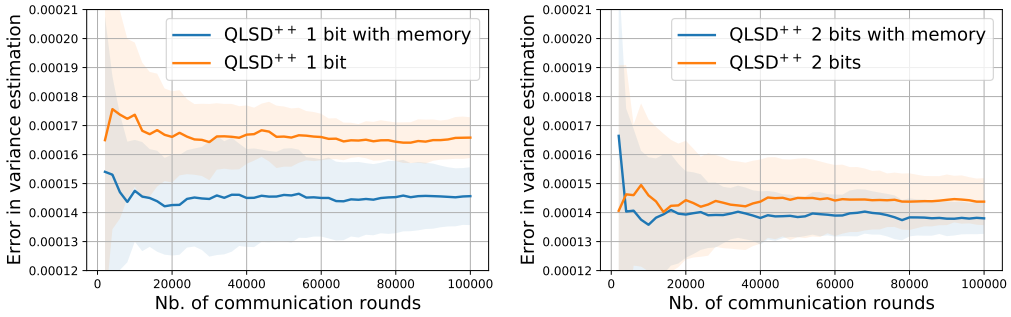


Figure 4.5 – Bayesian logistic regression on synthetic data.

Additional experimental details. For the Bayesian logistic regression experiment detailed in the main chapter, we ran the MCMC algorithms with $K = 500,000$ outer iterations and considered a burn-in period of length $50,000$.

Benefits of the memory mechanism. We also run an additional experiment on a low-dimensional synthetic dataset to highlight the benefits brought by the memory mechanism involved in QLSD^{++} when the dataset is highly heterogeneous. To this end, we consider the SYNTHETIC(α, β) dataset (Li et al., 2020b) with $\alpha = \beta = 1$, $d = 2$ and $n = 50$. We run QLSD^{++} with and without memory terms using $l = 100$, $\alpha = 1/(\omega + 1)$, $\gamma = 10^{-5}$ and for huge compression parameters, namely $s \in \{2^1, 2^2\}$. We use $K = 100,000$ outer iterations without considering a burn-in period. In order to have access to some ground truth, we also implement the Metropolis-adjusted Langevin algorithm (MALA) (Robert and Casella, 2004).

Figure 4.5 shows the Euclidean norm of the error between the true variance under π estimated with MALA and the empirical variance computed using samples generated by QLSD^{++}. As expected, we can notice that the memory mechanism reduces the impact of the compression on the asymptotic bias of QLSD^{++} when ω is large.

Algorithm 4.9 Variance-reduced Langevin Stochastic Dynamics (LSD^{++})

Input: minibatch sizes $\{b_i\}_{i\in[n]}$, number of iterations K, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, initial point θ_0 and $\alpha \in (0, \bar{\alpha}]$ with $\bar{\alpha} > 0$.

// Memory mechanism initialization

Initialize $\{\eta_0^{(1)}, \ldots, \eta_0^{(n)}\}$ and $\eta_0 = \sum_{i=1}^n \eta_0^{(i)}$.

for $k = 0$ **to** $K - 1$ **do**

  // Update of the control variates

  if $k \equiv 0 \pmod{l}$ **then**

    Set $\zeta_k = \theta_k$.

  else

    Set $\zeta_k = \zeta_{k-1}$

  for $i \in \mathcal{A}_{k+1}$ // On active clients **do**

    Draw $\mathcal{S}_{k+1}^{(i)} \sim \text{Uniform}\left(\wp_{N_i,b_i}\right)$.

    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/b_i) \sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\zeta_k)] + \nabla U_i(\zeta_k)$.

    Compute $g_{i,k+1} = H_{k+1}^{(i)}(\theta_k) - \eta_k^{(i)}$.

    Send $g_{i,k+1}$ to the central server.

    Set $\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha g_{i,k+1}$.

  // On the central server

  Compute $g_{k+1} = \eta_k + \frac{n}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.

  Set $\eta_{k+1} = \eta_k + \alpha \sum_{i\in\mathcal{A}_{k+1}}^n g_{i,k+1}$.

  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.

  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.

  Send θ_{k+1} to the n clients.

Output: samples $\{\theta_k\}_{k=0}^K$.

Table 4.5 – Bayesian Logistic Regression on *covtype* dataset.

Algorithm	99% HPD error
DG-SGLD	1.8e-2
QLSD^{++} 4 bits	2.2e-3
QLSD^{++} 8 bits	2.0e-2
QLSD^{++} 16 bits	1.9e-2

Results on a non-image dataset. In order to complement our results on an image dataset (FEMNIST), we also implement our methodology and one competitor (DG-SGLD) on the *covtype*[1] dataset. Again, the ground truth has been obtained by implementing a long-run Metropolis-adjusted Langevin algorithm. The results we obtained are gathered in Table 4.5.

- **License of the assets:** We use the Synthetic dataset whose associated code is under the MIT license, and the FEMNIST dataset whose data are publicy available and associated code is under MIT license.

[1] https://archive.ics.uci.edu/ml/datasets/covertype

- **Total amount of compute and type of resources used:** This experiment has been run on a laptop running Windows 10 and equipped with Intel(R) Core(TM) i7_8565U CPU 1.80GHz with 16Go of RAM. The total amount of compute is roughly 30 hours.
- **Training details:** Hyperparameter values are detailed in Section 4.4 in the main chapter. Regarding our experiment on real data, we use a random subset of the initial training data (for computational reasons).

4.E.3 Bayesian neural networks

- **License of the assets:** We use the MNIST, FMNIST, CIFAR10 and SVHN datasets which are publicly downloadable with the torchvision.datasets package.
- **Total amount of compute and type of resources used:** The total computational cost depends on the dataset, but is roughly 40 hours in the worst case.
- **Training details:** We consider the same hyperparameter values detailed in Table 4.6 for both training on MNIST and CIFAR10 except for the initialization and the sampling period. For the MNIST dataset, we use the default random weights given by pytorch whereas for CIFAR-10 we use the warm-start provided by the pytorchcv library and consider a burn-in period of half the sampling period ($K = 10^4$ iterations) with a thinning of 10.

In the following, we denote $\mathcal{D}_{\text{test}}$ the test dataset and for any data $(x, y) \in \mathcal{D}_{\text{test}}$, we define the preditive density by

$$p(y \mid x) = \int p(y \mid x, \theta)\ \pi(\theta \mid \mathcal{D})\ \mathrm{d}\theta, \tag{4.73}$$

where $p(y \mid x, \theta)$ is the conditional likelihood. For any input x, the predicted label is denoted by $y_{\text{pred}}(x) = \arg\max_y p(y \mid x)$.

Metrics used for the Bayesian neural network experiment in the main chapter. In the main chapter, we consider three metrics to compare the different Bayesian FL algorithms, namely *Accuracy*, *Agreement* and *TV*. They are defined in the following.

- **Accuracy:** Based on samples from the approximate posterior distribution, we compute the minimum mean-square estimator (*i.e.* corresponding to the posterior mean) and use it to make predictions on the test dataset. The *Accuracy* metric corresponds to the percentage of well-predicted labels.
- **Agreement:** Let denote p_{ref} and p the predictive densities associated to HMC and an approximate simulation-based algorithm, respectively. Similar to Izmailov et al. (2021), we define the agreement between p_{ref} and p as the fraction of the test datapoints for which the top-1 predictions of p_{ref} and p, *i.e.*

$$\text{agreement}(p_{\text{ref}}, p) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \mathbf{1}\left[\arg\max_{y'} p_{\text{ref}}(y' \mid x) = \arg\max_{y'} p(y' \mid x)\right].$$

- **Total variation (TV):** By denoting $\mathcal{Y}$ the set of possible labels, we consider the total variation metric between p_{ref} and p, *i.e.*

$$\text{TV}(p_{\text{ref}}, p) = \frac{1}{2|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \sum_{y' \in \mathcal{Y}} \left| p_{\text{ref}}(y' \mid x) - p(y' \mid x) \right|.$$

Performance results on a highly heterogeneous dataset. We train LeNet5 (LeCun et al., 1998) architecture on the MNIST dataset (Deng, 2012), and we consider the FMNIST (Xiao et al., 2017) as the out-of-distribution dataset. To obtain a highly heterogeneous setting, we split the data among $n = 20$ clients so that each client has a dominant label representing 40% of the total amount in the training set and 1% of the other labels as described in Figure 4.6.

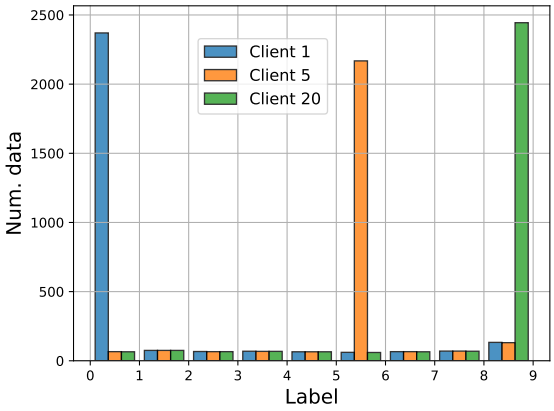


Figure 4.6 – Number of labels owned by different clients.

Inspired by the scores defined in Guo et al. (2017), we measure the performance of the different algorithms and report those results in Table 4.6. These statistics aim to better understand the predictions in order to calibrate the models (Rahaman and Thiery, 2021).

Method	SGLD	pSGLD	QLSD	QLSD PP	QLSD^{++}	QLSD^{++} PP	FedBe-Gauss.	FedBe-Dirich.	FSGLD
Accuracy	99.1	99.2	98.8	98.3	98.8	98.7	43.5	79.3	98.5
$10^2\times$ ECE	0.577	1.25	0.916	1.57	0.692	0.930	7.51	21.3	2.65
$10^2\times$ BS	1.38	1.39	1.98	2.23	1.91	2.18	66.6	36.1	2.64
$10^2\times$ nNLL	2.86	3.16	4.15	4.82	4.11	4.65	139	78.0	6.19
Weight Decay	5	5	5	5	5	5	0	0	5
Batch Size	64	64	64	64	64	64	64	64	64
Learning rate	1e-07	1e-08	1e-07	1e-07	1e-07	1e-07	1e-02	1e-02	1e-07
Local steps	N/A	N/A	1	1	1	1	250	250	16
Burn-in	100epch.	100epch.	1e04	1e04	1e04	1e04	N/A	N/A	1e04
Thinning	1	1	500	500	500	500	N/A	N/A	500
Training	1e03epch.	1e03epch.	1e05it.	1e05it.	1e05it.	1e05it.	N/A	N/A	1e05it.

Table 4.6 – Performance of Bayesian FL algorithms trained on the highly-heterogeneous dataset.

Expected Calibration Error (ECE). To measure the difference between the accuracy and confidence of the predictions, we group the data into $M \geq 1$ buckets defined for any $m \in [M]$ by $\text{B}_\text{m} = \{(x, y) \in \mathcal{D}_{\text{test}} : p(y_{\text{pred}}(x)|x) \in \left](m-1)/M, m/M\right]\}$. As in

the previous work of Ovadia et al. (2019), we denote the model accuracy on $\mathrm{B_m}$ by

$$\mathrm{acc}\left(\mathrm{B_m}\right) = \frac{1}{\left|\mathrm{B_m}\right|} \sum_{(x,y)\in \mathrm{B_m}} \mathbf{1}_{y_{\mathrm{pred}}(x)=y}$$

and define the confidence on $\mathrm{B_m}$ by

$$\mathrm{conf}\left(\mathrm{B_m}\right) = \frac{1}{\left|\mathrm{B_m}\right|} \sum_{(x,y)\in \mathrm{B_m}} p(y_{\mathrm{pred}}(x)|x).$$

As stressed in Guo et al. (2017), for any $m \in [M]$ the accuracy $\mathrm{acc}(\mathrm{B_m})$ is an unbiased and consistent estimator of $\mathbb{P}\left(y_{\mathrm{pred}}(x) = y \mid (m-1)/M < p(y_{\mathrm{pred}}(x)|x) \leq m/M\right)$. Therefore, the ECE defined by

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{\left|\mathrm{B_m}\right|}{\left|\mathcal{D}_{\mathrm{test}}\right|} \left|\mathrm{acc}\left(\mathrm{B_m}\right) - \mathrm{conf}\left(\mathrm{B_m}\right)\right|$$

is an estimator of

$$\mathbb{E}_{(x,y)}\left[\left|\,\mathbb{P}\left(y_{\mathrm{pred}}(x) = y \mid p(y_{\mathrm{pred}}(x)|x)\right) - p(y_{\mathrm{pred}}(x)|x)\right|\right].$$

Thus, ECE measures the absolute difference between the confidence level of a prediction and its accuracy.

Brier Score (BS). The BS is a proper scoring rule (see for example Dawid and Musio (2014)) that can only evaluate random variables taking a finite number of values. Denote by $\mathcal{Y}$ the finite set of possible labels, the BS measures the model's confidence in its predictions and is defined by

$$\mathrm{BS} = \frac{1}{\left|\mathcal{D}_{\mathrm{test}}\right|} \sum_{(x,y)\in\mathcal{D}_{\mathrm{test}}} \sum_{c\in\mathcal{Y}} (p(y=c|x) - \mathbf{1}_{y=c})^2.$$

Normalized negative log-likelihood (nNLL). This classical score defined by

$$\mathrm{nNLL} = -\frac{1}{\left|\mathcal{D}_{\mathrm{test}}\right|} \sum_{(x,y)\in\mathcal{D}_{\mathrm{test}}} \log p(y|x)$$

measures the model ability to predict good labels with high probability.

Out of distribution detection. Here we study the behavior of our proposed algorithms in the out-of-distribution (OOD) framework, we consider the pairs MNIST/FMNIST and CIFAR10/SVHN, comparing the densities of the predictive entropies on the ID vs OOD data. These densities denoted by p_{in} and p_{out} respectively, are approximated using a kernel estimator based on of the histogram associated with $\{\mathrm{Ent}(x) : x \in \mathcal{D}_{\mathrm{test}}^x\}$ for $\mathcal{D}_{\mathrm{test}} \in \{\mathrm{MNIST}, \mathrm{FMNIST}\}$ or $\{\mathrm{CIFAR10}, \mathrm{SVHN}\}$, where $\mathrm{Ent}(x)$ is the predictive entropy defined by:

$$\mathrm{Ent}(x) = \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x),$$

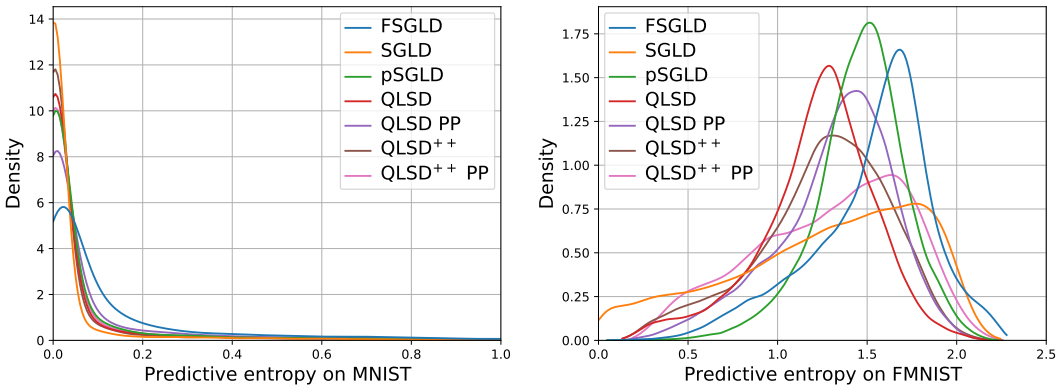


Figure 4.7 – Predictive entropies comparison between MNIST and FMNIST.

and $p(y|x)$ is defined by (4.73) and estimated by the different methods that we consider. The resulting densities from the different methods that we consider are displayed in Figure 4.7.

A new data point x is then labeled in the original dataset (MNIST or CIFAR10) if $p_{\text{in}}(\text{Ent}(x)) > p_{\text{out}}(\text{Ent}(x))$ and out-of-distribution otherwise.

Calibration results. Interpreting the predicted outputs as probabilities is only correct for well a calibrated model. Indeed, when a model is calibrated, the confidence is closed to the accuracy of the predictions. In order to evaluate the calibration of the models, we display the reliability diagram on the left-hand side of Figure 4.8. It represents the evolution of $\text{acc}(\text{B}_\text{m}) - \text{conf}(\text{B}_\text{m})$ in function of $\text{conf}(\text{B}_\text{m})$, closer the values are to zero better the model is calibrated.

For the second sub-experiment, we consider for any $\tau \in [0,1]$, the set $\mathcal{D}^{(\tau)}_{\text{pred}} = \{x \in \mathcal{D}^{x}_{\text{test}} : p(y|x) \geq \tau\}$ of classified data with credibility greater than τ. We define the test accuracy on $\mathcal{D}^{(\tau)}_{\text{pred}}$ by

$$\frac{\text{Card}\Big(\{x \in \mathcal{D}^{(\tau)}_{\text{pred}} : y_{\text{true}}(x) = y_{\text{pred}}(x)\}\Big)}{\text{Card}(\mathcal{D}^{(\tau)}_{\text{pred}})}.$$

The right-hand side of Figure 4.8 shows the evolution of the test accuracy on $\mathcal{D}^{(\tau)}_{\text{pred}}$ with respect to the credibility threshold τ. It can be noted that in both plots of Figure 4.8, the accuracy tends to 100% for confident predictions.

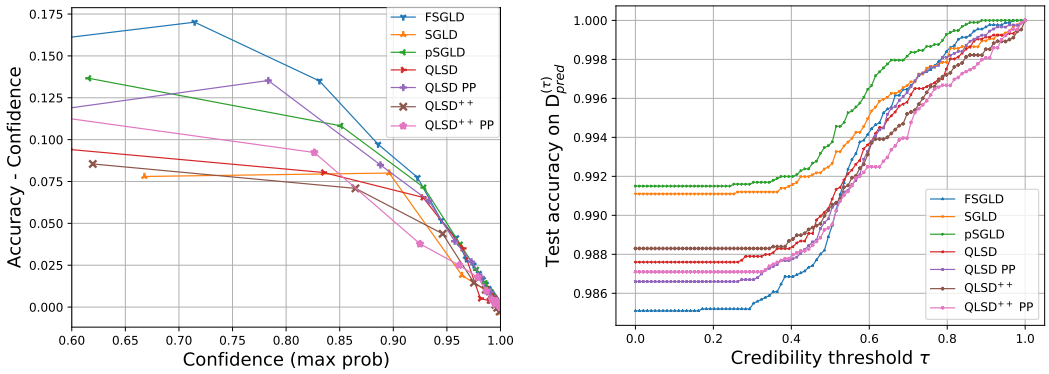


Figure 4.8 – Left: Calibration test from reliability diagrams – Right: Test accuracy on $\mathcal{D}_{\text{pred}}^{(\tau)}$ with respect to the threshold τ.

Chapter 5

Federated Conformal Prediction under Label Shift

Contents

Federated Learning (FL) is a machine learning framework where many clients collaboratively train models while keeping the training data decentralized. Despite recent advances in FL, the uncertainty quantification topic (UQ) remains partially addressed. Among UQ methods, conformal prediction (CP) approaches provides distribution-free guarantees under minimal assumptions. We develop a new federated conformal prediction method based on quantile regression and take into account privacy constraints. This method takes advantage of importance weighting to effectively address the label shift between agents and provides theoretical guarantees for both valid coverage of the prediction sets and differential privacy. Extensive experimental studies demonstrate that this method outperforms current competitors.

5.1 Introduction

Federated learning is an increasingly important framework for large-scale learning. FL allows many agents to train a model together under the coordination of a central server without ever transmitting the agents' data over the network, in an attempt to preserve privacy. There has been a considerable amount of FL work over the past 5 years, see e.g. Bonawitz et al. (2019); Yang et al. (2019); Kairouz et al. (2021); Li et al. (2020a). Compared to classical machine learning techniques, FL has two unique features. First, the networked agents are massively distributed, communication bandwidth is limited, and agents are not always available (*system heterogeneity*). Second, the data distribution at different agents can vary greatly (*statistical heterogeneity*); see Huang et al. (2022); Yoon et al. (2022). These features lead to serious challenges for both training and inference in federated systems. The focus of this work is on federated

inference procedures that allow to build prediction sets for each agent with a confidence level that can be guaranteed.

Conformal Prediction, originally introduced in Vovk et al. (1999); Shafer and Vovk (2008); Balasubramanian et al. (2014), has recently gained popularity. It generates prediction sets with guaranteed error rates. Conformal algorithms are shown to be always valid: the actual confidence level is the nominal one, without requiring any specific assumption about the distribution of the data beyond exchangeability; see Lei et al. (2013); Fontana et al. (2023) and references therein. With few exceptions, CP methods were developed for centralized environments.

We consider below a supervised learning problem with features x taking values in $\mathcal{X}$ and labels y taking values in $\mathcal{Y}$. Let $(X_k, Y_k)_{k=1}^{N_{\text{train}}+N}$ be an independent and identically distributed (i.i.d.) dataset. We divide the data into a *training* and a *calibration* dataset. Formally, let $\{\mathcal{K}_{\text{train}}, \mathcal{K}_{\text{cal}}\}$ be a partition of $\{1, \ldots, N_{\text{train}} + N\}$, and let $N = |\mathcal{K}_{\text{cal}}|$. Without loss of generality, we take $\mathcal{K}_{\text{cal}} = \{1, \ldots, N\}$. We learn a predictor $\hat{f}\colon \mathcal{X} \to \Delta_{|\mathcal{Y}|}$ on the training set $\mathcal{K}_{\text{train}}$, where $|\mathcal{Y}|$ is the number of classes and $\Delta_{|\mathcal{Y}|}$ is the $|\mathcal{Y}|$-dimensional probability simplex. For any covariate $x \in \mathcal{X}$ associated with a label $y \in \mathcal{Y}$, consider a classification score function $\mathcal{S}\colon \mathcal{Y} \times \Delta_{|\mathcal{Y}|} \to [0, 1]$, independent of other covariates and labels, which yields a non-conformity score given by $V(x, y) = \mathcal{S}(y, \hat{f}(x))$. This non-conformity measure estimates how unusual an example looks. Based on these non-conformity scores, standard CP procedure constructs, for each *significance level* $\alpha \in [0, 1]$, a (measurable) *set-valued* predictor $\mathcal{C}_\alpha(x)$ using $\{(X_k, Y_k)\}_{k=1}^{N}$ that satisfies the following conditions

$$\mathbb{P}\left(Y_{N+1} \in \mathcal{C}_\alpha(X_{N+1})\right) \geq 1 - \alpha, \tag{5.1}$$

where (X_{N+1}, Y_{N+1}) is a *test point* that is independent of $\mathcal{K}_{\text{train}}$ and $\mathcal{K}_{\text{cal}}$. The quantity $1 - \alpha$ is called the *confidence level*. The guarantee (5.1) is set up in a centralized environment – all data are available at a central node and usually assuming that the distributions of calibration and test data satisfy $P^{\text{cal}} = P^\star$. If there is a mismatch between the distributions P^{cal} and $P^\star$, then corrections should be made to ensure an appropriate confidence level; see Tibshirani et al. (2019); Podkopaev and Ramdas (2021a); Barber et al. (2022) and references therein.

Setup. In this work, we consider a federated learning system with n agents. We assume that, instead of storing the entire dataset on a centralized node, each agent $i \in [n]$ owns a local calibration set $\mathcal{D}_i = \{(X_k^i, Y_k^i)\}_{k=1}^{N^i}$, where N^i is the number of calibration samples for the agent i. We further assume that the calibration data are i.i.d. and that the statistical heterogeneity is due to *label shifts*:

$$(X_k^i, Y_k^i) \sim P^i = P_{X|Y} \times P_Y^i,$$

where $P_{X|Y}$, the conditional distribution of the feature given the label, is assumed identical among agents but P_Y^i, the prior label distribution, may differ across agents. In federated learning, statistical heterogeneity is the rule rather than the exception, and it is essential to take into account the presence of label shift at the agent level. We assume that a predictive model $\hat{f}$ has been learned by federated learning. The results we present are agnostic to the learning procedure.

For an agent $\star \in [n]$, and each $\alpha \in (0, 1)$, we are willing to compute a set-valued predictor, $\mathcal{C}_\alpha$ with confidence level $1 - \alpha$, which depends on the calibration data of *all the agents*. The goal is to construct informative conformal prediction sets for each agent,

even when its calibration set is limited in size, by using the calibration data of all the agents participating in the FL; we stress that the calibration data must always remain local to the networked agents. Most importantly, the resulting algorithm should attain both conformal and theoretical privacy guarantees – matched to the privacy guarantees that can be obtained in the FL training procedure.

Our main **contributions** to solving this challenging problem can be summarized as follows.

- We introduce a new method, `DP-FedCP`, to construct conformal prediction sets in a federated learning context that addresses label shift between agents; see Section 5.2. `DP-FedCP` is a federated learning algorithm based on federated computation of weighted quantiles of agent's non-conformity scores, where the weights reflect the label shift of each client with respect to the population. The quantiles are obtained by regularizing the pinball loss using Moreau-Yosida inf-convolution and a version of federated averaging procedure; see Section 5.3.
- We establish conformal prediction guarantees, ensuring the validity of the resulting prediction sets. Additionally, we provide differential private guarantees for `DP-FedCP`; see Section 5.4.
- We show that `DP-FedCP` provides valid confidence sets and outperforms standard approaches in a series of experiments on simulated data and image classification datasets; see Section 5.5.

Related Works. The construction of predictions sets with confidence guarantees has been the subject of much work, mostly in a centralized framework. The conformal framework, introduced in the pioneering works of Vovk et al. (1999) is appealing in its simplicity/flexibility; see e.g. (Angelopoulos et al., 2021; Fontana et al., 2023) and the references therein. For exchangeable data, this framework provides a model-free methodology for constructing prediction sets that satisfy the desired coverage (Shafer and Vovk, 2008; Papadopoulos et al., 2002; Fannjiang et al., 2022; Angelopoulos et al., 2022b).

These results can also be extended to non-exchangeable data. A method has been developed for dealing with covariate shift (Tibshirani et al., 2019). This method is based on evaluating the discrepancy between the distribution of the calibration data set P^{cal} and the test point distributed according to $P^\star$. Using an estimate of the Radon-Nikodym derivative $\mathrm{d}P^\star/\mathrm{d}P^{\text{cal}}$, a valid prediction set can be obtained by weighting the non-conformity scores. The seminal work of (Tibshirani et al., 2019) led to several improvements, either to form valid prediction sets as long as the f-divergence of the discrepancy remains small (Cauchois et al., 2020), or to formulate hypothesis tests under covariate shifts (Hu and Lei, 2020). In addition, Gibbs and Candes (2021) examine the shift in an online environment; and Lei and Candès (2021) show the validity of the prediction sets even when the distributional shift is only approximated. Since many real-world data sets do not satisfy exchangeability, valid prediction sets are developed in (Barber et al., 2022) that put more mass around the point of interest.

Conformal methods adapted to label shift are considered in (Podkopaev and Ramdas, 2021a,b) and have similar guarantees to those in (Tibshirani et al., 2019, Corollary 1). Methods for detecting and quantifying label shift have been proposed in (Lipton et al., 2018; Garg et al., 2020).

Differentially private quantiles can be derived based either on the exponential or Gaussian mechanisms (Gillenwater et al., 2021; Pillutla et al., 2022). Using the exponential mechanism, valid prediction sets are generated in (Angelopoulos et al., 2022a). However, quantile computation in a federated learning environment remains a challenge. A first federated approach based on quantile averaging was proposed in (Lu and Kalpathy-Cramer, 2021). However, this work does not provide theoretical guarantees, and the proposed method is vulnerable to distribution shifts. For federated deep learning, the differentially private versions are based on various techniques combination like gradient clipping and the addition of random noise Triastcyn and Faltings (2019); Wei et al. (2020).

Notation. Denote by $[n]$ the set $\{1, \dots, n\}$ and consider a finite number of labels, i.e., $|\mathcal{Y}| < \infty$. Each agent $i \in [n]$ has N_y^i calibration samples of label $y \in \mathcal{Y}$, and denote $N_y = \sum_{i=1}^n N_y^i$ their total number over all the calibration examples. Recall that N^i the total number of calibration samples on agent i, i.e., $N^i = \sum_{y \in \mathcal{Y}} N_y^i$. Define the total number of calibration data points $N := \sum_{i=1}^n \sum_{y \in \mathcal{Y}} N_y^i$. For a cumulative distribution function F and $\beta \in [0, 1]$, define by $Q_\beta(F) := \inf\{z \colon F(z) \geq \beta\}$ the β-quantile. Finally, for $v \in \mathbb{R}$ denote by δ_v the point-mass distribution.

5.2 Conformal Prediction for Federated Systems under Label Shift

Non-exchangeable data. In this section, we explain how to take advantage of calibration data to obtain a valid $(1 - \alpha)$-prediction set. Consider the calibration dataset $\{(X_k^i, Y_k^i) \colon k \in [N^i]\}_{i \in [n]}$ with data distributed according to $\{P^i\}_{i \in [n]}$. For $\{\pi_i\}_{i \in [n]} \in \Delta_n$ we define the mixture distribution of labels given for $y \in \mathcal{Y}$ by

$$P_Y^{\text{cal}}(y) = \sum_{i=1}^n \pi_i P_Y^i(y).$$

Our goal is to determine a set of likely outputs for a new data point $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star)$ drawn on agent $\star \in \mathbb{C}$ from the distribution $P^\star$. The conformal approach relies on non-conformity scores $V_k^i = V(X_k^i, Y_k^i) \in [0, 1]$, $i \in [n]$, $k \in [N^i]$ to determine the prediction set – see (Shafer and Vovk, 2008). These non-conformity scores are uniformly weighted to generate the conventional prediction set

$$\begin{aligned} \mathcal{C}_{\alpha,\bar{\mu}}(\mathbf{x}) &= \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\bar{\mu}\right) \right\}, \\ \bar{\mu} &= (N+1)^{-1} (\textstyle\sum_{i=1}^n \sum_{k=1}^{N^i} \delta_{V_k^i} + \delta_1). \end{aligned} \tag{5.2}$$

However, this method can lead to significant under-coverage in the presence of label shift (Podkopaev and Ramdas, 2021a). In fact, since the data $\{(X_k^i, Y_k^i) \colon k \in N^i\}_{i \in [n]}$ are often not exchangeable, it is required to correct the quantile to account for label shift to obtain valid prediction sets (Tibshirani et al., 2019). As proposed by Podkopaev and Ramdas (2021a), we begin by assuming that, for all $i \in [n]$ and $y \in \mathcal{Y}$, we have access to the likelihood ratios:

$$w_y^i = P_Y^i(y) / P_Y^{\text{cal}}(y). \tag{5.3}$$

Denote by $\mathcal{I} = \{(i,k)\colon i \in [n], k \in [N^i]\} \cup \{(\star, N^\star + 1)\}$. Using the weights $\{W_k^i\colon (i,k) \in \mathcal{I}\}$ provided in (5.44), the non-exchangeability correction of Tibshirani et al. (2019) is given for any $\mathbf{y} \in \mathcal{Y}$ by

$$p^\star_{Y_k^i,\mathbf{y}} = \frac{W_k^i}{W^\star_{N^\star+1} + \sum_{j=1}^n \sum_{l=1}^{N^j} W_l^j}, \tag{5.4}$$
$$\mu^\star_{\mathbf{y}} = p^\star_{\mathbf{y},\mathbf{y}}\delta_1 + \sum_{i=1}^n \sum_{k=1}^{N^i} p^\star_{Y_k^i,\mathbf{y}} \delta_{V_k^i}.$$

For any covariate $\mathbf{x} \in \mathcal{X}$, define the $(1-\alpha)$-prediction set with *oracle weights*

$$\mathcal{C}_{\alpha,\mu^\star}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y}\colon V(\mathbf{x},\mathbf{y}) \le Q_{1-\alpha}(\mu^\star_{\mathbf{y}})\right\}.$$

In contrast to the exchangeable setting, the quantile is calculated based on a weighted empirical distribution depending on $\mathbf{y}$. The validity of the prediction set is based on the concept of weighted exchangeability, which was introduced in (Tibshirani et al., 2019, Definition 1); see also (Podkopaev and Ramdas, 2021a, Theorem 2). In the following, we will suppose that the next assumption holds.

Assumption 5.1. *The calibration data points $\{(X_k^i, Y_k^i)\colon (i,k) \in \mathcal{I}\}$ are pairwise independent, and there are no ties between $\{V_k^i\colon (i,k) \in \mathcal{I}\}$ almost surely.*

Theorem 5.2. *If Assumption 5.1 holds, then for any $\alpha \in [0,1)$, we have*

$$1-\alpha \le \mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\mu^\star}(X^\star_{N^\star+1})\right) \le 1-\alpha + \mathbb{E}\left[\max_{(i,k)\in\mathcal{I}}\left\{p^\star_{Y_k^i,Y^\star_{N^\star+1}}\right\}\right], \tag{5.5}$$

where $p^\star_{Y_k^i,Y^\star_{N^\star+1}}$ is defined in (5.4).

This theorem is directly adapted from (Tibshirani et al., 2019, Corollary 1). For completeness, a formal proof is postponed to Section 5.C.1. It is important to note that the lower bound in (5.5) holds even in the presence of ties between non-conformity scores. Although Theorem 5.2 guarantees the validity of $\mathcal{C}_{\alpha,\mu^\star}(X^\star_{N^\star+1})$, this prediction set requires the challenging computation of the weights $p^\star_{y,\mathbf{y}}$. Indeed, the calculation of $W_{y,\mathbf{y}}$ requires the summation over $N!$ elements. The first key contribution of our work is given in Theorem 5.3, where we show that alternative weights, which are easier to compute, can lead to valid prediction sets. Specifically, the new weights $\bar{p}^\star_{y,\mathbf{y}}$ are computed on a smaller number of data points $\bar{N} \le N$, which are randomly selected based on a multinomial random variable with parameter $(\bar{N}, \{\pi_i\}_{i\in[n]})$. Actually, we denote by $\bar{N}^i$ the multinomial count associated with agent i. We take $\bar{N}^i \wedge N^i$ calibration data from agent i and denote $V_k^i = V(X_k^i, Y_k^i)$. For any label $y \in \mathcal{Y}$, the weight $\bar{p}^\star_{y,\mathbf{y}}$ is given by:

$$\bar{p}^\star_{y,\mathbf{y}} = \frac{w^\star_y}{w^\star_{\mathbf{y}} + \sum_{i=1}^n \sum_{k=1}^{N^i\wedge\bar{N}^i} w^\star_{Y_k^i}}. \tag{5.6}$$

In addition, consider the following prediction set

$$\bar{\mu}^\star_{\mathbf{y}} = \bar{p}^\star_{\mathbf{y},\mathbf{y}}\delta_1 + \textstyle\sum_{i=1}^n \sum_{k=1}^{N^i\wedge\bar{N}^i} \bar{p}^\star_{Y_k^i,\mathbf{y}}\delta_{V_k^i},$$
$$\mathcal{C}_{\alpha,\bar{\mu}^\star}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y}\colon V(\mathbf{x},\mathbf{y}) \le Q_{1-\alpha}(\bar{\mu}^\star_{\mathbf{y}})\right\}. \tag{5.7}$$

Denote by $\|w^\star\|_\infty = \max_{y\in\mathcal{Y}}\{w^\star_y\}$. Using the new prediction set $\mathcal{C}_{\alpha,\bar{\mu}^\star}$, we obtain the following result.

Theorem 5.3. *Assume Assumption 5.1. Set* $\bar{N} = \lfloor N/2 \rfloor$ *and* $\pi_i = N^i/N$, *for any* $i \in [n]$. *Then,*

$$\left| \mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1}) \right) - 1 + \alpha \right| \leq \frac{6}{N} + \frac{36 + 6\log N}{N} \|w^\star\|_\infty^2 + \frac{14 \log N}{N} \sum_{i:\ \frac{N^i}{12} < \log N} \sqrt{N^i}.$$

The preceding theorem shows that $\mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1})$ contains the true label $Y^\star_{N^\star+1}$ with probability *close to* $1-\alpha$. If $n = 1$ and $N \geq 46$, the set $\{i \in [n] \colon N^i < 12 \log N\}$ is empty. In this case, the convergence rate reduces to $N^{-1} \log N$. More precisely, if each agent has the same number of calibration data, the convergence rate $N^{-1} \log N$ is ensured when $N \geq 12n \log N$. This is for example the case when $N^i = 200$ and $n \leq 86538$. On the other hand, if $n = N$, each agent has only one data point, and in this case the bound becomes $N^{-1} n (\log N)^{3/2}$.

Approximate Weights. Ideally, we would like to use the weights defined in equation (5.6) to compute valid prediction sets. However, these weights depend on the probability distribution of the labels for each agent, which in many scenarios must be estimated (and therefore known up to an error). Based on empirical estimation of these label probability distributions $\{\widehat{P}^i_Y\}_{i\in[n]}$, for each label y, define the likelihood ratio as follows:

$$\widehat{w}^\star_y = \frac{\widehat{P}^\star_Y(y)}{\sum_{i=1}^n \pi_i \widehat{P}^i_Y(y)}, \tag{5.8}$$

and denote by $\widehat{p}^\star_{y,\mathbf{y}}$ the weight defined in (5.6) with $w^\star_y$ replaced by $\widehat{w}^\star_y$. We also consider $\widehat{\mu}_{\mathbf{y}}$ defined as in (5.7) with $\bar{p}^\star_{y,\mathbf{y}}$ replaced by $\widehat{p}^\star_{y,\mathbf{y}}$. The prediction set becomes

$$\mathcal{C}_{\alpha,\widehat{\mu}}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}\right) \right\}. \tag{5.9}$$

Since computing the exact weights $\bar{p}^\star_{y,\mathbf{y}}$ in (5.6) may not be feasible, we consider the approximation $\widehat{p}^\star_{y,\mathbf{y}}$ given in (5.8). We also construct a random variable $(\widehat{X}^\star_{N^\star+1}, \widehat{Y}^\star_{N^\star+1})$ as in (Lei and Candès, 2021) such that $\mathbb{P}(\widehat{Y}^\star_{N^\star+1} = y) = [\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}^\star_{\tilde{y}} P^{\mathrm{cal}}_Y(\tilde{y})]^{-1} \widehat{w}^\star_y P^{\mathrm{cal}}_Y(y)$, where $\widehat{w}^\star_y$ is defined in (5.8); and $\widehat{X}^\star_{N^\star+1} | \widehat{Y}^\star_{N^\star+1}$ is drawn according to $P_{X|Y}$. The validity of the resulting prediction set is established in Lemma 5.4. Note that this approach makes the weights' computation feasible, at the cost of introducing one additional approximation.

Lemma 5.4. *For any* $\alpha \in (0,1)$, *we have*

$$\left| \mathbb{P}(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})) - \mathbb{P}(\widehat{Y}^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^\star_{N^\star+1})) \right| \leq \frac{1}{2} \sum_{y\in\mathcal{Y}} \left| P^\star_Y(y) - \frac{\widehat{w}^\star_y P^{\mathrm{cal}}_Y(y)}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}^\star_{\tilde{y}} P^{\mathrm{cal}}_Y(\tilde{y})} \right| := \mathrm{R}, \tag{5.10}$$

where $\widehat{w}^\star_y$, $\mathcal{C}_{\alpha,\widehat{\mu}}$ *are defined in* (5.8) *and* (5.9), *respectively.*

When $\widehat{w}^\star_y$ is sufficiently close to $w^\star_y$, Lemma 5.4 shows that the approximate weights generate accurate prediction sets (as discussed in Section 5.C.1). The error disappears entirely when $\widehat{w}^\star_y = w^\star_y$ for all $y \in \mathcal{Y}$. Furthermore, using (Tibshirani et al., 2019,

Corollary 1), we can establish that $\widehat{Y}^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^\star_{N^\star+1})$ with probability nearly $1-\alpha$. Finally, similar ideas that developed for Theorem 5.3 on $Y^\star_{N^\star+1}$, in conjunction with Lemma 5.4, give a more accurate bound on the coverage validity.

Theorem 5.5. *Assume Assumption 5.1. For any $i \in [n]$, set $\pi_i = N^i/N$ and take $\bar{N} = \lfloor N/2 \rfloor$. Then,*

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| \leq \frac{36\|\widehat{w}^\star\|_\infty^2}{N(\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}})^2} + \mathrm{R} + \frac{6}{N} + \frac{2\log N}{N}\left(\frac{3\|\widehat{w}^\star\|_\infty^2}{(\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}})^2} \vee 7\sum_{i:\ \frac{N^i}{12}<\log N}\sqrt{N^i}\right),$$

where $\widehat{w}^\star_{Y^i_k}$, R are defined in (5.8)-(5.10) and $Y^{\mathrm{cal}} \sim P_Y^{\mathrm{cal}}$.

This theorem provides a lower bound on the probability of coverage that is independent of the data distribution. A formal proof can be found in Section 5.C.5. This result demonstrates that it is essential to include all agents with the most data. However, it also highlights a counterproductive effect when incorporating agents with few data.

Maximum Likelihood Estimation Weights. Denote by M^i_y the number of training data on agent i associated to label y. Consider the total number of local data $\mathsf{M}^\star = \sum_{y\in\mathcal{Y}} \mathsf{M}^i_y$, the number of training data with label y written by $\mathsf{M}_y = \sum_{i=1}^n \mathsf{M}^i_y$, and the total number of samples on all agents by $\mathsf{M} = \sum_{y\in\mathcal{Y}} \mathsf{M}_y$. When each agent independently learns its approximate label distribution based on counting the number of label in its training datasets, the empirical counterpart of (5.8) is given for any labels $(y, \mathbf{y}) \in \mathcal{Y}^2$ by

$$\widehat{w}^\star_y = \frac{\mathsf{M}\mathsf{M}^\star_y}{\mathsf{M}^\star\mathsf{M}_y}\mathbb{1}_{\mathsf{M}_y\geq 1}. \tag{5.11}$$

All the results in this article are given conditionally to the training dataset, meaning that they hold regardless of the specific training data. In order to determine the order of magnitude of the bound of Theorem 5.5, we analyze the average value of R. Given the number of training samples $\{\mathsf{M}^i\}_{i\in[n]}$, if we assume that each training point (X^i_k, Y^i_k) is distributed according to P^i, then taking the expectation over the training set yields:

$$\mathbb{E}\left[\mathrm{R}\right] \leq \frac{6}{\sqrt{\mathsf{M}^\star}} + 12\sqrt{\frac{\log|\mathcal{Y}| + \log\mathsf{M}^\star}{\mathsf{M}\min_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)}}. \tag{5.12}$$

The proof is given in Section 5.C.3. Interestingly, if the previous upper bound is plugged in Theorem 5.5 instead of Lemma 5.4, then the leading error of order $\mathrm{O}(\mathsf{M}^{\star-1/2} \vee N^{-1}\log N)$ is due to the weights' estimates $\{\widehat{p}^\star_{y,\mathbf{y}}\}_{y\in\mathcal{Y}}$. This bound shows that we should not attempt to estimate the likelihood ratios for a single agent, especially when the square root number of local training data on agent $\star$ is small compared to the number of calibration data. Rather, we need to do this for a group of agents that have approximately the same distribution, which will give us more stable estimators. The agent can benefit from learning simultaneous tasks by exploiting common structures (Caruana, 1998).

5.3 Privacy Preserving Federated CP

In the previous section, we constructed prediction sets that were valid in theory. However, their practical implementation in a federated environment posed challenges due to the reliance on estimations that are difficult to evaluate. In particular, estimating $Q_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}\right)$ in order to derive the prediction set $\mathcal{C}_{\alpha,\widehat{\mu}}(\mathbf{x})$, defined in (5.9), is challenging because it requires knowledge of the global distribution $\widehat{\mu}_y$. This section is divided into two parts: (1) a new method is developed, called `DP-FedCP`, for estimating quantiles under the federated constraints; (2) then, a method for computing probabilities $\{\widehat{p}^{\star}_{y,\mathbf{y}}\}_{y,\mathbf{y}\in\mathcal{Y}}$ with differential privacy (DP) guarantees is presented.

Quantile Regression and Moreau-Yosida Regularization. Let $\alpha \in (0,1)$, we now propose to estimate the weighted $(1-\alpha)$-quantile of $\widehat{\mu}_{\mathbf{y}}$ defined in (5.16). To this end, we develop a federated optimization algorithm based on "pinball loss" minimization, a quantile regression techniques with asymmetric penalties (Koenker and Hallock, 2001). For $v \in \mathbb{R}$ and $q \in \mathbb{R}$ define the pinball loss as

$$S_{\alpha,v}(q) = (1-\alpha)(v-q)\mathbb{1}_{v\geq q} + \alpha(q-v)\mathbb{1}_{q>v}.$$

For any $\mathbf{y} \in \mathcal{Y}$, the $(1-\alpha)$-quantile of $\widehat{\mu}_{\mathbf{y}}$ is given by

$$Q_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}\right) \in \operatorname*{arg\,min}_{q\in\mathbb{R}} \left\{ \mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}\left[S_{\alpha,V}(q)\right] \right\}; \tag{5.13}$$

e.g. see (Buhai, 2005). The pinball loss $S_{\alpha,v}$ is lower semi-continuous but not differentiable on $\mathbb{R}$. Hence, we consider the Moreau-Yosida inf-convolution (or *envelope*) $S^{\gamma}_{\alpha,v}$ instead of $S_{\alpha,v}$ – where γ is the regularization parameter; see e.g. (Moreau, 1963) and (Parikh et al., 2014, Chapter 3), whose expression is given by

$$S^{\gamma}_{\alpha,v}(q) = \min_{\tilde{q}\in\mathbb{R}} \left\{ S_{\alpha,v}(\tilde{q}) + \frac{1}{2\gamma}(\tilde{q}-q)^2 \right\}. \tag{5.14}$$

The function $S^{\gamma}_{\alpha,v}(\cdot)$ has an explicit expression given in (5.20). Note that the minima of $S_{\alpha,v}$ and $S^{\gamma}_{\alpha,v}$ coincide. We obtain the weighted quantile by considering $S^{\gamma}_{\alpha,v}$ instead of $S_{\alpha,v}$. An important property is that the inf-convolution of a proper lower semicontinuous convex function is a differentiable function whose derivative is Lipschitz; see (Rockafellar and Wets, 2009, Theorem 2.26). The original optimization problem given in (5.13) is replaced by a convex/smooth loss:

$$Q^{\gamma}_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}\right) \in \operatorname*{arg\,min}_{\mathbb{R}}\{\mathrm{S}^{\gamma}_{\alpha}(q)\}, \tag{5.15}$$

where $\mathrm{S}^{\gamma}_{\alpha}\colon \mathbb{R} \to \mathbb{R}_+$ is the function given by

$$\mathrm{S}^{\gamma}_{\alpha}\colon q \mapsto \mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S^{\gamma}_{\alpha,V}(q)].$$

For almost every value of $\alpha \in (0,1)$, there exists a unique minimizer of $\mathrm{S}^{\gamma}_{\alpha}$. This minimizer $Q^{\gamma}_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})$ of the regularized loss function deviates from the true quantile. However, the error is controlled by the regularization parameter γ and is asymptotically exact when $\gamma \to 0$. More precisely (see Section 5.A.2 for a proof) it holds that:

Algorithm 5.10 `DP-FedAvgQE`

Input: initial quantile q_0, target significance level α, number of rounds T, learning rate η, Moreau regularization parameter γ, local gradients $\{\nabla \mathrm{S}_\alpha^{i,\gamma}\}_{i\in[n]}$, local non-conformity scores $\{V_k^i\}_{k\in[N^i+1]}$, mixture weights $\{\lambda_{\mathbf{y}}^i\}_{i\in[n]}$, standard deviation of Gaussian mechanism noise σ_g, K number of local iteration.

for $t = 0$ **to** $T-1$ **do**

$S_{t+1} \leftarrow$ random subset of $[n]$ // Server side

for each agent $i \in S_{t+1}$ **do** // In parallel

Initialize quantile $q_{t,0}^i \leftarrow q_t$

for $k = 0$ **to** $K-1$ **do**

// Gradient with DP noise

$g_{t,k}^i \leftarrow \nabla \mathrm{S}_\alpha^{i,\gamma}(q_{t,k}^i) + z_{t,k}^i$, $z_{t,k}^i \sim \mathcal{N}(0, \sigma_g^2)$

// Update local quantile

$q_{t,k+1}^i \leftarrow q_{t,k}^i - \eta g_{t,k}^i$

$(\Delta q_{t+1}^i, \Delta \bar{q}_{t+1}^i) \leftarrow (q_{t,K}^i - q_{t,0}^i, \sum_{k\in[K]} \frac{q_{t,k}^i}{K})$

// On the central server

$q_{t+1} \leftarrow q_t + \frac{n}{|S_{t+1}|} \sum_{i\in S_{t+1}} \Delta q_{t+1}^i$

$\bar{q}_{t+1} \leftarrow \frac{t}{t+1}\bar{q}_t + \frac{n}{|S_{t+1}|} \sum_{i\in S_{t+1}} \frac{\lambda_{\mathbf{y}}^i \Delta \bar{q}_{t+1}^i}{t+1}$

Output: $\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) \leftarrow \bar{q}_T$.

Theorem 5.6. *Let $\gamma > 0$ and $\alpha \in (0,1)$. Assume that for all $\{y_\ell\}_{\ell\in[N+1]} \in \mathcal{Y}^{[N+1]}$, $1-\alpha \notin \{W_k/W_{N+1}\}_{k\in[N+1]}$, where $W_k = \sum_{\ell=1}^k \widehat{w}_{y_\ell}^\star$. Then, we have $|Q_{1-\alpha}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})| \leq \gamma$.*

The condition on α assumed in Theorem 5.6 ensures the uniqueness of the minimizer of S_α^γ.

Federated quantile computation. We now describe the Differentially Private Federated Average Quantile Estimation (`DP-FedAvgQE`) algorithm (see Algorithm 5.10), a novel method to compute quantile in a federated learning setting, with DP guarantees. We briefly described this method below. For each query $\mathbf{y} \in \mathcal{Y}$, we consider the distributions $\widehat{\mu}_{\mathbf{y}} = \sum_{i=1}^n \lambda_{\mathbf{y}}^i \widehat{\mu}_{\mathbf{y}}^i$, where $\lambda_{\mathbf{y}}^i$ and $\widehat{\mu}_{\mathbf{y}}^i$ are given by

$$\begin{aligned} \lambda_{\mathbf{y}}^i &= \tfrac{N^i}{N}\widehat{p}_{\mathbf{y},\mathbf{y}}^\star + \textstyle\sum_{k=1}^{N^i \wedge \bar{N}^i} \widehat{p}_{Y_k^i,\mathbf{y}}^\star, \\ \widehat{\mu}_{\mathbf{y}}^i &= \tfrac{N^i \widehat{p}_{\mathbf{y},\mathbf{y}}^\star}{\lambda_{\mathbf{y}}^i N}\delta_1 + \textstyle\sum_{k=1}^{N^i \wedge \bar{N}^i} \frac{\widehat{p}_{Y_k^i,\mathbf{y}}^\star}{\lambda_{\mathbf{y}}^i}\delta_{V_k^i}. \end{aligned} \tag{5.16}$$

To simplify the notation, for any client $i \in [n]$, we introduce the local loss function $\mathrm{S}_\alpha^{i,\gamma} \colon q \in \mathbb{R} \mapsto \mathbb{E}_{V\sim\widehat{\mu}_y^i}[S_{\alpha,V}^\gamma(q)]$; see (5.22) for explicit expression.

At each iteration $t \in [T]$, the server subsamples the participating agents $S_{t+1} \subseteq [n]$ independently of the past. Each selected agent $i \in S_{t+1}$ performs K local updates: (1) they independently compute their local gradient; (2) a Gaussian noise is added as in (5.17) to ensure the differential privacy. More precisely, for agent $i \in S_{t+1}$, at local iteration $k \in \{0, \ldots, K-1\}$, we define:

$$g_{t,k}^i = \nabla \mathrm{S}_\alpha^{i,\gamma}(q_{t,k}^i) + z_{t,k}^i, \tag{5.17}$$

where $\{z^i_{t,k}\colon (t,k) \in \{0,\dots,T-1\} \times [K]\}_{i\in[n]}$ are i.i.d. Gaussian random variables with zero mean and variance σ_g^2. For any agent $i \in S_{t+1}$, $g^i_{t,k}$ is an unbiased estimate of $\nabla \mathrm{S}^{i,\gamma}_\alpha(q^i_{t,k})$. (3) The participating agents update their local quantiles $q^i_{t,k+1} \leftarrow q^i_{t,k} - \eta g^i_{t,k}$, where η is a positive step-size; (4) then transmit $(\Delta q^i_{t+1}, \Delta \bar{q}^i_{t+1}) = (q^i_{t,K} - q^i_{t,0}, \sum_{k\in[K]} q^i_{t,k}/K)$ to the central server. The parameter Δq^i_{t+1} is used to update the common parameter q_t, while $\Delta \bar{q}^i_{t+1}$ is necessary to keep track of the average of the sampled parameters denoted $\bar{q}_t$; see Nemirovski et al. (2009); Bubeck et al. (2015). (5) Finally, the server performs an online average to update $\bar{q}_t$ and computes the new parameter following

$$q_{t+1} = q_t + (n/|S_{t+1}|) \sum_{i\in S_{t+1}} \Delta q^i_{t+1}.$$

At the final stage, the central server output the quantile estimate is given by

$$\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}) = \sum_{t=1}^T (n/|S_t|) \sum_{i\in S_t} \lambda^i_{\mathbf{y}} \Delta \bar{q}^i_t / T. \tag{5.18}$$

Algorithm 5.10 is a Federated Averaging procedure (McMahan et al., 2017) applied to the Moreau envelope of the pinball loss. As we will see in Section 5.4, the addition of an independent Gaussian noise on the parameter at each update round provides differential privacy guarantees; see Theorem 5.13 for more details.

Remark 5.7. *Privacy is also at risk when computing probabilities $\{\widehat{p}^\star_{y,\mathbf{y}}\}_{y,\mathbf{y}\in\mathcal{Y}}$. To compute the probabilities $\{\widehat{p}^\star_{y,\mathbf{y}}\}_{y,\mathbf{y}\in\mathcal{Y}}$ while preserving privacy, we need specific mechanisms to transmit the number of training labels $(\mathsf{M}^i_y)_{y\in\mathcal{Y}}$ from each agent i to the server. For this purpose, we use the method proposed in (Canonne et al., 2020). The idea is to add a discrete noise to the counts $\{\mathsf{M}^i_y\colon i \in [n]\}_{y\in\mathcal{Y}}$ and then transmit these noisy proxies. The resulting algorithm that combines the differentially-private count queries and federated quantile computation is given in Algorithm 5.11.*

Remark 5.8. *Algorithm 5.11 is designed to build a confidence set for the single agent $\star$. By vectorizing all computations, the algorithm can be scaled to compute a confidence set for each agent. This would result in an algorithm that remains linear in the number of clients but would be more efficient than computing several independent runs. From a practical perspective, complexity can be further improved by clustering clients into groups based on their label distributions and performing conformal prediction on a group level.*

Remark 5.9. *The local loss functions $\mathrm{S}^{i,\gamma}_\alpha$ are expressed as the expectation of pinball loss functions. Since the sensitivity of these pinball loss functions is 1, there is no need to clip the gradient. It is sufficient adding Gaussian noise $\mathcal{N}(0,\sigma_g^2)$ to guarantee differential privacy. The value of σ_g is chosen to provide a suitable tradeoff between privacy and utility, balancing the need for strong privacy protection with useful outputs. For an explicit setting of σ_g, refer to Theorem 5.13.*

5.4 Theoretical Guarantees

Convergence guarantee. We provide a convergence guarantee for `DP-FedAvgQE`. Details of the proofs can be found in the supplementary chapter. We show the convergence of $\{\widehat{Q}^\gamma_{1-\alpha,t}(\widehat{\mu}_{\mathbf{y}})\}_{t\in\mathbb{N}}$ to a minimizer which is unique under the assumptions discussed in Section 5.A.2. We briefly sketch key steps from the theoretical derivations, since the local loss functions $\{\mathrm{S}^{i,\gamma}_\alpha\}_{i\in[n]}$ have different minimizers, this client drift/heterogeneity

Algorithm 5.11 `DP-FedCP`

Input: calibration dataset $\{(X_k^i, Y_k^i)\colon k \in [N^i]\}_{i\in[n]}$, covariate $\mathbf{x}$, communication round number T, subsampling number $\bar{N}$, Gaussian noise parameters $\sigma_g, \bar{\sigma} \geq 0$.
for each agent $i \in [n] \cup \{\star\}$ **do** // In parallel
  Set $\forall y \in \mathcal{Y}, \mathsf{M}_y^i \leftarrow$ number train data with label y
  Generate $\{\eta_y^i\}_{y\in\mathcal{Y}}$ i.i.d. according to $\mathcal{N}_{\mathbb{Z}}\left(0, \bar{\sigma}^2\right)$
  Send $\forall y \in \mathcal{Y}, \widehat{N}_y^i \leftarrow \max(1, \mathsf{M}_y^i + \eta_y^i)$
  Compute & Send $\{V(X_k^i, Y_k^i)\colon k \in [N^i]\}_{i\in[n]}$
// On the central server
Aggregate $\widehat{N}_y \leftarrow \sum_{i\in[n]} \widehat{N}_y^i, \forall y \in \mathcal{Y}$
Aggregate $\widehat{N} \leftarrow \sum_{y\in\mathcal{Y}} \widehat{N}_y$
for each query $\mathbf{y} \in \mathcal{Y}$ **do**
  Sample $\{\bar{N}^i\}_{i\in[n]} \sim \mathcal{M}ulti(\bar{N}, \{N^i/N\}_{i\in[n]})$
  Compute $\widehat{p}^{\star}_{y,\mathbf{y}}$ as in (5.6) with $\widehat{w}_y^{\star}$ given in (5.11)
  Compute $\widehat{Q}^{\gamma}_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}) \leftarrow$`DP-FedAvgQE`
Output: $\widehat{\mathcal{C}}^{\gamma}_{\alpha,\widehat{\mu}}(\mathbf{x}) \leftarrow \left\{\mathbf{y}\colon V(\mathbf{x},\mathbf{y}) \leq \widehat{Q}^{\gamma}_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})\right\}$.

may slow down the convergence (Li et al., 2019). This dissimilarity is evaluated by the parameter $\zeta \geq 0$, which is given by

$$\zeta = \max_{i\in[n]} \|\nabla \mathrm{S}_{\alpha}^{i,\gamma} - \nabla \mathrm{S}_{\alpha}^{\gamma}\|_{\infty}^{1/2}.$$

The convergence analysis is performed for the estimate parameter $\widehat{Q}^{\gamma}_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})$ given in (5.18). We provide below the statements without subsampling, i.e. $S_t = [n]$, given in Section 5.B. Recall that $Q^{\gamma}_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})$ is provided in (5.15) and denote $\Delta = \mathbb{E}_{q_0}\|q_0 - Q^{\gamma}_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\|^2$. The following results hold with fixed train/calibration datasets $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}})$, and define their union by $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$.

Theorem 5.10. *Let $\gamma \in (0,1]$, $S_t = [n]$ and consider the step-size $\eta \in (0, \gamma/10]$. Then, for $t \in \{0,\dots,T-1\}$, $k \in \{0,\dots,K-1\}$, we have*

$$\mathbb{E}\left[\mathrm{S}_{\alpha}^{\gamma}(\widehat{Q}^{\gamma}_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}))\,|\,\mathcal{D}\right] - \mathrm{S}_{\alpha}^{\gamma}(Q^{\gamma}_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) \leq (\eta KT)^{-1}\Delta + 14\gamma^{-1}\eta^2 K(\sigma_g^2 + K\zeta^2).$$

The presence of heterogeneity among local datasets significantly influences convergence dynamics, particularly when the number of targets K, is significantly larger than 1. In such cases, the term $K^2\zeta^2$ poses challenges by potentially hindering the effectiveness of numerous local steps. Consider the step-size $\eta_\star$ defined by

$$\eta_\star = \min\left\{\frac{\gamma}{10}, \left(\frac{\gamma\Delta}{13K^2T(\sigma_g^2 + \zeta^2 K)}\right)^{1/3}\right\}.$$

Setting $\eta = \eta_\star$, we obtain the following result.

Corollary 5.11. *Let $\gamma \in (0,1]$, $S_t = [n]$ and consider the step-size $\eta_\star$. Then, for any $t \in \{0,\dots,T-1\}$, $k \in \{0,\dots,K-1\}$, we have*

$$\epsilon_{\text{optim}}^{(\gamma)} = \mathbb{E}\Big[\mathrm{S}_\alpha^\gamma(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})) \,|\, \mathcal{D}\Big] - \mathrm{S}_\alpha^\gamma(Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}})) \leq \frac{10\Delta}{\gamma K T} + \frac{5\left(\sigma_g^2 + \zeta^2 K\right)^{1/3}\Delta^{2/3}}{(\gamma K T^2)^{1/3}}. \quad (5.19)$$

As shown in Corollary 5.11, $\epsilon_{\text{optim}}^{(\gamma)}$ increases inversely proportional to γ. The smaller the regularization parameter γ, the smaller the step-size η must be, and the more iterations are required to achieve the same accuracy. However, the error caused by the Moreau envelope vanishes for $\gamma \downarrow 0^+$, i.e. $Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}})$ approaches $Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})$. Thus, there is a tradeoff between the accuracy of the quantile approximation $\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})$ and the computational cost.

Conformal guarantees for `DP-FedCP`. We show that the confidence set $\widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star)$ provided by `DP-FedCP` constitutes valid coverage of $Y_{N^\star+1}^\star$. The theoretical derivations and complete statements are given in Section 5.C. For all $i \in [n]$, denote by P_V^i the distribution of $V(X^i, Y^i)$ where $(X_i, Y_i) \sim P^i$, and consider $Y^{\text{cal}} \sim P_Y^{\text{cal}}$.

Theorem 5.12. *Assume there exist $m, M > 0$ such that for any $i \in [n]$, P_V^i admits a density f_V^i with respect to the Lebesgue measure that satisfies $m \leq f_V^i \leq M$. For any $\alpha \in [0,1] \setminus \mathbb{Q}$, it holds*

$$\left|\mathbb{P}(Y_{N^\star+1}^\star \in \widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star)) - \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star))\right| \leq 6M\sqrt{\frac{\log(N)\sum_{y\in\mathcal{Y}} P_Y^{\text{cal}}(y)\widehat{w}_y^\star}{m \min_{y\in\mathcal{Y}} \widehat{w}_y^\star}\left(\mathbb{E}\left[\epsilon_{\text{optim}}^{(\gamma)} | \mathcal{D}_{\text{train}}\right] + \gamma\right)} + \frac{2M\log N}{mN} + \frac{4\,\mathrm{Var}(\widehat{w}_{Y^{\text{cal}}}^\star)}{N(\mathbb{E}\widehat{w}_{Y^{\text{cal}}}^\star)^2} + \frac{2\mathbb{E}\widehat{w}_{Y_{N^\star+1}^\star}^\star}{N\mathbb{E}\widehat{w}_{Y^{\text{cal}}}^\star} + \frac{m}{2N\log N} + \frac{1}{N^2},$$

where $\epsilon_{\text{optim}}^{(\gamma)}$ is defined in (5.19).

This result illustrates an interesting tradeoff introduced by the regularization parameter γ. As shown in Corollary 5.11, $\epsilon_{\text{optim}}^{(\gamma)}$ increases inversely proportional to γ. Therefore, setting $\gamma \approx T^{-1/2}$ ensures a convergence rate of order $T^{-1/4}$ for the optimization procedure. In this case, the error term of order $\mathrm{O}(N^{-1}\log N)$ is guaranteed by choosing the number of iterations $T \approx N^4$. The condition $\alpha \in [0,1] \setminus \mathbb{Q}$ is a strong but unnecessary assumption. However, it provides a simple way to ensure that $\widehat{\mu}_{\mathbf{y}}$ has no jump at level $1-\alpha$. Interestingly, the same condition on α is used in (Podkopaev and Ramdas, 2021a, Corollary 1), where the authors explain why this condition cannot be avoided to ensure the consistency of the empirical quantile estimator.

Differential privacy guarantees. The (ϵ,δ)-differentially private nature of `DP-FedAvgQE` relies on two components: the additional Gaussian noise, combined with the bounded gradient which avoids extreme values/outliers. The parameter ϵ controls the level of

privacy protection provided by a differentially private algorithm, by limiting the probability of inferring any information about an individual in a given dataset. However, there is a small chance that the algorithm may leak some information, even though this probability is kept under control by the parameter δ. Based on the Rényi differential privacy (Mironov, 2017), joined to agent subsampling mechanism (Balle et al., 2018), we establish the (ϵ, δ)-DP property following similar ideas to those of (Noble et al., 2022, Theorem 4.1). Detailed proof and definitions are provided in Section 5.D.

Theorem 5.13. *If there is a constant number $S \in [n]$ of sampled agents, i.e., $S_t = S$, for all $t \in [T]$. Then, for all $\epsilon > 0$ and $\delta \in (0, 1-(1+\sqrt{\epsilon})(1-S/n)^T)$, the Algorithm 5.10 is (ϵ, δ)-DP towards a third party when*

$$\sigma_g \geq 2\sqrt{\frac{K \max_{i\in[n]} \lambda^i_{\mathbf{y}}}{\epsilon}\left(1 + \frac{24S\sqrt{T}\log(1/\bar{\delta})}{\epsilon n}\right)},$$

$$\text{where} \quad \bar{\delta} = \frac{n}{S}\left[1 - \left(\frac{1-\delta}{1+\sqrt{\epsilon}}\right)^{1/T}\right].$$

5.5 Numerical experiments

We conducted the experimental study of `DP-FedCP` using both synthetic toy examples and real datasets. To perform a comprehensive evaluation, we compared our method with relevant baselines, namely `Unweighted Local` and `Unweighted Global` (see Section 5.E for details). The `Unweighted Local` method computes the quantile based on the local validation data of the agent $\star$ and derives the local unweighted prediction set with $(1-\alpha)$ confidence level, given by

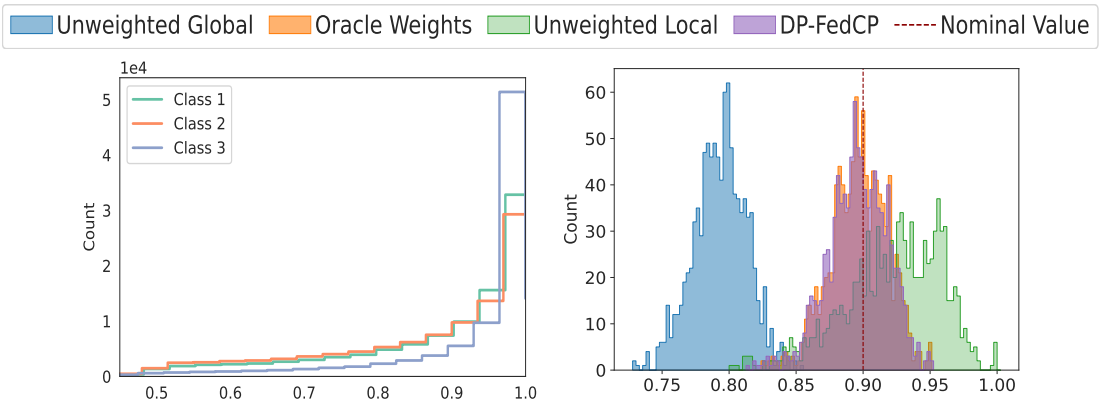
$$\mathcal{C}_{\alpha,\bar{\mu}^{\text{loc},\star}}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\bar{\mu}^{\text{loc},\star}_{\mathbf{y}}\right)\right\},$$

where $\bar{\mu}^{\text{loc},\star}_{\mathbf{y}} = \frac{1}{N^\star+1}\sum_{k=1}^{N^\star} \delta_{V^\star_k} + \frac{1}{N^\star+1}\delta_1$. This method is the adaptive classification technique with split-conformal calibration applied to agent $\star$, as introduced in Romano et al. (2020) and also described in Angelopoulos et al. (2021). On the other hand, the `Unweighted Global` method estimates the quantile based on aggregated non-conformity scores from all the agents, without taking into account the shift between calibration and target distributions. This method computes the $(1-\alpha)$-quantile in an analogous way to the "classical" conformal method recalled in (5.2).

For our experiments, we apply split-conformal calibration on the entire dataset, which requires all agents to report their non-conformity scores to a central server. We use the same non-conformity score $V(x, y)$ as considered in Romano et al. (2020); Angelopoulos et al. (2021). Given the covariate x, the predictor $\hat{f} \colon \mathcal{X} \to \Delta_{|\mathcal{Y}|}$ estimates the probability of each class, and orders them from the most to the least likely label. The non-conformity score is then computed as the sum of all the probabilities greater than the true label y. Formally, the non-conformity scores are given by

$$\rho(X^i_k, Y^i_k) = \textstyle\sum_{y\in\mathcal{Y}} \hat{f}(X^i_k)[y]\, \mathbb{1}\hat{f}(X^i_k)[y] > \hat{f}(X^i_k)[Y^i_k],$$
$$V(X^i_k, Y^i_k) = \rho(X^i_k, Y^i_k) + U^i_k \times \hat{f}(X^i_k)[Y^i_k],$$

where $U^i_k \in [0, 1]$ is a uniform random variable.



(a) Non-conformity scores

(b) Empirical coverage

Figure 5.1 – Simulated data experiment with 2D data. Target confidence level $(1-\alpha) = 0.9$.

Simulated Data Experiment. In the first experiment, we demonstrate that it is necessary to consider label shifts between agents to obtain valid coverage of prediction sets. We consider a simple classification problem with 3 labels. The conditional distributions of the features given the class label are 3 two-dimensional Gaussian distributions with means $\boldsymbol{\theta}_1 = [-1, 0], \boldsymbol{\theta}_2 = [1, 0], \boldsymbol{\theta}_3 = [1, 3]$ and with identity covariance matrices. We consider $n = 2$ agents with the distribution of labels $\{P_Y^1(y)\}_{y\in[3]} = \{0.8, 0.1, 0.1\}$ and $\{P_Y^2(y)\}_{y\in[3]} = \{0.1, 0.1, 0.8\}$. We use the Bayes classifier and consider calibration data with $(N^1, N^2) = (1000, 50)$. The inference is performed for agent 2.

We run independently 1000 experiments with different splits and record the obtained empirical coverage each time. Figure 5.1a shows the distribution of non-conformity scores for the different labels, and Figure 5.1b shows the empirical coverage of $(1-\alpha)$ prediction sets with $\alpha = 0.1$ using the `DP-FedCP` method (Algorithm 5.11) compared to `Unweighted Local` and `Unweighted Global`. We also included results obtained with `oracle-weights`, in which the conformal prediction sets are obtained using (5.89), i.e., assuming that the exact ratios $\{w_y^\star\}_{y\in\mathcal{Y}}$ are known.

The quantiles calculated via the `Unweighted Global` method are mostly due to the non-conformity scores from agent 1. This is due to the larger local dataset of agent 1, whose label distribution is very different from that of the target; see Figure 5.1a. The `Unweighted Local` method computes the quantiles based on the local data of agent 2, which has too little data to produce robust prediction sets. Therefore, `DP-FedCP` yields much better conformal prediction sets (see Figure 5.1b), which are little different from those obtained using the adaptive prediction set methods with oracle weights of Podkopaev and Ramdas (2021b).

CIFAR-10 Experiments. We investigate the performance of `DP-FedCP` on the CIFAR-10 dataset. We use a ResNet-56 (He et al., 2016) pre-trained on the CIFAR-10 training dataset as the underlying classifier with temperature scaling $T = 1.6$. We also randomly split the CIFAR-10 test dataset into a calibration dataset and a test dataset, each containing 5000 points, and repeat the experiment 1000 times. The number of agents is $n = 10$, and the prediction set is learned for the agent $\star = 4$ that has the smallest number of data points. The distribution of labels for agent i is $P_Y^i(i) = 0.55$

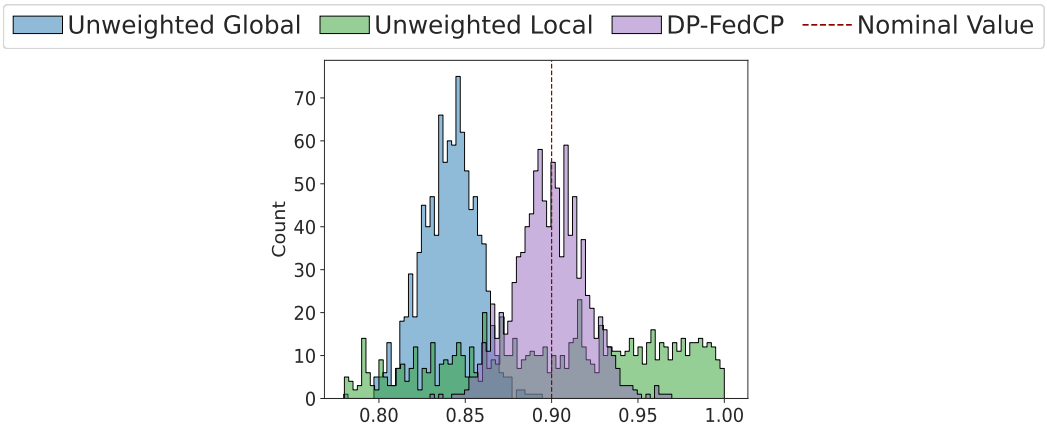
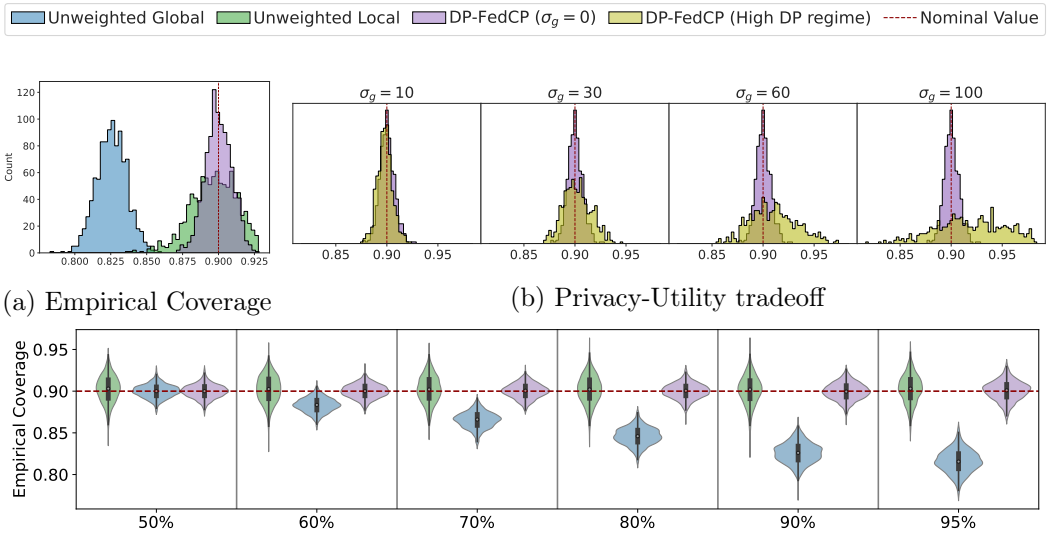


Figure 5.2 – Empirical coverage on the CIFAR-10 data.



(a) Empirical Coverage

(b) Privacy-Utility tradeoff

(c) Distribution of the empirical coverage for the methods as a function of the shift

Figure 5.3 – ImageNet experimental results: (a) Empirical coverage comparison of `DP-FedCP` with unweighted baselines (b) Empirical coverage comparison of `DP-FedCP` with non-DP version at different privacy parameter values (c) Effect of distribution shifts on empirical coverage for `DP-FedCP` and unweighted baselines.

and $P_Y^i(y) = 0.05$ for all $y \in [10] \backslash \{i\}$. We set the validation size for agent $\star$ to $N^\star = 50$, and for agent 2 the validation size is $N^2 = 2150$. The remaining agents have the same validation size of $N^i = 350$ for all $i \in [10] \backslash \{2, 4\}$. The significance level α is set to 0.1. In this configuration, both `Unweighted Local` and `Unweighted Global` methods perform significantly worse than `DP-FedCP`; see Figure 5.2.

ImageNet Experiments. We use a pre-trained ResNet-152 (He et al., 2016) as a base model with temperature scaling $T = 10$. We perform 1000 runs with different splits of the 50K ImageNet test dataset into calibration and test datasets of size 40K and 10K samples, respectively. The calibration data is split into 11 agents. For agent $i \in [10]$, the size of the calibration dataset is $N^i = 3950$, while we $N^{11} = 500$. For ImageNet, the distribution of non-conformity scores $V(\hat{f}(X), Y)$ varies significantly as

a function of the given label $Y = y$. In this experiment, we distribute the data between agents to ensure distinct non-conformity score distributions across agents, illustrated in Figures 5.6a and 5.6b. For this, we compute the mean of the non-conformity scores in function of the given label. We call G_1 the set of the 500 labels with the lowest means and G_2 the set of the remaining 500 labels. Agents $i \in [10]$ (`low-score` group) take 90% of their data from G_1 and the remaining 10% from G_2. Agent 11 takes 90% of its calibration data from G_2 and the remaining 10% from G_1.

We construct a prediction set with significance level $\alpha = 0.1$ for the distribution of the 11-th agent. Figure 5.3a shows the empirical coverage of the prediction sets. In contrast to unweighted alternatives, `DP-FedCP` achieves valid coverage. In Figure 5.3c, we evaluate the sensitivity of the different methods to the shift between G_1 and G_2. We repeat the previous experiment varying the shift parameter (90% in the first experiment) with 100 runs for each coefficient and show the Violin plot of the obtained empirical coverage. The experimental results show that `DP-FedCP` overcomes the challenge of obtaining valid conformal predictions in the presence of label shifts at a federated level compared to alternative methods.

Differential Privacy Experiments. We explore the tradeoff between privacy and coverage quality. We conducted the ImageNet experiment with different values of σ_g in the set $\{10, 30, 60, 100\}$. The results of the experiment are shown in Figure 5.3b, which illustrates the tradeoff between the differential privacy parameter σ_g and the robustness of the method. In particular, we observe that as σ_g increases, the robustness of the method decreases.

5.6 Conclusion

We present a novel method called `DP-FedCP`, which is designed to construct personalized conformal prediction sets in a federated learning scenario. Unlike existing algorithms, the proposed method takes into account the label shifts between different agents, and computes prediction sets with a prescribed confidence level. The resulting sets are theoretically guaranteed to provide valid coverage, while ensuring differential privacy. Finally, we illustrate the strong performance of `DP-FedCP` in a series of benchmarks.

5.A Moreau Envelope for Quantile Computation

5.A.1 Federated quantile using the Moreau envelope

Lemma 5.14. *Let $\alpha \in [0,1]$ and $(v,q) \in \mathbb{R}^2$, the Moreau envelope of the pinball loss with regularization parameter $\gamma > 0$ is given by*

$$S^{\gamma}_{\alpha,v}(q) = \begin{cases} (1-\alpha)(v-q) - \frac{\gamma(1-\alpha)^2}{2}; & \frac{v-q}{\gamma} > 1-\alpha, \\ \frac{(q-v)^2}{2\gamma}; & 0 \leq \frac{q-v}{\gamma} + 1 - \alpha \leq 1, \\ \alpha(q-v) - \frac{\gamma\alpha^2}{2}; & \frac{q-v}{\gamma} > \alpha. \end{cases} \tag{5.20}$$

Moreover, its gradient is given by

$$\nabla S^{\gamma}_{\alpha,v}(q) = -(1-\alpha)\mathbb{1}_{\{q<v-\gamma(1-\alpha)\}} + \alpha\mathbb{1}_{\{q>v+\gamma\alpha\}} + \frac{1}{\gamma}(q-v)\mathbb{1}_{\{v-\gamma(1-\alpha)<q<v+\gamma\alpha\}}.$$

Proof For all $\alpha \in [0,1]$, $(v,q) \in \mathbb{R}^2$, recall that the pinball loss and its subgradient are given by

$$S_{\alpha,v}(q) = (1-\alpha)(v-q)\mathbb{1}_{\{v\geq q\}} + \alpha(q-v)\mathbb{1}_{\{q>v\}}, \tag{5.21}$$

$$\partial S_{\alpha,v}(q) = \begin{cases} -(1-\alpha), & q < v \\ [-(1-\alpha), \alpha], & q = v \\ \alpha, & q > v \end{cases}.$$

Note that, by construction

$$\begin{aligned} S^{\gamma}_{\alpha,v}(q) &= \min_{\tilde{q}\in\mathbb{R}} \left\{ S_{\alpha,v}(\tilde{q}) + \frac{1}{2\gamma}(\tilde{q}-q)^2 \right\} \\ &= \min_{\tilde{q}\in\mathbb{R}} \left\{ (1-\alpha)(v-\tilde{q})\mathbb{1}_{\{v\geq\tilde{q}\}} + \alpha(\tilde{q}-v)\mathbb{1}_{\{v<\tilde{q}\}} + \frac{1}{2\gamma}(\tilde{q}-q)^2 \right\}. \end{aligned}$$

Denote $q_\star = \arg\min_{\tilde{q}\in\mathbb{R}}\{S_{\alpha,v}(\tilde{q}) + \frac{1}{2\gamma}(\tilde{q}-q)^2\}$ which exists and is unique (the function to be minimized is coercive and strongly convex). The stationary condition for the Moreau envelope is given by:

$$0 \in \partial S_{\alpha,v}(q_\star) + \frac{1}{\gamma}(q_\star - q), \text{ with } \quad \partial S_{\alpha,v}(q) = \begin{cases} -(1-\alpha), & q < v \\ [-(1-\alpha), \alpha], & q = v \\ \alpha, & q > v \end{cases}.$$

Considering the 3 different cases, we find that:

$$q_\star = \begin{cases} q + \gamma(1-\alpha), & q < v - \gamma(1-\alpha) \\ v, & q \in [v-\gamma(1-\alpha), v+\gamma\alpha] \\ q - \gamma\alpha, & q > v + \gamma\alpha \end{cases}.$$

We conclude the derivation by using the identity from Moreau envelope: $S^{\gamma}_{\alpha,v}(q) = S_{\alpha,v}\left(q_\star\right) + \frac{1}{2\gamma}\left(q_\star - q\right)^2$ and plugging in $q_\star$. ∎

To simplify the manuscript presentation, we now provide the definition of the local loss function $\mathrm{S}^{i,\gamma}_{\alpha}\colon q \in \mathbb{R} \mapsto \mathbb{E}_{V\sim\widehat{\mu}^i_y}[S^{\gamma}_{\alpha,V}(q)] \in \mathbb{R}_+$. Recall the weights $\widehat{p}^{\star}_{y,\mathbf{y}}$ are given in (5.8), and also that

$$\lambda^i_{\mathbf{y}} = \frac{N^i}{N}\widehat{p}^{\star}_{\mathbf{y},\mathbf{y}} + \sum_{k=1}^{N^i} \widehat{p}^{\star}_{Y^i_k,\mathbf{y}}.$$

Therefore, for $q \in \mathbb{R}$, we have

$$\mathrm{S}^{i,\gamma}_{\alpha}(q) = \frac{N^i \widehat{p}^{\star}_{\mathbf{y},\mathbf{y}}}{\lambda^i_{\mathbf{y}} N} S^{\gamma}_{\alpha,1}(q) + \sum_{k=1}^{N^i} \frac{\widehat{p}^{\star}_{Y^i_k,\mathbf{y}}}{\lambda^i_{\mathbf{y}}} S^{\gamma}_{\alpha,V^i_k}(q). \tag{5.22}$$

5.A.2 Moreau's approximation error

In this section, we consider fixed parameters $\alpha \in (0,1)$, $\gamma > 0$, $\{v_k\}_{k\in[N]}$ and $\{p_k\}_{k\in[N]} \in [0,1]^N$ satisfying $\sum_k p_k = 1$. We define $F := \sum_{k=1}^N p_k S_{\alpha,v_k}$ and $F_\gamma := \sum_{k=1}^N p_k S^{\gamma}_{\alpha,v_k}$, where S_{α,v_k} and S^{γ}_{α,v_k} are the pinball loss and its Moreau envelope defined for $v, q \in \mathbb{R}$ by (5.21)-(5.20). Without loss of generality, it is assumed that $\{v_k\}_{k\in[N]}$ is increasing since we can re-index $\{v_k\}_{k\in[N]}$ and if there exist $(j,j') \in [N]^2$ such that $j \neq j'$ and $v_j = v_{j'}$, we have $p_j S_{\alpha,v_j} + p_{j'} S_{\alpha,v_{j'}} = (p_j + p'_j) S_{\alpha,v_j}$. Finally, for $k \in [N]$, denote

$$I_k = \Big[v_k - \gamma(1-\alpha), v_k + \gamma\alpha\Big]. \tag{5.23}$$

Lemma 5.15. *If $(1-\alpha) \notin \{\sum_{l=1}^k p_l\}_{k\in[n]}$, then F admits a unique minimizer. Moreover, this minimizer belongs to $\{v_k\}_{k\in[n]}$ and we denote $k_\star \in [n]$ its index, i.e., $v_{k_\star} = \arg\min F$. In addition, F is decreasing on $(-\infty, v_{k_\star}]$ and increasing on $[v_{k_\star}, \infty)$. The function F_γ also admits a unique minimizer denoted $Q^{\gamma}_{1-\alpha} \in \mathbb{R}$, and F_γ is decreasing on $(-\infty, Q^{\gamma}_{1-\alpha}]$ and increasing on $[Q^{\gamma}_{1-\alpha}, \infty)$.*

Proof Note that F is differentiable on $\mathbb{R} \setminus \{v_k\}_{k\in[N]}$, and for all $q \in \mathbb{R} \setminus \{v_k\}_{k\in[N]}$, we have

$$F'(q) = \alpha \sum_{k\colon v_k<q} p_k - (1-\alpha) \sum_{k\colon v_k\geq q} p_k = \sum_{k\colon v_k<q} p_k - (1-\alpha).$$

Since $\alpha \in (0,1)$ with $(1-\alpha) \notin \{\sum_{l=1}^k p_l\}_{k\in[N]}$, we deduce that there exists a unique $v_{k_\star} \in \{v_k\}_{k\in[N]}$ such that, for $q \in \mathbb{R}$,

$$\sum_{k\colon v_k<q} p_k - (1-\alpha) \quad \text{is} \quad \begin{cases} < 0 & \text{if } q < v_{k_\star}, \\ > 0 & \text{if } q > v_{k_\star}. \end{cases}$$

Thus, from the continuity of F it follows that F is decreasing on $(-\infty, v_{k_\star}]$ and increasing on $[v_{k_\star}, \infty)$. Moreover, since $F'_\gamma = F'$ on $\mathbb{R}\setminus\{\cup_{k\in[N]} I_k\}$, its minimizer lies in $\cup_{k\in[N]} I_k$, where I_k is defined in (5.23). Finally, the strong convexity of F_γ on $\cup_{k\in[N]} I_k$ shows the uniqueness of $Q^{\gamma}_{1-\alpha} = \arg\min F_\gamma$ and also that F_γ is decreasing on $(-\infty, Q^{\gamma}_{1-\alpha}]$ and increasing on $[Q^{\gamma}_{1-\alpha}, \infty)$. ■

Denote by ∂F subgradient of F. If F is differentiable at $q \in \mathbb{R}$, then $\partial F(q) = \{F'(q)\}$. When $\partial F(q)$ is a singleton, by an abuse of notation, we use the same notation for the set and the unique element it contains.

Theorem 5.16. *Assume* $(1-\alpha) \notin \{\sum_{l=1}^{k} p_l\}_{k\in[N]}$, *then the unique minimizers* $(v_{k_\star}, Q^{\gamma}_{1-\alpha})$ *resp. of* (F, F_γ) *resp. satisfy* $|Q^{\gamma}_{1-\alpha} - v_{k_\star}| \leq \gamma$.

Proof First, for any $k \in [N]$ such that $v_{k_\star} - \gamma(1-\alpha) \in I_k$, we obtain

$$\frac{v_{k_\star} - v_k - \gamma(1-\alpha)}{\gamma} \leq \begin{cases} -(1-\alpha) & \text{if } v_k \geq v_{k_\star}, \\ \alpha & \text{else } v_k < v_{k_\star}. \end{cases} \tag{5.24}$$

Since Lemma 5.15 shows that F is decreasing on $(-\infty, v_{k_\star}]$ and increasing on $[v_{k_\star}, \infty)$ where $v_{k_\star}$ is the unique minimizer of $F(v_{k_\star})$, the convexity of F implies that:

$$\partial F(q) \subset \begin{cases} (-\infty, 0) & \text{if } q < v_{k_\star}, \\ (0, \infty) & \text{if } q > v_{k_\star}. \end{cases} \tag{5.25}$$

Thus, (5.24) combined with (5.25) give that

$$\begin{aligned} &F'_\gamma\Big(v_{k_\star} - \gamma(1-\alpha)\Big) \\ &= \sum_{k\colon v_{k_\star} - \gamma(1-\alpha) \notin I_k} p_k S'_{\alpha, v_k}\Big(v_{k_\star} - \gamma(1-\alpha)\Big) + \sum_{k\colon v_{k_\star} - \gamma(1-\alpha) \in I_k} p_k \frac{\Big(v_{k_\star} - v_k - \gamma(1-\alpha)\Big)}{\gamma} \\ &\leq \alpha \sum_{k\colon v_k < v_{k_\star}} p_k - (1-\alpha) \sum_{k\colon v_k \geq v_{k_\star}} p_k = \partial F(v_{k_\star} - \epsilon) < 0, \end{aligned}$$

where $\epsilon = 2^{-1} \min_{k=1}^{N-1}\{v_{k+1} - v_k\}$. A similar reasoning shows that $F'_\gamma(v_{k_\star} + \gamma\alpha) \geq \partial F(v_{k_\star} + \epsilon) > 0$. Since F_γ is decreasing on $(-\infty, Q^{\gamma}_{1-\alpha}]$ and increasing on $[Q^{\gamma}_{1-\alpha}, \infty)$ by Lemma 5.15, we have $F'_\gamma < 0$ on $(-\infty, v_{k_\star} - \gamma(1-\alpha)]$ and $F'_\gamma > 0$ on $[v_{k_\star} - \gamma\alpha, \infty)$. Therefore, we deduce that $Q^{\gamma}_{1-\alpha} \in I_{k_\star}$. Using that the interval $I_{k_\star}$ is of length γ, this implies that $|Q^{\gamma}_{1-\alpha} - v_{k_\star}| \leq \gamma$. ■

5.B FL convergence guarantee: proof of Theorem 5.10

In this section, we suppose that $\{S_t\}_{t\in[T]}$ is a sequence of i.i.d. random variables, such that, for any $(i, i') \in [n]^2$, $i \in S_t$ and $i' \in S_t$ are independent if $i \neq i'$. For any $i \in [n]$, let $\{z^i_{t,k}\colon k \in \{0, \ldots, K\}\}_{t=0}^{T}$ be a sequence of i.i.d. standard Gaussian variables. Moreover, consider the local loss function $F^i : \mathbb{R} \to \mathbb{R}$ and denote, for $t \in \{0, \ldots, T\}, k \in \{0, \ldots, K\}$

$$F = \sum_{i=1}^{n} \lambda^i_{\mathbf{y}} F^i, \qquad\qquad g^i_{t,k} = \nabla F^i(q^i_{t,k}) + z^i_{t,k}.$$

In this section, we establish the convergence of the iterates given by Theorem 5.24 under the following assumptions:

Assumption 5.17. *The function* $\sum_{i=1}^{n} \lambda^i_{\mathbf{y}} F^i$ *admits at least a minimizer in* $\mathbb{R}$, *we denote* $q_\star$ *one of them, i.e.,* $q_\star \in \arg\min\{\sum_{i=1}^{n} \lambda^i_{\mathbf{y}} F^i\}$.

Assumption 5.18. *For any $i \in [n]$, F^i is continuously differentiable and convex, i.e., for any $q, \tilde{q} \in \mathbb{R}$,*

$$F^i(\tilde{q}) \leq F^i(q) + \langle \nabla F^i(q), \tilde{q} - q \rangle.$$

Assumption 5.19. *For any $i \in [n], t \in \{0, \dots, T\}, K \in \{0, \dots K\}$, ∇F^i is continuously differentiable. In addition, there exist $H^i \geq 0$ such that the function ∇F^i is H^i-smooth, i.e., for any $q, \tilde{q} \in \mathbb{R}$,*

$$\nabla F^i(\tilde{q}) \leq \nabla F^i(q) + \left\langle \nabla F^i(q), \tilde{q} - q \right\rangle + (H^i/2) \left\| \tilde{q} - q \right\|^2.$$

Moreover, denote $H = \max_{i \in [n]} \{H^i\}$.

We introduce the key assumptions appearing in the theoretical derivations below.

Assumption 5.20. *For any $i \in [n]$, the variance of the gradients is uniformly bounded, for all $q \in \mathbb{R}$, we have*

$$\mathbb{E}\left[\left\| \frac{\mathbb{1}_{i \in S_t}}{\mathbb{P}(i \in S_t)} \nabla F^i(q) - \nabla F^i(q) \right\|^2 \right] \leq \xi^2,$$

$$\mathbb{E}\left[\left\| \sum_{i \in S_t} \frac{\lambda_{\mathbf{y}}^i}{\mathbb{P}(i \in S_t)} \nabla F^i(q_\star) - \nabla F(q_\star) \right\|^2 \right] \leq \xi_\star^2.$$

Assumption 5.21. *The heterogeneity denoted ζ is bounded everywhere*

$$\max_{i \in [n]} \left\{ \|\nabla F^i - \nabla F\|_\infty \right\} \leq \zeta^2.$$

We prove Theorem 5.24 using Lemma 5.22 and Lemma 5.23. Note that these results are close to Woodworth et al. (2020, Appendix C). However, we treat partial participation, i.e., $S_t \subseteq [n]$ and consider an objective function defined by importance weights $\{\lambda_{\mathbf{y}}^i\}_{i \in [n]}$. At time $t \in \{0, \dots, T\}$, denote

$$\bar{q}_{t,k} = \sum_{i=1}^n \lambda_{\mathbf{y}}^i (\mathbb{1}_{S_{t+1}}(i) / \mathbb{P}(i \in S_{t+1})) q_{t,k}^i$$

the average of the local parameters defined in Algorithm 5.10. Finally, we introduce the following step-size:

$$\eta_0 = \frac{1}{10} \min \left(\frac{1}{H}, \min_{i=1}^n \left\{ \frac{\mathbb{P}(i \in S_t)}{\lambda_{\mathbf{y}}^i H^i} \right\} \right). \tag{5.26}$$

Lemma 5.22. *Assume Assumption 5.17-Assumption 5.18-Assumption 5.19-Assumption 5.20 and consider $\eta \in (0, \eta_0]$. Then, for any $t \in \{0, \dots, T-1\}$, $k \in \{0, \dots, K-1\}$, we have*

$$\mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F(q_\star) \right] \leq \frac{1}{\eta} \mathbb{E} \left\| \bar{q}_{t,k} - q_\star \right\|^2 - \frac{1}{\eta} \mathbb{E} \left\| \bar{q}_{t,k+1} - q_\star \right\|^2 + 2H \sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E} \left\| \bar{q}_{t,k} - q_{t,k}^i \right\|^2 + 3\eta \xi_\star^2 + \eta^2 \sigma^2 \sum_{i=1}^n \frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}.$$

Proof Developing the squared norm, we find

$$\mathbb{E}\left\|\bar{q}_{t,k+1}-q_\star\right\|^2=\mathbb{E}\left\|\bar{q}_{t,k}-\eta\sum_{i=1}^n\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right)-q_\star\right\|^2 + \eta^2\mathbb{E}\left\|\sum_{i=1}^n\lambda_{\mathbf{y}}^i\left\{\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i\in S_{t+1})}g_{t,k}^i-\nabla F^i\left(q_{t,k}^i\right)\right\}\right\|^2. \quad (5.27)$$

We start by upper bounding the first term, we have

$$\mathbb{E}\left\|\bar{q}_{t,k}-\eta\sum_{i=1}^n\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right)-q_\star\right\|^2=\mathbb{E}\left\|\bar{q}_{t,k}-q_\star\right\|^2 - 2\eta\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left\langle\bar{q}_{t,k}-q_\star,\nabla F^i\left(q_{t,k}^i\right)\right\rangle+\eta^2\mathbb{E}\left\|\sum_{i=1}^n\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right)\right\|^2. \quad (5.28)$$

Using Assumption 5.19, we know that F^i is H^i-smooth and thus $F=\sum_{i=1}^n\lambda_{\mathbf{y}}^iF^i$ is $\bar{H}$-smooth, where $\bar{H}=\sum_{i=1}^n\lambda_{\mathbf{y}}^iH^i$. Following (Nesterov, 2003), the smoothness and convexity of F imply that

$$\left\|\nabla F\left(\bar{q}_{t,k}\right)-\nabla F(q_\star)\right\|^2\leq 2\bar{H}\left(F\left(\bar{q}_{t,k}\right)-F(q_\star)\right).$$

For any $a,b\in\mathbb{R}$, using that $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$, the last term of (5.28) can be upper bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{i=1}^n\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right)\right\|^2\\
&\leq 2\mathbb{E}\left\|\sum_{i=1}^n\lambda_{\mathbf{y}}^i\left\{\nabla F^i\left(q_{t,k}^i\right)-\nabla F^i\left(\bar{q}_{t,k}\right)\right\}\right\|^2+2\mathbb{E}\left\|\sum_{i=1}^n\lambda_{\mathbf{y}}^i\left\{\nabla F^i\left(\bar{q}_{t,k}\right)-\nabla F^i\left(q_\star\right)\right\}\right\|^2\\
&\leq 2\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left\|\nabla F^i\left(q_{t,k}^i\right)-\nabla F^i\left(\bar{q}_{t,k}\right)\right\|^2+2\mathbb{E}\left\|\nabla F\left(\bar{q}_{t,k}\right)-\nabla F\left(q_\star\right)\right\|^2\\
&\leq 2\sum_{i=1}^n(H^i)^2\lambda_{\mathbf{y}}^i\mathbb{E}\left\|q_{t,k}^i-\bar{q}_{t,k}\right\|^2+4\bar{H}\mathbb{E}\left[F\left(\bar{q}_{t,k}\right)-F\left(q_\star\right)\right]. \quad (5.29)
\end{aligned}$$

Regarding the inner product in (5.28), Assumption 5.18 and Assumption 5.19 show

$$\begin{aligned}
&-\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left\langle\bar{q}_{t,k}-q_\star,\nabla F^i\left(q_{t,k}^i\right)\right\rangle\\
&=-\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left\langle q_{t,k}^i-q_\star,\nabla F^i\left(q_{t,k}^i\right)\right\rangle+\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left\langle q_{t,k}^i-\bar{q}_{t,k},\nabla F^i\left(q_{t,k}^i\right)\right\rangle\\
&\leq-\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left[F^i\left(q_{t,k}^i\right)-F^i\left(q_\star\right)\right]+\sum_{i=1}^n\lambda_{\mathbf{y}}^i\mathbb{E}\left[F^i\left(q_{t,k}^i\right)-F^i\left(\bar{q}_{t,k}\right)+\frac{H^i}{2}\mathbb{E}\left\|q_{t,k}^i-\bar{q}_{t,k}\right\|^2\right]
\end{aligned}$$

$$\leq -\mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F\left(q_\star\right)\right] + \frac{1}{2}\sum_{i=1}^{n} H^i \lambda_{\mathbf{y}}^i \left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2. \tag{5.30}$$

Moreover, recall that the estimator $(\mathbb{1}_{S_{t+1}}(i)/\mathbb{P}(i \in S_{t+1}))\nabla F^i$ is unbiased, i.e.,

$$\mathbb{E}\left[\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i(q_{t,k}^i) - \nabla F^i\left(q_{t,k}^i\right)\right] = 0, \tag{5.31}$$

and we also have

$$\begin{aligned}
\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(q_{t,k}^i\right) - \nabla F^i\left(q_{t,k}^i\right) &= \left\{\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(q_\star\right) - \nabla F^i\left(q_\star\right)\right\} \\
&+ \left\{\frac{\mathbb{1}_{i \in S_{t+1}}}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(\bar{q}_{t,k}\right) - \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(q_\star\right) - \nabla F^i\left(\bar{q}_{t,k}\right) + \nabla F^i\left(q_\star\right)\right\} \\
&+ \left\{\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(q_{t,k}^i\right) - \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(\bar{q}_{t,k}\right) - \nabla F^i\left(q_{t,k}^i\right) + \nabla F^i\left(\bar{q}_{t,k}\right)\right\}.
\end{aligned} \tag{5.32}$$

We now upper bound the second term of (5.27). Since for all random variable X, $E[\|X - EX\|^2] \leq \mathbb{E}\|X\|^2$, combining (5.31) and (5.32) gives

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\left\{\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}g_{t,k}^i - \nabla F^i\left(q_{t,k}^i\right)\right\}\right\|^2 \\
&= \sum_{i=1}^{n}(\lambda_{\mathbf{y}}^i)^2\mathbb{E}\left\|\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(q_{t,k}^i\right) - \nabla F^i\left(q_{t,k}^i\right)\right\|^2 + \sum_{i=1}^{n}(\lambda_{\mathbf{y}}^i)^2\mathbb{E}\left\|\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}z_{t,k}^i\right\|^2 \\
&\leq 3\sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2(1 - \mathbb{P}(i \in S_{t+1}))}{\mathbb{P}(i \in S_{t+1})}\left[\mathbb{E}\left\|\nabla F^i\left(q_{t,k}^i\right) - \nabla F^i\left(\bar{q}_{t,k}\right)\right\|^2 + \mathbb{E}\left\|\nabla F^i\left(\bar{q}_{t,k}\right) - \nabla F^i\left(q_\star\right)\right\|^2\right] \\
&\quad + 3\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}\nabla F^i\left(q_\star\right) - \nabla F\left(q_\star\right)\right\|^2 + \sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}\mathbb{E}\left\|z_{t,k}^i\right\|^2 \\
&\leq 3\sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}\left\{(H^i)^2\mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 + 2H^i\mathbb{E}\left[F^i\left(\bar{q}_{t,k}\right) - F^i\left(q_\star\right)\right]\right\} \\
&\quad + 3\xi_\star^2 + \sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}\sigma^2 \\
&\leq 3\sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}(H^i)^2\mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 + 6\sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}H^i\mathbb{E}\left[F^i\left(\bar{q}_{t,k}\right) - F^i\left(q_\star\right)\right] \\
&\quad + 3\xi_\star^2 + \sigma^2\sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}.
\end{aligned} \tag{5.33}$$

Plugging (5.29)-(5.30)-(5.33) back into (5.27) with $\eta \leq \eta_0$, it holds

$$\begin{aligned}
&\mathbb{E}\left\|\bar{q}_{t,k+1} - q_\star\right\|^2 \leq \mathbb{E}\left\|\bar{q}_{t,k} - q_\star\right\|^2 + \eta \sum_{i=1}^n \left(1 + 2\eta H^i + 3\eta \frac{\lambda_{\mathbf{y}}^i H^i}{\mathbb{P}(i \in S_{t+1})}\right) \lambda_{\mathbf{y}}^i H^i \mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 \\
&+ \eta \left(4\eta \bar{H} + 6\eta \max_{i=1}^n \left\{\frac{(\lambda_{\mathbf{y}}^i)^2 H^i}{\mathbb{P}(i \in S_{t+1})}\right\} - 2\right) \mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F\left(q_\star\right)\right] \\
&\quad + 3\eta^2 \xi_\star^2 + \eta^2 \sigma^2 \sum_{i=1}^n \frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})} \\
&\leq \mathbb{E}\left\|\bar{q}_{t,k} - q_\star\right\|^2 + 2H\eta \sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 - \eta \mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F\left(q_\star\right)\right] \\
&\quad + 3\eta^2 \xi_\star^2 + \eta^2 \sigma^2 \sum_{i=1}^n \frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}.
\end{aligned}$$

■

Lemma 5.23. *Assume Assumption 5.17-Assumption 5.18-Assumption 5.19-Assumption 5.20-Assumption 5.21, and for all $t \in [T]$, suppose that $S_t = [n]$. We consider $\eta \in (0, 2/\sum_{i=1}^n \lambda_{\mathbf{y}}^i H^i]$. Then, for any $t \in \{0, \ldots, T-1\}$, $k \in \{0, \ldots, K-1\}$, we have*

$$\sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 \leq 6K\eta^2 \left(\sigma^2 + \xi^2 + K\zeta^2\right).$$

Proof Let $\epsilon > 0$, for any $i, i' \in [n]$ and any $k \in [K]$,

$$\begin{aligned}
&\mathbb{E}\left\|q_{t,k}^i - q_{t,k}^{i'}\right\|^2 - 2\eta^2(\sigma^2 + \xi^2) = \mathbb{E}\left\|q_{t,k-1}^i - q_{t,k-1}^{i'} - \eta\left(g_{t,k-1}^i - g_{t,k-1}^{i'}\right)\right\|^2 - 2\eta^2(\sigma^2 + \xi^2) \\
&= \mathbb{E}\left\|q_{t,k-1}^i - q_{t,k-1}^{i'} - \eta\left(\nabla F^i(q_{t,k-1}^i) - \nabla F^{i'}(q_{t,k-1}^{i'})\right)\right\|^2 \\
&\quad + \eta^2 \mathbb{E}\left\|\left(\nabla F^i(q_{t,k-1}^i) - \nabla F^{i'}(q_{t,k-1}^{i'})\right) - \left(g_{t,k-1}^i - g_{t,k-1}^{i'}\right)\right\|^2 - 2\eta^2(\sigma^2 + \xi^2) \\
&\leq \mathbb{E}\left\|q_{t-1}^i - q_{t-1}^{i'} - \eta\left(\nabla F(q_{t-1}^i) - \nabla F(q_{t-1}^{i'})\right)\right. \\
&\quad \left. + \eta\left(\nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i) - \nabla F(q_{t-1}^{i'}) + \nabla F^{i'}(q_{t-1}^{i'})\right)\right\|^2 \\
&\leq \left(1 + \frac{1}{\epsilon}\right) \mathbb{E}\left\|q_{t-1}^i - q_{t-1}^{i'} - \eta\left(\nabla F(q_{t-1}^i) - \nabla F(q_{t-1}^{i'})\right)\right\|^2 \\
&\quad + (1+\epsilon)\eta^2 \mathbb{E}\left\|\nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i) - \nabla F(q_{t-1}^{i'}) + \nabla F^{i'}(q_{t-1}^{i'})\right\|^2 \\
&\leq \left(1 + \frac{1}{\epsilon}\right) \mathbb{E}\left\|q_{t-1}^i - q_{t-1}^{i'}\right\|^2 + (1+\epsilon)\eta^2 \mathbb{E}\left\|\nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i)\right\|^2 \\
&\quad + (1+\epsilon)\eta^2 \mathbb{E}\left\|\nabla F(q_{t-1}^{i'}) - \nabla F^{i'}(q_{t-1}^{i'})\right\|^2
\end{aligned}$$

$$- 2(1+\epsilon)\eta^2\mathbb{E}\left\langle \nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i), \nabla F(q_{t-1}^{i'}) - \nabla F^{i'}(q_{t-1}^{i'})\right\rangle. \tag{5.34}$$

The third inequality is implied by the co-coercivity: $\eta \in (0, 2/\sum_{i=1}^n \lambda_{\mathbf{y}}^i H^i]$, $\forall (q, \tilde{q}) \in \mathbb{R}^2$,

$$\begin{aligned}
&\left\|q - \tilde{q} - \eta\left(\nabla F(q) - \nabla F(\tilde{q})\right)\right\|^2 \\
&= \left\|\tilde{q} - q\right\|^2 - \eta\left[2\left\langle q - \tilde{q}, \nabla F(q) - \nabla F(\tilde{q})\right\rangle + \eta\left\|\nabla F(q) - \nabla F(\tilde{q})\right\|^2\right] \leq \left\|\tilde{q} - q\right\|^2,
\end{aligned}$$

and we also have

$$\begin{aligned}
&\sum_{i=1}^n\sum_{i'=1}^n \lambda_{\mathbf{y}}^i\lambda_{\mathbf{y}}^{i'}\mathbb{E}\left\langle \nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i), \nabla F(q_{t-1}^{i'}) - \nabla F^{i'}(q_{t-1}^{i'})\right\rangle \\
&\qquad = \mathbb{E}\left\|\sum_{i=1}^n \lambda_{\mathbf{y}}^i\left(\nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i)\right)\right\|^2 \geq 0.
\end{aligned}$$

Therefore, summing (5.34) gives that

$$\begin{aligned}
\sum_{i=1}^n\sum_{i'=1}^n \lambda_{\mathbf{y}}^i\lambda_{\mathbf{y}}^{i'}\mathbb{E}\left\|q_{t,k}^i - q_{t,k}^{i'}\right\|^2 \leq \left(1 + \frac{1}{\epsilon}\right)\sum_{i=1}^n\sum_{i'=1}^n \lambda_{\mathbf{y}}^i\lambda_{\mathbf{y}}^{i'}\mathbb{E}\left\|q_{t-1}^i - q_{t-1}^n\right\|^2 \\
+ 2\eta^2\left(\sigma^2 + \xi^2 + (1+\epsilon)\zeta^2\right).
\end{aligned}$$

Set $\epsilon = K - 1$, since for any $i, i' \in [n]$, $x_{t,0}^i = x_{t,0}^{i'}$, we get

$$\begin{aligned}
\sum_{i=1}^n\sum_{i'=1}^n \lambda_{\mathbf{y}}^i\lambda_{\mathbf{y}}^{i'}\mathbb{E}\left\|q_{t,k}^i - q_{t,k}^{i'}\right\|^2 &\leq 2\eta^2\left(\sigma^2 + \xi^2 + (1+\epsilon)\zeta^2\right)\sum_{k'=0}^{K-1}\left(1 + \frac{1}{\epsilon}\right)^{k'} \\
&\leq 6K\eta^2\left(\sigma^2 + \xi^2 + (1+\epsilon)\zeta^2\right).
\end{aligned}$$

Since $\sum_{i=1}^n \lambda_{\mathbf{y}}^i = 1$, the Jensen's inequality yields that

$$\sum_{i=1}^n \lambda_{\mathbf{y}}^i\mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 \leq \sum_{i=1}^n\sum_{i'=1}^n \lambda_{\mathbf{y}}^i\lambda_{\mathbf{y}}^{i'}\mathbb{E}\left\|q_{t,k}^i - q_{t,k}^{i'}\right\|^2,$$

which concludes the proof. ■

In addition, with the previous notations consider the step-size

$$\eta_\star = \min\left\{\eta_0, \left(\frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{14HK^2T\left[\sigma^2 + K\zeta^2\right]}\right)^{1/3}\right\}, \tag{5.35}$$

and define the average parameter

$$\hat{q}_T = \frac{1}{T}\sum_{t=0}^{T-1}\left\{\sum_{i=1}^n \lambda_{\mathbf{y}}^i\left[\frac{1}{K}\sum_{k=0}^{K-1} q_{t,k}^i\right]\right\}. \tag{5.36}$$

Theorem 5.24. *Assume Assumption 5.17-Assumption 5.18-Assumption 5.19-Assumption 5.20-Assumption 5.21. We consider $\eta \in (0, \eta_0]$ with $S_t = [n]$, for all $t \in [T]$. Then, for any $t \in \{0, \dots, T-1\}$, $k \in \{0, \dots, K-1\}$, we have*

$$\mathbb{E}F\left(\hat{q}_T\right) - F(q_\star) \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta KT} + 2\eta^2\left[6H(K\zeta)^2 + \sigma^2\left(6HK + \max_{i\in[n]}\lambda_{\mathbf{y}}^i\right)\right],$$

where $\eta_0, \hat{q}_T$ are given in (5.26) and (5.36). Moreover, for $\eta = \eta_\star$ defined in (5.35) and $H \geq K^{-1}\max_{i\in[n]}\lambda_{\mathbf{y}}^i$, it follows

$$\mathbb{E}F\left(\hat{q}_T\right) - F(q_\star) \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta_0 KT} + \frac{5(\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2)^{2/3}\left[H\left(\sigma^2 + K\zeta^2\right)\right]^{1/3}}{(KT^2)^{1/3}}.$$

Proof For any $\eta \leq \eta_0$, using Lemma 5.22 we have

$$\mathbb{E}\left[F\left(\bar{q}_{t,k}\right)\right] - F(q_\star) \leq \frac{1}{\eta}\mathbb{E}\left\|\bar{q}_{t,k} - q_\star\right\|^2 - \frac{1}{\eta}\mathbb{E}\left\|\bar{q}_{t,k+1} - q_\star\right\|^2 + 2H\sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E}\left\|\bar{q}_{t,k} - q_{t,k}^i\right\|^2 + 2\eta^2\sigma^2\max_{i\in[n]}\{\lambda_{\mathbf{y}}^i\}. \quad (5.37)$$

Moreover, by Lemma 5.23 it follows that

$$\sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E}\left\|q_{t,k}^i - \bar{q}_{t,k}\right\|^2 \leq 6K\eta^2(\sigma^2 + K\zeta^2). \quad (5.38)$$

Combining (5.37) and (5.38), we obtain

$$\mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F(q_\star)\right] \leq \frac{1}{\eta}\mathbb{E}\left\|\bar{q}_{t,k} - q_\star\right\|^2 - \frac{1}{\eta}\mathbb{E}\left\|\bar{q}_{t,k+1} - q_\star\right\|^2 + 12H(K\eta\zeta)^2 + 2(\eta\sigma)^2\left(6HK + \max_{i\in[n]}\{\lambda_{\mathbf{y}}^i\}\right).$$

Moreover, telescoping proves that

$$\sum_{t=0}^{T-1}\sum_{k=0}^{K-1}\left[\mathbb{E}\left\|\bar{q}_{t,k} - q_\star\right\|^2 - \mathbb{E}\left\|\bar{q}_{t,k+1} - q_\star\right\|^2\right] \leq \mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2.$$

Therefore, the convexity Assumption 5.18 gives that

$$\mathbb{E}\left[F\left(\frac{1}{KT}\sum_{t=1}^{KT}\bar{q}_{t,k}\right) - F(q_\star)\right] \leq \frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=0}^{K-1}\mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F(q_\star)\right] \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta KT} + 2\eta^2\left[6H(K\zeta)^2 + \sigma^2\left(6HK + \max_{i\in[n]}\{\lambda_{\mathbf{y}}^i\}\right)\right].$$

Finally, the choice of η provided in (5.35) ensures that

$$\mathbb{E}\left[F\left(\hat{q}_T\right)\right] - F(q_\star) \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta_0 KT} + \frac{5(\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2)^{2/3}\left[H\left(\sigma^2 + K\zeta^2\right)\right]^{1/3}}{(KT^2)^{1/3}}.$$

■

Now, we denote $\alpha \in (0,1)$ the confidence level, and consider the functions defined for $t \in \{0,\dots,T\}, k \in \{0,\dots,K\}$ by

$$F = S^\gamma_{\alpha,\widehat{\mu}_{\mathbf{y}}}, \qquad F^i = S^\gamma_{\alpha,\widehat{\mu}^i_{\mathbf{y}}}.$$

Denote by $q_\star$ the minimizer of $S^\gamma_{\alpha,\widehat{\mu}_{\mathbf{y}}} = \sum_{i=1}^n \lambda^i_{\mathbf{y}} S^\gamma_{\alpha,\widehat{\mu}^i_{\mathbf{y}}}$, which always exists. In addition, note that Assumption 5.19 and Assumption 5.21 are satisfied with:

$$H^i = \frac{1}{\gamma}, \qquad \zeta = \max_{i\in[n]} \|\nabla \mathrm{S}^{i,\gamma}_\alpha - \nabla \mathrm{S}^\gamma_\alpha\|^{1/2}_\infty. \tag{5.39}$$

Corollary 5.25. *Let $\gamma \in (0, (\max_{i\in[n]} \lambda^i_{\mathbf{y}})^{-1}K]$ and consider the step-size $\eta_0 = \gamma/10$. Then, for any $t \in \{0,\dots,T-1\}$, $k \in \{0,\dots,K-1\}$, we have*

$$\mathbb{E}S^\gamma_{\alpha,\widehat{\mu}_{\mathbf{y}}}\left(\hat{q}_T\right) - S^\gamma_{\alpha,\widehat{\mu}_{\mathbf{y}}}(q_\star) \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta_0 KT} + \frac{5(\sigma^2 + K\zeta^2)^{1/3}(\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2)^{2/3}}{(\gamma KT^2)^{1/3}},$$

where $\hat{q}_T$ is provided in (5.36).

Proof

Since Assumption 5.17-Assumption 5.18-Assumption 5.19-Assumption 5.20-Assumption 5.21 are satisfied with $\{H^i\}_{i\in[n]}, \zeta$ provided in (5.39), applying Theorem 5.24 concludes the proof. ■

5.C Theoretical Coverage Guarantee

5.C.1 General coverage guarantee

Consider an increasing sequence $\{v_k\}_{k\in[N+1]} \in (\mathbb{R}\cup\{+\infty\})^{N+1}$ and $\{p_k\}_{k\in[N+1]} \in \Delta_{N+1}$. For any $\alpha \in [0,1]$, recall that

$$Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k\delta_{v_k}\right) = \inf\left\{t\in[-\infty,\infty]\colon\ \mathbb{P}\left(V\leq t\right)\geq 1-\alpha,\ \text{where } V\sim\sum_{k=1}^{N+1} p_k\delta_{v_k}\right\}.$$

Lemma 5.26. *Let $\{v_\ell\}_{\ell\in[N+1]}$ be an increasing sequence and $\{p_\ell\}_{\ell\in[N+1]} \in \Delta_{N+1}$. If $V \sim \sum_{l=1}^{N+1} p_l\delta_{v_l}$, then, for all $\alpha \in [0,1)$, we have*

$$1-\alpha \leq \mathbb{P}\left(V \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k\delta_{v_k}\right)\right) < 1-\alpha+\max_{k=1}^{N+1}\left\{p_k\right\}.$$

Proof Fix $\alpha \in [0,1)$, and by convention set $\sum_{k=1}^{0} p_k = 0$. There exists $k \in [N+1]$, such that $1-\alpha \in (\sum_{l=1}^{k-1} p_l, \sum_{l=1}^{k} p_l]$, hence $Q_{1-\alpha}(\sum_{l=1}^{N+1} p_l \delta_{v_l}) = v_k$. This last identity implies that

$$1-\alpha \leq \mathbb{P}\left(V \leq Q_{1-\alpha}\left(\sum_{l=1}^{N+1} p_l \delta_{v_l}\right)\right) = \sum_{l=1}^{k} p_l < 1-\alpha + \max_{k=1}^{N+1}\left\{p_k\right\}.$$

■

Denote $\{(X_k, Y_k)\}_{k\in[N+1]}$ a set of pairwise independent random variables and for $k \in [N+1]$, denote $Z_k = (X_k, Y_k)$, $V_k = V(X_k, Y_k)$. Let $\mathfrak{S}_{N+1}$ be the set of all permutations of $[N+1]$ and consider $\mathfrak{S}_{N+1}^k = \{\sigma \in \mathfrak{S}_{N+1} \colon \sigma(N+1) = k\}$. Moreover, write f the joint density of $\{Z_k\}_{k\in[N+1]}$, and for all $k \in [N+1]$ define

$$p_k^{z_{1:N+1}} = \begin{cases} \frac{1}{N+1} & \text{if } \sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) = 0 \\ \frac{\sum_{\sigma\in\mathfrak{S}_{N+1}^k} f(z_{\sigma(1)},\ldots,z_{\sigma(N+1)})}{\sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)},\ldots,z_{\sigma(N+1)})} & \text{otherwise} \end{cases}. \quad (5.40)$$

Lemma 5.27. *For any $\alpha \in [0,1)$, we have*

$$\begin{aligned} &\int \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1} \\ &= \int \left[\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)}\right] \left[\sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}. \end{aligned}$$

Proof First, let's show the invariance of $\sigma \in \mathfrak{S}_{N+1} \mapsto Q_{1-\alpha}(\sum_{k=1}^{N+1} p_{\sigma(k)}^{z_{\sigma(1):\sigma(N+1)}} \delta_{v_{\sigma(k)}}) \in \mathbb{R}$. For that, fix $\tilde{\sigma} \in \mathfrak{S}_{N+1}$. The invariance is immediate when $\sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) = 0$. Therefore, assume that $\sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \neq 0$. We get

$$\begin{aligned} \sum_{k=1}^{N+1} p_{\tilde{\sigma}(k)}^{z_{\tilde{\sigma}(1):\tilde{\sigma}(N+1)}} \delta_{v_{\tilde{\sigma}(k)}} &= \sum_{k=1}^{N+1} \frac{\sum_{\sigma\in\mathfrak{S}_{N+1}^{\tilde{\sigma}(k)}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})}{\sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})} \delta_{v_{\tilde{\sigma}(k)}} \\ &= \sum_{k=1}^{N+1} \frac{\sum_{\sigma\in\mathfrak{S}_{N+1}^{k}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})}{\sum_{\sigma\in\mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})} \delta_{v_k} = \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}. \end{aligned}$$

Moreover, we can write

$$\begin{aligned} &\int \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1} \\ &= \sum_{\sigma\in\mathfrak{S}_{N+1}} \int \mathbb{1}_{v_{\sigma(N+1)} \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\sigma(\bar{k})}^{z_{\sigma(1):\sigma(N+1)}} \delta_{v_{\sigma(\bar{k})}}\right)} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \frac{\mathrm{d}z_{\sigma(1)} \cdots \mathrm{d}z_{\sigma(N+1)}}{(N+1)!} \\ &= \sum_{k=1}^{N+1} \sum_{\sigma\in\mathfrak{S}_{N+1}^k} \int \mathbb{1}_{v_{\sigma(N+1)} \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\sigma(\bar{k})}^{z_{\sigma(1):\sigma(N+1)}} \delta_{v_{\sigma(\bar{k})}}\right)} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \frac{\mathrm{d}z_{\sigma(1)} \cdots \mathrm{d}z_{\sigma(N+1)}}{(N+1)!} \\ &= \sum_{k=1}^{N+1} \int \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} \left[\sum_{\sigma\in\mathfrak{S}_{N+1}^k} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!} \end{aligned}$$

$$= \sum_{k=1}^{N+1} \int \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} p_k^{z_{1:N+1}} \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}.$$

■

Given $\mathrm{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ define

$$D_N^{\mathrm{z}} = \left(z_1, \ldots, z_N, \mathrm{z}\right), \qquad \mu_{\mathbf{y}}^N = p_{N+1}^{D_N^{\mathrm{z}}} \delta_1 + \sum_{k=1}^{N} p_k^{D_N^{\mathrm{z}}} \delta_{V_k},$$

and consider the prediction set given by

$$\mathcal{C}_{\alpha,\mu^N}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\mu_{\mathbf{y}}^N) \right\}.$$

Theorem 5.28. *Assume there are no ties between $\{V_k\}_{k \in [N+1]}$ almost surely. Then, for any $\alpha \in [0, 1)$, we have*

$$1 - \alpha \leq \mathbb{P}\left(Y_{N+1} \in \mathcal{C}_{\alpha,\mu^N}(X_{N+1})\right) \leq 1 - \alpha + \mathbb{E}\left[\max_{k=1}^{N+1} \{p_k^{Z_{1:N+1}}\}\right],$$

where $p_k^{Z_{1:N+1}}$ is defined in (5.40).

Proof Let α be in $[0, 1)$ and for any $(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}$, denote $z_k = (x_k, y_k)$, $v_k = V(x_k, y_k)$. First, we can write

$$\begin{aligned}
&\mathbb{P}\left(Y_{N+1} \in \mathcal{C}_{\alpha,\mu^N}(X_{N+1})\right) = \mathbb{P}\left(Y_{N+1} \in \left\{\mathbf{y} \in \mathcal{Y} \colon V(X_{N+1}, \mathbf{y}) \leq Q_{1-\alpha}(\mu_{\mathbf{y}}^N)\right\}\right) \\
&= \mathbb{P}\left(V(X_{N+1}, Y_{N+1}) \leq Q_{1-\alpha}(\mu_{Y_{N+1}}^N)\right) \\
&\overset{(\star)}{=} \mathbb{P}\left(V(X_{N+1}, Y_{N+1}) \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N+1} p_k^{Z_{1:N+1}} \delta_{V_k}\right)\right) \\
&= \int_{(\mathcal{X}\times\mathcal{Y})^{N+1}} \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{V_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}.
\end{aligned}$$

Where $(\star)$ holds since

$$\begin{aligned}
&\left(V_{N+1} \leq Q_{1-\alpha}(\mu_{Y_{N+1}}^N)\right) \\
&\iff \left(p_{N+1}^{D_N^{(X_{N+1},Y_{N+1})}} \delta_{1 \leq V_{N+1}} + \textstyle\sum_{k=1}^{N} p_k^{D_N^{(X_{N+1},Y_{N+1})}} \delta_{V_k \leq V_{N+1}} \leq \alpha\right) \\
&\iff \left(p_{N+1}^{D_N^{(X_{N+1},Y_{N+1})}} \delta_{V_{N+1} \leq V_{N+1}} + \textstyle\sum_{k=1}^{N} p_k^{D_N^{(X_{N+1},Y_{N+1})}} \delta_{V_k \leq V_{N+1}} \leq \alpha\right) \\
&\iff \left(V(X_{N+1}, Y_{N+1}) \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N+1} p_k^{Z_{1:N+1}} \delta_{V_k}\right)\right).
\end{aligned}$$

Define the set $E \subset (\mathcal{X} \times \mathcal{Y})^{N+1}$ of points such that the non-conformity scores are pairwise distinct:

$$E = \{(z_1, \ldots, z_{N+1}) \in (\mathcal{X} \times \mathcal{Y})^{N+1} \colon \textstyle\prod_{k<\ell} \left(v(x_k, y_k) - v(x_\ell, y_\ell)\right) \neq 0\},$$

$$E^c = (\mathcal{X} \times \mathcal{Y})^{N+1} \setminus E,$$
$$F = \{(z_1, \dots, z_{N+1}) \in (\mathcal{X} \times \mathcal{Y})^{N+1} \colon \textstyle\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \dots, z_{\sigma(N+1)}) \neq 0\}$$

In addition, combining Lemma 5.27 with the no-tie assumption on $\{V_k\}_{k \in [N+1]}$ gives that

$$\int \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}\right)} f(z_1, \dots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1} \tag{5.41}$$
$$= \int_{E \cap F} \left[\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)}\right] \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \dots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}.$$

Consider the random variable $V \sim \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}$, we have

$$\mathbb{P}\left(V \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)\right) = \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)}.$$

Therefore, applying Lemma 5.26 on $(z_1, \dots, z_{N+1}) \in E \cap F$ implies that

$$1 - \alpha \leq \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} \leq 1 - \alpha + \max_{k=1}^{N+1} \{p_k^{z_{1:N+1}}\}. \tag{5.42}$$

Lastly, using that

$$\int_{E \cap F} \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \dots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}$$
$$= \int_E \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \dots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!} = 1, \tag{5.43}$$

and combining the bounds (5.42)-(5.43) with (5.41) yields the result. ■

5.C.2 Proof of Theorem 5.2

First, recall that

$$\mathcal{I} = \left\{(\star, N^\star + 1)\right\} \cup \left\{(i, k) \colon i \in [n], k \in [N^i]\right\}.$$

For any $\{(\mathrm{x}_k^i, \mathrm{y}_k^i)\}_{(i,k) \in \mathcal{I}} \in (\mathcal{X} \times \mathcal{Y})^{N+1}$, we define the set

$$D_N^{(\mathrm{x}^\star_{N^\star+1}, \mathrm{y}^\star_{N^\star+1})} = \{(\mathrm{x}_k^i, \mathrm{y}_k^i) \colon i \in [n], k \in [N^i]\} \cup \{(\mathrm{x}^\star_{N^\star+1}, \mathrm{y}^\star_{N^\star+1})\}.$$

We consider a bijection (ϕ, φ) between the set $[N+1]$ and $\mathcal{I}$. This bijection is defined for any $k \in [N]$ as follows:

$$(\phi(k), \varphi(k)) = \begin{cases} (j, \ell) & \text{if } 1 \leq \ell := k - \sum_{i=1}^{j-1} N^i \leq \sum_{i=1}^{j} N^i \\ (\star, N^\star + 1) & \text{otherwise} \end{cases}.$$

Recall that $\forall i \in [n]$ and $y \in \mathcal{Y}$, the likelihood ratio is given by

$$w_y^i = \frac{P_Y^i(y)}{P_Y^{\text{cal}}(y)},$$

and for all $(i,k) \in \mathcal{I}$ we write

$$\begin{aligned} \mathfrak{P}_k^i &= \left\{\sigma \in \mathfrak{S}_{N+1} \colon \phi(\sigma(N+1)) = i, \varphi(\sigma(N+1)) = k\right\}, \\ W_k^i(D_N^{(\mathrm{x}_{N^\star+1}^\star, \mathrm{y}_{N^\star+1}^\star)}) &= w_{\mathrm{y}_k^i}^\star \sum_{\sigma \in \mathfrak{P}_k^i} \prod_{\ell=1}^{N} w_{\mathrm{y}_{\varphi(\ell)}^{\phi(\ell)}}^{\phi(\ell)}. \end{aligned} \tag{5.44}$$

Given the set of points $D_N^{(\mathbf{x},\mathbf{y})}$, for all $(i,k) \in \mathcal{I}$ define

$$p_{i,k}^\star = \frac{W_k^i(D_N^{(\mathbf{x},\mathbf{y})})}{\sum_{\ell=1}^{N+1} W_{\varphi(\ell)}^{\phi(\ell)}(D_N^{(\mathbf{x},\mathbf{y})})}. \tag{5.45}$$

Finally, define the probability measure and the prediction set given by

$$\begin{aligned} \mu_{\mathbf{y}}^\star &= p_{\star,N^\star+1}^\star \delta_1 + \sum_{i=1}^{n} \sum_{k=1}^{N^i} p_{i,k}^\star \delta_{V_k^i}, \\ \mathcal{C}_{\alpha,\mu^\star}(\mathbf{x}) &= \left\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x},\mathbf{y}) \le Q_{1-\alpha}(\mu_{\mathbf{y}}^\star)\right\}. \end{aligned}$$

Theorem 5.29. *If Assumption 5.1 holds, then for any $\alpha \in [0,1)$, we have*

$$1 - \alpha \le \mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\mu^\star}(X_{N^\star+1}^\star)\right) \le 1 - \alpha + \mathbb{E}\left[\max_{(i,k)\in\mathcal{I}} \{p_{k,i}^\star\}\right],$$

where $p_{i,k}^\star$ is defined in (5.45).

Proof By independence, the joint density f of $\{(X_\ell^j, Y_\ell^j) \colon (j,\ell) \in \mathcal{I}\}$ with respect to $(P_{X|Y} \times P_Y^{\text{cal}})^{\otimes(N+1)}$ is given for $\{(x_k^i, y_k^i) \colon (i,k) \in \mathcal{I}\} \in (\mathcal{X} \times \mathcal{Y})^{N+1}$ by

$$\begin{aligned} f\left((x_1^1, y_1^1), \ldots, (x_{N^1}^1, y_{N^1}^1), \ldots, (x_1^n, y_1^n), \ldots, (x_{N^n}^n, y_{N^n}^n), (x_{N^\star+1}^\star, y_{N^\star+1}^\star)\right) \\ = w_{y_{N^\star+1}^\star}^\star \prod_{j=1}^{n} \prod_{\ell=1}^{N^j} w_{y_\ell^j}^j. \end{aligned}$$

Using the definition of $p_{i,k}^\star$ (5.45), for all $(i,k) \in \mathcal{I}$ we have

$$\begin{aligned} p_{i,k}^\star &= \frac{W_k^i(D_{N+1})}{\sum_{\ell=1}^{N+1} W_{\varphi(\ell)}^{\phi(\ell)}(D_{N+1})} \\ &= \frac{\sum_{\sigma \in \mathfrak{P}_k^i} f(z_{\varphi\circ\sigma(1)}^{\phi\circ\sigma(1)}, \ldots, z_{\varphi\circ\sigma(N+1)}^{\phi\circ\sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\varphi\circ\sigma(1)}^{\phi\circ\sigma(1)}, \ldots, z_{\varphi\circ\sigma(N+1)}^{\phi\circ\sigma(N+1)})} \\ &= \frac{\sum_{\sigma \in \mathfrak{S}_{N+1} \colon \sigma(N+1) = (\phi,\varphi)^{-1}(i,k)} f(z_{\varphi\circ\sigma(1)}^{\phi\circ\sigma(1)}, \ldots, z_{\varphi\circ\sigma(N+1)}^{\phi\circ\sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\varphi\circ\sigma(1)}^{\phi\circ\sigma(1)}, \ldots, z_{\varphi\circ\sigma(N+1)}^{\phi\circ\sigma(N+1)})}. \end{aligned}$$

Therefore, applying Theorem 5.28 concludes the proof. ■

5.C.3 Proof of Lemma 5.4 and Equation (5.12)

In this section, we consider a probability measure P_Y^{cal} dominating $P_Y^\star$ and suppose that the likelihood ratios are given by $w_y^\star = P_Y^\star(y)/P_Y^{\mathrm{cal}}(y)$. Let $\{\widehat{w}_y^\star\}_{y\in\mathcal{Y}}$ be fixed, and $\forall y \in \mathcal{Y}$, define $\nu_y = [\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})]^{-1} \widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)$. Denote $\widehat{Y}_{N^\star+1}^\star$ a multinomial random variable of parameter ν independent of the calibration dataset $\left\{(X_k^i, Y_k^i) \colon k \in [N^i]\right\}_{i\in[n]}$ and write $\mathcal{P}(\mathcal{Y})$ the partition of the set $\mathcal{Y}$. Moreover, denote $\widehat{P}_Y^\star$ the probability distribution of $\widehat{Y}_{N^\star+1}^\star$, and consider $\widehat{X}_{N^\star+1}^\star$ such that $(\widehat{X}_{N^\star+1}^\star, \widehat{Y}_{N^\star+1}^\star) \sim P_{X|Y} \times \widehat{P}_Y^\star$.

Lemma 5.30. *For any prediction set $\mathcal{C}\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ independent of $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star)$ and $(\widehat{X}_{N^\star+1}^\star, \widehat{Y}_{N^\star+1}^\star)$, we have*

$$\left|\mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star)) - \mathbb{P}(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star))\right| \leq \frac{1}{2}\sum_{y\in\mathcal{Y}} \left| P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}\right|.$$

Proof Developing the left-hand side as follows, we get

$$\begin{aligned}
&\left|\mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star)\right) - \mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star)\right)\right| \\
&= \left|\mathbb{E}\left[\mathbb{1}_{Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star)}\right] - \mathbb{E}\left[\mathbb{1}_{\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star)}\right]\right| \\
&= \left|\sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\left(w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}\right)\int \mathbb{E}\mathbb{1}_{y\in\mathcal{C}(x)} \mathrm{d}P_{X|Y=y}(x)\right| \\
&\leq \frac{1}{2}\sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\left|w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}\right|.
\end{aligned}$$

Finally, using that $P_Y^{\mathrm{cal}}(y) w_y^\star = P_Y^\star(y)$ concludes the proof. ∎

Remark 5.31. *If for some probability atoms $\{m_i\}_{i\in[n]} \in \Delta_n$, we know good approximations $\widehat{w}_y^\star$ of the likelihood ratios $[(P_Y^{\mathrm{cal}})^{-1}(\sum_{i=1}^n m_i P_Y^i)](y)$. Then, Lemma 5.30 implies the following result:*

$$\begin{aligned}
&\left|\mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star)\right) - \mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star)\right)\right| \\
&\qquad \leq \mathrm{d_{TV}}\left(P_Y^\star, \sum_{i=1}^n m_i P_Y^i\right) + \frac{1}{2}\sum_{y\in\mathcal{Y}}\left|\left(\sum_{i=1}^n m_i P_Y^i\right)(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}\right|.
\end{aligned}$$

Lemma 5.32. *If $|\mathcal{Y}| \geq 2$ and $\mathsf{M} \in \mathbb{N}^\star$, then we have*

$$|\mathcal{Y}| \exp\left(-\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}\right) \wedge 1 \leq \sqrt{\frac{2\log|\mathcal{Y}|}{(\log 2)\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}}}.$$

Proof Introduce the set $E = \{2\log|\mathcal{Y}| \leq \mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}\}$, we obtain

$$|\mathcal{Y}|\exp\left(-\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}\right)\wedge 1 \leq \mathbb{1}_E\exp\left(-\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}/2\right) + \mathbb{1}_{E^c}. \tag{5.46}$$

We also have that for all $x \geq 0$, that $\mathrm{e}^{-x} \leq 1/\sqrt{x}$. Using this inequality on the first right-side term implies that

$$\exp\left(-\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}/2\right) \leq \sqrt{\frac{2}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}}} \leq \sqrt{\frac{\log|\mathcal{Y}|}{\log 2}}\sqrt{\frac{2}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}}}.$$

Moreover, remark that

$$\mathbb{1}_{E^c} \leq \mathbb{1}_{E^c}\sqrt{\frac{2\log|\mathcal{Y}|}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}}}.$$

Finally, plugging the two previous inequalities in (5.46) concludes the proof. ■

Recall that M_y^i denotes the number of training data on agent i associated to label $y \in \mathcal{Y}$. Consider the total number of local data $\mathsf{M}^\star = \sum_{y\in\mathcal{Y}}\mathsf{M}_y^i$, the number of training data with label y is given by $\mathsf{M}_y = \sum_{i=1}^n \mathsf{M}_y^i$, and the total number of samples on all agents is written by $\mathsf{M} = \sum_{y\in\mathcal{Y}}\mathsf{M}_y$. Recall that the likelihood ratios and the weights are given for any labels $(y,\mathbf{y}) \in \mathcal{Y}^2$ by

$$\widehat{w}_y^\star = \begin{cases} \frac{\mathsf{M}\mathsf{M}_y^\star}{\mathsf{M}^\star\mathsf{M}_y} & \text{if } \mathsf{M}_y \geq 1 \\ 0 & \text{otherwise} \end{cases}, \qquad \widehat{p}_{y,\mathbf{y}}^\star = \frac{(\mathsf{M}_y^\star/\mathsf{M}_y)\cdot\mathbb{1}_{\mathsf{M}_y\geq 1}}{\mathsf{M}^\star + (\mathsf{M}_{\mathbf{y}}^\star/\mathsf{M}_{\mathbf{y}})\cdot\mathbb{1}_{\mathsf{M}_{\mathbf{y}}\geq 1}}.$$

For any $y \in \mathcal{Y}$, we also consider $\nu \in \Delta_{|\mathcal{Y}|}^{\mathbb{Q}} = \{p' \in \mathbb{Q}_+^{|\mathcal{Y}|} \colon \sum_{y\in\mathcal{Y}} p'_y = 1\}$ defined by

$$\nu_y = \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}.$$

For any parameter $p \in \Delta_{|\mathcal{Y}|}^{\mathbb{Q}}$, denote M_p a multinomial random variable independent of the training/calibration datasets, and define $\widehat{Y}_{N^\star+1}^\star = M_\nu$. For any set A in the partition of $\mathcal{Y}$, we have $M_\nu^{-1}(A) = \cup_{p\in\Delta_{|\mathcal{Y}|}^{\mathbb{Q}}}\{\nu^{-1}(\{p\})\cap M_p^{-1}(A)\}$. Therefore, $\widehat{Y}_{N^\star+1}^\star$ is a valid random variable. Given the target coverage level $1-\alpha$, recall that the prediction set is defined for any $\mathbf{x} \in \mathcal{X}$ by

$$\widehat{\mu}_{\mathbf{y}}^{\text{MLE}} = \widehat{p}_{\mathbf{y},\mathbf{y}}^\star\delta_1 + \textstyle\sum_{i=1}^n\sum_{k=1}^{N^i\wedge\bar{N}^i}\widehat{p}_{Y_k^i,\mathbf{y}}^\star\delta_{V_k^i},,$$
$$\mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(\mathbf{x}) = \left\{\mathbf{y}\colon V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}^{\text{MLE}}\right)\right\}.$$

Since the considered likelihood ratios are now depending on the training dataset, it is no longer possible to apply Lemma 5.30. However, conditioning by the training dataset, a similar reasoning shows that

$$\left|\mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(X_{N^\star+1}^\star)) - \mathbb{P}(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(\widehat{X}_{N^\star+1}^\star))\right|$$
$$\leq \frac{1}{2}\sum_{y\in\mathcal{Y}}\mathbb{E}\left|P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}\right|. \tag{5.47}$$

By utilizing the following lemma combined with (5.47), we can control the difference between the probabilities of the events $Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(X^\star_{N^\star+1})$ and $\widehat{Y}^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(\widehat{X}^\star_{N^\star+1})$.

Theorem 5.33. *For any $\alpha \in (0,1)$, we have*

$$\sum_{y\in\mathcal{Y}} \mathbb{E}\left| P^\star_Y(y) - \frac{\widehat{w}^\star_y P^{\text{cal}}_Y(y)}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}^\star_{\tilde{y}} P^{\text{cal}}_Y(\tilde{y})}\right| \leq \frac{6}{\sqrt{\mathsf{M}^\star}} + 12\sqrt{\frac{\log|\mathcal{Y}| + \log \mathsf{M}^\star}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P^{\text{cal}}_Y(y)\}}}.$$

Proof For any $y \in \mathcal{Y}$, introduce the following quantities: $\hat{f}^\star_y = \mathsf{M}^\star_y/\mathsf{M}^\star$, $f^\star = \{P^\star_Y(y)\}_{y\in\mathcal{Y}}$, $\hat{f}^\star = \{\hat{f}^\star_y\}_{y\in\mathcal{Y}}$, and $\widehat{r} = \{(P^{\text{cal}}_Y(y)\mathsf{M}/\mathsf{M}_y)\mathbb{1}_{\mathsf{M}_y>0}\}_{y\in\mathcal{Y}}$. We denote $\odot$ the Hadamard product, i.e., for any vectors $a, b \in \mathbb{R}^{|\mathcal{Y}|}$, $a\odot b$ is the vector of the componentwise product between a and b. We now bound the quantity in the right-hand side of the previous inequality, we obtain

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \mathbb{E}\left| P^\star_Y(y) - \frac{\widehat{w}^\star_y P^{\text{cal}}_Y(y)}{\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}^\star_{\tilde{y}} P^{\text{cal}}_Y(\tilde{y})}\right| &= \sum_{y\in\mathcal{Y}} \mathbb{E}\left| f^\star_y - \hat{f}^\star_y + \hat{f}^\star_y - \frac{\hat{f}^\star_y \widehat{r}_y}{\sum_{\tilde{y}\in\mathcal{Y}} \hat{f}^\star_{\tilde{y}}\widehat{r}_{\tilde{y}}}\right| \\
&\leq \mathbb{E}\left\| f^\star - \hat{f}^\star + \hat{f}^\star - \frac{\hat{f}^\star \odot \widehat{r}}{1 + \langle \hat{f}^\star, \widehat{r} - \mathbb{1}\rangle}\right\|_1 \\
&\leq \mathbb{E}\|f^\star - \hat{f}^\star\|_1 + \mathbb{E}\left\| \frac{\hat{f}^\star\langle \hat{f}^\star, \widehat{r} - \mathbb{1}\rangle - \hat{f}^\star \odot (\widehat{r} - \mathbb{1})}{1 + \langle \hat{f}^\star, \widehat{r} - \mathbb{1}\rangle}\right\|_1.
\end{aligned}$$

First, we establish the following equality that will be injected in the computation of $\mathbb{E}\|f^\star - \hat{f}^\star\|_1$.

$$\|f^\star - \hat{f}^\star\|_1 = \max_{u\in[-1/2,1/2]^{|\mathcal{Y}|}} \langle f^\star - \hat{f}^\star, u + \mathbb{1}/2\rangle = \max_{u\in[0,1]^{|\mathcal{Y}|}} \langle f^\star - \hat{f}^\star, u\rangle. \tag{5.48}$$

Then, using the result provided by (Agrawal and Jia, 2017, Lemma C.2), for any $\delta \in (0,1)$, we get

$$\mathbb{P}\left(\max_{u\in[0,1]^{|\mathcal{Y}|}} \langle f^\star - \hat{f}^\star, u\rangle \geq \sqrt{\frac{-2\log\delta}{\mathsf{M}^\star}}\right) \leq \delta. \tag{5.49}$$

Looking back to $\mathbb{E}\|f^\star - \hat{f}^\star\|_1$ and using the two previous identities, the following lines hold

$$\begin{aligned}
\mathbb{E}\|f^\star - \hat{f}^\star\|_1 &= \int_{t\geq 0} \mathbb{P}\left(\|f^\star - \hat{f}^\star\|_1 \geq t\right) \mathrm{d}t \\
&\leq \delta + \int_{t\geq\delta} \mathbb{P}\left(\|f^\star - \hat{f}^\star\|_1 \geq t\right) \mathrm{d}t \\
&\leq \delta + \int_{t\geq\delta} \exp\left(-\mathsf{M}^\star t^2/2\right) \mathrm{d}t \qquad \text{using (5.48)-(5.49)} \\
&\leq \delta + \frac{1}{\sqrt{\mathsf{M}^\star}} \int_{t\geq\delta\sqrt{\mathsf{M}^\star}} \exp\left(-t^2/2\right) \mathrm{d}t.
\end{aligned}$$

After optimizing for $\delta > 0$, we can retrieve the following upper bound

$$\mathbb{E}\|f^\star - \hat{f}^\star\|_1 \leq \frac{1.4}{\sqrt{\mathsf{M}^\star}}.$$

Let $\epsilon \in (0, 1/2]$, we have that

$$\left\| \frac{\hat{f}^\star \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle - \hat{f}^\star \odot (\widehat{r} - \mathbb{1})}{1 + \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle} \right\|_1 \leq 2\mathbb{1}_{\left\| \widehat{r} - \mathbb{1} \right\|_\infty > \epsilon} + \frac{2\epsilon}{1 - \epsilon}.$$

Taking the expectation for both sides, it shows

$$\mathbb{E} \left\| \frac{\hat{f}^\star \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle - \hat{f}^\star \odot (\widehat{r} - \mathbb{1})}{1 + \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle} \right\|_1 \leq 2\,\mathbb{P}\left(\left\| \widehat{r} - \mathbb{1} \right\|_\infty > \epsilon \right) + \frac{2\epsilon}{1 - \epsilon}.$$

We now upper bound the first term in the right-hand side of the inequality, we obtain

$$\begin{aligned}
\mathbb{P}\left(\left\| \widehat{r} - \mathbb{1} \right\|_\infty > \epsilon \right) &= \mathbb{P}\left(\max_{y \in \mathcal{Y}} \left| \frac{P_Y^{\mathrm{cal}}(y) \mathbb{1}_{\mathsf{M}_y > 0}}{\mathsf{M}_y / \mathsf{M}} - 1 \right| > \epsilon \right) \\
&\leq \sum_{y \in \mathcal{Y}} \mathbb{P}\left(\left| \frac{P_Y^{\mathrm{cal}}(y) \mathbb{1}_{\mathsf{M}_y > 0}}{\mathsf{M}_y / \mathsf{M}} - 1 \right| > \epsilon \right) \\
&\leq \sum_{y \in \mathcal{Y}} \left\{ \mathbb{P}\left(\frac{\mathsf{M}_y}{\mathsf{M}} < \frac{P_Y^{\mathrm{cal}}(y) \mathbb{1}_{\mathsf{M}_y > 0}}{1 + \epsilon} \right) + \mathbb{P}\left(\frac{\mathsf{M}_y}{\mathsf{M}} > \frac{P_Y^{\mathrm{cal}}(y)}{1 - \epsilon} \right) \right\}.
\end{aligned} \tag{5.50}$$

Since the random variable M_y is the sum of independent Bernoulli random variables, using the Chernoff bound it follows that

$$\begin{aligned}
\mathbb{P}\left(\mathsf{M}_y / \mathsf{M} < P_Y^{\mathrm{cal}}(y)/(1 + \epsilon) \right) &\leq \exp\left(-2\epsilon^2 N P_Y^{\mathrm{cal}}(y)/9 \right), \\
\mathbb{P}\left(\mathsf{M}_y / \mathsf{M} > P_Y^{\mathrm{cal}}(y)/(1 - \epsilon) \right) &\leq \exp\left(-4\epsilon^2 N P_Y^{\mathrm{cal}}(y)/3 \right).
\end{aligned} \tag{5.51}$$

Therefore, combining (5.50) with (5.51) gives

$$\mathbb{P}\left(\left\| \widehat{r} - \mathbb{1} \right\|_\infty > \epsilon \right) \leq \sum_{y \in \mathcal{Y}} \left[\mathbb{P}\left(\mathsf{M}_y = 0 \right) + 2 \exp\left(-\epsilon^2 N P_Y^{\mathrm{cal}}(y)/5 \right) \right].$$

Putting all the previous results together, we obtain

$$\begin{aligned}
\sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y) \mathbb{E} \left| w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right| &\leq \frac{1.4}{\sqrt{\mathsf{M}^\star}} + 4\epsilon \\
&\quad + 4 \sum_{y \in \mathcal{Y}} \exp\left(-2\epsilon^2 N P_Y^{\mathrm{cal}}(y)/9 \right) + \sum_{y \in \mathcal{Y}} (1 - P_Y^{\mathrm{cal}}(y))^{\mathsf{M}}.
\end{aligned} \tag{5.52}$$

Consider the following quantity

$$\epsilon = \frac{3}{2} \sqrt{\frac{2 \log |\mathcal{Y}| + \log \mathsf{M}^\star}{\mathsf{M} \min_{y \in \mathcal{Y}} \{ P_Y^{\mathrm{cal}}(y) \}}}.$$

If $\epsilon \leq 1/2$, it yields that

$$\sum_{y\in\mathcal{Y}} \exp\left(-2\epsilon^2 N P_Y^{\mathrm{cal}}(y)/9\right) \leq \sum_{y\in\mathcal{Y}} \exp\left(-\log|\mathcal{Y}| - (1/2)\log \mathsf{M}^\star\right) = \frac{1}{\sqrt{\mathsf{M}^\star}}.$$

Therefore, combining this last inequality with (5.47)-(5.52) implies that

$$\left|\mathbb{P}(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(X^\star_{N^\star+1})) - \mathbb{P}(\widehat{Y} \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(X^\star_{N^\star+1}))\right| \leq \frac{3}{\sqrt{\mathsf{M}^\star}} + 3\sqrt{\frac{2\log|\mathcal{Y}| + \log \mathsf{M}^\star}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}}} + |\mathcal{Y}|\left(1 - \min_{y\in\mathcal{Y}}\{P_Y^{\mathrm{cal}}(y)\}\right)^{\mathsf{M}} \wedge 1.$$

Otherwise, if $\epsilon > 1/2$, then, the last inequality immediately holds since the right-hand term is greater than 1. Lastly, applying Lemma 5.32 concludes the proof ■

5.C.4 Proof of Theorem 5.12

First, for all $i \in [n]$, denote by $F_V^i : u \in [0,1] \mapsto \mathbb{P}(V(X,Y) \leq u) \in [0,1]$ the cumulative distribution function of $V(X^i, Y^i)$, where $(X^i, Y^i) \sim P^{\mathrm{i}}$. Recall that $N = \sum_{i=1}^n N^i$, $\mathcal{I} = (i,k)\colon i \in [n], k \in [N^i] \cup \{(\star, N^\star+1)\}$, and also that there is almost surely no ties between the $\{V(X_k^i, Y_k^i)\}_{(i,k)\in\mathcal{I}}$. To simplify the notation, we re-index $\{(X_k^i, Y_k^i, V_k^i)\}_{(i,k)\in\mathcal{I}}$ into $\{X_k, Y_k, V_k\}_{k\in[N+1]}$, sorted such that $\{V_k\}_{k\in[N+1]}$ is non-decreasing.

Now, we consider $\alpha \in [0,1]\backslash\mathbb{Q}$. Using Theorem 5.16, this condition ensures the existence and uniqueness of $k_{\mathrm{opt}} \in [N+1]$ such that $V_{k_{\mathrm{opt}}} = \arg\min_{q\in\mathbb{R}} \left\{\mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}\left[S_{\alpha,V}(q)\right]\right\}$, and this condition also proves the existence and uniqueness of $Q^\gamma_{1-\alpha}$ minimizing $\{\mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S^\gamma_{\alpha,V}(q)]\colon q \in \mathbb{R}\}$. Moreover, for any $k \in [N+1] \setminus \{k_{\mathrm{opt}}\}$, define

$$\widehat{\rho}_{k_{\mathrm{c}}} = \begin{cases} \sum_{\ell\colon V_\ell \in (V_{k_{\mathrm{opt}}}, V_k]} \widehat{p}^\star_{Y_\ell, Y^\star_{N^\star+1}}, & \text{if } k > k_{\mathrm{opt}} \\ \sum_{\ell\colon V_\ell \in [V_k, V_{k_{\mathrm{opt}}})} \widehat{p}^\star_{Y_\ell, Y^\star_{N^\star+1}}, & \text{if } k < k_{\mathrm{opt}} \end{cases}. \tag{5.53}$$

Lemma 5.34. *Let $\alpha \in [0,1]\backslash\mathbb{Q}$, for any $V_k \in [\min(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}), V_{k_{\mathrm{opt}}}), \max(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}), V_{k_{\mathrm{opt}}})]$, we have*

$$|\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}) - V_k| \leq \widehat{\rho}^{-1}_{k_{\mathrm{c}}}\left(\mathrm{S}^\gamma_\alpha(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}^\gamma_\alpha(Q^\gamma_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))\right) + \widehat{\rho}^{-1}_{k_{\mathrm{c}}}\gamma,$$

where $\widehat{\rho}_{k_{\mathrm{c}}}$ is given in (5.53).

Proof First, suppose that $V_{k_{\mathrm{opt}}} \leq V_k < \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})$. Since $\partial S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(V_k) = \widehat{\rho}_{k_{\mathrm{c}}} + \partial S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(V_{k_{\mathrm{opt}}})$, the convexity of $S_{\alpha,\widehat{\mu}_{\mathbf{y}}}$ implies that

$$\begin{aligned}\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}) - V_k &\leq \widehat{\rho}^{-1}_{k_{\mathrm{c}}}\left(S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(V_k)\right) \\ &\leq \widehat{\rho}^{-1}_{k_{\mathrm{c}}}\left(\mathrm{S}^\gamma_\alpha(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}^\gamma_\alpha(Q^\gamma_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))\right) + \widehat{\rho}^{-1}_{k_{\mathrm{c}}}\gamma. \end{aligned} \tag{5.54}$$

The last inequality holds since $\|\mathrm{S}^\gamma_\alpha - S_{\alpha,\widehat{\mu}_{\mathbf{y}}}\|_\infty \leq \gamma/2$. Moreover, (5.54) is immediately satisfied when $V_k = \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})$. Therefore, (5.54) holds for all $V_k \in [V_{k_{\mathrm{opt}}}, \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})]$. Finally, the same lines show that (5.54) is also satisfied when $\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}) \leq V_k \leq V_{k_{\mathrm{opt}}}$.

■

For any $\gamma > 0$ and $T \in \mathbb{N}^\star$, recall that $Q^\gamma_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})$ and $\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})$ are defined in (5.14), (5.18). Consider

$$C^\gamma_T = \frac{2N \sum_{y\in\mathcal{Y}} P^{\mathrm{cal}}_Y(y)\widehat{w}^\star_y}{\min_{y\in\mathcal{Y}} \widehat{w}^\star_y} \left[\mathbb{E}\, \mathrm{S}^\gamma_\alpha(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}^\gamma_\alpha(Q^\gamma_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) + \gamma\right] \tag{5.55}$$

and define

$$k_{\mathrm{c}} = \begin{cases} k_{\mathrm{opt}} + \mathrm{Ent}\left(\sqrt{\frac{mNC^\gamma_T}{2\log N}}\right) & \text{if } k_{\mathrm{opt}} + \mathrm{Ent}\left(\sqrt{\frac{mNC^\gamma_T}{2\log N}}\right) \leq N+1 \\ k_{\mathrm{opt}} - \mathrm{Ent}\left(\sqrt{\frac{mNC^\gamma_T}{2\log N}}\right) & \text{otherwise} \end{cases}. \tag{5.56}$$

Lemma 5.35. *Assume there exists $m > 0$ such that for any $i \in [n]$, P^i_V admits a density f^i_V with respect to the Lebesgue measure that satisfies $m \leq f^i_V$. Fix $\alpha \in [0,1] \setminus \mathbb{Q}$, and suppose that $C^\gamma_T < 2m^{-1}N\log N$. We have $k_{\mathrm{c}} \in [N+1]$, and with probability at least $m(2N\log N)^{-1} + N^{-2}$ the next inequality holds*

$$|V_{k_{\mathrm{c}}} - V_{k_{\mathrm{opt}}}| \leq \sqrt{\frac{2C^\gamma_T \log N}{mN}} + \frac{2\log N}{mN}.$$

Proof Since we suppose $\alpha \in [0,1]\setminus\mathbb{Q}$, we can apply Theorem 5.16 to prove the existence and uniqueness of $k_{\mathrm{opt}} \in [N+1]$ such that $Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}) = V_{k_{\mathrm{opt}}}$. Since by assumption we have

$$\frac{mNC^\gamma_T}{2\log N} < (N+1)^2.$$

Therefore, it holds that

$$\mathrm{Ent}\left(\sqrt{\frac{mNC^\gamma_T}{2\log N}}\right) \leq N.$$

Hence, we deduce that $k_{\mathrm{c}} \in [N+1]$. Next, consider

$$L_N = \frac{2\log N}{mN}$$

and denote by P_N the partitioned obtained by splitting the interval $[0,1]$ into intervals of length L_N. Finally, define

$$A_N = \left\{\forall S \in P_N, \exists (i,k) \in \mathcal{I}, V^i_k \in S\right\}.$$

Note that $|P_N| = \lceil 1/L_N \rceil \leq mN/(2\log N) + 1$ and $\log(1 - mL_N) \leq -mL_N$, thus

$$\begin{aligned} \mathbb{P}\left(A_N\right) &\leq \sum_{S\in P_N} \prod_{(i,k)\in\mathcal{I}} \mathbb{P}\left(V^i_k \notin S\right) \\ &\leq \sum_{S\in P_N} \prod_{i=1}^{n} \mathbb{P}\left(V^i_1 \notin S\right)^{N^i + \mathbb{1}_\star(i)} \end{aligned}$$

$$\begin{aligned}
&\leq |P_N|(1-mL_N)^{N+1} \\
&\leq \Big(mN/(2\log N)+1\Big)\exp\Big(-m(N+1)L_N\Big) \\
&\leq \frac{m}{2N\log N}+\frac{1}{N^2}.
\end{aligned}$$

Without loss of generality, we can assume that $k_{\text{opt}} < k_{\text{c}}$. Denote $K = \{k_{\text{opt}}, \ldots, k_{\text{c}}\}$ the indices between k_{c} and k_{opt}. Consider $\mathcal{S} = \{I \in P_N \colon \exists k \in K, V_k \in I\}$, on the event A_N we get

$$\begin{aligned}
|V_{k_{\text{c}}} - V_{k_{\text{opt}}}| &\leq \sum_{I\in S}|I| \\
&\leq L_N(|k_{\text{c}} - k_{\text{opt}}|+1) \\
&\leq L_N\sqrt{\frac{mNC_T^{\gamma}}{2\log N}} + L_N = \sqrt{C_T^{\gamma}L_N} + L_N.
\end{aligned}$$

■

Lemma 5.36. *For any $(i,k) \in \mathcal{I}$, assume that (X_k^i, Y_k^i) is distributed according to $P_{X|Y} \times P_Y^i$ and suppose the random variables are pairwise independent. We have*

$$\mathbb{P}\left(\min_{y\in\mathcal{Y}} \widehat{p}^{\star}_{y,Y^{\star}_{N^{\star}+1}} < \frac{\min_{y\in\mathcal{Y}}\widehat{w}_y^{\star}}{2N\sum_{y\in\mathcal{Y}}P_Y^{\text{cal}}(y)\widehat{w}_y^{\star}}\right) \leq \frac{4\,\text{Var}(\widehat{w}^{\star}_{Y^{\text{cal}}})}{N(\mathbb{E}\widehat{w}^{\star}_{Y^{\text{cal}}})^2} + \frac{2\mathbb{E}\widehat{w}^{\star}_{Y^{\star}_{N^{\star}+1}}}{N\mathbb{E}\widehat{w}^{\star}_{Y^{\text{cal}}}}.$$

Proof First, recall that $I = \{(i,k) \colon i \in [n], k \in [N^i]\} \cup \{(\star, N^{\star}+1)\}$. We have

$$\begin{aligned}
\mathbb{E}\left[\sum_{(i,k)\in\mathcal{I}} \widehat{w}^{\star}_{Y_k^i}\right] &= \sum_{i=1}^{n}\sum_{y\in\mathcal{Y}}\Big(N^i + \mathbb{1}_{\star}(i)\Big)P_Y^i(y)\widehat{w}_y^{\star} \\
&= \sum_{y\in\mathcal{Y}} P_Y^{\star}(y)\widehat{w}_y^{\star} + N\sum_{y\in\mathcal{Y}}\left(\sum_{i=1}^{n}\pi_i P_Y^i(y)\right)\widehat{w}_y^{\star} \\
&= \sum_{y\in\mathcal{Y}}\Big[P_Y^{\star}(y) + NP_Y^{\text{cal}}(y)\Big]\widehat{w}_y^{\star}.
\end{aligned}$$

Therefore, using the Bienaymé-Tchebytchev inequality implies that

$$\begin{aligned}
&\mathbb{P}\left(\min_{(i,k)\in\mathcal{I}}\{\widehat{p}^{\star}_{Y_k^i, Y^{\star}_{N^{\star}+1}}\} < \frac{\min_{y\in\mathcal{Y}}\widehat{w}_y^{\star}}{2N\sum_{y\in\mathcal{Y}}P_Y^{\text{cal}}(y)\widehat{w}_y^{\star}}\right) \\
&= \mathbb{P}\left(\min_{(i,k)\in\mathcal{I}}\{\widehat{w}^{\star}_{Y_k^i}\} < \frac{\min_{y\in\mathcal{Y}}\widehat{w}_y^{\star}}{2N\sum_{y\in\mathcal{Y}}P_Y^{\text{cal}}(y)\widehat{w}_y^{\star}}\sum_{(i,k)\in\mathcal{I}}\widehat{w}^{\star}_{Y_k^i}\right) \\
&\leq \mathbb{P}\left(\sum_{(i,k)\in\mathcal{I}}\widehat{w}^{\star}_{Y_k^i} \geq 2N\sum_{y\in\mathcal{Y}}P_Y^{\text{cal}}(y)\widehat{w}_y^{\star}\right)
\end{aligned}$$

$$\leq \mathbb{P}\left(\sum_{i=1}^{n}\sum_{k=1}^{N^i}\left(\widehat{w}^\star_{Y^i_k} - \mathbb{E}\widehat{w}^\star_{Y^i_k}\right) \geq \frac{N}{2}\sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}^\star_y\right) + \mathbb{P}\left(\widehat{w}^\star_{Y^\star_{N^\star+1}} \geq \frac{N}{2}\sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}^\star_y\right)$$
$$\leq \frac{4\,\mathrm{Var}(\widehat{w}^\star_{Y^{\mathrm{cal}}})}{N(\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}})^2} + \frac{2\mathbb{E}\widehat{w}^\star_{Y^\star_{N^\star+1}}}{N\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}}}.$$

■

Theorem 5.37. *Assume there exist $m, M > 0$ such that for any $i \in [n]$, P_V^i admits a density f_V^i with respect to the Lebesgue measure that satisfies $m \leq f_V^i \leq M$. Let $\alpha \in [0,1] \setminus \mathbb{Q}$, and suppose that $C_T^\gamma < 2m^{-1}N\log N$. It holds*

$$\left|\mathbb{P}(Y^\star_{N^\star+1} \in \widehat{\mathcal{C}}^\gamma_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})) - \mathbb{P}(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1}))\right| \leq 3M\sqrt{\frac{2C_T^\gamma \log N}{mN}}$$
$$+ \frac{2M\log N}{mN} + \frac{4\,\mathrm{Var}(\widehat{w}^\star_{Y^{\mathrm{cal}}})}{N(\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}})^2} + \frac{2\mathbb{E}\widehat{w}^\star_{Y^\star_{N^\star+1}}}{N\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}}} + \frac{m}{2N\log N} + \frac{1}{N^2}, \quad (5.57)$$

where C_T^γ is defined in (5.55).

Proof Using the definitions of $\widehat{\mathcal{C}}^\gamma_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})$, $\mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})$ provided in Algorithm 5.11 and (5.9), we can write

$$\mathbb{P}(Y^\star_{N^\star+1} \in \widehat{\mathcal{C}}^\gamma_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})) - \mathbb{P}(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1}))$$
$$= \mathbb{P}\left(V(X^\star_{N^\star+1}, Y^\star_{N^\star+1}) \leq \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})\right) - \mathbb{P}\left(V(X^\star_{N^\star+1}, Y^\star_{N^\star+1}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\right). \quad (5.58)$$

Recall that $k_{\mathrm{opt}} = \arg\min_{k\in[N+1]} \mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S^\gamma_{\alpha,V}(V_k)]$ is a random variable and consider the event

$$B_N = \left\{|V_{k_{\mathrm{c}}} - V_{k_{\mathrm{opt}}}| \leq \sqrt{C_T^\gamma L_N} + L_N\right\} \cap \left\{\min_{y\in\mathcal{Y}} \widehat{p}^\star_{y,Y^\star_{N^\star+1}} \geq \frac{\min_{y\in\mathcal{Y}} \widehat{w}^\star_y}{2N\sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}^\star_y}\right\}$$

and denote by B_N^c its complement. Since $V(X^\star_{N^\star+1}, Y^\star_{N^\star+1})$ and $\{\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}), Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\}$ are independent, we have

$$\left|\mathbb{P}(V(X^\star_{N^\star+1}, Y^\star_{N^\star+1}) \leq \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - \mathbb{P}(V(X^\star_{N^\star+1}, Y^\star_{N^\star+1}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))\right|$$
$$= \left|\mathbb{E}\left[\mathbb{1}_{V(X^\star_{N^\star+1}, Y^\star_{N^\star+1}) \leq \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})} - \mathbb{1}_{V(X^\star_{N^\star+1}, Y^\star_{N^\star+1}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})}\right]\right|$$
$$\leq \mathbb{P}\left(B_N^c\right) + \mathbb{E}\left[\mathbb{1}_{B_N}\left|F_{V^\star}(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - F_{V^\star}(Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))\right|\right]. \quad (5.59)$$

The following inequality holds

$$\left|F_{V^\star}(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})) - F_{V^\star}(Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))\right| \leq \|F_{V^\star}(\cdot + \widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) - F_{V^\star}\|_\infty.$$

Thus, using that $F_{V^\star}$ is M-Lipschitz, we get

$$\mathbb{E}\left[\mathbb{1}_{B_N}\|F_{V^\star}(\cdot+\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})-Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))-F_{V^\star}\|_\infty\right]\leq M\mathbb{E}\left[\mathbb{1}_{B_N}|\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})-Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})|\right]. \tag{5.60}$$

Furthermore, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{B_N}|\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})-Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})|\right]&\leq\mathbb{E}\Big[\mathbb{1}_{B_N}\left|V_{k_{\mathrm{c}}}-V_{k_{\mathrm{opt}}}\right|\Big]\\&\quad+\mathbb{E}\left[\mathbb{1}_{V_{k_{\mathrm{c}}}\in[\min(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}),V_{k_{\mathrm{opt}}}),\max(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}),V_{k_{\mathrm{opt}}})]}|\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})-V_{k_{\mathrm{c}}}|\right].\end{aligned}$$

Applying Lemma 5.34, this implies that

$$\begin{aligned}&\mathbb{1}_{V_{k_{\mathrm{c}}}\in[\min(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}),V_{k_{\mathrm{opt}}}),\max(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}),V_{k_{\mathrm{opt}}})]}|\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})-V_{k_{\mathrm{c}}}|\\&\qquad\leq\begin{cases}\widehat{\rho}^{-1}_{k_{\mathrm{c}}}\left(\mathrm{S}^\gamma_\alpha(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}))-\mathrm{S}^\gamma_\alpha(Q^\gamma_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))\right)+\widehat{\rho}^{-1}_{k_{\mathrm{c}}}\gamma & \text{if } k_{\mathrm{c}}\neq k_{\mathrm{opt}}\\ 0 & \text{otherwise}\end{cases},\end{aligned}$$

where recall that $\widehat{\rho}_{k_{\mathrm{c}}}$ is defined in (5.53) and also that $\mathcal{I}=\{(i,k)\colon i\in[n],k\in[N^i]\}\cup\{(\star,N^\star+1)\}$. Moreover, on the event B_N, we immediately have that

$$\widehat{\rho}_{k_{\mathrm{c}}}\geq|k_{\mathrm{c}}-k_{\mathrm{opt}}|\min_{(i,k)\in\mathcal{I}}\{\widehat{p}^\star_{Y^i_k,Y^\star_{N^\star+1}}\}\geq\frac{|k_{\mathrm{c}}-k_{\mathrm{opt}}|\min_{y\in\mathcal{Y}}\widehat{w}^\star_y}{2N\sum_{y\in\mathcal{Y}}P^{\mathrm{cal}}_Y(y)\widehat{w}^\star_y}.$$

Finally, recall that k_{c} is given in (5.56) and suppose that $C^\gamma_T<2m^{-1}N\log N$. Therefore, using the bound provided in Lemma 5.35 implies that

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{B_N}|\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}})-Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})|\right]\\&\leq\mathbb{E}\Big[\mathbb{1}_{B_N}\left|V_{k_{\mathrm{c}}}-V_{k_{\mathrm{opt}}}\right|\Big]+\mathbb{E}\left[\frac{\left(\mathbb{E}\,\mathrm{S}^\gamma_\alpha(\widehat{Q}^\gamma_{1-\alpha,T}(\widehat{\mu}_{\mathbf{y}}))-\mathrm{S}^\gamma_\alpha(Q^\gamma_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}))+\gamma\right)\mathbb{1}_{k_{\mathrm{c}}\neq k_{\mathrm{opt}}}}{(2N\sum_{y\in\mathcal{Y}}P^{\mathrm{cal}}_Y(y)\widehat{w}^\star_y)^{-1}|k_{\mathrm{c}}-k_{\mathrm{opt}}|\min_{y\in\mathcal{Y}}\widehat{w}^\star_y}\right]\\&\leq\sqrt{\frac{2C^\gamma_T\log N}{mN}}+\frac{2\log N}{mN}+\frac{C^\gamma_T\mathbb{1}_{mNC^\gamma_T\geq2\log N}}{\mathrm{Ent}\left(\sqrt{\frac{mNC^\gamma_T}{2\log N}}\right)}.\end{aligned}\tag{5.61}$$

Combining (5.58)-(5.59)-(5.60)-(5.61) shows that

$$\begin{aligned}&\left|\mathbb{P}(Y^\star_{N^\star+1}\in\widehat{\mathcal{C}}^\gamma_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1}))-\mathbb{P}(Y^\star_{N^\star+1}\in\mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1}))\right|\\&\qquad\leq\mathbb{P}\left(B^c_N\right)+M\left(3\sqrt{\frac{2C^\gamma_T\log N}{mN}}+\frac{2\log N}{mN}\right).\end{aligned}\tag{5.62}$$

Using Lemma 5.36 gives that

$$\mathbb{P}\left(B^c_N\right)\leq\frac{4\,\mathrm{Var}(\widehat{w}^\star_{Y^{\mathrm{cal}}})}{N(\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}})^2}+\frac{2\mathbb{E}\widehat{w}^\star_{Y^\star_{N^\star+1}}}{N\mathbb{E}\widehat{w}^\star_{Y^{\mathrm{cal}}}}+\frac{m}{2N\log N}+\frac{1}{N^2}.\tag{5.63}$$

Lastly, plugging (5.63) into (5.62) concludes the proof when $C^\gamma_T<2m^{-1}N\log N$. However, if $C^\gamma_T\geq2m^{-1}N\log N$ then (5.57) immediately holds. Thus, (5.57) always holds. ■

5.C.5 Proofs of Theorem 5.3 and Theorem 5.5

Recall that $\{\pi_i\}_{i\in[n]} \in \Delta_n$ and $P^{\mathrm{cal}} = \sum_{i=1}^n \pi_i P^i$. Moreover, draw $(\widehat{X}^\star_{N^\star+1}, \widehat{Y}^\star_{N^\star+1})$ according to $P_{X|Y} \times \widehat{P}^\star_Y$, where $\widehat{P}^\star_Y$ is defined in Section 5.C.3 and denote $\widehat{V}_{N+1} = V(\widehat{X}^\star_{N^\star+1}, \widehat{Y}^\star_{N^\star+1})$. In the following paragraph, we explain how to construct a sequence $\{V_k\}_{k\in[N]}$ of i.i.d. random variables distributed according to $P^{\mathrm{cal}}(V)$ – see Lemma 5.38, where $P^{\mathrm{cal}}(V)$ denotes the distribution of $V(X,Y)$ with $(X,Y) \sim P^{\mathrm{cal}}$. We also explain the construction of a bijection $\psi : \{(i,k)\colon i \in [n], k \in [N^i]\} \to [N]$. For all $k \in [N]$, draw M_k according to a categorical random variable with parameter $\{\pi_i\}_{i\in[n]}$ and define $\bar{N}^i_k = \sum_{l=1}^k \mathbb{1}_{\{i\}}(M_l)$. If $\bar{N}^{M_k}_k \leq N^{M_k}$ then define $(X_k, Y_k) \leftarrow (X^{M_k}_{\bar{N}^{M_k}_k}, Y^{M_k}_{\bar{N}^{M_k}_k})$ and $\psi(M_k, \bar{N}^{M_k}_k) = k$. Else $\bar{N}^{M_k}_k > N^{M_k}$, then draw (X_k, Y_k) according to P^{M_k} – where we recall that P^{M_k} is the distribution of calibration of agent $M_k \in [n]$. If there exists $k \in [N]$ such that $\bar{N}^{M_k}_k > N^{M_k}$, consider

$$J_0 = \left\{k \in [N]\colon \bar{N}^{M_k}_k > N^{M_k}\right\}, \qquad J_1 = \left\{(i,k) \in [n] \times \mathbb{N}\colon \bar{N}^i_N < k \leq N^i\right\}.$$

Since the following inequalities hold:

$$\mathrm{Card}(J_0) + \sum_{i=1}^n \min(N^i, \bar{N}^i_N) = N, \qquad \mathrm{Card}(J_1) + \sum_{i=1}^n \min(N^i, \bar{N}^i_N) = N,$$

we deduce that $\mathrm{Card}(J_0) = \mathrm{Card}(J_1)$. Moreover, using the existence of $k \in [N]$ such that $\bar{N}^{M_k}_k > N^{M_k}$, we deduce that $J_0 \neq \emptyset$. Therefore, there exists a bijection $\varphi : J_0 \to J_1$. We have previously defined ψ on $\{(i,k)\colon i \in [n], k \in [N^i]\}\backslash J_1$. For any $k \in J_0$, define $\psi(\varphi(k)) = k$. Remark, ψ is now correctly defined on $\{(i,k)\colon i \in [n], k \in [N^i]\} \to [N]$.

Lemma 5.38. *Denote $P^{\mathrm{cal}}(V)$ the distribution of $V(X,Y)$ with $(X,Y) \sim P^{\mathrm{cal}}$. The sequence $\{V_k\}_{k\in[N]}$ is a sequence of i.i.d. random variables distributed according to $P^{\mathrm{cal}}(V)$.*

Proof Let $h : \mathbb{R}^N \to \mathbb{R}$ be a continuous and bounded function, we have

$$\begin{aligned}
\mathbb{E}\left[h(V_1, \ldots, V_N)\right] &= \sum_{i_1,\ldots,i_N \in [n]} \left(\prod_{k'=1}^N \pi_{i_{k'}}\right) \int h(v_1, \ldots, v_N) \prod_{k=1}^N \mathrm{d}P^{i_k}(v_k) \\
&= \sum_{i_1,\ldots,i_{N-1} \in [n]} \left(\prod_{k'=1}^{N-1} \pi_{i_{k'}}\right) \int h(v_1, \ldots, v_N) \prod_{k=1}^{N-1} \mathrm{d}P^{i_k}(v_k) \left(\sum_{i=1}^n \pi_i \mathrm{d}P^i\right)(v_N) \\
&= \sum_{i_1,\ldots,i_{N-1} \in [n]} \left(\prod_{k'=1}^{N-1} \pi_{i_{k'}}\right) \int h(v_1, \ldots, v_N) \prod_{k=1}^{N-1} \mathrm{d}P^{i_k}(v_k) \mathrm{d}P^{\mathrm{cal}}(v_N) \\
&= \cdots = \int h(v_1, \ldots, v_N) \prod_{k=1}^N \mathrm{d}P^{\mathrm{cal}}(v_k).
\end{aligned}$$

This last line concludes the proof. ∎

In the following, we consider general likelihood ratios $\{\widehat{w}^\star_y\}_{y\in\mathcal{Y}}$ and recall that $w^\star_y = P^\star_Y(y)/P^{\mathrm{cal}}_Y(y)$. In addition, let $\bar{N} \in [N]$, denote for any $i \in [n], k \in [N^i \wedge \bar{N}^i_{\bar{N}}], k' \in [\bar{N}]$, $V_{k'} = V(X_{k'}, Y_{k'})$ and define

$$\begin{aligned}
&\bar{D}_{\widehat{Y}^\star_{N^\star+1}} = \{\widehat{Y}^\star_{N^\star+1}\} \cup \{Y_k \colon k \in [\bar{N}]\}, \quad \widehat{D}_{\widehat{Y}^\star_{N^\star+1}} = \{\widehat{Y}^\star_{N^\star+1}\} \cup \{Y_k \colon i \in [n], k \in [N^i \wedge \bar{N}^i_{\bar{N}}]\} \\
&\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_{k'}} = \frac{\widehat{w}^\star_{Y_{k'}}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l}}, \qquad \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y^i_k} = \frac{\widehat{w}^\star_{Y^i_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{j=1}^{n} \sum_{l=1}^{N^j \wedge \bar{N}^j_{\bar{N}}} \widehat{w}^\star_{Y^j_l}}.
\end{aligned} \tag{5.64}$$

Moreover, consider $\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} = \widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} (\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l})^{-1}$ and $\widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} = \widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} (\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{j=1}^{n} \sum_{l=1}^{N^j \wedge \bar{N}^j_{\bar{N}}} \widehat{w}^\star_{Y^j_l})^{-1}$. Lastly, define

$$\begin{aligned}
X &= \sum_{k=1}^{\bar{N}} \bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k} \mathbb{1}_{V_k < \widehat{V}_{N+1}}, \\
\delta &= \sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y^i_k} \mathbb{1}_{V^i_k < \widehat{V}_{N+1}} - \sum_{k=1}^{\bar{N}} \bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k} \mathbb{1}_{V_k < \widehat{V}_{N+1}}.
\end{aligned} \tag{5.65}$$

Lemma 5.39. *For any $\{\pi_i\}_{i\in[n]} \in \Delta_n$, it holds*

$$\begin{aligned}
&-\epsilon - \mathbb{P}\left(\delta \le -\epsilon\right) \\
&\quad \le \mathbb{P}\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\textstyle\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y^i_k} \delta_{V^i_k} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1 \right) \right) - 1 + \alpha \\
&\qquad\qquad \le \mathbb{E}\left[\max_{k=1}^{\bar{N}+1} \{ \bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k} \} \right] + \epsilon + \mathbb{P}\left(\delta \ge \epsilon\right),
\end{aligned}$$

where $\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k}$ and $\widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y^i_k}$ are defined in (5.64).

Proof By the definition of the quantile combined with (5.65), we get

$$\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\textstyle\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y^i_k} \delta_{V^i_k} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1 \right) \right) \iff \left(X + \delta > \alpha\right).$$

Therefore, it holds that

$$\begin{aligned}
&\mathbb{P}\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\textstyle\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y^i_k} \delta_{V^i_k} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1 \right) \right) \\
&\qquad\qquad = \mathbb{E}\left[\mathbb{1}_{X>\alpha}\right] + \mathbb{E}\left[\mathbb{1}_{X+\delta>\alpha} - \mathbb{1}_{X>\alpha}\right].
\end{aligned} \tag{5.66}$$

Remark that

$$\mathbb{1}_{X>\alpha+\epsilon} - \mathbb{1}_{X>\alpha} - \mathbb{1}_{\delta\le-\epsilon} \le \mathbb{1}_{X+\delta>\alpha} - \mathbb{1}_{X>\alpha} \le \mathbb{1}_{X>\alpha-\epsilon} - \mathbb{1}_{X>\alpha} + \mathbb{1}_{\delta\ge\epsilon}. \tag{5.67}$$

Thus, combining (5.67) with (5.66) gives

$$\mathbb{P}\left(X > \alpha + \epsilon\right) - \mathbb{P}\left(\delta \leq -\epsilon\right) \leq \mathbb{P}\left(\widehat{V}_{N+1} \leq Q_{1-\alpha}\left(\sum_{i=1}^{n}\sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_{V_k^i} + \widehat{p}_{\widehat{Y}^\star_{N^\star+1}}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_1\right)\right) \leq \mathbb{P}\left(X > \alpha - \epsilon\right) + \mathbb{P}\left(\delta \geq \epsilon\right). \quad (5.68)$$

Consider $\tilde{\alpha} \in (0,1)$,

$$(X > \tilde{\alpha}) \iff \left(\widehat{V}_{N+1} \leq Q_{1-\tilde{\alpha}}\left(\bar{p}_{\widehat{Y}^\star_{N^\star+1}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_{V_k}\right)\right).$$

Hence $\mathbb{P}(X > \tilde{\alpha}) = \mathbb{P}(\widehat{V}_{N+1} \leq Q_{1-\tilde{\alpha}}(\bar{p}_{\widehat{Y}^\star_{N^\star+1}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_{V_k}))$. Applying Theorem 5.28 gives that

$$0 \leq \mathbb{P}\left(\widehat{V}_{N+1} \leq Q_{1-\tilde{\alpha}}\left(\bar{p}_{\widehat{Y}^\star_{N^\star+1}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_{V_k}\right)\right) - 1 + \tilde{\alpha} \leq \mathbb{E}\left[\max_{k=1}^{\bar{N}+1} \{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\}\right].$$

Therefore, plugging the previous inequality into (5.68) concludes the proof. ■

Recall that $(\widehat{w}^\star_{Y_1} - \mathbb{E}\widehat{w}^\star_{Y_1})$ is σ-sub Gaussian if for any $s \in \mathbb{R}$, the following inequality holds

$$\int_{y \in \mathcal{Y}} \exp\left(s\left[\widehat{w}^\star_y - \int_{y' \in \mathcal{Y}} \widehat{w}^\star_{y'} \mathrm{d}P_Y^{\mathrm{cal}}(y')\right]\right) \mathrm{d}P_Y^{\mathrm{cal}}(y) \leq \exp\left(\frac{\sigma^2 s^2}{2}\right).$$

Lemma 5.40. *Assume the random variable $(\widehat{w}^\star_{Y_1} - \mathbb{E}\widehat{w}^\star_{Y_1})$ is σ-sub Gaussian with parameter $\sigma \geq 0$. For any $\epsilon \geq 8\sigma^2 \log N/(\bar{N}\mathbb{E}[\widehat{w}^\star_{Y_1}]^2)$ and $\{\pi_i\}_{i\in[n]} \in \Delta_n$, we have*

$$\mathbb{P}\left(|\delta| > \epsilon\right) \leq \mathbb{P}\left(\sum_{i=1}^{n}\left(\bar{N}^i_{\bar{N}} - N^i\right)_+ > \frac{N\epsilon}{4}\right) + \frac{4\,\mathrm{Var}\left(\widehat{w}^\star_{Y_1}\right)}{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]^2} + \frac{1}{N},$$

where δ is defined in (5.65).

Proof First, denote $\mathcal{J}_0 = J_0 \cap [\bar{N}]$ and remark that $\mathrm{Card}(\mathcal{J}_0) = \sum_{i=1}^{n}(\bar{N}^i_{\bar{N}} - N^i)_+$. Moreover, recall that δ is defined by

$$\delta = \sum_{i=1}^{n}\sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}}\left(\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}} - \bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\right)\mathbb{1}_{V_k^i < V_{N+1}} - \sum_{k\in\mathcal{J}_0} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\mathbb{1}_{V_k < \widehat{V}_{N+1}}.$$

Using definition of the weighs $\bar{p}^{\bar{D}}$ given in (5.64) and the definition of ψ provide at the beginning of this section, note that

$$\sum_{k\in\mathcal{J}_0} \bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\mathbb{1}_{V_k < \widehat{V}_{N+1}} \leq \frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}. \quad (5.69)$$

From the definition of $\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}$ and $\bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}$ provided in (5.64), we obtain

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{k=1}^{N^i\wedge\bar{N}^i_N}\left(\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}-\bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\right)\mathbb{1}_{V_k^i<V_{N+1}}\\
&\qquad=\sum_{i=1}^{n}\sum_{k=1}^{N^i\wedge\bar{N}^i_N}\left(\frac{\widehat{w}^\star_{Y_k^i}\mathbb{1}_{V_k^i<V_{N+1}}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}}+\sum_{j=1}^{n}\sum_{l\in J_i}\widehat{w}^\star_{Y_l^j}}-\frac{\widehat{w}^\star_{Y_k^i}\mathbb{1}_{V_k^i<V_{N+1}}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}\right),\\
&\qquad=\left(\frac{1}{\sum_{l\in[\bar{N}+1]\setminus\mathcal{J}_0}\widehat{w}^\star_{Y_l}}-\frac{1}{\sum_{l\in[\bar{N}+1]\setminus\mathcal{J}_0}\widehat{w}^\star_{Y_l}+\sum_{l\in\mathcal{J}_0}\widehat{w}^\star_{Y_l}}\right)\sum_{k\in[\bar{N}]\setminus\mathcal{J}_0}\widehat{w}^\star_{Y_k}\mathbb{1}_{V_k^i<V_{N+1}}.
\end{aligned}$$

Therefore, we know that

$$\left|\sum_{i=1}^{n}\sum_{k=1}^{N^i\wedge\bar{N}^i_N}\left(\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}-\bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\right)\mathbb{1}_{V_k^i<V_{N+1}}\right|\leq\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}.$$

Hence, plugging the previous line into (5.65) combined with (5.69) gives

$$\big|\delta\big|\leq\frac{2\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}. \tag{5.70}$$

Moreover, define the following event:

$$E_N=\left\{\sum_{i=1}^{n}\left(\bar{N}^i_{\bar{N}}-N^i\right)_+\leq\frac{N\epsilon}{4}\right\}.$$

Next, using the event E_N, we decompose the following probability

$$\begin{aligned}
\mathbb{P}\left(\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}\geq\frac{\epsilon}{2};E_N\right)&\leq\mathbb{P}\left(\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}\geq\frac{\epsilon\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{4};E_N\right)\\
&\quad+\mathbb{P}\left(\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}<\frac{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{2}\right).
\end{aligned}\tag{5.71}$$

Since the $\{\widehat{w}^\star_{Y_k}\}_{k\in[N]}$ are σ-sub Gaussian, the first term of the previous right-hand side inequality is upper bounded thanks to Hoeffding's inequality

$$\begin{aligned}
\mathbb{P}\left(\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}\geq\frac{\epsilon\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{4};E_N\right)&=\mathbb{E}\left[\mathbb{P}\left(\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}\geq\frac{\epsilon\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{4}\,\middle|\,\mathrm{Card}(\mathcal{J}_0)\right)\mathbb{1}_{E_N}\right]\\
&\leq\exp\left(-\frac{(\epsilon\bar{N}/4)^2\mathbb{E}[\widehat{w}^\star_{Y_1}]^2}{2\mathrm{Card}(\mathcal{J}_0)\sigma^2}\right)\leq\frac{1}{N},
\end{aligned}\tag{5.72}$$

where the last inequality holds by setting $\epsilon \geq 8\sigma^2 \log N/(\bar{N}\mathbb{E}[\widehat{w}^\star_{Y_1}]^2)$. Moreover, since $\widehat{w}^\star_{Y_1} \geq 0$ almost surely, from the Chebyshev inequality we deduce that

$$\mathbb{P}\left(\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l} < \frac{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{2}\right) \leq \frac{4\,\mathrm{Var}\left(\widehat{w}^\star_{Y_1}\right)}{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]^2}. \tag{5.73}$$

Therefore, combining (5.71), (5.72) and (5.73) shows

$$\mathbb{P}\left(\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}} \geq \frac{\epsilon}{4}; E_N\right) \leq \frac{1}{N} + \frac{4\,\mathrm{Var}\left(\widehat{w}^\star_{Y_1}\right)}{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]^2}. \tag{5.74}$$

Lastly, using (5.70) we derive the next inequality:

$$\mathbb{P}\left(\left|\delta\right| > \epsilon\right) \leq 1 - \mathbb{P}\left(E_N\right) + \mathbb{P}\left(\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}} \geq \frac{\epsilon}{2}; E_N\right).$$

Plugging (5.74) into the previous line completes the proof. ■

Lemma 5.41. *For any $i \in [n]$, consider $\pi_i = N^i/N$ and $2\bar{N} \leq N$. We have*

$$\mathbb{P}\left(\sum_{i=1}^{n}\left(\bar{N}^i_{\bar{N}} - N^i\right)_+ > \frac{7}{4}\log(nN)\sum_{j:\,\frac{Nj}{6}<\log(nN)}\sqrt{N^j}\right) \leq \frac{1}{N}. \tag{5.75}$$

Proof First, define $\epsilon = 7[\log(nN)/N]\sum_{j\in A}\sqrt{N^j}$ where we consider the following set

$$A = \left\{i \in [n]\colon N^i < 6\log(nN)\right\}.$$

If $A \neq \emptyset$, for all $i \in A$ take

$$\alpha_i = \frac{\sqrt{\pi_i}}{\sum_{j\in A}\sqrt{\pi_j}}.$$

Using the union bound, we get

$$\mathbb{P}\left(\sum_{i=1}^{n}\left(\bar{N}^i_{\bar{N}} - N^i\right)_+ > \frac{N\epsilon}{4}\right) \leq \sum_{i\in[n]\setminus A}\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i\right) + \sum_{i\in A}\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i + \frac{\alpha_i N\epsilon}{4}\right). \tag{5.76}$$

If $i \in [n] \setminus A$, then applying the Chernoff bound gives

$$\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i\right) \leq \exp\left(-\frac{\pi_i(N-\bar{N})}{1+2\bar{N}/(N-\bar{N})}\right) \leq \exp\left(-\frac{\pi_i N}{6}\right) \leq \frac{1}{nN}. \tag{5.77}$$

If $i \in A$, then applying the Bernstein inequality gives

$$\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i + \frac{\alpha_i N\epsilon}{4}\right) \leq \exp\left(-\frac{(\alpha_i N\epsilon/4)^2}{2\bar{N}\pi_i(1-\pi_i) + \alpha_i N\epsilon/6}\right). \tag{5.78}$$

Moreover, we can suppose that $n \geq 2$ otherwise (5.75) immediately holds. Thus, $nN \geq 4$ and using the fact that $2\bar{N} \leq N$ we deduce that

$$4\left(\frac{2}{3} + \sqrt{\frac{2\bar{N}}{N\log(nN)}}\right) \leq 7.$$

Therefore, it follows

$$N\epsilon \geq 4\log(nN)\left(\frac{2}{3} + \sqrt{\frac{2\bar{N}}{N\log(nN)}}\right)\sum_{j\in A}\sqrt{N^j}.$$

The last inequality is enough to ensure that

$$\frac{(N\epsilon)^2/8}{4(\bar{N}/N)(\sum_{j\in A}\sqrt{N^j})^2 + N\epsilon\sum_{j\in A}\sqrt{N^j}/3} \geq \log(nN). \tag{5.79}$$

Moreover, using the definition of α_i implies

$$\frac{(\alpha_i N\epsilon/4)^2}{2\bar{N}\pi_i(1-\pi_i) + \alpha_i N\epsilon/6} \geq \frac{(N\epsilon)^2/8}{4(\bar{N}/N)(\sum_{j\in A}\sqrt{N^j})^2 + N\epsilon\sum_{j\in A}\sqrt{N^j}/3}.$$

Hence, combining (5.78) with (5.79) shows that

$$\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i + \frac{\alpha_i N\epsilon}{4}\right) \leq \frac{1}{nN}. \tag{5.80}$$

Finally, plugging (5.77) and (5.80) into (5.76) yields the result. ■

Lemma 5.42. *Recall that* $\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{\bar{N}^\star+1}}}\}_{k\in[\bar{N}]}$ *is defined in* (5.64) *and for the sake of simplicity define* $Y_{\bar{N}+1} = \widehat{Y}^\star_{\bar{N}^\star+1}$. *If* $\|\widehat{w}\|_\infty < \infty$, *then*

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\right] \leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\left(\mathrm{Var}(\widehat{w}_{Y_1}) + \frac{\mathrm{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}}\right).$$

Moreover, if the random variables $(\widehat{w}_{Y_k} - \mathbb{E}\widehat{w}_{Y_k})_{k\in[\bar{N}+1]}$ *are* σ*-subGaussian with parameter* $\sigma \geq 0$, *then*

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\right] \leq \frac{\sigma\sqrt{8\log(\bar{N}+1)}}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{2}{\bar{N}}\left(1 \vee \frac{\mathbb{E}\widehat{w}_{Y_{\bar{N}+1}}}{\mathbb{E}\widehat{w}_{Y_1}} + \frac{2\,\mathrm{Var}(\widehat{w}_{Y_1})}{(\mathbb{E}\widehat{w}_{Y_1})^2} + \frac{2\,\mathrm{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\right).$$

Proof By definition of the probabilities $\{\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\}_{k\in[\bar{N}+1]}$, we have

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\{\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\}\right] = \mathbb{E}\left[\frac{\max_{k=1}^{\bar{N}+1}\left\{\widehat{w}_{Y_k}\right\}}{\sum_{l=1}^{\bar{N}+1}\widehat{w}_{Y_l}}\right].$$

The expectation is split by introducing $\mathcal{A}_{N+1} = \{\sum_{l=1}^{\bar{N}+1} \widehat{w}_{Y_l} \leq (\bar{N}\mathbb{E}\widehat{w}_{Y_1} + \mathbb{E}\widehat{w}_{Y_{\bar{N}+1}})/2\}$, we obtain

$$\mathbb{E}\left[\frac{\max_{k=1}^{\bar{N}+1}\left\{\widehat{w}_{Y_k}\right\}}{\sum_{l=1}^{\bar{N}+1}\widehat{w}_{Y_l}}\right] \leq \frac{2}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_{N+1}^c}\cdot\max_{k=1}^{\bar{N}+1}\left\{\widehat{w}_{Y_k}\right\}\right] + \mathbb{P}\left(\mathcal{A}_{N+1}\right). \tag{5.81}$$

Using the Chebyshev's inequality it follows that

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}_{N+1}\right) &= \mathbb{P}\left(2\sum_{k=1}^{\bar{N}+1}\widehat{w}_{Y_k} < \bar{N}\mathbb{E}\widehat{w}_{Y_1} + \mathbb{E}\widehat{w}_{Y_{\bar{N}+1}}\right) \\ &\leq \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1} + \mathbb{E}\widehat{w}_{Y_{\bar{N}+1}}/\bar{N})^2}\left(\mathrm{Var}(\widehat{w}_{Y_1}) + \frac{\mathrm{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}}\right). \end{aligned} \tag{5.82}$$

When $\widehat{w}$ is bounded, combining (5.81) with (5.82) gives

$$\begin{aligned}\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\}\right] &\leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \mathbb{P}\left(\mathcal{A}_{N+1}\right) \\ &\leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\left(\mathrm{Var}(\widehat{w}_{Y_1}) + \frac{\mathrm{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}}\right).\end{aligned}$$

Otherwise, if we suppose that the $(\widehat{w}_{Y_k} - E\widehat{w}_{Y_k})_{k\in[\bar{N}+1]}$ are σ-sub-Gaussian, then applying the result given in (Boucheron et al., 2013, Section 2.5) combined with (5.81)-(5.82) concludes the proof since it follows that $\mathbb{E}[\max_{k=1}^{\bar{N}+1}\{\widehat{w}_{Y_k}\}] \leq \max_{k=1}^{\bar{N}+1}\{\mathbb{E}\widehat{w}_{Y_k}\} + \sigma\sqrt{2\log(\bar{N}+1)}$. ■

Finally, for any point $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, define the following measure and prediction set

$$\begin{aligned}\widehat{\mu}_{\mathbf{y}} &= \widehat{p}_{\mathbf{y}}^{\widehat{D}_{\mathbf{y}}}\delta_1 + \sum_{i=1}^n\sum_{k=1}^{N^i\wedge\bar{N}_{\bar{N}}^i}\widehat{p}_{Y_k^i}^{\widehat{D}_{\mathbf{y}}}\delta_{V_k^i} \\ \mathcal{C}_{\alpha,\widehat{\mu}}(\mathbf{x}) &= \left\{\mathbf{y}\in\mathcal{Y}\colon V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\right\}.\end{aligned}$$

Theorem 5.43. *For any $i \in [n]$, let $\pi_i = N^i/N$ and consider $\bar{N} = \lfloor N/2 \rfloor$. If $\|\widehat{w}^\star\|_\infty < \infty$, then it holds*

$$\begin{aligned}\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| &\leq \frac{1}{2}\sum_{y\in\mathcal{Y}}\left|P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}\right| \\ &\quad + \frac{6}{N} + \frac{36\|\widehat{w}\|_\infty^2}{N(\mathbb{E}\widehat{w}_{Y_1})^2} + \frac{2\log N}{N}\left(\frac{3\|\widehat{w}\|_\infty^2}{(\mathbb{E}\widehat{w}^\star_{Y_1})^2} \vee 7\sum_{j\colon \frac{N^j}{12}<\log N}\sqrt{N^j}\right).\end{aligned}$$

Proof Using Lemma 5.30 implies that

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})\right) - \mathbb{P}\left(\widehat{Y}^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^\star_{N^\star+1})\right)\right|$$

$$\leq \frac{1}{2}\sum_{y\in\mathcal{Y}} P_Y^{\text{cal}}(y)\left| w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\text{cal}}(\tilde{y})}\right|. \quad (5.83)$$

Since $\{\widehat{w}_y^\star\}_{y\in\mathcal{Y}}$ are bounded, we deduce that $\widehat{w}_{Y_1}^\star$ is σ-subGaussian with $\sigma = 2^{-1}(\max_{y\in\mathcal{Y}}\{\widehat{w}_y^\star\} - \min_{y\in\mathcal{Y}}\{\widehat{w}_y^\star\})$. Therefore, the inequality derived in Lemma 5.41 combined with Lemma 5.40 provide an upper bound on $\mathbb{P}(|\delta| > \epsilon)$ with

$$\epsilon = \frac{8\sigma^2\log N}{\bar{N}\mathbb{E}[\widehat{w}_{Y_1}^\star]^2} \vee \frac{7\log(nN)\sum_{j:\,\frac{Nj}{6}<\log(nN)}\sqrt{N^j}}{N}.$$

Plugging this result into the bound derived in Lemma 5.39 shows that

$$\left|\mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star)\right) - 1 + \alpha\right| \leq \mathbb{E}\left[\bar{p}_{\widehat{Y}_{N^\star+1}^\star}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}} \vee \max_{k=1}^{\bar{N}}\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\}\right] + \frac{2}{N}$$
$$+ \frac{4\,\text{Var}\left(\widehat{w}_{Y_1}^\star\right)}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1}^\star)^2} + \frac{8\sigma^2\log N}{\bar{N}\mathbb{E}[\widehat{w}_{Y_1}^\star]^2} \vee \frac{7\log(nN)\sum_{j:\,\frac{Nj}{6}<\log(nN)}\sqrt{N^j}}{N}. \quad (5.84)$$

Moreover, applying Lemma 5.42, we deduce that

$$\mathbb{E}\left[\bar{p}_{\widehat{Y}_{N^\star+1}^\star}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}} \vee \max_{k=1}^{\bar{N}}\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\}\right] \leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\left(\text{Var}(\widehat{w}_{Y_1}) + \frac{\text{Var}(\widehat{w}_{\widehat{Y}_{N^\star+1}^\star})}{\bar{N}}\right). \quad (5.85)$$

Therefore, combining (5.84) with (5.85) implies that

$$\left|\mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star)\right) - 1 + \alpha\right| \leq \frac{2}{\bar{N}}\left(\frac{\bar{N}}{N} + \frac{\|\widehat{w}\|_\infty}{\mathbb{E}\widehat{w}_{Y_1}} + \frac{4\,\text{Var}\left(\widehat{w}_{Y_1}^\star\right)}{(\mathbb{E}\widehat{w}_{Y_1}^\star)^2} + \frac{2\,\text{Var}(\widehat{w}_{\widehat{Y}_{N^\star+1}^\star})}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\right)$$
$$+ \frac{8\sigma^2\log N}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1}^\star)^2} \vee \frac{14\log N\sum_{j:\,\frac{Nj}{6}<\log(nN)}\sqrt{N^j}}{N}.$$

If $N \geq 6$, then remark that $N/\bar{N} \leq 3$. Thus, we deduce that

$$\left|\mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star)\right) - 1 + \alpha\right| \leq \frac{6}{N}\left(1 + \frac{\|\widehat{w}\|_\infty}{\mathbb{E}\widehat{w}_{Y_1}} + \frac{4\,\text{Var}\left(\widehat{w}_{Y_1}^\star\right)}{(\mathbb{E}\widehat{w}_{Y_1}^\star)^2} + \frac{\text{Var}(\widehat{w}_{\widehat{Y}_{N^\star+1}^\star})}{(\mathbb{E}\widehat{w}_{Y_1})^2}\right)$$
$$+ \frac{24\sigma^2\log N}{N(\mathbb{E}\widehat{w}_{Y_1}^\star)^2} \vee \frac{14\log N\sum_{j:\,\frac{Nj}{12}<\log N}\sqrt{N^j}}{N}. \quad (5.86)$$

Finally, using the next inequality combined with (5.83) and (5.86) concludes the proof

$$\mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star)) = \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star)) - \mathbb{P}(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star))$$
$$+ \mathbb{P}(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star)).$$

■

The proof of the following result is similar to that of Theorem 5.43.

Theorem 5.44. *For any $i \in [n]$, let $\pi_i = N^i/N$ and set $\bar{N} = \lfloor N/2 \rfloor$. Moreover, for $y \in \mathcal{Y}$ consider $\widehat{w}_y^\star = w_y^\star$. If $\|w^\star\|_\infty < \infty$, then it holds*

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| \leq \frac{6}{N} + \frac{36 + 6\log N}{N}\|w^\star\|_\infty^2 + \frac{14\log N}{N}\sum_{j\colon \frac{N^j}{12} < \log N}\sqrt{N^j}.$$

Proof First, for $y \in \mathcal{Y}$ recall that $w_y^\star = (P_Y^\star/P_Y^{\mathrm{cal}})(y)$. Thus, we remark that $\mathbb{E}w^\star_{Y_1} = 1$. Therefore, applying Theorem 5.43 concludes the proof. ■

5.D Differential privacy guarantee: proof of Theorem 5.13

In this section, we recall the definition of being (ϵ, δ)-DP (Dwork et al., 2014). The idea behind differential privacy is to ensure that no attacker can determine with high confidence whether a particular individual's data is included in the dataset or not. Often, a controlled amount of random noise is added to the data, so that any individual data point becomes indistinguishable from the noise. This ensures that the probability distribution of an algorithm's output does not change significantly when a single individual's data is added or removed.

Definition 5.45. *For any $\epsilon > 0$ and $\delta \in [0, 1)$, a randomized mechanism $\mathcal{A}$ is said to be (ϵ, δ)-DP, if for all neighboring datasets D, D', and for any event E:*

$$\mathbb{P}\left(\mathcal{A}(D) \in E\right) \leq \exp(\epsilon)\,\mathbb{P}\left(\mathcal{A}(D') \in E\right) + \delta.$$

The following result gives the noise level sufficient to ensure the (ϵ, δ)-DP regarding the third-party attacker. This type of attacker is an external entity who does not have access to the private data but can observe the algorithm outputs. This attacker tries to infer sensitive information about individuals by analyzing the output.

Theorem 5.46. *Assume there is a constant number $S \in [n]$ of sampled agents, i.e., $S_t = S$, for all $t \in [T]$. Then, for all $\epsilon > 0$ and $\delta \in (0, 1 - (1 + \sqrt{\epsilon})(1 - S/n)^T)$, the Algorithm 5.10 is (ϵ, δ)-DP towards a third party if*

$$\sigma_g \geq 2\sqrt{\frac{K \max_{i\in[n]} \lambda^i_{\mathbf{y}}}{\epsilon}\left(1 + \frac{24S\sqrt{T}\log(1/\bar{\delta})}{\epsilon n}\right)}, \quad \bar{\delta} = \frac{n}{S}\left[1 - \left(\frac{1-\delta}{1+\sqrt{\epsilon}}\right)^{1/T}\right].$$

Proof The loss function $\nabla S_\alpha^{i,\gamma}$ has a sensitivity of 1. Therefore, for any $\tilde{\alpha} > 1$, we know that $\nabla S_\alpha^{i,\gamma} + \mathcal{N}(0, \sigma_g^2)$ is $(\tilde{\alpha}, \tilde{\alpha}/2\sigma_g^2)$-RDP (Mironov, 2017, Corollary 3). By assumption, note that $\bar{\delta} \in (0, 1)$ and for any $t \in [T]$, consider

$$\epsilon_t = \frac{nK\tilde{\alpha}\max_{i\in[n]} \lambda^i_{\mathbf{y}}}{2(t+1)S\sigma_g^2} - \frac{\log\bar{\delta}}{\tilde{\alpha} - 1},$$

$$\tilde{\alpha} = 1 + \sigma_g\sqrt{\frac{2\sqrt{T}S\log(1/\bar{\delta})}{nK\max_{i\in[n]} \lambda^i_{\mathbf{y}}}}.$$

After K local iterations, using the RDP composition result, the mechanism becomes $(\tilde{\alpha}, K\tilde{\alpha}/2\sigma_g^2)$-RDP. Using the aggregation step on the server, the mechanism is now $(\tilde{\alpha}, nK\tilde{\alpha}\max_{i\in[n]}\lambda_{\mathbf{y}}^i/2(t+1)S\sigma_g^2)$-RDP. Based on the RDP to DP conversion, we know that the mechanism is $(\epsilon_t, \bar{\delta})$-DP. Define $f\colon x \in \mathbb{R} \mapsto \log\{1+(S/n)(\mathrm{e}^x-1)\} \in \mathbb{R}$. For any $x \in \mathbb{R}$, we have

$$f'(x) = \frac{S}{n + 2x\left(n-S\right)}.$$

In addition, since the agents are subsampled, it yields the $(\tilde{\epsilon}_t, \tilde{\delta}_t)$-DP (Balle et al., 2018, Theorem 9), where we denote

$$\tilde{\epsilon}_t = f\left(\epsilon_t\right), \qquad \tilde{\delta}_t = \frac{S\bar{\delta}}{n}.$$

Since $f'(x) \leq S/n$ on $\mathbb{R}_+$, we deduce that

$$\tilde{\epsilon}_t \leq \frac{K\tilde{\alpha}\max_{i\in[n]}\lambda_{\mathbf{y}}^i}{2(t+1)\sigma_g^2} - \frac{S\log\bar{\delta}}{n\left(\tilde{\alpha}-1\right)}.$$

For all $a, b \in \mathbb{R}$, using that $(a+b)^2 \leq 2a^2 + 2b^2$ gives that

$$\sum_{t=1}^{T}\tilde{\epsilon}_t^2 \leq \frac{2\left(K\tilde{\alpha}\max_{i\in[n]}\lambda_{\mathbf{y}}^i\right)^2}{4\sigma_g^4}\sum_{t=2}^{T+1}\frac{1}{t^2} + \frac{2TS^2\log(1/\bar{\delta})^2}{n^2(\tilde{\alpha}-1)^2}. \tag{5.87}$$

Plugging the definition of $\tilde{\alpha}$ into (5.87) combined with $\sum_{t=2}^{T+1} t^{-2} \leq 1$ show that

$$\begin{aligned}\sum_{t=1}^{T}\tilde{\epsilon}_t^2 &\leq \frac{\left(K\max_{i\in[n]}\lambda_{\mathbf{y}}^i\right)^2}{\sigma_g^4}\left(1 + \frac{3\sigma_g^2\sqrt{T}S\log(1/\bar{\delta})}{nK\max_{i\in[n]}\lambda_{\mathbf{y}}^i}\right) \\ &= \frac{\left(K\max_{i\in[n]}\lambda_{\mathbf{y}}^i\right)^2}{\sigma_g^4} + \frac{3\sqrt{T}KS\log(1/\bar{\delta})\max_{i\in[n]}\lambda_{\mathbf{y}}^i}{n\sigma_g^2}. \end{aligned} \tag{5.88}$$

By assumption, recall that

$$\sigma_g \geq 2\sqrt{\frac{K\max_{i\in[n]}\lambda_{\mathbf{y}}^i}{\epsilon}}\sqrt{1 + \frac{24\sqrt{T}S\log(1/\bar{\delta})}{\epsilon n}}.$$

Therefore, we obtain

$$\frac{\sigma_g^4\epsilon^2}{16} \geq \left(K\max_{i\in[n]}\lambda_{\mathbf{y}}^i\right)^2 + \frac{3\sigma_g^2\sqrt{T}SK\log(1/\bar{\delta})\max_{i\in[n]}\lambda_{\mathbf{y}}^i}{n}.$$

The previous inequality combined with (5.88) implies that

$$\sum_{t=1}^{T}\tilde{\epsilon}_t^2 \leq \epsilon^2/16.$$

Define the following quantity

$$\tilde{\delta} = \frac{1-\delta}{\prod_{t=1}^{T}(1-\tilde{\delta}_t)} - 1.$$

Since $1-(1-\delta)^{1/T} \leq \tilde{\delta}_t \leq 1-[(1-\delta)/2]^{1/T}$, we can verify that $\tilde{\delta} \in [0,1]$ and also $\tilde{\delta} = \sqrt{\epsilon}$. Using the assumption $\delta \leq 1-(1+\sqrt{\epsilon})(1-S/n)^T$, it yields

$$\mathrm{e} + \tilde{\delta}^{-1}\sqrt{\sum_{t=1}^{T}\epsilon_t^2} \leq \mathrm{e}+1 \leq \mathrm{e}^2.$$

Thus, we have

$$2\sum_{t=1}^{T}\tilde{\epsilon}_t^2 \log\left(\mathrm{e} + \tilde{\delta}^{-1}\sqrt{\sum_{t=1}^{T}\epsilon_t^2}\right) \leq \frac{\epsilon^2}{4}.$$

Since $(\exp(\tilde{\epsilon}_t)-1)/(\exp(\tilde{\epsilon}_t)+1) \leq \tilde{\epsilon}_t$, we deduce that

$$\sum_{t=1}^{T}\frac{\exp(\tilde{\epsilon}_t)-1}{\exp(\tilde{\epsilon}_t)+1} \leq \sum_{t=1}^{T}\tilde{\epsilon}_t^2 \leq \frac{\epsilon^2}{16}.$$

Finally, applying (Kairouz et al., 2015, Theorem 3.5) concludes the proof. ■

Note that the local loss function $\mathrm{S}_\alpha^{i,\gamma}\colon q \in \mathbb{R} \mapsto \mathbb{E}_{V\sim\widehat{\mu}_y^i}[S_{\alpha,V}^\gamma(q)] \in \mathbb{R}_+$ is expressed as the expectation of pinball loss functions. Since the sensitivity of these pinball loss functions is 1, we do not need to clip the gradient. It is sufficient to add additional Gaussian noise $\mathcal{N}(0,\sigma_g)$ to guarantee differential privacy. The value of σ_g is chosen to provide a suitable tradeoff between privacy and utility, balancing the need for strong privacy protection with the requirement for useful output.

5.E Additional numerical results

5.E.1 Algorithm design

The objective is to generate valid federated prediction sets for the testset by leveraging the calibration datasets of the agents. In this section, we present four different algorithms that were compared in our experiments. The first two are unweighted algorithms (`Unweighted Local` and `Unweighted Global`), while the other two are weighted benchmarks: `Oracle Weights` which uses true weights, and `Estimated Weights` which employs estimated weights. Additionally, we propose the `DP-FedCP` method, available in both basic and differentially private versions. The main differences between these approaches lie in the way the resulting quantile is computed, such as the importance given to the set of non-conformity scores and their corresponding weights. Moreover, the approaches' ability to maintain coverage and privacy guarantees varies.

Unweighted Local. The Unweighted method assigns equal weight to all non-conformity scores, regardless of the agent or label. As a result, the resulting quantile and prediction sets are only influenced by the local scores of the querying agent. Specifically, the `Unweighted Local` approach only employs the local calibration dataset of Agent $\star$ to compute the corresponding prediction set. This approach is easily computable

because it does not involve any exchange of information between agents. Formally, the confidence set can be expressed as follows:

$$\bar{\mu}^{\mathrm{loc},\star} = \frac{1}{N^\star + 1} \sum_{k=1}^{N^\star} \delta_{V_k^\star} + \frac{1}{N^\star + 1} \delta_1,$$
$$\mathcal{C}_{\alpha,\bar{\mu}^{\mathrm{loc},\star}}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\bar{\mu}^{\mathrm{loc},\star}\right) \right\}.$$

Unweighted Global. The `Unweighted Global` approach calculates the quantile by using all non-conformity scores gathered on the central server from all agents. This approach violates FL constraints since all non-conformity scores are shared with the central server. Furthermore, each collected score is given equal weight, without taking into account any label shift. The corresponding prediction set is represented as follows:

$$\bar{\mu} = \frac{1}{N + 1} \sum_{i=1}^{n} \sum_{k=1}^{N^i} \delta_{V_k^i} + \frac{1}{N + 1} \delta_1,$$
$$\mathcal{C}_{\alpha,\bar{\mu}}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\bar{\mu}\right) \right\}.$$

Oracle Weights. This approach utilize importance weights to compute the prediction set. The `Oracle Weights` method has access to the true distribution of all the agents, enabling it to calculate the exact likelihood ratios. For this method, a number of $\bar{N} = \lfloor N/2 \rfloor$ data points of the calibration dataset is randomly selected — denoted as $(X_k, Y_k)_{k \in [\bar{N}]}$. This sumbsampling is based on a multinomial random variable with parameter $(\bar{N}, \{N^i/N\}_{i \in [n]})$; more details are provided in Section 5.2 (see for example Theorem 5.3). For any label $y \in \mathcal{Y}$, the prediction set is determined by:

$$\begin{aligned} \bar{\mu}_{\mathbf{y}}^\star &= \bar{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k,\mathbf{y}}^\star \delta_{V_k}, \\ \mathcal{C}_{\alpha,\bar{\mu}^\star}(\mathbf{x}) &= \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu}_{\mathbf{y}}^\star) \right\}. \end{aligned} \tag{5.89}$$

Estimated Weights. The empirical equivalent of `Oracle Weights` based on client label counts is `Estimated Weights`. However, two sources of error are introduced: (1) the calibration subsampling and (2) the likelihood ratio estimations (refer to (5.11)). Similar to `Oracle Weights`, we draw a multinomial distribution $\{\bar{N}^i\}_{i \in [n]} \sim \mathcal{M}(\bar{N}, \{N^i/N\}_{i \in [n]})$ and subsample $N^i \wedge \bar{N}^i$ calibration data from each client i. The resulting prediction set is represented by:

$$\widehat{p}_{y,\mathbf{y}}^\star = \frac{(\mathsf{M}_y^\star / \mathsf{M}_y) \cdot \mathbb{1}_{\mathsf{M}_y \geq 1}}{(\mathsf{M}_{\mathbf{y}}^\star / \mathsf{M}_{\mathbf{y}}) \cdot \mathbb{1}_{\mathsf{M}_{\mathbf{y}} \geq 1} + \sum_{\tilde{y} \in \mathcal{Y}} \mathrm{Card}(\{(i,k) \in [n] \times [N^i] \colon k \leq \bar{N}^i, Y_k^i = y\}) \times (\mathsf{M}_{\tilde{y}}^\star / \mathsf{M}_{\tilde{y}}) \cdot \mathbb{1}_{\mathsf{M}_{\tilde{y}} \geq 1}},$$
$$\widehat{\mu}_{\mathbf{y}}^{\text{MLE}} = \widehat{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i} \widehat{p}_{Y_k^i,\mathbf{y}}^\star \delta_{V_k^i},$$
$$\mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(\mathbf{x}) = \left\{ \mathbf{y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}^{\text{MLE}}\right) \right\}.$$

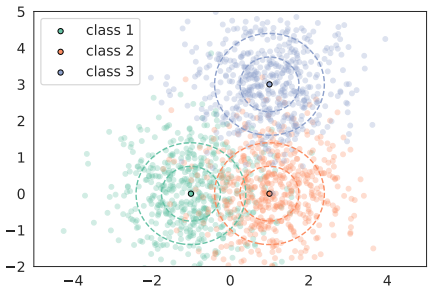


Figure 5.4 – Simulated data distribution: 3 two-dimensional Gaussians with means $\boldsymbol{\theta}_1 = [-1, 0], \boldsymbol{\theta}_2 = [1, 0], \boldsymbol{\theta}_3 = [1, 3]$ and identity covariance matrices.

5.F.2 Additional Details on Numerical Experiments

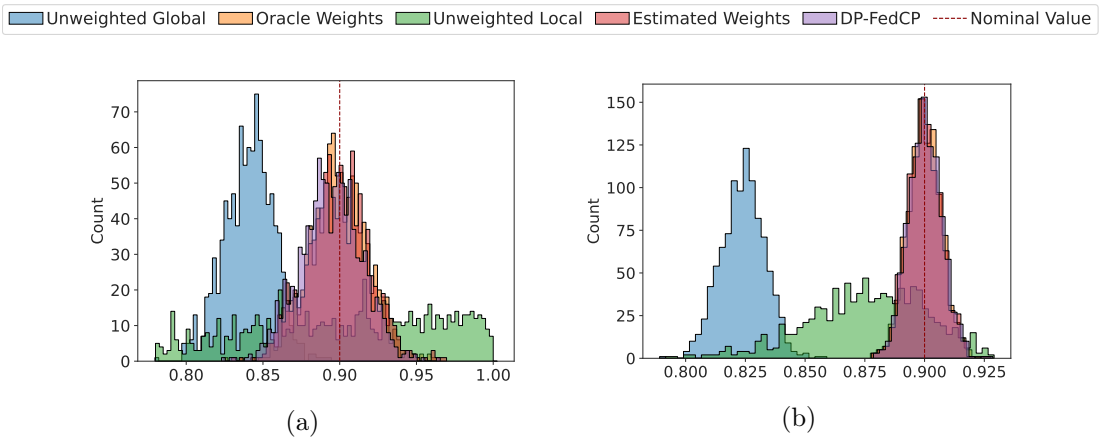
To provide a comprehensive overview of our experimental methodology, we present additional details about the federated optimization parameters, along with supplementary figures to complement those presented in Section 5.5.

Optimization parameters. For all experiments, we split the initial dataset $\mathcal{D}$ across the clients $(\mathcal{D}_i)_{i\in[n]}$ and the test dataset $\mathcal{D}_{\text{test}}$ as detailed in Section 5.5. Label shift is then simulated by resampling using the clients' local distributions $\left\{P^y\right\}_{y\in\mathcal{Y}}$. For ImageNet experiments, we also assume that we have labels sampled from the target client's distribution for weights' approximation. The optimization parameters taken for `DP-FedCP` experiments are $T = 200$ iterations, $\gamma = 1e^{-6}$ regularization parameter, $\eta = 1e^{-3}$ step-size, and $K = 20$ local iteration rounds. We sample all clients during each communication round with the server.

Simulated Data Experiments. The generated data consists of 3 Gaussians, two of which significantly overlap each other, see Figure 5.4. This data design is chosen such that we obtain different distributions of non-conformity scores for different classes, which is directly related to the different degrees of model confidence for different data samples.

CIFAR-10 Experiments. Using the CIFAR-10 data, we demonstrate a comparison of the empirical coverage for all considered methods: `Unweighted Local`, `Unweighted Global`, `Oracle Weights`, `Estimated Weights` and the proposed `DP-FedCP` method version with $(\sigma_g = 0)$, see Figure 5.5a. `DP-FedCP` along with weighted baselines, shows valid coverage results, unlike unweighted baselines. At the same time, both weighted algorithms are extremely similar in performance to `DP-FedCP`. Unlike weighted baselines, `DP-FedCP` is federated and privacy preserving.

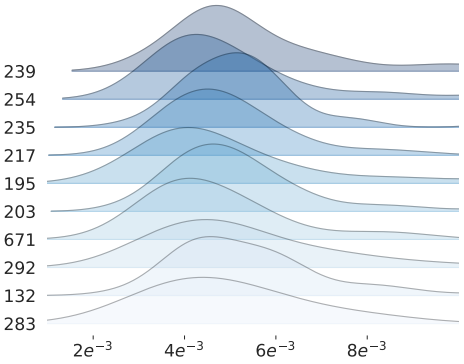
ImageNet Experiments. The ImageNet experiment is designed to have very different score distributions across agents. The grouping scheme of clients into a `low-score` group (Figure 5.6a) and a `high-score` group that consists of the querying agent (Figure 5.6b) creates adversial heterogeneity, possible in real-life scenarios, under which unweighted methods are more prone to perform very poorly. Comparing the `DP-FedCP` method with weighted and unweighted baselines on ImageNet data, we note the same



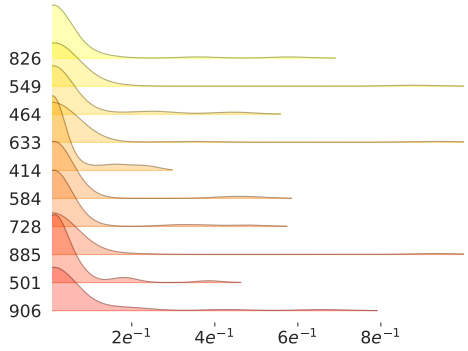
(a) (b)

Figure 5.5 – Experimental results with all benchmarks. (a) CIFAR-10 empirical coverage. (b) ImageNet empirical coverage.

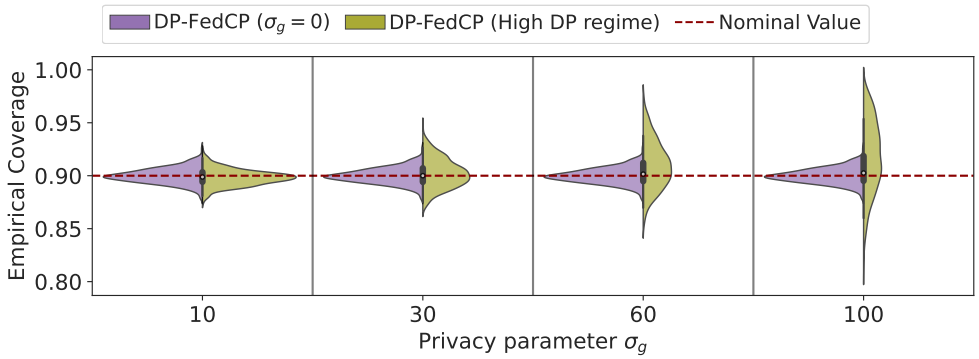
behavior as on CIFAR-10 dataset, see Figure 5.5b. Only algorithms that account for shifts between agents' data achieve the desired empirical coverage. There is an additional violin plot for the ImageNet differential privacy study that demonstrates the effect of the DP parameter σ_g on the resulting empirical coverage; see Figure 5.6c.



(a) Distribution of scores for labels in G1



(b) Distribution of scores for labels in G2



(c) Privacy-Utility tradeoff

Figure 5.6 – ImageNet complementary results. (a) Score distributions of the classes with the lowest non-conformity scores. (b) Score distributions of the classes with the highest non-conformity scores. (c) Violin plot of the empirical coverage distribution for `DP-FedCP`with different DP regimes.

Chapter 6

Conclusion and Perspectives

The final chapter summarizes the main findings of the research, highlighting the contributions to address the research questions. It provides a throughout discussion about limitations with some suggestions for future works.

6.1 Conclusion

We explored various machine learning aspects, with a particular focus on Bayesian methods and uncertainty quantification within the FL framework. Throughout the chapters, we identified key themes and proposed novel methodologies, aiming to improve the reliability and efficiency of machine learning algorithms.

Chapters 2 and 3 on Bayesian FL, addressed the challenges of reliable Bayesian inference with distributed datasets. In these chapters, we designed novel approaches to overcome communication bottlenecks and statistical heterogeneity. Our methods involved multiple local Langevin steps with a combination of global consensus updates performed at the central node. In `DG-LMC`, we explored a general sampling consensus step, while `VR-FALD`*'s specificity lied in its variance reduction scheme performed on the local agents. These developed methods showcased favorable convergence properties, establishing their efficiency. We also provided theoretical analyses and non-asymptotic bounds to support their performance.

Chapters 4 and 5 presented methodologies to address the crucial uncertainty quantification aspect. We focused on Bayesian approaches through `QLSD`. This algorithm provided a comprehensive methodology based on Langevin stochastic dynamics for Bayesian FL inference. It effectively handles communication bottlenecks through gradient compression and incorporates variance reduction techniques. However, (1) designing effective prior distributions for deep learning remains challenging and (2) the validity of the Bayesian model can be criticized. To complement uncertainty management, we also explored a frequentist method for constructing personalized prediction sets. We leveraged conformal prediction approaches to provide distribution-free guarantees under minimal assumptions. This methodology addresses the issue of distribution shift, which is crucial when the underlying data distribution varies across agents.

Our works open up promising avenues for future research by addressing the distributed sampling challenge. Furthermore, our research has provided comprehensive approaches for uncertainty quantification in FL by investigating both Bayesian and frequentist methodologies. Our research contributes to the development of methodologies to handle uncertainty while addressing distribution shift. Overall, our work brings us closer to reliable and efficient FL systems and promising avenues have been outlined for future researches.

6.2 Perspectives and Future work

Some avenues for further research are presented for all the different chapter and associated research questions of this thesis.

Chapter 2. This chapter addressed the crucial task of performing reliable Bayesian inference in the modern era of machine learning. Markov chain Monte Carlo (MCMC) algorithms have proven to be invaluable in large scale sampling, but designing them to handle distributed datasets has remained challenging. Existing methods often failed in terms of reliability and sampling quality. To bridge this gap, we proposed a novel approach called `DG-LMC`, specifically tailored for scenarios where datasets are partitioned on computing nodes under a master/slaves architecture. The convergence properties of `DG-LMC` have been rigorously analyzed, establishing its efficacy for reliable Bayesian inference. With its favorable convergence properties and empirical validation, future research directions may involve exploring extensions and refinement of the Gaussian consensus step performed by the central server.

Chapter 3. We addressed the challenges of Bayesian FL inference, considering communication bottlenecks and the statistical heterogeneity limitation. While distributed MCMC algorithms have been extensively studied, they are badly designed to handle FL-specifications. Our findings illustrated that statistical heterogeneity leads to a local drift, negatively affecting convergence. To overcome this issue, we presented a novel algorithm, `VR-FALD`*, which incorporates control variates to effectively mitigate the client drift and reduce stochastic gradient variance. Our approach combines ideas from Langevin Monte Carlo and Federated Averaging schemes. Importantly, our theoretical analysis encompasses a unified framework for Bayesian FL, including connections with the global consensus Monte Carlo method. Based on this theoretical framework, we have derived non-asymptotic bounds for both `FALD` and `VR-FALD`*. We also demonstrated the significance of variance reduction for statistical heterogeneity. Future research directions may involve further results on non-convex potential landscapes with other control variate schemes and investigating practical guidelines to improve FL performance.

Chapter 4. In this chapter, we presented a comprehensive methodology based on Langevin stochastic dynamics for Bayesian FL. We performed statistical inference on locally stored data across multiple clients while considering FL constraints. To overcome these challenges, we introduced a novel federated MCMC algorithm, named `QLSD`, extending the SGLD algorithm to the FL setting. `QLSD` efficiently handles the communication bottleneck through gradient compression. Furthermore, to enhance performance, we incorporated variance reduction techniques, resulting in two improved versions, namely `QLSD`* and `QLSD`$^{++}$. Our proposed algorithms were supported by non-asymptotic convergence guarantees, providing a solid foundation. Future research can incorporate Metropolis-Hastings schemes to remove asymptotic biases. More efficient communication can be pursued by alternative compression techniques such as biased or vectorial compressions.

Chapter 5. This chapter introduced a novel method called `DP-FedCP` for constructing personalized conformal prediction sets within the FL setting. Our proposed method leverages CP approaches to provide distribution-free guarantees under minimal assumptions. This work addresses the crucial aspect of uncertainty quantification, by incorporating importance weighting to effectively handle label shift, resulting in improved prediction sets with prescribed confidence This method provides personalized conformal prediction sets with valid coverage and differential privacy guarantees. To support our methodology, we developed non-asymptotic bounds and discussed the parameters' tuning to achieve an optimal accuracy/privacy tradeoff. Future research directions may involve further investigations into different types of distribution shifts and new methods for shift estimation.

Résumé des contributions

Motivée par les questions de recherche (RQ) mentionnées précédemment, cette thèse apporte plusieurs contributions, qui sont détaillées dans la section suivante. Chaque chapitre se concentre sur une direction de recherche spécifique et aborde les domaines clés suivants :

- ✦ Développement de méthodes avancées d'échantillonnage distribué ciblant une distribution postérieure globale.
- ✦ Construction de méthodes de simulation efficaces en grande dimension, pour des distributions connues à une constante de normalization près.
- ✦ Application de méthodes d'inférence approximative pour l'apprentissage profond.
- ✦ Développement de méthodes de gestion fédérée de l'incertitude, qui se basent sur des prédictions conformelles.

Chaque chapitre examine en détail l'une de ces principales orientations de recherche.

Part II: Echantillonnage distribué & Langevin MC

• Chapter 2: *`DG-LMC`: Un algorithme MCMC distribué synchrone clé en main et scalable via l'échantillonnage de Langevin Monte Carlo dans le cadre de Gibbs* (RQ#3-RQ#4)

Dans ce travail, nous proposons un algorithme d'échantillonnage efficace, qui s'adapte aux architectures centralisées. Notre méthode se concentre spécifiquement sur l'inférence bayésienne à partir des ensembles de données $\{\mathcal{D}_i\}_{i=1}^n$ observés sur n nœuds. Nous développons une procédure pour approcher les distributions a posteriori admettant une densité donnée par

$$\pi(\theta|\mathcal{D}_{1:n}) \propto \prod_{i=1}^{n} \exp(-U_i(\theta)), \tag{6.1}$$

où la fonction potentielle $U_i \colon \mathbb{R}^{d_i} \to \mathbb{R}$ dépend de l'ensemble d'apprentissage $\mathcal{D}_i$. L'idée centrale de notre nouvelle méthodologie, appelée Distributed Gibbs using Langevin Monte Carlo (`DG-LMC`), consiste à concevoir une distribution jointe $\Pi_{\boldsymbol{\rho}}$ avec des variables auxiliaires $z_1 \in \mathbb{R}^{d_1}, \ldots, z_n \in \mathbb{R}^{d_n}$ satisfaisant

$$\Pi_\rho(\mathcal{D}_{1:n}|z_{1:n},\theta) \propto \prod_{i=1}^{n} \Pi_\rho(\mathcal{D}_i|z_i), \qquad \Pi_\rho(z_{1:n}|\theta) = \prod_{i=1}^{n} \Pi_\rho(z_i|\theta), \tag{6.2}$$

où $\rho > 0$ est un paramètre de tolérance tel que $\lim_{\rho\to 0} \Pi_{\boldsymbol{\rho}}(\theta|\mathcal{D}) = \pi(\theta|\mathcal{D})$. Travailler avec Π_ρ présente un avantage significatif : les variables auxiliaires $\{z_i\}_{i=1}^n$ sont conditionnellement indépendantes étant donné θ. Par conséquent, en utilisant (6.2), on

obtient la décomposition suivante :

$$\begin{aligned}\Pi_\rho(\theta|\mathcal{D}_{1:n}) &= \int \Pi_\rho(\theta, z_{1:n}|\mathcal{D}_{1:n})\,\mathrm{d}z_{1:n} \\ &= \frac{1}{\Pi_\rho(\mathcal{D}_{1:n})}\int \Pi_\rho(\theta, z_{1:n})\Pi_\rho(\mathcal{D}_{1:n}|\theta, z_{1:n})\,\mathrm{d}z_{1:n} \\ &= \frac{1}{\Pi_\rho(\mathcal{D}_{1:n})}\int \Pi_\rho(\theta)\prod_{i=1}^{n}\left[\Pi_\rho(\mathcal{D}_i|z_i)\Pi_\rho(z_i|\theta)\right]\,\mathrm{d}z_{1:n}.\end{aligned}$$

En utilisant l'échantillonneur Gibbs, la distribution $\Pi_\rho(\theta, z_{1:n}|\mathcal{D}_{1:n})$ peut être échantillonnée efficacement en parallèle sans avoir besoin de transmettre de donnée.

Contributions. Les principales contributions peuvent être résumées comme suit :

(1) Nous introduisons une nouvelle méthodologie appelée Distributed Gibbs using Langevin Monte Carlo (`DG-LMC`) dans la Section 2.2. Cet algorithme demande à chaque travailleur d'échantillonner z_i à partir de la distribution conditionnelle $\Pi_\rho(z_i|\mathcal{D}_i, \theta)$ et de communiquer cet échantillon au nœud maître. Ensuite, le nœud central échantillonne θ selon $\Pi\rho(\theta|z_{1:n})$ et renvoie ce paramètre à chaque travailleur.

(2) Point important, nous présentons une analyse quantitative complète du biais induit et démontrons des résultats de convergence explicites dans la Section 2.3. Cela représente notre principale contribution, et il me semble que cette étude théorique est l'une des plus complètes parmi les travaux existants en apprentissage bayésien distribué avec une architecture centralisée. Plus précisément, nous discutons de la complexité de l'algorithme, de la sélection des hyperparamètres et offrons aux praticiens des lignes directrices simples pour les ajuster. De plus, nous effectuons une comparaison approfondie de notre méthode avec des approches existantes dans la Section 2.4.

(3) Enfin, dans la Section 2.5, nous démontrons les avantages de l'échantillonneur proposé par rapport aux algorithmes MCMC distribués populaires et récents à travers diverses expériences numériques.

Deux défis majeurs subsistent : l'échantillonnage efficace à partir de la distribution conditionnelle $\Pi_{\boldsymbol{\rho}}(z_i|\theta, \mathcal{D}_i)$ pour $i \in [n]$, et la réduction des cycles de communication fréquents avec le nœud maître. Nous abordons ces deux problèmes en utilisant l'algorithme Langevin Monte Carlo (LMC) pour approcher l'échantillonnage à partir de $\Pi_{\boldsymbol{\rho}}(z_i|\theta, \mathcal{D}_i)$ (Rossky et al., 1978; Roberts and Tweedie, 1996). Pour $i \in [n]$, nous introduisons $\Pi\rho$ dont les densités conditionnelles sont définies comme suit :

$$\begin{aligned}\Pi_{\boldsymbol{\rho}}(z_i|\mathcal{D}_i, \theta) &\propto \exp\left(-U_i(z_i) - \left\|z_i - \theta\right\|^2/(2\rho_i)\right), \\ \Pi_{\boldsymbol{\rho}}(\theta|z_{1:n}) &= \mathcal{N}\left(\boldsymbol{\mu}(z_{1:n}), \mathbf{Q}^{-1}\right)\end{aligned}$$

où la matrice de précision $\mathbf{Q} = (\sum_{i=1}^n {\rho_i}^{-1})\mathbf{I_d}$ et la moyenne $\boldsymbol{\mu}(z_{1:n}) = \mathbf{Q}^{-1}\sum_{i=1}^n z_i/\rho_i$. Lorsque le paramètre de tolérance $\rho \to 0$, en utilisant (Scheffé, 1947), on montre que ce schéma d'augmentation de données satisfait

$$\lim_{\rho\to 0}\Pi_{\boldsymbol{\rho}}(\theta|\mathcal{D}) = \lim_{\rho\to 0}\int \Pi_{\boldsymbol{\rho}}(\theta, z_{1:n})\mathrm{d}z_{1:n} = \pi(\theta|\mathcal{D}).$$

Basé sur l'équation différentielle stochastique de Langevin sur-amortie, à l'itération k, nous mettons à jour les paramètres comme suit :

$$z_i^{(k+1)} = \left(1 - \frac{\gamma_i}{\rho_i}\right) z_i^{(k)} + \frac{\gamma_i}{\rho_i}\theta^{(k)} - \gamma_i \nabla U_i(z_i^{(k)}) + \sqrt{2\gamma_i}\xi_i^{(k)},$$
$$\theta^{(k+1)} = \boldsymbol{\mu}(z_{1:n}^{(k)}) + \mathbf{Q}^{-1/2}\xi_0^{(k)} \text{ pendant les tours de communcation, autrement } \theta^{(k)},$$

où $\gamma_i > 0$ est une taille de pas fixe et $\{\xi_i^{(k)} : i \in [n], k \in \mathbb{N}\}$ est une séquence i.i.d. de variables aléatoires gaussiennes standard. Pour réduire les coûts de communication, nous permettons à chaque travailleur d'effectuer $N_i \geq 1$ étapes LMC locales (Dieuleveut and Patel, 2019). La variation de N_i entre les travailleurs empêche `DG-LMC` de subir des retards importants dus aux temps de réponse déséquilibrés des travailleurs (Ahn et al., 2014). Nous fournissons une analyse quantitative détaillée du biais et établissons des résultats de convergence explicites non asymptotiques. Notre analyse englobe la complexité de `DG-LMC`, la sélection des hyperparamètres, et offre aux praticiens des directives simples pour les ajuster. À notre connaissance, cette étude théorique est l'une des études les plus complètes sur l'apprentissage machine bayésien distribué avec une architecture maître/esclave.

Theorem 6.1 (Informel)**.** *Sous certaines hypothèses décrites dans le Chapter 2, il existe* $\kappa \in (0,1)$*,* $\gamma, \rho, C_0, C_1, C_2 > 0$ *tels que pour* $k \geq 0$*, la distribution* μ_k *de l'échantillon* θ_k *satisfait*

$$W_2\Big(\mu_k, \pi(\cdot|\mathcal{D}_{1:n})\Big) \leq C_0(1-\kappa)^k + C_1\sqrt{d\gamma(\rho^2 + \gamma/\rho^2)} + C_2 d\rho.$$

• Chapter 3: ***FALD:*** *Dynamique de Langevin avec moyenne fédérée* (RQ#1)

Dans ce chapitre, nous nous intéressons à l'échantillonnage à partir d'une distribution cible π dont la densité peut être décomposée comme dans l'équation (6.1). Pour résoudre ces problèmes, nous proposons un algorithme MCMC appelé `FALD`, qui combine la dynamique stochastique de gradient de Langevin (SGLD) avec l'idée de la moyennisation fédérée.

Contributions. Les principales contributions peuvent être résumées comme suit :

(1) Nous étudions une version en boucle aléatoire de l'algorithme `FALD` proposé dans Deng et al. (2021), et nous établissons des bornes supérieures non asymptotiques en distance de Wasserstein pour les potentiels fortement convexes U. Une analyse de `FALD` a été réalisée dans Deng et al. (2021, Theorème 5.7), cependant, la preuve est entachée d'une erreur ; voir Section 3.B.1.

(2) Nous donnons des bornes inférieures correspondantes pour montrer que même avec des gradients exacts, `FALD` peut être plus lent que `SGLD` en raison de la dérive des clients.

(3) Nous proposons une nouvelle méthode (`VR-FALD`*) qui contourne les limites de `FALD`. Cet algorithme étend la méthode Shifted Local-SVRG de Gorbunov et al. (2021) au contexte bayésien. `VR-FALD`* combine la méthode Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD) (Dubey et al., 2016) et adapte les techniques de réduction des biais de `SCAFFOLD` (Karimireddy et al., 2020).

(4) Nous obtenons des garanties théoriques pour `VR-FALD`* qui mettent en évidence son effet de réduction de la variance du gradient et sa capacité à traiter l'hétérogénéité des données.

(5) Les résultats sont basés sur un cadre général développé dans le supplément, qui englobe une large famille d'algorithmes bayésiens fédérés basés sur la dynamique de Langevin. Il s'agit de la première étude unificatrice parmi les travaux existants sur l'inférence bayésienne fédérée.

(6) Enfin, Section 3.4 illustre nos résultats sur des benchmarks classiques de FL et fournit une comparaison approfondie avec les méthodes bayésiennes FL existantes.

L'algorithme `FALD` échantillonne à partir de π en respectant une contrainte majeure : chaque potentiel U_i et son gradient ∇U_i ne peuvent être calculés que par le i-ème client. Dans cette méthode, chaque client possède un paramètre θ_k^i qui est mis à jour localement tandis que les paramètres globaux θ_k^s sont mis à jour sur le serveur central. À chaque tour, les clients exécutent des étapes de `SGLD` pour mettre à jour leurs paramètres locaux

$$\tilde{\theta}_{k+1}^i = \theta_k^i - \gamma \nabla U^i(\theta_k^i) + \sqrt{2\gamma} Z_{k+1}^i,$$

où Z_{k+1}^i est un vecteur gaussien de dimension d éventuellement corrélé entre les clients. Chaque client envoie $\tilde{\theta}_{k+1}^i$ au serveur central avec une probabilité $p_c \in \big(0, 1\big]$ correspondant à la réalisation d'une variable de Bernoulli B_{k+1}. Pendant les tours de communication, le serveur central fait la moyenne des paramètres reçus

$$\theta_{k+1}^s = (B_{k+1}/n) \sum_{i \in [n]} \tilde{\theta}_{k+1}^i + (1 - B_{k+1}) \theta_k^s.$$

Ensuite, ce paramètre du serveur θ_{k+1}^s est renvoyé aux clients locaux qui mettent à jour leurs paramètres locaux θ_k^i de la manière suivante

$$\theta_{k+1}^i = B_{k+1} \theta_{k+1}^s + (1 - B_{k+1}) \tilde{\theta}_{k+1}^i.$$

Comme indiqué dans le Theorem 1.2, les échantillons $\{\theta_k^s\}_{k \in \mathbb{N}}$ générés par le serveur central ciblent la distribution a posteriori π. Des explications supplémentaires sur les bornes de convergence de `FALD` sont fournies dans le Chapter 3. Bien que théoriquement solide, cette méthode peut souffrir d'une variance élevée en raison des gradients stochastiques utilisés lors du `SGLD` local et de l'hétérogénéité des données, ce qui entrave la convergence. Plus précisément, nous montrons l'impossibilité pour un algorithme qui ne traite pas l'hétérogénéité de fournir une erreur asymptotique de Wasserstein inférieure à la taille de discrétisation $\mathrm{O}(\gamma)$. Pour résoudre ce problème, nous proposons une alternative : `VR-FALD`* basée sur une combinaison de variables de contrôle et de techniques de réduction des biais. Des améliorations théoriques sont déduites et les comportements expérimentaux de nos algorithmes sont présentés.

Theorem 6.2 (Informel)**.** *Sous les hypothèses décrites dans le Chapter 4, il existe $\gamma\star > 0$, tel que pour $\gamma \in (0, \gamma_\star)$, il existe $\kappa \in (0, 1)$, $C_0, C_1, C_2, C_3 > 0$ tels que pour $k \geq 0$, la distribution μ_k de l'échantillon θ_k satisfait*

$$W_2^2\Big(\mu_k, \pi(\cdot|\mathcal{D}_{1:n})\Big) \leq (1-\kappa)^k C_0 + \gamma C_1 \mathbb{E}\Big(\textstyle\sum_{i=1}^n \widehat{U}_i(\theta_\star)\Big) + \frac{\gamma^2 C_2}{p_c^2} \sum_{i=1}^n \|\nabla U_i\left(\theta_\star\right)\|^2 + \gamma^2 C_3.$$

Part III: Apprentissage fédéré : Quantification de l'incertitude via des approches bayésiennes et fréquentistes

• Chapter 4: *QLSD: Dynamique stochastique quantifiée de Langevin pour l'apprentissage fédéré bayésien* (RQ#2)

Plusieurs travaux ont tenté d'améliorer l'efficacité de l'apprentissage distribué/fédéré en réduisant le coût de communication. Certaines méthodes se sont concentrées sur la quantification de chaque coordonnée des gradients calculés (Alistarh et al., 2017), de sorte qu'un nombre beaucoup plus faible de bits soit nécessaire pour la transmission. Des quantifications agressives, telles que la représentation binaire ou ternaire, ont également été étudiées. D'autres méthodes ont imposé la parcimonie aux gradients lors de la communication, où seulement une petite fraction des gradients est échangée entre les nœuds à chaque itération. Les idées sous-jacentes de ces méthodes consistent essentiellement à compresser les gradients, où chaque entrée peut être représentée par beaucoup moins de bits que le nombre flottant 32 bits d'origine. Une telle compression introduit des bruits stochastiques supplémentaires, c'est-à-dire une erreur de quantification, dans le processus d'optimisation, ce qui ralentit la convergence ou peut même conduire à la divergence (Alistarh et al., 2017). Les performances de ces approches reposent sur le compromis entre le nombre de bits communiqués par itération et la qualité de ces informations. Ainsi, des schémas agressifs peuvent n'envoyer qu'un seul bit par coordonnée (Bernstein et al., 2018; Tang et al., 2021) ou utiliser la quantification vectorielle (Leconte et al., 2021).

Contributions. Les principales contributions peuvent être résumées comme suit :

(1) Nous proposons `QLSD`, un algorithme de MCMC général spécifiquement conçu pour l'inférence bayésienne dans le cadre du FL, ainsi que deux alternatives à variance réduite, visant en particulier l'*hétérogénéité*, les *surcoûts de communication* et la *participation partielle*.

(2) Nous fournissons une analyse de convergence non asymptotique des algorithmes proposés. La partie théorique met en évidence l'impact de l'hétérogénéité statistique mesurée par l'écart entre les distributions a posteriori locales.

(3) Nous proposons des mécanismes efficaces pour atténuer l'impact de l'hétérogénéité statistique sur la convergence, soit en utilisant des gradients stochastiques biaisés, soit en introduisant un mécanisme de *mémoire* qui étend Horváth et al. (2022) au cadre bayésien. En particulier, nous constatons que la réduction de la variance permet en effet à l'algorithme MCMC proposé de converger vers la distribution a posteriori cible lorsque le nombre d'observations devient grand.

(4) Nous illustrons les avantages des méthodes proposées à l'aide de plusieurs benchmarks FL. Nous montrons que la méthodologie proposée fonctionne bien par rapport aux méthodes bayésiennes FL de pointe.

Dans ce travail, nous étendons ces idées au cadre bayésien. Nous développons un nouvel algorithme d'inférence bayésienne fédérée, appelé Quantized Langevin Stochastic Dynamics (`QLSD`), pour résoudre le goulot d'étranglement de communication des algorithmes distribués/fédérés. Ce cadre intègre le cas de n clients, chacun possédant un potentiel local $U_i : \mathbb{R}^d \to \mathbb{R}$ calculé en fonction de son ensemble de données local $\mathcal{D}_i$.

Les agents effectuent une inférence bayésienne pour cibler la distribution a posteriori proportionnelle à $\exp(-\sum_{i=1}^{n} U_i)$ tout en respectant les contraintes de l'apprentissage fédéré. En utilisant une séquence non biaisée $\{\mathcal{C}_k\}_{k\geq 1}$ d'opérateurs de compression (Alistarh et al., 2017), ces agents ne communiquent qu'une version quantifiée de leur gradient stochastique $\widehat{\nabla U_i}$ à chaque tour d'agrégation. Ensuite, le serveur central effectue une étape de dynamique de Langevin basée sur les gradients compressés reçus. Le paramètre θ_k est mis à jour en utilisant les informations des clients participants $\mathcal{A}_{k+1}$:

$$\theta_{k+1} = \theta_k - \gamma \frac{n}{|\mathcal{A}_{k+1}|} \sum_{i \in \mathcal{A}_{k+1}} \mathcal{C}_{k+1}(\widehat{\nabla U_i}(\theta_k)) + \sqrt{2\gamma} Z_{k+1}, \tag{6.3}$$

où Z_{k+1} est un bruit gaussien standard. Sous les hypothèses énoncées dans le Theorem 4.5, les échantillons θ_k générés par (6.3) sont approximativement distribués selon $\prod_{i\in[n]} \exp(U_i)$. Cependant, nous démontrons théoriquement et expérimentalement que cette méthode souffre d'hétérogénéité et de l'utilisation d'un gradient stochastique $\widehat{\nabla U_i}$. Pour améliorer les performances, nous introduisons donc des mécanismes conduisant à des versions améliorées appelées `QLSD`* et `QLSD`$^{++}$. Dans la première version `QLSD`*, le gradient stochastique $\widehat{\nabla U_i}$ dans (6.3) est remplacé par l'oracle $\widetilde{\nabla U_i}(\theta) = \widehat{\nabla U_i}(\theta) - \widehat{\nabla U_i}(\theta_\star)$, où $\theta_\star = \arg\min \sum_i U_i$; pour plus de détails, voir l'algorithme du point fixe de Langevin (Brosse et al., 2018). Il est intéressant de noter que $\widetilde{\nabla U_i}$ est une estimation biaisée de ∇U_i puisque l'espérance $\mathbb{E}[\widetilde{\nabla U_i}] \neq \nabla U_i$ malgré $\mathbb{E}[\sum_i \widetilde{\nabla U_i}] = \sum_i \nabla U_i$. Dans le Theorem 4.7, nous dérivons des garanties de convergence asymptotique et non asymptotique pour l'algorithme proposé. Cependant, obtenir le minimiseur $\theta_\star$ est compliqué en pratique. Nous développons donc une dernière alternative appelée `QLSD`$^{++}$ qui s'appuie sur la technique bien connue SVRG (Johnson and Zhang, 2013) pour réduire le bruit introduit par la variance du gradient stochastique combiné avec un mécanisme de mémoire pour résoudre le problème d'hétérogénéité (Horváth et al., 2022; Philippenko and Dieuleveut, 2020). Enfin, nous illustrons les performances de l'approche proposée par rapport à divers benchmarks d'apprentissage fédéré bayésien. De plus, nous mettons en évidence numériquement les avantages de la compression en obtenant une précision similaire aux méthodes classiques avec moins de bits.

Theorem 6.3 (Informel)**.** *Sous les hypothèses décrites dans le Chapter 4, il existe $\gamma_\star > 0$ tel que pour $\gamma \in (0, \gamma_\star)$, il existe $\kappa \in (0,1)$, $C_0, C_1 > 0$ tels que pour $k \geq 0$, la distribution μ_k de l'échantillon θ_k satisfait*

$$W_2^2\Big(\mu_k, \pi(\cdot|\mathcal{D}_{1:n})\Big) \leq (1-\kappa)^k C_0 + \gamma C_1.$$

• Chapter 5: *Prédiction conforme pour la quantification de l'incertitude fédérée avec des distributions locales différentes* (RQ#5-RQ#6-RQ#7)

Une quantification précise de l'incertitude est essentielle dans les applications modernes d'apprentissage automatique. Cela est crucial pour développer des méthodes fiables garantissant la validité des prédictions. Cependant, estimer des ensembles de prédictions valides peut être un défi dans des contextes distribués, et ce défi est encore exacerbé en cas de distribution de sortie différentes.

Contributions. Les principales contributions peuvent être résumées comme suit :

(1) Nous introduisons une nouvelle méthode, `DP-FedCP`, pour construire des ensembles de prédictions conformelles dans un contexte d'apprentissage fédéré qui prend en compte les distributions de sortie différentes entre les agents ; voir la Section 5.2.

`DP-FedCP` est un algorithme d'apprentissage fédéré basé sur le calcul fédéré des quantiles pondérés des scores de non-conformité des agents, les poids reflétant le décalage entre les distribtions de sortie de chaque client par rapport à la population. Les quantiles sont obtenus en régularisant la perte de pinball à l'aide de l'inf-convolution de Moreau-Yosida et d'une version de la procédure d'agrégation fédérée ; voir la Section 5.3.

(2) Nous établissons des garanties de prédictions conformelles, garantissant la validité des ensembles de prédictions obtenus. De plus, nous fournissons des garanties de confidentialité différentielle pour `DP-FedCP`; voir la Section 5.4.

(3) Nous montrons que `DP-FedCP` fournit des ensembles de confiance valides et surpasse les approches standards dans une série d'expériences sur des données simulées et des ensembles de données de classification d'images ; voir la Section 5.5.

Contrairement aux méthodes conformes habituelles, l'algorithme `DP-FedCP` ne calcule les scores de non-conformité que sur un sous-ensemble de $\bar{N}$ données de calibration. Par exemple, $\bar{N} = \lfloor N/2 \rfloor$ lorsque la moitié des points de calibration est utilisée. Un mécanisme clé de `DP-FedCP` consiste à évaluer la disparité entre les distributions de calibration et de test (P^{cal} et $P^\star$). Sur la base d'une estimation du rapport de vraisemblance de Radon-Nikodym $\widehat{w}_y^\star = \mathrm{d}P_Y^\star / \mathrm{d}P_Y^{\mathrm{cal}}$, un ensemble de prédictions valide peut être obtenu en pondérant les scores de non-conformité. Soit $\{(X_k, Y_k)\}_{k\in[\bar{N}]}$ les échantillons de calibration utilisés pour construire les ensembles de prédictions. Pour tout $\mathbf{y} \in \mathcal{Y}$, nous construisons une famille de poids $\{\widehat{p}_{y,\mathbf{y}}^\star\}_{y\in\mathcal{Y}}$ donnée par

$$\widehat{p}_{y,\mathbf{y}}^\star = \frac{\widehat{w}_y^\star}{\widehat{w}_{Y_{N^\star+1}^\star}^\star + \sum_{\ell=1}^{\bar{N}} \widehat{w}_{Y_\ell}^\star}.$$

Ensuite, en utilisant ces poids, `DP-FedCP` utilise les scores de non-conformité locaux pour dériver des ensembles de prédictions personnalisés pour un nouveau point de données $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star) \sim P^\star$, comme suit

$$\begin{aligned}
&\bar{\mu}_{\mathbf{y}}^\star = \widehat{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \textstyle\sum_{k=1}^{\bar{N}} \widehat{p}_{Y_k,\mathbf{y}}^\star \delta_{V_k}, \\
&\mathcal{C}_{\alpha,\bar{\mu}^\star}(X_{N^\star+1}^\star) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(X_{N^\star+1}^\star, \mathbf{y}) \le Q_{1-\alpha}(\bar{\mu}_{\mathbf{y}}^\star) \right\}.
\end{aligned}$$

Des bornes non asymptotiques garantissant la validité de ces ensembles de prédictions sont fournies dans la Section 5.4. En particulier, lorsque les rapports de vraisemblance sont connus, le résultat suivant est vérifié

$$\begin{aligned}
\left| \mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\bar{\mu}^\star}(X_{N^\star+1}^\star) \right) - 1 + \alpha \right| \le \frac{6}{N} + \frac{36 + 6\log N}{N} \|\widehat{w}^\star\|_\infty^2 \\
+ \frac{14 \log N}{N} \sum\nolimits_{j\colon \frac{Nj}{12} < \log N} \sqrt{N^j},
\end{aligned}$$

où N^i correspond aux données de calibration appartenant à l'agent $i \in [n]$. L'ensemble de prédictions $\mathcal{C}_{\alpha,\bar{\mu}^\star}(X_{N^\star+1}^\star)$ est généralement difficile à déterminer car calculer le quantile exact $Q_{1-\alpha}(\bar{\mu}_{\mathbf{y}}^\star)$ de manière fédérée est loin d'être évident. En fait, nous développons une méthode résolvant ce problème tout en garantissant qu'aucun attaquant ne peut déterminer avec une grande confiance si les données d'un individu particulier sont incluses dans l'ensemble de données ou non.

Bibliography

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. page 33

S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017. page 254

S. Ahn, B. Shahbaba, and M. Welling. Distributed Stochastic Gradient MCMC. In *International Conference on Machine Learning*, 2014. pages 25, 33, 36, 41, 42, 43, 97, 171, 281

A. F. Aji and K. Heafield. Sparse Communication for Distributed Gradient Descent. In *Conference on Empirical Methods in Natural Language Processing*, 2017. page 174

M. Al-Shedivat, J. Gillenwater, E. Xing, and A. Rostamizadeh. Federated Learning via posterior Inference: A new perspective and practical algorithms. In *ICLR 2021*, 2021. pages 96, 97

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 2017. pages 27, 28, 95, 96, 97, 171, 172, 174, 177, 214, 283, 284

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. page 21

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1–2):5–43, 2003. page 171

A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. page 22

A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik. Uncertainty sets for image classifiers using Conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9. pages 224, 234

A. N. Angelopoulos, S. Bates, T. Zrnic, and M. I. Jordan. Private Prediction Sets. *Harvard Data Science Review*, 4(2), apr 28 2022a. https://hdsr.mitpress.mit.edu/pub/deziirvg. page 225

A. N. Angelopoulos, K. Krauth, S. Bates, Y. Wang, and M. I. Jordan. Recommendation systems with distribution-free reliability guarantees. *arXiv preprint arXiv:2207.01609*, 2022b. page 224

J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, 29(3):599–615, 2019. pages 17, 175

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013. page 66

V. Balasubramanian, S.-S. Ho, and V. Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014. pages 22, 223

B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 31, 2018. pages 234, 270

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022. pages 223, 224

R. Bardenet, A. Doucet, and C. C. Holmes. On markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47), 2017. pages 16, 32

J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018. pages 27, 97, 283

P. A. Bernstein and E. Newcomer. *Principles of Transaction Processing*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2009. ISBN 1558606238. page 33

K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, B. McMahan, et al. Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1:374–388, 2019. page 222

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. page 32

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. page 267

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. pages 20, 97

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011. page 105

N. Brosse, E. Moulines, and A. Durmus. The Promises and Pitfalls of Stochastic Gradient Langevin Dynamics. In *Neural Information Processing Systems*, pages 8278–8288, 2018. pages 17, 28, 43, 172, 175, 180, 284

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. page 231

S. Buhai. Quantile regression: overview and selected applications. *Ad Astra*, 4(4):1–17, 2005. page 229

T. D. Bui, C. V. Nguyen, S. Swaroop, and R. E. Turner. Partitioned variational inference: A unified framework encompassing federated and continual learning. *arXiv preprint arXiv:1811.11206*, 2018. page 171

S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konecny, H. B. McMahan, V. Smith, and A. Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv preprint arXiv:1812.01097*, 2018. page 181

C. L. Canonne, G. Kamath, and T. Steinke. The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020. page 231

R. Caruana. *Multitask learning*. Springer, 1998. page 228

M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020. page 224

N. Chatterji, N. Flammarion, Y. Ma, P. Bartlett, and M. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018. pages 17, 102, 159

C. Chen, N. Ding, C. Li, Y. Zhang, and L. Carin. Stochastic Gradient MCMC with Stale Gradients. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2937–2945. Curran Associates, Inc., 2016. pages 42, 43

H.-Y. Chen and W.-L. Chao. Fedbe: Making Bayesian model ensemble applicable to Federated Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=dgtpE6gKjHn. pages 96, 171, 172, 182

H. H. S. Chittoor and O. Simeone. Coded consensus Monte Carlo: Robust one-shot distributed Bayesian learning with stragglers. *arXiv preprint arXiv:2112.09794*, 2021. page 97

A. Chowdhury and C. Jermaine. Parallel and Distributed MCMC via Shepherding Distributions. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1819–1827, 2018. pages 33, 42, 43

D. S. Clark. Short proof of a discrete gronwall inequality. *Discrete applied mathematics*, 16(3):279–281, 1987. page 159

C. Coglianese and D. Lehr. Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ*, 105:1147, 2016. page 96

L. Corinzia, A. Beuret, and J. M. Buhmann. Variational federated multi-task learning. *arXiv preprint arXiv:1906.06268*, 2019. page 171

H. Dai, M. Pollock, and G. Roberts. Monte Carlo fusion. *Journal of Applied Probability*, 56(1):174–191, 2019. doi: 10.1017/jpr.2019.12. page 42

H. Dai, M. Pollock, and G. Roberts. Bayesian fusion: Scalable unification of distributed statistical analyses. *arXiv preprint arXiv:2102.02123*, 2021. pages 96, 97, 98

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a. pages 115, 154

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society, Series B*, 79(3):651–676, 2017b. pages 16, 17, 179

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019. pages 17, 37, 99, 100, 103, 164, 179

A. P. Dawid and M. Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014. pages 166, 219

D. A. De Souza, D. Mesquita, S. Kaski, and L. Acerbi. Parallel MCMC without embarrassing failures. In *International Conference on Artificial Intelligence and Statistics*, pages 1786–1804. PMLR, 2022. page 97

J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004. page 42

G. Del Grosso, H. Jalalzai, G. Pichler, C. Palamidessi, and P. Piantanida. Leveraging Adversarial Examples to quantify membership information leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. page 21

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. pages 107, 167, 218

W. Deng, Y.-A. Ma, Z. Song, Q. Zhang, and G. Lin. On convergence of Federated averaging Langevin dynamics. *arXiv preprint arXiv:2112.05120*, 2021. pages 26, 96, 97, 98, 102, 108, 140, 281

A. Dieuleveut and K. K. Patel. Communication tradeoffs for Local-SGD with large step-size. In *Advances in Neural Information Processing Systems*, volume 32, pages 13601–13612, 2019. pages 25, 36, 281

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step-size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 06 2020. page 176

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer, 2018. pages 16, 52, 62, 160

K. A. Dubey, S. J Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing. Variance reduction in stochastic gradient Langevin dynamics. *Advances in neural information processing systems*, 29, 2016. pages 17, 26, 96, 281

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 06 2017. doi: 10.1214/16-AAP1238. pages 16, 41

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019. pages 17, 37, 40, 67, 69, 70, 73, 78, 99, 100, 103, 104, 114, 115, 164, 179, 190, 193, 211

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019. pages 17, 179

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. page 269

K. El Mekkaoui, D. Mesquita, P. Blomstedt, and S. Kaski. Federated stochastic gradient Langevin dynamics. In *Uncertainty in Artificial Intelligence*, pages 1703–1712. PMLR, 2021. pages 42, 43, 97, 105, 167, 171, 181

C. Fannjiang, S. Bates, A. Angelopoulos, J. Listgarten, and M. I. Jordan. Conformal prediction for the design problem. *arXiv preprint arXiv:2202.03613*, 2022. page 224

M. Fatima, M. Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017. page 96

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html. page 106

M. Fontana, G. Zeni, and S. Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023. pages 22, 223, 224

G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, and I. Bloch. Encoding the latent posterior of Bayesian Neural Networks for uncertainty quantification. *arXiv preprint arXiv:2012.02818*, 2020. page 171

J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020. page 18

S. Garg, Y. Wu, S. Balakrishnan, and Z. Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020. page 224

I. Gibbs and E. Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021. page 224

J. Gillenwater, M. Joseph, and A. Kulesza. Differentially private quantiles. In *International Conference on Machine Learning*, pages 3713–3722. PMLR, 2021. page 225

A. M. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. T. Suresh. Shuffled model of Federated learning: Privacy, communication and accuracy trade-offs. *arXiv preprint arXiv:2008.07180*, 2020. page 174

E. Gorbunov, F. Hanzely, and P. Richtárik. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2021. pages 19, 26, 96, 100, 122, 281

U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B*, 56(4):549–603, 1994. pages 98, 173

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. pages 21, 106, 166, 168, 218, 219

F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR, 2021. pages 97, 174

F. Hanzely and P. Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020. page 19

L. Hasenclever, S. Webb, T. Lienart, S. Vollmer, B. Lakshminarayanan, C. Blundell, and Y. W. Teh. Distributed Bayesian Learning with Stochastic Natural Gradient Expectation Propagation and the Posterior Server. *Journal of Machine Learning Research*, 18(106):1–37, 2017. page 171

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. pages 108, 167, 182, 235, 236

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013. pages 96, 171

J. M. Holte. Discrete gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7, 2009. page 134

S. Horváth and P. Richtarik. A better alternative to error feedback for communication-efficient distributed learning. In *International Conference on Learning Representations*, 2020. page 19

S. Horváth, D. Kovalev, K. Mishchenko, P. Richtárik, and S. Stich. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, pages 1–16, 2022. pages 28, 95, 96, 100, 171, 172, 176, 178, 283, 284

X. Hu and J. Lei. A distribution-free test of covariate shift using conformal prediction. *arXiv preprint arXiv:2010.07147*, 2020. page 224

W. Huang, M. Ye, and B. Du. Learn from others and be yourself in heterogeneous Federated learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10133–10143, 2022. doi: 10.1109/CVPR52688.2022.00990. page 222

P. Humbert, B. L. Bars, A. Bellet, and S. Arlot. One-shot Federated Conformal prediction. *arXiv preprint arXiv:2302.06322*, 2023. page 23

D. J. Hunter. Uncertainty in the Era of Precision Medicine. *New England Journal of Medicine*, 375(8):711–713, 2016. page 171

P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020. page 171

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, pages 4629–4640. PMLR, 2021. pages 96, 108, 165, 171, 182, 217

H. Jalalzai, E. Kadoche, R. Leluc, and V. Plassier. Membership Inference Attacks via Adversarial Examples. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022. URL https://openreview.net/forum?id=K_qZc3pbUAX. page 5

R. Johnson and T. Zhang. Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013. pages 17, 28, 147, 176, 284

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-Efficient Distributed Statistical Inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019. pages 33, 42, 171

P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015. page 271

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. pages 18, 95, 170, 222

S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, 2019. page 19

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for Federated Learning. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020. pages 18, 26, 96, 100, 147, 171, 178, 281

R. Kassab and O. Simeone. Federated generalized bayesian learning via distributed stein variational gradient descent. *IEEE Transactions on Signal Processing*, 2022. page 171

J. Kent. Time-reversible diffusions. *Advances in Applied Probability*, 10:819–835, 12 1978. ISSN 1475-6064. doi: 10.1017/S0001867800031396. page 71

R. Koenker and K. F. Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001. page 229

J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. page 19

D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020. page 100

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009. Available at https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. pages 108, 167, 182

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. page 22

L. Leconte, A. Dieuleveut, E. Oyallon, E. Moulines, and G. PAGES. DoStoVoQ: Doubly stochastic voronoi vector quantization SGD for Federated Learning. 2021. pages 27, 283

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. pages 107, 167, 218

M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. page 68

S. Lee, C. Park, S.-N. Hong, Y. C. Eldar, and N. Lee. Bayesian Federated Learning over wireless networks. *IEEE Journal on Selected Areas in Communications*, 2020. page 97

J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013. pages 22, 223

L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021. pages 224, 227

C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. page 105

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a. page 222

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated Optimization in Heterogeneous Networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020b. pages 18, 171, 215

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019. pages 18, 96, 232

T. Lin, L. Kong, S. U. Stich, and M. Jaggi. Ensemble distillation for robust model fusion in federated learning. *arXiv preprint arXiv:2006.07242*, 2020. page 18

Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In *International Conference on Learning Representations*, 2018. page 174

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3122–3130. PMLR, 2018. page 224

D. Liu and O. Simeone. Channel-Driven Monte Carlo Sampling for Bayesian Distributed Learning in Wireless Data Centers. *IEEE Journal on Selected Areas in Communications*, 2021a. page 171

D. Liu and O. Simeone. Wireless Federated Langevin Monte Carlo: Repurposing channel noise for Bayesian sampling and privacy. *arXiv preprint arXiv:2108.07644*, 2021b. page 172

Q. Liu and A. T. Ihler. Distributed estimation, information loss and exponential families. *Advances in neural information processing systems*, 27, 2014. page 97

C. Lu and J. Kalpathy-Cramer. Distribution-free Federated learning with Conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021. pages 23, 225

W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. page 167

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. pages 95, 96, 98, 140, 170, 171, 172, 231

D. Mesquita, P. Blomstedt, and S. Kaski. Embarrassingly Parallel MCMC using Deep Invertible Transformations. In *Uncertainty in Artificial Intelligence*, pages 1244–1252. PMLR, 2020. pages 42, 97

S. Minsker, S. Srivastava, L. Lin, and D. Dunson. Scalable and robust Bayesian Inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning*, 2014. pages 19, 42, 96, 97

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017. pages 234, 269

J. J. Moreau. Propriétés des applications «prox». *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 256:1069–1071, 1963. page 229

T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. The true cost of stochastic gradient Langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017. page 17

W. Neiswanger, C. Wang, and E. P. Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 623–632, 2014. pages 19, 20, 33, 42, 96, 97, 102, 173

C. Nemeth and C. Sherlock. Merging MCMC subposteriors through Gaussian-process approximations. *Bayesian Analysis*, 13(2):507–530, 2018. pages 33, 42, 97, 171

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009. page 231

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003. pages 113, 178, 184, 242

M. Noble, A. Bellet, and A. Dieuleveut. Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 10110–10145. PMLR, 2022. page 234

F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. ISSN 0022-1236. page 68

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. pages 166, 219

H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002. pages 23, 224

N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014. page 229

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3): 378–384, 1981. page 173

C. Philippenko and A. Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees . *arXiv preprint arXiv:2006.14591*, 2020. pages 28, 171, 180, 284

K. Pillutla, Y. Laguel, J. Malick, and Z. Harchaoui. Differentially private Federated quantiles with the distributed discrete gaussian mechanism. In *International Workshop on Federated Learning: Recent Advances and New Challenges*, 2022. page 225

V. Plassier, M. Vono, A. Durmus, and E. Moulines. DG-LMC: A turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within gibbs. In *International Conference on Machine Learning*, pages 8577–8587. PMLR, 2021. pages 5, 8, 97, 105, 171, 172, 181

V. Plassier, M. Makni, A. Rubashevskii, and E. Moulines. Prediction for Federated Uncertainty Quantification Under Label Shift. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2023a. pages 5, 9

V. Plassier, E. Moulines, and A. Durmus. Federated averaging Langevin dynamics: Toward a unified theory and new algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 5299–5356. PMLR, 2023b. pages 5, 8

V. Plassier, F. Portier, and J. Segers. Risk bounds when learning infinitely many response functions by ordinary linear regression. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59, pages 53–78. Institut Henri Poincaré, 2023c. page 5

A. Podkopaev and A. Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021a. pages 223, 224, 225, 226, 233

A. Podkopaev and A. Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. *arXiv preprint arXiv:2110.06177*, 2021b. pages 224, 235

M. Rabinovich, E. Angelino, and M. I. Jordan. Variational Consensus Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 28, pages 1207–1215, 2015. pages 33, 42, 43

R. Rahaman and A. H. Thiery. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34, 2021. pages 21, 218

I. Raicu, I. Foster, A. Szalay, and G. Turcu. AstroPortal: A Science Gateway for Large-scale Astronomy Data Analysis. In *TeraGrid Conference*, pages 12–15, 2006. page 33

L. J. Rendell, A. M. Johansen, A. Lee, and N. Whiteley. Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, 30(2):249–259, 2020. pages 20, 21, 33, 34, 35, 36, 39, 42, 43, 44, 97, 98, 108, 171

D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science, 2013. pages 70, 190

J. Ro, M. Chen, R. Mathews, M. Mohri, and A. T. Suresh. Communication-efficient agnostic federated averaging. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 1753–1757. International Speech Communication Association, 2021. pages 18, 96

C. P. Robert. *The Bayesian Choice: from decision-theoretic foundations to computational implementation*. Springer, New York, 2 edition, 2001. pages 33, 171

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, Berlin, 2 edition, 2004. page 215

G. O. Roberts and J. S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Annals of Applied Probability*, 16(4):2123–2139, 11 2006. doi: 10.1214/105051606000000510. page 37

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996. pages 16, 25, 34, 35, 71, 98, 104, 111, 173, 190, 280

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. page 229

Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020. page 234

P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978. doi: http://dx.doi.org/10.1063/1.436415. pages 25, 34, 280

F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2020. pages 18, 174

H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947. pages 21, 25, 35, 280

S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016. pages 19, 33, 42, 97

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-Bit Stochastic Gradient Descent and Application to Data-Parallel Distributed Training of Speech DNNs. In *Interspeech*, 2014. page 174

G. Shafer and V. Vovk. A tutorial on Conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008. pages 22, 223, 224, 225

N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui. Uveqfed: Universal vector quantization for federated learning. *IEEE Transactions on Signal Processing*, 69:500–514, 2020. page 97

R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 909–910, 2015. doi: 10.1109/ALLERTON.2015.7447103. page 33

S. P. Singh and M. Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33, 2020. page 18

L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. page 167

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pages 912–920, 2015. page 42

S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019. page 19

S. U. Stich, J.-B. Cordonnier, and M. Jaggi. Sparsified SGD with Memory. In *Advances in Neural Information Processing Systems*, 2018. pages 99, 174

L. Sun, A. Salim, and P. Richtárik. Federated Learning with a sampling algorithm under isoperimetry. *arXiv preprint arXiv:2206.00920*, 2022. pages 97, 102

H. Tang, S. Gan, A. A. Awan, S. Rajbhandari, C. Li, X. Lian, J. Liu, C. Zhang, and Y. He. 1-bit adam: Communication efficient large-scale training with adam's convergence speed. In *International Conference on Machine Learning*, pages 10118–10129. PMLR, 2021. pages 27, 97, 283

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019. pages 23, 223, 224, 225, 226, 227

A. Triastcyn and B. Faltings. Federated learning with bayesian differential privacy. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2587–2596. IEEE, 2019. page 225

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. page 19

J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer. A survey on distributed machine learning. *ACM Comput. Surv.*, 53(2), Mar. 2020. ISSN 0360-0300. doi: 10.1145/3377454. page 33

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Berlin Heidelberg, 2008. pages 120, 190

M. Vono, N. Dobigeon, and P. Chainais. Split-and-augmented Gibbs sampler - Application to large-scale inference problems. *IEEE Transactions on Signal Processing*, 67 (6):1648–1661, 2019. doi: 10.1109/TSP.2019.2894825. page 43

M. Vono, N. Dobigeon, and P. Chainais. Asymptotically exact data augmentation: Models, properties, and algorithms. *Journal of Computational and Graphical Statistics*, 30(2):335–348, 2020. pages 33, 39, 97

M. Vono, D. Paulin, and A. Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *Journal of Machine Learning Research*, 23(25), 2022a. pages 20, 21, 34, 35, 64, 66, 97, 108, 171

M. Vono, V. Plassier, A. Durmus, A. Dieuleveut, and E. Moulines. Qlsd: Quantised Langevin Stochastic Dynamics for Bayesian federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6459–6500. PMLR, 2022b. pages 5, 9, 97, 102, 105, 138, 139

V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. 1999. pages 22, 223, 224

C. Wang, X. Chen, A. J. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. *Advances in neural information processing systems*, 26, 2013. page 100

H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020a. page 18

J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous Federated optimization. *Advances in neural information processing systems*, 2020b. pages 18, 96

J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. pages 19, 95, 171, 172

X. Wang and D. B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013. pages 19, 42, 97, 171

X. Wang, F. Guo, K. A. Heller, and D. B. Dunson. Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems*, 2015. pages 19, 42, 96, 171

K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020. page 225

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on International Conference on Machine Learning*, pages 681–688, 2011. Available at https://www.ics.uci.edu/~welling/publications/papers/stoc{L}angevin_v6.pdf. pages 16, 17, 96, 98, 105, 172, 173

S. White, T. Kypraios, and S. Preston. Piecewise approximate Bayesian computation: fast inference for discretely observed markov models using a factorised posterior distribution. *Statistics and Computing*, 25(2):289–301, 2015. page 20

A. G. Wilson, P. Izmailov, M. D. Hoffman, Y. Gal, Y. Li, M. F. Pradier, S. Vikram, A. Foong, S. Lotfi, and S. Farquhar. Evaluating approximate Inference in Bayesian deep learning. 2021. pages 22, 96, 182

B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020. pages 18, 241

C. Wu and C. P. Robert. Average of recentered parallel mcmc for big data. *arXiv preprint arXiv:1706.04780*, 2017. page 98

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. page 218

Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10 (2):1–19, 2019. pages 95, 222

J. Yoon, G. Park, W. Jeong, and S. J. Hwang. Bitwidth heterogeneous Federated learning with progressive weight dequantization. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25552–25565. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yoon22a.html. page 222

M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni. Probabilistic Federated neural matching. 2018. page 18

M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019. page 96

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. page 21

Y. Zhang, D. Liu, and O. Simeone. Leveraging channel noise for sampling and privacy via quantized Federated Langevin Monte Carlo, 2022. URL https://arxiv.org/abs/2202.13932. page 97

M. Zinkevich, M. Weimer, L. Li, and A. Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010. page 19

Titre : Simulation de Monte Carlo distribuée avec apprentissage statistique à grande échelle : Inférence bayésienne et prédiction conformelle

Mots clés : Monte Carlo, Apprentissage Fédéré, Inférence Bayésienne, Prediction Conformelle

Résumé : Cette thèse dévelope des approches dans les secteurs de l'inférence bayésienne et la quantification des incertitudes. Les méthodes de Monte Carlo fédéré permettent à plusieurs agents/nœuds d'effectuer des calculs localement, tandis qu'un serveur central agrège les résultats pour échantillonner selon la posteriori globale. Ces techniques d'échantillonnage bénéficient de l'incorporation de connaissances à travers la priori, conduisant à l'amélioration des résultats. Cette capacité est d'autant plus nécessaire pour de petits jeux de données ou lorsque celles-ci sont bruitées. De plus, l'incertitude associée aux paramètres est naturellement quantifiée. Dans une première partie, nous introduisons deux méthodes basées sur les chaînes de Markov. Chacunes de ces méthodes reposent sur un serveur central orchestrant les entités locales. Celui-ci agrège l'information provenant de chaque agent afin de produire des paramètres adaptées tout en évitant le transfert de données. Cette approche réduit la quantité d'information transférée entre agents, ce qui la rend particulièrement avantageuse lorsque les communications sont limitées. En outre, les méthodes développées abordent les enjeux de confidentialité et étendent des algorithmes d'optimisation à l'inférence bayésienne. La deuxième partie de la thèse se concentre sur la gestion de l'incertitude. Initialement, nous présentons l'approche bayésienne, basée sur des opérateurs de compression afin de résoudre les problèmes de bande passante. Dans la dernière partie de cette thèse, nous introduisons une méthode fréquentiste basée sur les prédictions conformelles. Contrairement aux méthodes bayésiennes, aucune hypothèse sur la distribution des paramètres n'est requise. Nous développons une approche indépendante du prédicteur entraîné. Plus précisément, cette méthode utilise la technique de régression quantile pour générer des ensembles de prédictions personnalisées. Celle-ci aborde efficacement l'hétérogénéité entre agents pour déterminer des pondérations d'importance. Un aspect crucial de notre approche est la préservation des informations sensibles de chaque utilisateur, et nous veillons à protéger la confidentialité.

Title : Distributed Monte Carlo simulation with large-scale Machine Learning: Bayesian Inference and Conformal Prediction

Keywords : Monte Carlo, Federated Learning, Bayesian Inference, Conformal Prediction

Abstract : Distributed methods have emerged as powerful tools for addressing the data centralization challenge. This thesis introduces innovative approaches to tackle large-scale Bayesian inference and uncertainty quantification, aiming to provide effective solutions in distributed data environments. Federated Monte Carlo methods allow multiple agents/nodes to conduct computations locally and securely, with a central server combining the results to obtain samples from the global posterior distribution. These sampling techniques benefit from the incorporation of prior knowledge, leading to improved results. Additionally, the uncertainty associated with the parameters and the predictions are naturally quantified, which is crucial for decision-making. Especially with limited or noisy data, the ability to quantify uncertainty becomes even more essential. In the first part, we introduce two methods based on Markov chains. Both methods are designed to target a global posterior and relied on a central server to orchestrate multiple local entities. The server aggregates information from each agent to produce statistical solutions, while limiting the amount of transferred data. This approach reduces the data transfer between participating agents, making it particularly advantageous when communications are limited. These developed methods not only address data privacy concerns but also extend existing learning algorithms to Bayesian inference problems. This contributes to the development of more robust and efficient machine learning algorithms, and holds potential applications in various domains, including epidemiology and finance. The second part of the thesis focuses on uncertainty management. Initially, we present the Bayesian approach, which employs compression operators to overcome bandwidth limitations. In the final part, we introduce a frequentist method based on conformal predictions. Unlike other methods, our approach is model-agnostic and can be applied to any predictive model. Specifically, this method leverages quantile regression techniques to generate personalized prediction sets while maintaining robustness to outliers. The label shift between agents is effectively addressed by determining quantiles based on importance weights. A crucial aspect of our approach is the preservation of privacy, as we ensure the protection of sensitive user information.

Institut Polytechnique de Paris
91120 Palaiseau, France