Variational Inference : theory and large scale applications.

Tom Huix

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAX071

Thèse de doctorat

Variational Inference: theory and large scale applications.

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

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

Thèse présentée et soutenue à Palaiseau, le 2 octobre 2024, par

TOM HUIX

Composition du Jury :

Olivier Cappé Directeur de recherche CNRS	Président
Julyan Arbel Chargé de recherche Inria, University of Grenoble	Rapporteur
Aurélien Garivier Professor at Ecole Normale Supérieure de Lyon (UMPA)	Rapporteur
Kamélia Daudel Assistant Professor at ESSEC Business School	Examinateur
Randal Douc Professor at Telecom Sudparis	Examinateur
Eric Moulines Professor at Ecole Polytechnique (CMAP)	Directeur de thèse
Anna Korba Assistant professor at ENSAE (CREST)	Co-directeur de thèse
Alain Durmus Professor at Ecole Polytechnique (CMAP)	Invité

Remerciements

C'est avec une grande émotion que je commence la rédaction de ces remerciements. C'est vrai que la thèse n'a pas toujours été un exercice simple pour moi, mais aujourd'hui, j'écris ces quelques lignes avec un sentiment de fierté et d'accomplissement. À travers ces mots, je tiens à remercier les personnes sans qui cette aventure n'aurait jamais été possible.

Tout d'abord, j'aimerais commencer par remercier mes trois directeurs de thèse : Eric, Alain et Anna. Eric, je te remercie sincèrement pour ton accompagnement et ta grande disponibilité tout au long de ces trois années. Ta passion pour notre domaine, tes conseils précieux, ainsi que tes histoires et anecdotes ont rendu ce chemin enrichissant, mais aussi inspirant. Alain, un grand merci pour ton encadrement tout au long de cette thèse. Travailler sur les bandits contextuels avec Pierre et toi a été un réel plaisir. Même si notre papier n'a pas rencontré le succès que l'on espérait, il a quand même fini par trouver son public. Je te remercie pour ta patience et excuses moi pour toutes les « atrocités » mathématiques que Pierre et moi t'avons infligées. Et enfin, Anna, un immense merci pour ces trois années de collaboration. Merci d'avoir été soucieuse de mon bien-être lorsque la thèse devenait difficile pour moi. Je suis fier d'avoir eu l'opportunité de travailler avec une personne aussi brillante que toi. Tu m'as appris à défendre mes idées et mes travaux avec conviction, qu'il s'agisse de répondre à une review bancale ou de faire face à ceux qui doutent des Wasserstein Gradient Flows. Travailler à tes côtés a été un immense plaisir, que ce soit en passant des heures devant un tableau noir ou en partageant un verre à l'ICML. Merci pour tout ce que tu m'as apporté au cours de ces années. J'espère avoir été un premier thésard à la hauteur de tes attentes.

Je remercie aussi tous les membres de mon jury. Merci à mes rapporteurs, Julyan et Aurélien, d'avoir pris le temps de relire attentivement mon travail. Merci pour votre temps et votre engagement pour la recherche. Merci aussi à Olivier, Randal et Kamélia d'avoir accepté d'être membres de ce jury.

Je tiens ensuite à remercier ma famille pour leur soutien durant ces trois années. Merci à mes parents pour tout l'amour, la confiance, le dévouement et les encouragements que j'ai reçus tout au long de ma vie. À ma grande soeur, merci d'avoir été, depuis toujours, mon modèle, mais aussi ma première supportrice. Merci également à Maxime, mon beau-frère (unique et préféré), dont la bonne humeur et la joie de vivre ont su égailler tous nos repas de famille. Et bien sûr, je ne peux pas remercier ma famille sans parler de mon/ma futur(e) neveu/nièce, que j'aime déjà de tout mon coeur. Un immense merci aussi à mes grands-parents, oncles, tantes, cousins et à toute ma famille pour leur soutien extraordinaire tout au long de ces trois années.

Merci à toi, Pierre, mon co-auteur, mon collègue, mais surtout mon ami. Je garde en mémoire toutes ces nuits blanches passées à modifier notre preuve pour la énième fois, ou à chercher désespérément pourquoi notre algorithme ne convergeait pas. Merci pour tous ces moments (relativement) agréables passés ensemble à travailler, mais surtout pour tous les super moments que nous avons partagés en dehors des maths. Je te souhaite le meilleur pour la fin de ta thèse et pour tous tes futurs projets.

Ensuite, je tiens à remercier tous mes amis du Centre Lagrange, avec qui j'ai partagé mon quotidien pendant ces trois dernières années. Un grand merci à notre Shrek, Yazid, pour avoir illuminé toutes nos soirées. Merci à Gabriel pour nos passionnantes discussions sur les problèmes de la recherche, que ce soit autour d'un café au Beans ou d'une caïpirinha chez toi. Vincent, merci pour ta gentillesse infinie, ta résilience et ta sociabilité, tu incarnes le Centre à toi tout seul. Merci, Louis, pour toutes tes idées et tes convictions, ces années à tes côtés ont été incroyablement enrichissantes. J'ai commencé par être ton petit stagiaire, aujourd'hui, j'espère être devenu ton ami. Merci à toi Mehdi pour la bonne humeur et l'ambiance que tu apportes au Centre. Merci à toi, Lisa, la dernière arrivée au labo, pour ton enthousiasme et ta motivation. Merci Maxence pour tes apparitions, malheureusement trop rares,

au Lagrange, mais surtout pour cette incroyable semaine à Vienne pour ICML. Un grand merci aux tontons Pablo et Achille pour nous avoir intégré dans ce labo. Enfin, merci aux Suisses, Valentin et Thomas, c'est toujours un plaisir de découvrir que vous êtes en France pour la semaine. N'hésitez pas à nous prévenir à l'avance ! Merci également pour ces quelques jours à Zurich et pour toute la bonne humeur que vous apportez au Centre. Merci à tous pour ces discussions enrichissantes pendant nos pauses-café. Mais surtout, merci pour les moments magiques passés au Maroc, à Font-Romeu, ou simplement au bar à Paris pour célébrer la fin d'une deadline. Je ne vois pas comment j'aurais pu terminer cette thèse sans vous. Je vous souhaite à tous le meilleur, que ce soit dans la recherche ou ailleurs.

Je tiens à remercier Nicolas, une superbe rencontre faite à Singapour il y a déjà plus de quatre ans. C'est grâce à toi que je suis là aujourd'hui, sans tes conseils, je n'aurais jamais fait le MVA ni entrepris cette thèse. Merci pour ta bienveillance, ta patience et pour tous tes précieux conseils quand je me sentais perdu. Bon courage pour la fin de ta thèse !

Je remercie ensuite mes amis du Liryc à Bordeaux. J'ai passé deux étés très agréables à vos côtés. Un merci tout particulier à toi, Mariette, pour ton énergie, tous tes potins et pour tous les points qu'on t'a volés au Quizz Room. Mais surtout, merci pour ta gentillesse et pour tous les moments de rire partagés dans notre bureau. Je te souhaite, à toi aussi, bon courage pour la fin de ta thèse.

Un grand merci à tous les Jaunes de Supelec Rennes. C'est un véritable plaisir, même si trop rare à mon goût, de continuer à nous retrouver ensemble. Merci à chacun d'entre vous pour tous les super souvenirs, que ce soit à l'école, lors des repas de Noël ou pendant un Heb-dromadaire. Je vous souhaite à tous le meilleur.

Merci à mes potes de prépa, d'être toujours là après toutes ces années. Je tiens particulièrement à te remercier, Bastien, toi qui as été mon colocataire pendant la moitié de ma thèse. Merci de m'avoir supporté et soutenu dès le début de cette aventure. Je garde en mémoire tous ces moments magiques passés ensemble, surtout quand on partait surfer juste après le boulot. Même si, à mes yeux, tu restes toujours ce jeune de 17 ans, tu es devenu une personne extraordinaire, et je suis fier de t'avoir toujours à mes côtés.

Un immense merci à mes amis d'enfance : Paul, Alexis, Corentin et Maxime. Merci de m'accompagner depuis mon plus jeune âge, d'être constamment présents à mes côtés et d'être devenus de véritables membres de ma famille. Je pourrais difficilement compter tous les moments précieux que nous avons partagés, tant ils sont nombreux. Votre amitié est inestimable.

CONTENTS

List of publications

Gaussian Variational inference for Bayesian Neural Networks:

➢ **Huix, T.**, Majweski, S., Durmus, A., &Moulines, É. (2024, May). Variational Inference of overparameterized Bayesian Neural Networks: a theoretical and empirical study. Submitted at Transactions on Machine Learning Research (TMLR).

➢ Descours, A., **Huix, T.**, Guillin, A., Michel, M., Moulines, É., & Nectoux, B. (2023, July). Law of large numbers for bayesian two-layer neural network trained with variational inference. Accepted at Conference on Learning Theory (COLT).

➢ Descours, A., **Huix, T.**, Guillin, A., Michel, M., Moulines, É., & Nectoux, B. (2024). Central Limit Theorem for Bayesian Neural Network trained with Variational Inference. Submitted at Mathematical Statistics and Learning (MSL).

Mixture of Gaussian Variational inference:

➢ **Huix, T.**, Korba, A., Durmus, A., & Moulines, É. (2024, January). Theoretical Guarantees for Variational Inference with Fixed-Variance Mixture of Gaussians. Accepted at International Conference on Machine Learning (ICML).

Thompson Sampling for contextual Bandit problems:

➢ **Huix, T.**, Zhang, M., & Durmus, A. (2023, April). Tight regret and complexity bounds for thompson sampling via langevin monte carlo. Accepted at Artificial Intelligence and Statistics (AISTAT).

➢ **Huix, T.**, Clavier, P., & Durmus, A. (2024, january). VITS: Variational Inference Thomson Sampling for contextual bandits. Accepted at International Conference on Machine Learning (ICML).

LIST OF SYMBOLS

$\mathbb{R}$	Set of real numbers.
$\mathbb{N}$	Set of integers.
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors.
I_d	Identity matrix of size $d \times d$.
$\det(M)$	Determinant of matrix M.
tr	The trace function of a matrix.
$M^\top$	The transpose of a matrix M.
$\lambda_{\min}(A)$	The minimum eigen value of a symmetric-real matrix A.
$\lambda_{\max}(A)$	The maximum eigen value of a symmetric-real matrix A.
$\bar{E}$	The complementary of the event E.
$\mathbb{1}$	The indicator function.
$\mathbb{P}(\cdot)$	Probability of an event.
$\mathbb{E}[\cdot]$	Expectation of a random variable.
$\mathcal{P}(E)$	The set of probability measures on E.
$\mathcal{P}_2(E)$	The set of probability measures on E with bounded second moments.
$L^2(\mu)$	The space of functions such that $\int \|f\|^2 \mathrm{d}\mu \leq +\infty$.
$\mathcal{T}_\# \nu$	The pushforward measure of ν by $\mathcal{T}$.
$\|\cdot\|$	Norm.
$\langle \cdot, \cdot \rangle$	Inner product.
∇f	Gradient of $f : \mathbb{R}^d \to \mathbb{R}$.
$\nabla^2 \psi$	Hessian of $f : \mathbb{R}^d \to \mathbb{R}$.
$\mathrm{J}f$	The Jacobian of $f : \mathbb{R}^d \to \mathbb{R}^m$.
Δf	The Laplacian of f.
$\mathcal{N}(\mu, \Sigma)$	The d-multidimensional Gaussian distribution with mean μ and covariance Σ.
$\mathcal{S}_+^\star$	The set of symmetric positive definite matrices.
$\odot$	The component wise product.
$\tilde{O}$	The big O notation with logarithmic factors omitted.

CHAPTER 1

INTRODUCTION

1 Background on Bayesian Machine Learning

1.1 Classical Machine Learning

In this section we describe the classical Machine Learning (ML) framework for supervised tasks. First, let's start by defining the input data

$$x \in \mathrm{X}\,, \quad x \sim \mathcal{P}_X\,.$$

$\mathcal{P}_X$ is an unknown distribution on $(\mathrm{X}, \mathcal{B}(\mathrm{X}))$, where $\mathcal{B}(\mathrm{X})$ is the Borel σ-algebra on X. Similarly, the output data, associated to the input data x, is defined by

$$y \in \mathrm{Y}\,, \quad y \sim \mathcal{P}_{Y|X}(x)\,,$$

where y is generated conditionally on x. In this case, $\mathcal{P}_{Y|X}$ is a Markov kernel with source $(X, \mathcal{B}(\mathrm{X}))$ and target $(\mathrm{Y}, \mathcal{B}(\mathrm{Y}))$.

The main objective of supervised learning is to predict the target y given an input x. More precisely, to learn the Markov kernel $\mathcal{P}_{Y|X}$. However, this Markov kernel is obviously unknown. In parametric Machine Learning, this Markov kernel is aproximated by a parametric model $\{\mathcal{P}_\theta : \theta \in \Theta\}$ where $\mathcal{P}_\theta$ is also a Markov kernel on $(\mathrm{X}, \mathcal{B}(\mathrm{X})) \times (\mathrm{Y}, \mathcal{B}(\mathrm{Y}))$, parametrized by $\theta \in \Theta$. We assume that $\mathcal{P}_\theta$ admits a density with respect to some dominating measure λ_{ref}.

Example: Exponential Family

A well-known example of such parametric models is the Exponential family. Many standard distributions are members of the Exponential family, including the Normal, Poisson, and Binomial distributions. This family of distributions is defined by the following density function:

$$\frac{\mathrm{d}\mathcal{P}_\theta}{\mathrm{d}\lambda_{\text{ref}}}(y|x) := h(y)\ \exp(g(\theta,x)T(y) - C(\theta,x))$$

where $h : \mathbb{R} \to \mathbb{R}^+$ is called the base measure, $C : \Theta \times \mathrm{X} \to \mathbb{R}$ the log-partition function, $g : \Theta \times \mathrm{X} \to \mathbb{R}$ the natural parameter and $T : \mathbb{R} \to \mathbb{R}$ the natural sufficient statistic.

Consequently, the objective is now to learn the parametric Markov kernel that provides the closest approximation to the true distribution, based on a set of labeled examples. More precisely, given a set of n data points, called a dataset and denoted by $\mathrm{D}_n = (x_i, y_i) i \leq n$, we will try to find an estimator of $\theta_\star$ such that the Markov kernel $\mathcal{P}_{\theta^\star}$ "approximates" the true $\mathcal{P}_{Y|X}$. In the following paragraph we describe the most commonly used estimator, the so-called Maximum Likelihood Estimator.

Maximum Likelihood Estimator Let's first define the likelihood function associated to the observations D_n

$$\mathrm{L}_n(\theta|\mathrm{D}_n) \propto \exp\left\{\sum_{i=1}^{n} \ell(\theta|x_i, y_i)\right\} ,$$

where the log-likelihood is given by $\ell(\theta|x_i, y_i) = \log(\mathrm{d}\mathcal{P}_\theta/\mathrm{d}\lambda_{\text{ref}})$ and we have assumed that the data are iid. Then the MLE estimator is denoted by $\hat{\theta}_n$ and is defined as

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta\in\Theta} \mathrm{L}_n(\theta|\mathrm{D}_n) .$$

This optimization task may be solved by many approaches such as gradient descent for example. This estimator is consistent and asymptotically normal under mild conditions.

Main limitation of classical Machine Learning One limitation of traditional machine learning is the absence of uncertainty quantification. These methods do not account for epistemic uncertainty, which arises from uncertainty in the model parameters. While this uncertainty can be reduced with sufficient data, it is not eliminated at fixed n. According to [Guo et al., 2017], neural networks trained with maximum likelihood estimation are often miscalibrated and overconfident, meaning that their predicted probabilities do not accurately reflect the true likelihood. In many real-world applications, such as medical diagnosis or weather forecasting for example, uncertainty quantification is essential for making informed decisions.

Another limitation of traditional machine learning is its inability to incorporate prior knowledge. In many real-world applications, such as astronomy or agriculture, we have some prior knowledge about the parameters of the model. Incorporating this prior knowledge can help to constrain the model, making it more robust and improving learning efficiency. This can lead to better performance, especially in situations where data is limited or noisy.

The last limitation of traditional machine learning is the challenge of distribution shift. In many real-world applications, the distribution of data is not iid. This means that the distribution of data can differ between training and testing sets, a phenomenon known as distribution shift. According to [Hein et al., 2019], ReLU networks tend to be overly confident when faced with out-of-distribution data. Additionally, research by [Modas et al., 2022] and [Fawzi et al., 2016] has shown that neural networks used for classification can experience a significant drop in accuracy when faced with common corruptions.

To address all these limitations, the next section will introduce the Bayesian Machine Learning framework, which provides a more robust approach to handling uncertainty and distribution shift.

1.2 Bayesian Machine Learning

In Bayesian Machine Learning (BML), the setting considered is similar to the one of ML. The input data $x \in \mathrm{X}$ is generated by $x \sim \mathcal{P}_X$, the output data $y \in \mathrm{Y}$ is generated by $y \sim \mathcal{P}_{Y|X}$. The parametric model is also $\{\mathcal{P}_\theta : \theta \in \Theta\}$, and the dataset is $\mathrm{D}_n := \{(x_i, y_i)\}_{i \leq n}$.

However, the objective is not to find $\theta^\star$ such that $\mathcal{P}_{\theta^\star}$ "approximates" $\mathcal{P}_{Y|X}$. Instead, the goal is to find the distribution of the parameters θ given the data D_n, known as the **posterior distribution**.

The first step of Bayesian Machine Learning is to integrate a prior knowledge p_0 which represents initial beliefs about the parameters before seeing the data.

By specifying both a prior p_0 and a likelihood L_n, we obtain the joint distribution of the parameters θ and the data D_n using the product rule of probability: $\mathbb{P}(\theta, \mathrm{D}_n) = \mathrm{p}_0(\theta) \times \mathrm{L}_n(\theta|\mathrm{D}_n)$. This combination of the prior and the likelihood, when applied through Bayes' rule, results in the posterior distribution denoted $\hat{\mathrm{p}}$ and given by

$$\hat{\mathrm{p}}(\theta|\mathrm{D}_n) = \frac{\mathrm{L}_n(\theta|\mathrm{D}_n)\mathrm{p}_0(\theta)}{\mathrm{p}(\mathrm{D}_n)}$$

The denominator $\mathrm{p}(\mathrm{D}_n)$ in the Bayes' rule is called the normalizing constant, the model evidence or the marginal likelihood and is given by

$$\mathrm{p}(\mathrm{D}_n) = \int \mathrm{L}_n(\theta|\mathrm{D}_n)\mathrm{d}\mathrm{p}_0(\theta)$$

Note that this term does not depend on θ, meaning it provides no information for the optimization task. Furthermore, this integral is often intractable, which is a known issue in Bayesian Machine Learning.

Finally, in Bayesian Machine Learning the prediction is made by integrating over the posterior distribution. The expectation of a function $f(\theta)$ under the posterior is computed as follow:

$$\mathbb{E}[f(\theta)|\mathrm{D}_n] = \int f(\theta)\hat{\mathrm{p}}(\theta|\mathrm{D}_n)\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)\,,$$

where θ_i are samples from the posterior distribution and this integral approximation is known as Monte Carlo approximation. For example, given a new test data point $(x^\star, y^\star)$, the predictive distribution is given by

$$p(y^\star|x^\star, \mathrm{D}_n) = \int \mathrm{L}(\theta|y^\star, x^\star)\hat{\mathrm{p}}(\theta|\mathrm{D}_n)\mathrm{d}\theta\,.$$

Comparaison with classical ML Firstly, in Bayesian Machine Learning the epistemic uncertainty is directly taken into account through the posterior distribution. In [Kristiadi et al., 2020], it was demonstrated that the Bayesian paradigm reduces the overconfidence problem. This method allows to obtain well-calibrated estimators. Then, prior knowledge is directly incorporated through the prior distribution. This prior is particularly valuable in settings with limited data. The prior serves to regularize the model, resulting in more robust estimators. Finally, the Bayesian Machine Learning framework addresses distribution shifts by naturally incorporating uncertainty. The posterior distribution tends to concentrate in regions corresponding to the training data. As a result, regions of out-of-distribution data correspond to areas of high uncertainty, mitigating issues of overconfidence in such regions.

Main issue: The posterior distribution is often intractable, meaning it cannot be computed exactly or efficiently. This is due to the intractability of the normalizing constant, which involves a high-dimensional integral that is often impossible to solve analytically. In such cases, we need to find a distribution that closely approximates the posterior distribution, this method is known as **approximate inference**. In the following sections, we will introduce the two most commonly used approximate inference methods in Bayesian Machine Learning: the Monte Carlo Markov Chain (MCMC) and the Variational Inference (VI).

1.2.1 Approximate Bayesian inference with MCMC

One of the most commonly used approximate inference method is the so-called Markov Chain Monte Carlo (MCMC). Let's start by defining a Markov Chain with values in $\mathbb{R}^d$ endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ which is a sequence of random variables $(X_k)_{k\in\mathbb{N}}$ defined on a filtered space $(\mathcal{F}_k)_{k\in\mathbb{N}}$ such that

$$\mathbb{P}(X_{k+1} \in A|\mathcal{F}_k) = \mathbb{P}(X_{k+1} \in A|X_k) \quad \mathbb{P}-\text{a.s.}$$

A homogeneous Markov Chain $(X_k)_{k\in\mathbb{N}}$ is characterized by a Markov kernel $\mathrm{K} : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to [0,1]$. The objective of MCMC methods is to construct a Markov kernel K having a unique invariant distribution $\hat{\mathrm{p}}(\cdot|\mathrm{D}_n)$, ie,

$$\int_{\mathbb{R}^d} K(x,A)\hat{\mathrm{p}}(\mathrm{d}x|\mathrm{D}_n) = \hat{\mathrm{p}}(A|\mathrm{D}_n) \quad \forall A \in \mathcal{B}(\mathbb{R}^d)$$

It means that applying K to a sample X distributed according to the posterior will result in a sample Y also distributed according to the posterior. Knowing this Markov kernel K, we can iteratively sample a new parameter X_{k+1} according to $K(X_k,\cdot)$. After a sufficiently large number of iterations, the parameters generated by the Markov Chain will be distributed according to a distribution that is close the posterior distribution $\hat{\mathrm{p}}(\cdot|\mathrm{D}_n)$. Thus, this method allows us to sample parameters from the posterior, even though it is known only up to a normalizing constant.

Example: Langevin Monte Carlo

One well-known MCMC method is the Langevin Monte Carlo (LMC). This scheme is derived from the Euler-Maruyama approximation of the Langevin diffusion [Roberts and Tweedie, 1996a] and is given by

$$X_{k+1} = X_k + \gamma\nabla\log(\mathrm{L}_n(X_k|\mathrm{D}_n)) + \sqrt{2\gamma}\epsilon_{k+1}, \tag{1.1}$$

where $\epsilon_{k\geq 0}$ are i.i.d. standard Gaussian noises, and $\gamma \geq 0$ denotes the time discretization step-size. The scheme allows to define a Markov chain with a transition kernel K_γ given by

$$K_\gamma(X,A) = \frac{1}{(4\pi\gamma)^{d/2}}\int_A \exp\left(-\frac{1}{4\gamma}\|\tilde{X} - X + \gamma\nabla\log(\mathrm{L}_n(X|\mathrm{D}_n))\|^2\right)\mathrm{d}\tilde{\mathrm{X}}$$

Sampling from the Markov kernel $\mathrm{K}_\gamma(X,\cdot)$ is equivalent to updating the parameter X following the Langevin Monte Carlo scheme (1.1). Under certain assumptions on γ and the Likelihood (see [Dalalyan, 2017, Durmus and Moulines, 2017] for additional details), the distribution of $(X_k)_{k\in\mathbb{N}}$ converges to samples of the biased stationary distribution $\hat{\mathrm{p}}_\gamma$ as k goes to infinity. Moreover, $\hat{\mathrm{p}}_\gamma$ approaches the posterior distribution $\hat{\mathrm{p}}$ as γ tends to zero. The Langevin Monte Carlo method is adapted to high-dimensional problems.

MCMC methods, can be quite expensive in terms of computation, especially when dealing with large datasets (see [Cobb and Jalaian, 2021]) or large models (see [Izmailov et al., 2021]). Additionally, in non-convex setting, achieving convergence of the Markov Chain may require a significant number of steps. In these cases, variational inference provides a good alternative approach to approximate the posterior distribution.

1.2.2 Variational Inference

Variational inference (VI) [Hinton and Camp, 1993, MacKay, 1995, MacKay et al., 1995, Blei et al., 2017] has emerged as a powerful alternative in Bayesian inference. By framing the problem as an optimization task, VI aims to find an approximate candidate distribution within a parametric family of distributions $\mathcal{G}$ that minimizes the (reverse) Kullback-Leibler (KL) divergence to the target:

$$\tilde{\mathrm{q}} = \underset{\mathrm{p}\in\mathcal{G}}{\operatorname{argmin}}\,\mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}})$$

where $\mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}}) = \int\log(\mathrm{dp}/\mathrm{d}\hat{\mathrm{p}})\mathrm{dp}$ if p is absolutely continuous with respect to $\hat{\mathrm{p}}$ and $+\infty$ else.

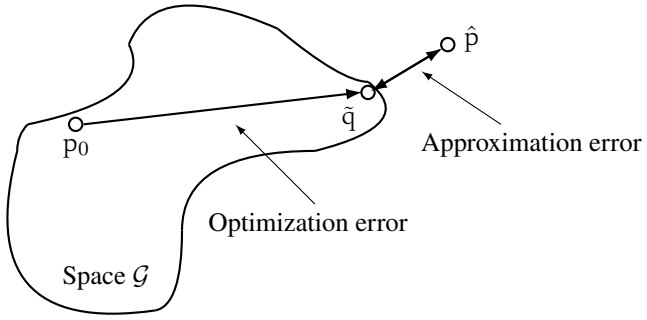


Fig. 1.1 Variational Inference problem.

An important choice in VI is the variational family $\mathcal{G}$. The flexibility of the variational family is crucial to capture the true posterior distribution. However, the complexity of the variational family should be kept low to ensure efficient optimization. One important example is the non-degenerate Gaussian variational family [Lambert et al., 2022a, Diao et al., 2023], $\mathcal{G} = \{\mathcal{N}(\mu, \Sigma) | \mu \in \mathbb{R}^d, \Sigma \in \mathscr{S}_+^\star\}$. This choice of VI family is supported by the Bernstein-Von Mises theorem [Van der Vaart, 2000]. This theorem, subject to specific regularity conditions, asserts that a properly scaled version of the posterior converges to a Gaussian as the sample size grows. Consequently, in the regime of large data, the Gaussian variational family is well-suited to approximate the posterior distribution.

However, in the setting of limited data, the posterior distribution may be multimodal. In this case, the Gaussian Variational family may not be able to capture the multimodal structure of the posterior distribution. In such cases, a larger family of distributions, such as the mixture of Gaussians, can be considered [Gershman et al., 2012, Arenz et al., 2018, Lin et al., 2019]. This illustrates the first source of error in Variational Inference: the approximation error. This error quantifies how far $\tilde{\mathrm{q}}$ is from $\hat{\mathrm{p}}$. The second error is the optimization error, which is the error introduced by the optimization algorithm used to minimize the KL divergence to approach $\tilde{\mathrm{q}}$.

In the next paragraphs, we will present two popular algorithms used to solve the Variational Inference problem, both of which have been studied in this thesis: Bayes by Backprop and Riemannian Gradient Descent.

Bayes By Backprop: The algorithm derived in [Blundell et al., 2015] and called Bayes by Backprop is commonly used to train Bayesian Neural Networks (more details about BNN can be found in Section 1.3). The variational family considered for the weights of the Neural Network is

$$\mathcal{G} = \left\{\mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) | \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}^{+d}\right\}$$

This setting is called the mean-field Gaussian variational family and is particularly suited for large Bayesian Neural Networks. As the number of neurons increases, they tend to become more independent. Consequently, the covariance matrix of the posterior becomes diagonal. The parametric variational distribution can be rewritten as

$$\tilde{\mathrm{q}}(\theta|w) = \prod_{i=1}^{d} \tilde{\mathrm{q}}^1(\theta_i|w_i) = \prod_{i=1}^{d} \mathcal{N}(\theta_i|\mu_i, \sigma_i^2)\,,$$

where $\theta = (\theta_1, \ldots, \theta_d)$, $\mu = (\mu_1, \ldots, \mu_d)$, $\sigma = (\sigma_1, \ldots, \sigma_d)$ and $w = (\mu, \sigma)$ are the variational parameters. The optimization problem is then defined as

$$\begin{aligned} w^\star &= \underset{w \in (\mathbb{R}^d \times \mathbb{R}^{+d})}{\mathrm{argmin}} \; \mathrm{KL}(\tilde{\mathrm{q}}(\theta|w) | \hat{\mathrm{p}}(\theta|\mathrm{D}_n)) \\ &= \underset{w \in (\mathbb{R}^d \times \mathbb{R}^{+d})}{\mathrm{argmin}} \; \mathrm{KL}(\tilde{\mathrm{q}}(\theta|w) | \mathrm{p}_0(\theta)) - \mathbb{E}_{\theta \sim \tilde{\mathrm{q}}(\theta|w)}[\log(\mathrm{L}_n(\theta|\mathrm{D}_n))] \\ &= \underset{w \in (\mathbb{R}^d \times \mathbb{R}^{+d})}{\mathrm{argmax}} \; \mathrm{ELBO}(w)\,, \end{aligned}$$

where ELBO is called Evidence Lower Bound or variational free energy. This cost function is a sum of a regularization term (the KL between the variational posterior and the prior) and a data-driven term (the expected log-likelihood). A useful trick used to optimize the ELBO is the reparametrization trick. It allows to rewrite the variational distribution $\tilde{\mathrm{q}}^1(\cdot|w_i)$ as the pushforward of a reference probability measure with density γ by a deterministic map $\mathcal{T}_{w_i}$. In this case, we choose $\gamma = \mathcal{N}(0,1)$, the standard Gaussian distribution. Then, we have

$$\tilde{\mathrm{q}}^1(\cdot|w_i) = \mathcal{T}_{w_i}\#\gamma \quad , \qquad \mathcal{T}_{w_i} : z \to \mu_i + \sigma_i \odot z$$

In [Blundell et al., 2015], the authors show that with this reparametrization trick, we can exchange the derivative and the expectation:

$$\frac{\partial}{\partial w_i}\mathbb{E}_{\tilde{\mathrm{q}}^1(\theta_i|w_i)}[f(\theta_i, w_i)] = \frac{\partial}{\partial w_i}\mathbb{E}_{z\sim\gamma}[f(\mathcal{T}_{w_i}(z), w_i)] = \mathbb{E}_\gamma[\frac{\partial f(\theta_i, w_i)}{\partial \theta_i}\frac{\partial \mathcal{T}_{w_i}(z)}{\partial w_i} + \frac{\partial f(\mathcal{T}_{w_i}(z), w_i)}{\partial w_i}]$$

This result allows us to optimize the ELBO using Stochastic Gradient Descent. However, computing the expectation with respect to γ is often intractable. In practice, we approximate this expectation using Monte Carlo methods, which involve drawing samples from the distribution γ and averaging the results.

Riemannian gradient descent: Another approach used to solve the Variational Inference problem is the Riemannian gradient descent and is described in [Lambert et al., 2022b]. In this case we consider as variational family the set of non-degenerate Gaussian distributions

$$\mathcal{G} = \{\mathcal{N}(\mu, \Sigma) | \mu \in \mathbb{R}^d, \Sigma \in \mathscr{S}_+^\star\}$$

As explained previously, this choice of Gaussian variational family is justified by Bernstein-Von Mises theorem. As noted in [Lambert et al., 2022b], $\mathcal{G}$ equipped with the Wasserstein distance of order 2 is a complete metric space as a closed subset of $\mathcal{P}_2(\mathbb{R}^d)$. Recall that for two Gaussian distributions $p_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $p_1 = \mathcal{N}(\mu_1, \Sigma_1)$, their Wasserstein distance has a closed form:

$$W_2^2(p_0, p_1) = \|\mu_0 - \mu_1\|^2 + \mathrm{tr}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}) \, .$$

This Wasserstein distance on $\mathcal{G}$ allows to derive a Riemannian metric denoted $\mathfrak{g}$. The corresponding geodesic is given through the exponential map. More precisely, for a Gaussian distribution $p = \mathcal{N}(\mu_p, \Sigma_p)$, this map is defined as follows:

$$\begin{aligned} \exp_p(\mu_v, \Sigma_v) =& (\mu_p + \mu_v + (\Sigma_v + \mathrm{I}_d)(\cdot - \mu_p))_\# p \\ =& \mathcal{N}(\mu_p + \mu_v, \, (\Sigma_v + \mathrm{I}_d)\, \Sigma_p \, (\Sigma_v + \mathrm{I}_d)) \, . \end{aligned} \tag{1.2}$$

With all these preliminaries, we can now present and motivate the algorithm developed in [Lambert et al., 2022b] to efficiently solve the Variational Inference problem. This method can be formalized as a Riemannian gradient descent scheme on $\mathcal{G}$. Firstly, we define the loss function $\mathcal{F} : \mathrm{p} \to \mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}})$. Moreover we also define the potential function $U(\theta) \propto -\log \hat{\mathrm{p}}(\theta|\mathrm{D}_n)$. Then, following [Lambert et al., 2022b], we derive the gradient operator of $\mathcal{F}$ on $\mathcal{G}$ equipped with $\mathfrak{g}$ as

$$\nabla_{\mathfrak{g}}\mathcal{F}(\mathrm{p}) = (\int \nabla U(\theta) \mathrm{dp}(\theta), \, \int \nabla^2 U(\theta) \mathrm{dp}(\theta) - \Sigma_{\mathrm{p}}^{-1}) \tag{1.3}$$

where Σ_{p} is the covariance matrix of the Gaussian distribution p. From this expression, the corresponding Riemannian gradient descent [Bonnabel, 2013] using a step size $\gamma > 0$ defines the sequence of iterates $\{\mathrm{q}_k\}$ recursively as:

$$\mathrm{q}_{k+1} = \exp_{\mathrm{q}_k}(-\gamma \nabla_{\mathfrak{g}}\mathcal{F}(\mathrm{q}_k)) \, .$$

Combining (1.2) and (1.3), this recursion allows to define a sequence of means $\{\mu_k\}$ and covariance matrices $\{\Sigma_k\}$ by the recursions

$$\begin{aligned}
\mu_{k+1} &= \mu_k - \gamma \int \nabla U(\theta) \mathrm{d}\mathrm{q}_k(\theta) \\
\Sigma_{k+1} &= A_k \Sigma_k A_k \\
A_k &= \mathrm{I}_d - \gamma \Big(\int \nabla^2 U(\theta) \mathrm{d}\mathrm{q}_k(\theta) - \Sigma_k^{-1} \Big) \\
\mathrm{q}_{k+1} &= \mathcal{N}(\mu_{k+1}, \Sigma_{k+1})
\end{aligned}$$

The main computational challenge in this recursion stems is that the integrals involved are typically intractable. To overcome this issue, we employ a Monte Carlo procedure to approximate these integrals. Subsequently, we consider a sequence of mean values denoted as $\{\tilde{\mu}_k\}$ and covariance matrices $\{\tilde{\Sigma}_k\}$ such that:

$$\begin{aligned}
\tilde{\mu}_{k+1} &= \tilde{\mu}_k - \gamma \nabla U(\tilde{\theta}_k) \\
\tilde{\Sigma}_{k+1} &= \tilde{A}_k \tilde{\Sigma}_k \tilde{A}_k \\
\tilde{A}_k &= \mathrm{I}_d - \gamma (\nabla^2 U(\tilde{\theta}_k) - \tilde{\Sigma}_k^{-1}) \\
\tilde{\theta}_k &\sim \mathcal{N}(\tilde{\mu}_k, \tilde{\Sigma}_k) \, .
\end{aligned}$$

1.3 Bayesian Neural Networks

A nice application of Bayesian Machine Learning is the Bayesian Neural Networks (BNN). In recent years, neural networks, and specifically deep learning, have emerged as the leading approach for various regression and classification tasks across multiple domains, including computer vision and natural language processing. The L-layers Feed-Forward Neural Network $f_\theta : \mathbb{R}^{\mathsf{d_X}} \to \mathbb{R}^{\mathsf{d_Y}}$ can be written as

Input layer:	$y_1 = \sigma(a_1 x + b_1)$	
Hidden layers:	$y_l = \sigma(a_l y_{l-1} + b_l)$	for any $l = 2, \dots, L-1$
Output layer:	$f(x) = a_L y_{L-1} + b_L$	

where σ is the activation function and $\theta = \{(a_i, b_i)\}_{i=1}^L$ are the NNs parameters. We have $(a_1, b_1) \in \mathbb{R}^{M \times \mathsf{d_X}} \times \mathbb{R}^M$, $(a_1, b_1) \in \mathbb{R}^{M \times M} \times \mathbb{R}^M$ for any $l = 2, \dots, L-1$ and $(a_L, b_L) \in \mathbb{R}^{\mathsf{d_Y} \times M} \times \mathbb{R}^{\mathsf{d_Y}}$. However, this kind of networks are typically trained using maximum likelihood estimation, which does not provide a measure of uncertainty in the model predictions. Furthermore, as reported in [Guo et al., 2017], neural networks are overconfident in their predictions, meaning that they assign overly high probabilities to their predictions. To address this issue, Bayesian Machine Learning techniques can be employed. This approach involves selecting a prior distribution for the neural network parameters, with the aim of approximating the posterior distribution. By doing so, we can incorporate prior knowledge, quantify uncertainty, and improve the robustness and accuracy of our models.

However, Deep Neural Networks are high dimensional models, often reaching hundreds of millions of parameters. Furthermore, the weights of neural networks can be permuted across layers, resulting in multiple solutions that yield the same outcomes for any given minimum. Consequently, neural networks cannot be convex. Due to these complexities, traditional Bayesian methods, which rely on MCMC, are not directly applicable [Izmailov et al., 2021]. A more suited approach to train BNNs is Variational Inference [Blei et al., 2017, Graves, 2011, Hinton and Camp, 1993].

2 Background on Bandit problems

2.1 Bandit problem

2.1.1 Presentation of the Bandit problem

Bandit problem is a sequential decision making problem where an agent has to choose iteratively among sevelar possible actions, called “arms”. Each action has an associated reward distribution which is unknown. The objective of the agent is to maximize the total expected rewards obtained. It was firstly introduced in [Thompson, 1933] by William R. Thompson to study medical trials.

The 2-armed bandit problem The simpliest form of bandit problem is the 2-armed bandit problem, imagine a client in a casino faced with a choice between two slot machines. The customer can either play the left machine or the right machine. Each slot machine has an unknown distribution of rewards. The customer's goal is to choose which machine to play at each round to maximize their total rewards.

Fig. 1.2 Illustration of the 2-armed bandit problem.

The main challenge in this task is to effectively manage a suitable **exploitation** and **exploration** trade-off [Robbins, 1952, Katehakis and Veinott, 1987, Berry and Fristedt, 1985, Auer et al., 2002, Lattimore and Szepesvári, 2020, Kveton et al., 2020a].

- **Exploitation:** refers to selecting an arm that is currently believed to be the best based on past observation
- **Exploration:** refers to selecting arms that have not been selected frequently in the past in order to gather more information on the reward distribution of these arms.

2.1.2 Mathematical framework

The Bandit problem is a game played over T rounds (called the **horizon**). Let's denote by A the set of arms. A bandit process can be defined as follows: at each iteration $t \in [T]$ and given the history $\mathrm{D}_{t-1} = \{(a_s, r_s)\}_{s<t}$

- The agent chooses an action $a_t \sim \mathbb{Q}_t(\cdot|\mathrm{D}_{t-1})$.
- The environment reveals a reward $r_t \sim \mathrm{R}(\cdot|a_t)$. Here, $\mathrm{R}(\cdot|a_t)$ is an unknown Markov Kernel on $\mathrm{A} \times \mathbb{R}$.

The sequence of conditional distributions $\mathbb{Q}_{1:T} = \{\mathbb{Q}_t\}_{t\leq T}$ describes the strategy used by the agent to choose the action at each round. The objective of the agent is to find the sequence $\mathbb{Q}_{1:T}$ that maximizes the sum of rewards defined as follow

$$\mathrm{SReward}(\mathbb{Q}_{1:T}) = \sum_{t\leq T} f(a_t) \,,$$

where $f(a) = \int r\mathrm{R}(\mathrm{d}r|a)$ is the expected reward associated to an action $a \in \mathrm{A}$. Maximizing the sum of rewards is equivalent to minimizing the cumulative regret, a concept that is more commonly studied in the bandit literature. The cumulative regret is defined as

$$\mathrm{CREG}(\mathbb{Q}_{1:T}) = \sum_{t\leq T} f(a^\star) - f(a_t) \,,$$

where $a^\star = \mathrm{argmax}_{a\in\mathrm{A}} f(a)$ is the best action.

2.1.3 Main real-world applications

A/B Testing: A/B testing is a statistical method used to compare two versions of a webpage to determine which one performs better. Users are randomly assigned to either the control group (website A) or the treatment group (website B), and their behavior is tracked to see which version produces better results. 2-armed Bandit algorithms can be used to optimize A/B testing by dynamically allocating traffic to the best-performing web-site. More precisely, at each time step t, a user connects to the website, a bandit algorithm is used to choose an action $a_t \in \{\text{website A}, \text{website B}\}$, and the reward is $r_t = 1$ if the user purchases the product (or make other determined action) and $r_t = 0$ otherwise.

Recommendation systems: Bandit algorithms are also widely used in recommendation systems to optimize the selection of items to display to users. In this context, the items are the arms, and the reward is the user's engagement with the item (e.g., clicks, purchases, likes). The goal is to maximize the user's engagement by selecting the most relevant items.

Clinical trials: Bandit algorithms are also used in clinical trials to optimize the allocation of treatments to patients. In a clinical trial, patients are randomly assigned to different treatments, and their outcomes are observed to determine which treatment is most effective. Bandit algorithms can be used to dynamically allocate treatments based on the observed outcomes to maximize the overall benefit to patients. For instance see [Réda, 2022].

While bandit problems are useful for some real-world applications, they have limitations and cannot model all scenarios. One significant limitation is that the reward distribution for a given action is fixed. However, in many situations, the reward distribution may depend on the state of the environment, which can be represented by a state vector. For example, in clinical trials the effectiveness of a treatment may depend on some patients' characteristics. This setting is called contextual bandit problem and will be explored in the following section and will be the focus of this thesis

2.2 Contextual Bandit problems

Contextual bandit problem is a particular instance of Multi-armed Bandit problem, which supposes, at each round, that the set of arms and the corresponding reward depend on a d-dimensional feature vector called a contextual vector or context. This setting allows the agent to take into account the state of the environment when choosing an action. The classic bandit problem can be seen as a special case of the contextual bandit problem where the context is a constant vector. For instance, in a recommender system, this approach enables personalized recommendations for each user based on their characteristics. Similarly, in a clinical trial, the treatment can be tailored to the patient's features. By taking into account contextual information, these systems can improve their accuracy, efficiency, and overall performance.

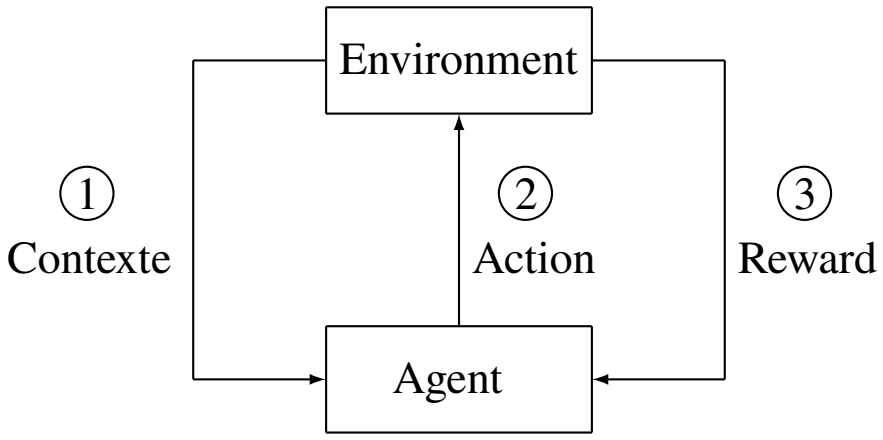


Fig. 1.3 Illustration of the contextual bandit problem.

This scenario has been extensively studied over the past decades and learning algorithms have been developed to address this problem [Langford and Zhang, 2007a, Abbasi-Yadkori et al., 2011a, Agrawal and Goyal, 2013a, Kveton et al., 2020a], and they have been successfully applied in several real-world problem such as recommender systems, mobile health and finance [Li et al., 2010, Agarwal et al., 2016a, Tewari and Murphy, 2017, Bouneffouf et al., 2020].

2.2.1 Mathematical framework

We now present in more details the contextual bandit framework. Let X be a contextual space and consider $\mathcal{A} : \mathrm{X} \to 2^{\mathrm{A}}$ a set-valued action map, where 2^{A} stands for the power set of the action space A. For simplicity, we assume here that $\sup_{x\in\mathrm{X}} \mathrm{Card}(\mathcal{A}(x)) < +\infty$. A (deterministic or random) function $\pi : \mathrm{X} \to \mathrm{A}$ is said to be a policy if for any $x \in \mathsf{X}$, $\pi(x) \in \mathcal{A}(x)$. Then, for a fixed horizon $T \in \mathbb{N}$, a contextual bandit process can be defined as follows: at each iteration $t \in [T]$ and given the past observations $\mathrm{D}_{t-1} = \{(x_s, a_s, r_s)\}_{s<t}$:

- The agent receives a contextual feature $x_t \in \mathsf{X}$;
- The agent chooses an action $a_t = \pi_t(x_t)$ where π_t is a policy sampled from $\mathbb{Q}_t(\cdot|\mathrm{D}_{t-1})$;
- Finally, the agent receives a reward r_t sampled from $\mathrm{R}(\cdot|x_t, a_t)$ given D_{t-1}. Here, R is a Markov kernel on $(\mathrm{A} \times \mathrm{X}) \times \mathsf{R}$, where $\mathsf{R} \subset \mathbb{R}$

For a fixed family of conditional distributions $\mathbb{Q}_{1:T} = \{\mathbb{Q}_t\}_{t\leq T}$, this process defines a random sequence of policies, $\pi_{1:T} = \{\pi_t\}_{t\leq T}$ with distribution still denoted by $\mathbb{Q}_{1:T}$ by abuse of notation. Let's defined the optimal expected reward for a contextual vector $x \in \mathrm{X}$ and the expected reward given x and any action $a \in \mathcal{A}(x)$ as follow

$$f_\star(x) = \max_{a\in\mathrm{A}(x)} f(x,a)\,, \quad f(x,a) = \int r\mathrm{R}(\mathrm{d}r|x,a)\,.$$

The main challenge of a contextual bandit problem is to find the distribution $\mathbb{Q}_{1:T}$ that minimizes the cumulative regret defined as

$$\begin{aligned} &\mathrm{CREG}(\mathbb{Q}_{1:T}) = \textstyle\sum_{s\leq T} \mathrm{Regret}_s^{\pi_s} \\ &\text{with} \quad \mathrm{Regret}_s^{\pi_s} = f_\star(x_s) - f(x_s, \pi_s(x_s)). \end{aligned} \tag{1.4}$$

Similar to the classic bandit problem, Contextual Bandit problems also face the challenge of balancing the exploitation-exploration trade-off. The existing algorithms for addressing these problems can be broadly categorized into two groups. The first category is based on maximum likelihood and the principle of optimism in the face of uncertainty (OFU) and has been studied in [Auer et al., 2002, Chu et al., 2011, Abbasi-Yadkori et al., 2011b, Li et al., 2017a, Ménard and Garivier, 2017, Zhou et al., 2020, Foster and Rakhlin, 2020, Zenati et al., 2022]. The second category consists in randomized probability matching algorithms, which is based on Bayesian belief and posterior sampling. Thompson Sampling (TS) is one of the most famous algorithms that fall into this latter category. Since its introduction by [Thompson, 1933], it has been widely studied, both theoretically and empirically [Agrawal and Goyal, 2012, Kaufmann et al., 2012a, Agrawal and Goyal, 2013a, Russo and Van Roy, 2014, 2016, Lu and Van Roy, 2017, Riquelme et al., 2018, Jin et al., 2021a]. Despite the fact that OFU algorithms offer better theoretical guarantees compared to classic TS-based algorithms, traditional TS methodologies still appeal to us due to their straightforward implementation and empirical advantages. In [Agrawal and Goyal, 2012], the authors claimed that: "In applications like display advertising and news article recommendation, TS is competitive with or better than popular methods such as UCB". Similarly, [Chapelle and Li, 2011] has examined the empirical performances of TS on both simulated and real data. Their experiments demonstrate that TS outperforms OFU methods, leading them to conclude: "In any case, TS is very easy to implement and should thus be considered as a standard baseline". Taking all these factors into account, we have decided to focus on TS-based algorithms for addressing contextual bandit problems.

2.2.2 Thompson Sampling for contextual bandit

In this subsection, we will discuss the Thompson Sampling algorithm for contextual bandit problems. The key idea behind Thompson Sampling is to use the Bayesian approach to model the uncertainty of the reward distribution. At each time step $t \in [T]$, the algorithm samples a policy from the posterior distribution and selects the action according to this policy. After observing the reward, the posterior distribution is updated to reflect the new information.

More precisely, let's stat by choosing a parametric model $\{\mathrm{R}_\theta : \theta \in \mathbb{R}^d\}$ for the reward distribution, where for any θ, R_θ is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$ parameterized by $\theta \in \mathbb{R}^d$. We assume that R_θ admits a density with respect to some dominating measure λ_{ref}. For instance, let's consider the exponential family presented in Section 1.1. With the introduced notations, the likelihood function associated to the observations D_t at step $t > 1$ is given by

$$\mathrm{L}_t(\theta) \propto \exp\left\{\sum_{s=1}^{t-1} \ell(\theta|x_s, a_s, r_s)\right\} ,$$

where the log-likelihood is given by $\ell(\theta|x_s, a_s, r_s) = \log(\mathrm{dR}_\theta/\mathrm{d}\lambda_{\mathrm{ref}})(r_s|x_s, a_s)$. Choosing a prior on θ with density p_0, and applying Bayes formula, the posterior distribution at round $t \in [T]$ is given by

$$\hat{\mathrm{p}}_t = \mathrm{L}_t(\theta)\mathrm{p}_0(\theta)/\mathfrak{Z}_t$$

where $\mathfrak{Z}_t = \int \mathrm{L}_t(\theta)\mathrm{p}_0(\theta)\mathrm{d}\theta$ denotes the normalizing constant and we used the convention that $\hat{\mathrm{p}}_1 = \mathrm{p}_0$. Moreover we define the potential function $U(\theta) \propto -\log \hat{\mathrm{p}}_t(\theta)$. Then, at each iteration $t \in [T]$, TS consists in sampling a parameter θ_t from the posterior $\hat{\mathrm{p}}_t$ and from it, use as a policy, $\pi_t^{(\mathrm{TS})}(x)$ defined for any x by

$$\pi_t^{(\mathrm{TS})}(x) = a^{\theta_t}(x) , \quad a^\theta(x) = \operatorname*{argmax}_{a \in \mathcal{A}(x)} \int r \mathrm{R}_\theta(\mathrm{d}r|x, a)$$

The pseudo-code associated to Thompson Sampling is given in Algorithm 1.

Algorithm 1 Thompson Sampling

for $t = 1, \ldots, T$ **do**
  receive a context $x_t \in \mathsf{X}$
  sample θ_t from $\hat{\mathrm{p}}_t$
  choose $a_t = \pi_t^{(\mathrm{TS})}(x_t)$
  receive $r_t \sim \mathrm{R}(\cdot|x_t, a_t)$
  update the posterior $\hat{\mathrm{p}}_{t+1}$ with the new data point (x_t, a_t, r_t).
end for

The main challenge in Thompson Sampling is to sample from the posterior distribution. Since $\mathfrak{Z}_t$ is generally intractable, sampling from the posterior distribution is not in general an option. In this case, we need to use approximation methods, such as Variational Inference or Langevin Monte Carlo.

CHAPTER 2

RÉSUMÉ DE LA THÈSE

Dans cette section nous résumons les différentes contributions de cette thèse. Celle-ci s'articule autours de deux axes principaux: les garanties théoriques pour l'Inférence Variationnelle et les algorithmes de Thompson Sampling pour les problèmes de bandits.

1 Partie I: Les garanties théoriques pour l'Inférence Variationnelle.

Pour les trois premiers chapitres (chapitre 4-5-6) de cette thèse nous allons étudier théoriquement les Réseaux de Neurones Bayesiens surparamétrés entrainés par Inférence Variationnelle Gaussienne. Dans le dernier chapitre (chapitre 7) de cette première partie, nous considérons une famille Variationnelle plus large, capable notament de modéliser les distributions multimodales, j'ai nommé les Mixtures de Gaussiennes.

1.1 Chapitre 4: Inférence Variationnelle pour les Réseaux de Neurones Bayésiens surparamétrés: étude théorique et pratique du Tempering.

Dans ce travail nous allons considérer un réseau de neurone bayesien $f_{\bar{w}} : \mathbb{R}^{\mathsf{dx}} \to \mathbb{R}^{\mathsf{dy}}$ avec une seule couche cachée qui, pour tout $x \in \mathbb{R}^{\mathsf{dx}}$ est décrit par:

$$f_{\bar{w}}(x) = \frac{1}{N}\sum_{j=1}^{N} s(w_j, x) \;, \quad s(w_j, x) = a_j h(b_j, x) \;,$$

où $h(\cdot,\cdot)$ est la fonction d'activation et $\bar{w}$ désigne les paramètres du réseau. Ils se décomposent en N différents neurones chacun avec un paramètre $w_j = (a_j, b_j) \in \mathbb{R}^{\mathsf{dx}} \times \mathbb{R}^{\mathsf{dy}}$. L'ensemble des paramètres du modèle est $\bar{w} = \{w_j\}_{j \leq N}$. Nous allons ensuite appliquer, à ce Réseau de Neurones, l'algorithm Bayes-By-Backprop [Blundell et al., 2015] qui a été présenté dans la Section 1.2.2 du Chapitre 1. Pour rappel, cette méthode utilise une approche Bayesienne d'Inférence Variationnelle sur le Réseau de Neurones. On définit alors la vraisemblance $\mathrm{L}(\bar{w}|\mathrm{D}_p) \propto \prod_{i=1}^{p} \exp(-\ell(f_{\bar{w}}(x_i), y_i))$, la distribution à priori $\mathrm{p}_0(\bar{w})$ et on obtient la distribution à posteriori $\hat{\mathrm{p}}(\bar{w}|\mathrm{D}_p) \propto \mathrm{L}(\bar{w}|\mathrm{D}_p) \times \mathrm{p}_0(\bar{w})$. Pour faciliter la compréhension, dans la suite de ce document nous appellerons la distribution à posteriori : posterior. L'objectif de cet algorithme est de résoudre le problème d'optimisation suivant:

$$\tilde{\mathrm{q}} = \operatorname*{argmin}_{\mathrm{q} \in \mathcal{G}} \mathrm{KL}(\mathrm{q}|\hat{\mathrm{p}}) \;, \tag{2.1}$$

avec $\mathcal{G} = \left\{ \mathrm{q}_{(\mu,\sigma)} = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) \right\}$ est la famille Variationnelle Gaussienne Mean-Field. Le dernier rappel concernant l'algorithme Bayes-By-Backprop est que l'optimisation de (2.1) est équivalente à l'optimisation de l'ELBO définit par

$$\begin{aligned} &\tilde{\mathrm{q}} = \mathrm{q}_{\boldsymbol{\theta}^{\star,N}} \\ &\boldsymbol{\theta}^{\star,N} = \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmax}}\, \mathrm{ELBO}^N(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmax}} -\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})|\mathrm{p}_0(\bar{w})) + \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})}[\sum_{i=1}^{p} \log \mathrm{L}(y_i|x_i,\bar{w})] \,. \end{aligned} \tag{2.2}$$

On peut noter que la fonction ELBO^N se décompose en deux termes: le premier est un terme de régularisation qui pénalise les distributions variationnelles qui s'éloignent de la distribution à priori et le deuxième est un terme qui mesure la qualité de la distribution variationnelle par rapport aux données. Dans ce travail nous allons étudier le régime surparamétré, c'est-à-dire lorsque le nombre de neurones N tends vers l'infini.

Première contribution Nous avons montré que dans le régime surparamétré si nous considérons l'ELBO définie dans (2.2) alors le terme de régularisation ($\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})|\mathrm{p}_0(\bar{w}))$) va devenir prédominant comparé au terme relatif aux données. En d'autres termes, nous allons optimiser uniquement que le premier terme de l'ELBO et donc obtenir une distribution variationnelle qui colle la distribution à priori. Cette contribution est résumée dans la proposition suivante:

Proposition 1. *(Informel). Si on suppose que $\mathcal{G}_\Theta$ est la famille des distributions Gaussiennes avec une covariance diagonale, que la distribution à priori $\mathrm{p}_0 \in \mathcal{G}_\Theta$ et que X est compact. Si on choisit ℓ comme la perte quadratique ou la cross-entropy et que la fonction d'activation h est Lipschitz. Alors, $\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}(\bar{w})|\mathrm{p}_0(\bar{w})) \to 0$ lorsque $N \to \infty$.*

Deuxième contribution Nous avons proposé une version tempérée de l'ELBO qui permet de pallier au problème d'équilibrage entre les deux termes de l'ELBO. Cette version tempérée est définie par:

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = -\eta \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}} \,|\, \mathrm{p}_0) + \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})}[\sum_{i=1}^{p} \log \mathrm{L}(y_i|x_i,\bar{w})] \,. \tag{2.3}$$

C'est le paramètre η qui permet d'équilibrer les deux termes de l'ELBO. Le cas $\eta = 1$ correspond au bayesien classique, $\eta > 1$ est appelé le régime de "warm posterior" et enfin le cas $\eta < 1$ est appelé le régime de "cold posterior". En choisissant $\eta = \tau p/N$ avec τ un paramètre fixé alors on peut montrer qu'on n'a plus ce problème de collapse de la distribution variationnelle sur la distribution à priori. Cette contribution est résumée dans la proposition suivante:

Proposition 2. *(Informel). Si on suppose que $\eta = \tau p/N$ et que la fonction de perte ℓ est la perte quadratique. Alors, $\limsup_{N\to\infty} \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}, \mathrm{p}_0) > 0$ quand $N \to \infty$.*

Troisième contribution Enfin, nous avons montré que cette version tempérée de l'ELBO converge vers une fonctionnelle bien définie lorsque le nombre de neurones et de données tend vers l'infini. Cette approche s'inspire des papiers récents qui étudient la convergence des algorithmes de descente de gradient pour les réseaux de neurones à une couche cachée dans le régime surparamétré, [Chizat and Bach, 2018, Rotskoff et al., 2019, Mei et al., 2018, Tzen and Raginsky, 2019, De Bortoli et al., 2020a]. Cette contribution est résumée dans le théorème suivant:

Theorem 3. *(Informel). Sous certaines hypothèses décrites dans le chapitre 4, si on considère la version tempérée de l'ELBO définie dans* (2.3) *on obtient alors la borne suivante:*

$$\left| \frac{1}{p} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \mathrm{R}_\tau(\nu_N^{\boldsymbol{\theta}}) \right| \leq \frac{C}{N} + M_G \sqrt{\log(\delta/2)/(2p)} \,,$$

où $\nu_N^{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ est la distribution empirique des paramètres, C et M_G sont des constantes indépendantes du problème et la fonctionnelle de risque est définie par:

$$\mathrm{R}_\tau(\nu) = -\int \ell\left(y, \iint \phi(\theta,z,x)\mathrm{d}\nu(\theta)\mathrm{d}\gamma(z)\right)\mathrm{d}\pi(x,y) - \tau \int \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta|\mathrm{p}_0{}^1)\mathrm{d}\nu(\theta) \,,$$

Ce résultat nous montre que dans le régime surparamétré ($N \to \infty$) et le régime large data ($p \to \infty$) la version tempérée de l'ELBO converge vers la fonctionnelle de risque R_τ.

1.2 Chapitre 5: Loi des Grands Nombres pour les Réseaux de Neurones Bayésiens à deux couches entraînés avec Inférence Variationnelle.

Dans ce chapitre nous considérons une configuration identique à celle du chapitre précédent (Chapitre 4). Cependant, certains éléments de la configuration ont été simplifiés pour faciliter l'étude théorique du réseau. Premièrement le réseau de neurones considéré possède uniquement des poids en entrée

$$f_{\bar{w}}(x) = \frac{1}{N}\sum_{i=1}^{N} h(w_i, x)\,, \tag{2.4}$$

où on rappelle que h est la fonction d'activation. On voit donc que la sortie du réseau de neurones est de dimension 1. Enfin, nous considérons uniquement les problèmes de régression, donc la fonction de perte étudiée est la perte quadratique, définie par:

$$l(y_1, y_2) = \frac{1}{2}|y_1 - y_2|^2\,.$$

Maintenant, nous allons rappeler le "reparameterisation trick" de [Blundell et al., 2015]. Soit $\tilde{\mathrm{q}}^1_{\theta_i}(w_i) = \mathcal{N}(w_i|\mu_i, \sigma_i)$ la distribution variationnelle associée au neurone i. Alors échantillonner w_i selon $\tilde{\mathrm{q}}^1_{\theta_i}(w_i)$ est équivalent à échantillonner z selon $\gamma = \mathcal{N}(0,1)$ et à appliquer la transformation suivante: $w_i = \mu_i + \sigma_i \odot z$. On peut donc définir la fonction $\phi : (\theta, z, x) \to h(\mu + \sigma \odot z, x)$ qui est la fonction d'activation combinée au "reparameterisation trick'. Nous avons maintenant définie toutes les notations nécéssaires pour developper nos algorithmes qui sont étudiés dans ce chapitre.

Le premier algorithme est la version de Bayes-By-Backprop (plus de détails de cette approche dans le Chapitre 1) avant l'approximation de Monte Carlo. Cet algorithme est appelé **Idealized SGD** dans la suite de ce travail. Le principe de Bayes-By-Backprop est d'utiliser directement une descente de gradient sur l'ELBO. Nous désignons par ${\theta_k}^i$ les poids variationnels associés au i^{eme} neurone après k étapes de descente de gradient. Nous obtenons alors la récursion suivante:

$$\begin{aligned}\theta^i_{k+1} = \theta^i_k &- \frac{\eta}{N^2}\sum_{j=1, j\neq i}^{N}\Big(\langle\phi(\theta^j_k, \cdot, x_k), \gamma\rangle - y_k\Big)\langle\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\rangle \\ &- \frac{\eta}{N^2}\Big\langle(\phi(\theta^i_k, \cdot, x_k) - y_k)\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\Big\rangle - \frac{\eta}{N}\nabla_\theta \mathrm{KL}(q^1_{\theta^i_k}|P^1_0)\end{aligned} \tag{2.5}$$

Cependant cette version est inutilisable en pratique car les intégrales de la forme: $\langle U, \gamma\rangle = \int U(z)\mathrm{d}\gamma(z)$ ne sont pas calculables. Nous allons alors les approximer par du Monte Carlo, en prenant B échantillons de Monte Carlo, ie, $\langle U, \gamma\rangle \approx \frac{1}{B}\sum_{l=1}^{B} U(\mathsf{z}^l)$, avec z^l échantillonné selon γ. On obtient alors l'algorithme suivant:

$$\theta^i_{k+1} = \theta^i_k - \frac{\eta}{N^2 B}\sum_{j=1}^{N}\sum_{\ell=1}^{B}\big(\phi(\theta^j_k, \mathsf{Z}^{j,\ell}_k, x_k) - y_k\big)\nabla_\theta\phi(\theta^i_k, \mathsf{Z}^{i,\ell}_k, x_k) - \frac{\eta}{N}\nabla_\theta \mathrm{KL}(q^1_{\theta^i_k}|P^1_0) \tag{2.6}$$

Avec $(\mathsf{Z}^{j,\ell}_k, 1 \le j \le N, 1 \le \ell \le B, k \ge 0)$ une suite de variables aléatoires échantillonnées selon γ de manière i.i.d. Cet algorithme est appelé **Bayes-By-Backprop SGD** dans la suite de ce travail.

Pour chacun des algorithmes, il est utile de définir la distribution empirique des paramètres du réseau après $\lfloor Nt \rfloor$ étapes de descente de gradient. Cette distribution est définie par:

$$\mu^N_t = \frac{1}{N}\sum_{i=1}^{N}\delta_{\theta^N_{\lfloor Nt\rfloor}}$$

Soit $T > 0$ un paramètre fixé, on définit $\mu^N = \{\mu^N_t, t \in [T]\}$ la trajectoire des paramètres.

Première contribution: Nous avons premièrement démontré que la suite des trajectoires μ^N suit la même loi des grands nombres pour les algorithmes **Idealized SGD** et **Bayes-By-Backprop SGD**. Cette contribution est résumée dans le théorème suivant:

Theorem 4. *(Informel). Sous certaines hypothèses décrites dans le chapitre 5, si la suite $\{\mu^N\}_{N\geq 1}$ est décrite par **Idealized SGD** ou **Bayes-By-Backprop SGD** alors elle converge en probabilité vers une unique solution déterministe $\bar{\mu}$ qui est caractérisée par l'équation suivante: pour toute fonction de test $f \in \mathcal{C}^\infty(\Theta)$ et tout $t \in [T]$ on a:*

$$\begin{aligned}\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle = -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \bar{\mu}_s\big\rangle \mathrm{d}s. \end{aligned} \tag{2.7}$$

Deuxième contribution: Nous avons ensuite analysé la structure de l'équation limite (2.7), et nous avons constaté qu'elle peut être réécrite de manière plus simple. Cela nous a permis de développer un nouvel algorithme, appelé **Minimal-VI SGD**, qui respecte la même équation limite. Cet algorithme est décrit par l'équation suivante:

$$\theta^i_{k+1} = \theta^i_k - \frac{\eta}{N^2}\sum_{j=1}^N \big(\phi(\theta^j_k, \mathsf{Z}^1_k, x_k) - y_k\big)\nabla_\theta \phi(\theta^i_k, \mathsf{Z}^2_k, x_k) - \frac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k} | P^1_0) \tag{2.8}$$

Notons qu'à chaque étape uniquement deux variables aléatoires Gaussiennes (Z^1_k, Z^2_k) sont échantillonnées pour cet algorithme. Alors que chaque étape des algorithmes précédent nécessite la simulation de $O(N)$ variables aléatoires gaussiennes. Cette version **Minimal-VI SGD** est donc beaucoup moins coûteuse que se soit théoriquement ou en pratique.

Troisième contribution: Nous avons montré que la suite des trajectoires μ^N suit la même loi des grands nombres pour l'algorithme **Minimal-VI SGD**. Cette contribution est résumée dans le théorème suivant:

Theorem 5. *(Informel). Sous certaines hypothèses décrites dans le chapitre 5, si la suite $(\mu^N)_{N\geq 1}$ est décrite par **Minimal-VI SGD** alors elle satisfait toutes les affirmations du Théorème 4.*

Cet algorithme est donc beaucoup moins coûteux en terme de complexité de calcul, et en plus il a le même comportement limite.

1.3 Chapitre 6: Théorème Central Limite pour les Réseaux de Neurones Bayésiens entraînés avec Inférence Variationnelle.

Dans ce chapitre nous suivons le travail réalisé précédement. Nous considèrons la même configuration de réseau définie en (2.4), la même famille variationnelle et les trois même algorithmes **Idealized SGD** (2.5), **Bayes-By-Backprop** (2.6) et **Minimal VI** (2.8). Dans le chapitre précédent nous avons démontré que ces trois différents algorithmes suivent une même loi des grands nombres et que la distribution des leurs paramètres convergent toutes vers une même distribution limite $\bar{\mu}_t$ qui est déterministe. Dans ce chapitre nous allons étudier la vitesse de convergence de ces distributions en dérivant un théorème Centrale Limite. Premièrement, nous allons définir le processus de fluctuation:

$$\eta^N : t \in \mathbb{R}_+ \mapsto \sqrt{N}(\mu^N_t - \bar{\mu}_t)\,,$$

Première contribution Nous avons premièrement démontré un Théorème Central Limite pour l'algorithme **Idealized SGD** (2.5). Cette contribution est résumée dans le théorème suivant:

Theorem 6. *(Informel). Sous certaines hypothèses décrites dans le Chapitre 6, si on considère l'algorithm **Idealized SGD**, la suite $(\eta^N)_{N\geq 1}$ converge en loi vers un processus $\eta^\star$. Ce processus est l'unique solution d'une certaine équation. Celui-ci est entièrement caractérisée par un G-processus et plus précisément par sa structure*

de covariance. Vous pouvez trouver plus de détails sur cette équation et sur la definition du G-processus dans le Chapitre 6. Pour cet algorithme, la structure de covariance est donnée par:

$$Cov(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f](x, y, \bar{\mu}_v), \mathscr{Q}[g](x, y, \bar{\mu}_v)) \mathrm{d}v,$$

où $\mathscr{Q}[f](x, y, \bar{\mu}_v) = \langle \phi(\cdot, \cdot, x) - y, \bar{\mu}_v \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \bar{\mu}_v \otimes \gamma \rangle$.

Deuxième contribution Nous avons ensuite démontré un Théorème Central Limite pour l'algorithme **Bayes-By-Backprop SGD** (2.6). Plus précisément, nous avons démontré que cet algorithme suit exactement le même Théorème Central Limite que l'algorithme **Idealized SGD** et donc respecte le Théorème 6. Cela justifie encore plus l'utilisation de cette approximation de Monte Carlo car on voit qu'il n'y a ni différence de convergence ni de vitesse de convergence entre l'approximation Monte Carlo et la version idéalisée.

Troisième contribution Enfin, nous avons démontré un Théorème Central Limite pour le dernier algorithme **Minimal-VI SGD** (2.8). Cette contribution est résumée dans le théorème suivant:

Theorem 7. *(Informel). Sous certaines hypothèses décrites dans le Chapitre 6, si on considère l'algorithme **Minimal-VI SGD**, la suite* $(\eta^N)_{N \geq 1}$ *converge en loi vers un processus* $\eta^\star$*. Ce processus est caractérisé par un G-process dont la structure de covariance est donnée par:*

$$Cov(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f](x, y, z^1, z^2, \bar{\mu}_v), \mathscr{Q}[g](x, y, z^1, z^2, \bar{\mu}_v)) \mathrm{d}v,$$

où $\mathscr{Q}[f](x, y, z^1, z^2, \bar{\mu}_v) = \langle \phi(\cdot, z^1, x) - y, \bar{\mu}_v \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z^2, x), \bar{\mu}_v \rangle$.

On peut voir que pour cet algorithme le G-processus a une structure de covariance différente. Par exemple, si on considère une fonction de test scalaire, on peut montrer que la variance du G-processus associé à **Minimal-VI SGD** est plus grande que celles associées aux autres algorithmes. Cependant, rappelons que cet algorithme est beaucoup plus efficace computationnellement.

1.4 Chapitre 7: Garanties théoriques pour l'Inférence Variationnelle avec des Mélanges de Gaussiennes à Variance fixée

Dans les trois premiers chapitres de cette thèse, nous avons étudié la famille Variationnelle Gaussienne. Cependant, lorsque la distribution que nous souhaitons approcher est multimodale, une simple gaussienne ne permet pas d'obtenir une approximation satisfaisante. L'utilisation d'une famille variationnelle plus flexible, telle que les mélanges de Gaussiennes, permet de résoudre ce problème et de mieux approcher les distributions multimodales. De plus, la famille des mélanges de Gaussiennes est dense dans l'espace des distributions de probabilité avec des moments d'ordre p bornés dans la métrique de Wasserstein-p [Delon and Desolneux, 2020]. Cette famille de distributions est donc extrêmement pertinente pour notre problème d'Inférence Variationnelle. Cependant dans ce chapitre, nous proposons de considérer un cadre simplifié où les gaussiennes ont des poids égaux et partagent la même matrice de covariance diagonale $\epsilon^2 \mathrm{I}_d$. Ce cadre permet de réduire la complexité du problème, tout en restant théoriquement difficile et pertinent en pratique. La famille Variationnelle est alors définie par:

$$\mathcal{G}_n = \left\{ \frac{1}{n} \sum_{i=1}^n \mathcal{N}(x^i, \epsilon^2 \mathrm{I}_d),\ x^i \in \mathbb{R}^d \right\} = \left\{ k_\epsilon \star \mu,\ \mu = \frac{1}{n} \sum_{i=1}^n \delta_{x^i},\ x^i \in \mathbb{R}^d \right\}.$$

où k_ϵ est le noyau Gaussien normalisé, ie, $k_\epsilon \propto \exp\left(-\|x\|^2/(2\epsilon^2)\right)$.

Rappelons que l'objectif de notre approche est d'approximer la distribution cible $\mu^\star \propto \exp(-V)$, plus précisément le problème d'optimisation de l'Inférence Variationnelle s'écrit de la manière suivante

$$\hat{\nu} = \underset{\nu \mathcal{G}_n}{\operatorname{argmin}} \mathrm{KL}(\nu, \mu^\star) \tag{2.9}$$

On peut donc maintenant définir la fonction objective

$$\begin{aligned}\mathcal{F}_\epsilon(\mu) &= \mathrm{KL}(k_\epsilon \star \mu, \mu^\star) \\ &= \int V d(k_\epsilon \star \mu) + \int \log(k_\epsilon \star \mu) d(k_\epsilon \star \mu)\end{aligned}$$

On peut voir que puisque les variances de nos Gaussiennes sont fixées, l'Inférence Variationnelle vise à optimiser les emplacements des moyennes du mélange gaussien $(x^i)_{i=1}^n$ pour approcher la distribution cible. On peut donc voir nos mélanges de Gaussiennes comme des systèmes de particules que nous pouvons faire évoluer le long du flow qui diminue la fonction objective $\mathcal{F}_\epsilon$. Pour faire cela, nous allons utiliser un algorithme de descente de gradient Wasserstein (plus de détails sur l'algorithme dans le Chapitre 7). Soit $\mu_l = \frac{1}{n}\sum_{i=1}^n \delta_{x_l^i}$ la distribution empirique des particules après l étapes de descente de gradient Wasserstein. Cette distribution empirique est mise à jour récursivement selon la récurrence suivante

$$\mu_{l+1} = (\mathrm{I}_d - \gamma \nabla \mathcal{F}_\epsilon'(\mu_l))_{\#}\mu_l \tag{2.10}$$

où $\gamma > 0$ est le pas de discrétisation de l'algorithme et $\nabla \mathcal{F}_\epsilon'(\mu_l)$ est le flow de gradient Wasserstein de $\mathcal{F}_\epsilon$. l'Equation (2.10) permet de faire évoluer les particules de notre système suivant la récurrence:

$$x_{l+1}^j = x_l^i - \gamma \left(\int_{\mathbb{R}^d} \nabla V(y) k_\epsilon(y - x^j) \mathrm{d}y + \int_{\mathbb{R}^d} \frac{\sum_{i=1}^n \nabla k_\epsilon(y - x^i)}{\sum_{i=1}^n k_\epsilon(y - x^i)} k_\epsilon(y - x^j) \mathrm{d}y \right),$$

Première contribution: Nous avons premièrement étudié l'erreur d'optimisation, c'est-à-dire l'erreur obtenue lorsque que nous optimisons (2.9). Nous avons démontré la regularité (smoothness) de notre fonction objectif $\mathcal{F}_\epsilon$ puis cela nous a permit d'établir un lemme de descente qui est résumé dans la proposition suivante:

Proposition 8. *(Informel) Sous certaines hypothèses décrites dans le chapitre 7, l'inégalité suivante est vérifiée:*

$$\mathcal{F}_\epsilon(\mu_{l+1}) - \mathcal{F}_\epsilon(\mu_l) \leq -\gamma \left(1 - \frac{\gamma}{2} M\right) \|\nabla \mathcal{F}_\epsilon'(\mu_l)\|^2_{L^2(\mu_l)}.$$

avec M une constante.

Ainsi, pour un γ suffisamment petit, cette proposition nous montre que la fonction objectif $\mathcal{F}_\epsilon$ décroit à chaque itération l.

Deuxième contribution: Ensuite, nous nous sommes intéressés à l'erreur d'approximation, qui quantifie à quel point $\hat{\nu}$ est loin de $\mu^\star$. Plus précisément nous supposons que pour chaque $n \in \mathbb{N}$ nous avons trouvé l'optimiseur de (2.9) que nous notons $\hat{\nu}_n = k_\epsilon \star \mu_n$ avec $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i}$. Nous avons démontré que si la distribution cible est un mélange, potentiellement infini, de Gaussiennes alors l'erreur d'approximation est bornée par une fonction qui décroit en $\frac{\log(n)}{n}$, cette contribution est résumée dans le Théorème suivant:

Theorem 9. *(Informel) Sous certaines hypothèses décrites dans le chapitre 7, l'inégalité suivante est vérifiée:*

$$\mathrm{KL}(\mu_n, \mu^\star) \leq C_{\mu^\star}^2 \frac{\log(n) + 1}{n}$$

où $C_{\mu^\star}$ est une constante qui dépend de la distribution cible $\mu^\star$.

2 Partie II: Les algorithmes de Thompson Sampling pour les problèmes de bandits.

Dans cette partie de la thèse, nous allons étudier les problèmes de Bandit contextuel en utilisant une approche Bayésienne. Plus précisément, nous allons considérer les algorithmes de Thompson Sampling. Une description complète du problème étudié, ainsi que de l'algorithme de Thompson Sampling est donnée dans la Section 2 du Chapitre 1.

Les algorithmes de Thompson Sampling classiques présentent deux limites majeures. La première est que l'échantillonnage selon la posterior est souvent difficile, voire impossible. Et qu'il faut donc utiliser des méthodes

d'approximation. La deuxième est que ces algorithmes possèdent de moins bonnes garanties théoriques que celles obtenues par les approches fréquentistes, de type UCB. Dans cette partie de la thèse, nous proposons des solutions à ces deux problèmes, dans le Chapitre 8 nous allons étudier le problème de Bandit contextuel avec une approche d'Inférence Variationnelle Gaussienne. Et enfin dans le Chapitre 9 nous étudions un algorithme appelé Feel-Good Thompson Sampling presenté dans Zhang [2022a], qui permet d'améliorer les garanties théoriques de Thompson Sampling classique en modifiant légèrement la posterior.

2.1 Chapitre 8: Thompson Sampling avec Inférence Variationnelle pour les problèmes de bandits contextuels.

Pour rappel, le problème de Bandit contextuel est défini de la manière suivante: à chaque itération $t \in [T]$ un agent observe un contexte $x_t \in \mathrm{X}$ qui représente l'état de l'environnement, puis il choisit une action $a_t = \pi_t(x_t)$, où π_t est appelé politique, et enfin il reçoit une récompense $r_t \sim \mathrm{R}(\cdot|x_t, a_t)$. L'objectif de l'agent est de trouver une politique qui permet de minimiser le regret cumulé définie

$$\mathrm{CREG}(\mathbb{Q}_{1:T}) = \sum_{t \leq T} f_\star(x_t) - f(x_t, \pi_t(x_t)) \,,$$

où $f(x,a) = \int r \mathrm{R}(\mathrm{d}r|x,a)$ et $f_\star(x) = \max_{a \in \mathcal{A}(x)} f(x,a)$.

Pour résoudre ce problème, nous allons utiliser l'algorithme de Thompson Sampling. On commence par choisir un modèle paramétrique R_θ pour modéliser la distribution de récompense R. Dans ce chapitre, nous choisissons de considérer comme modèle la famille exponentielle (voir Chapitre 8 pour plus de détails). On applique ensuite l'approche Bayesienne, on définit la posterior $\hat{\mathrm{p}}_t$. Le principe de Thompson Sampling est d'échantillonner pour chaque itération t, un paramètre θ_t selon la posterior, et de choisir l'action de la manière suivante:

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}(x_t)} \int r \mathrm{R}_{\theta_t}(\mathrm{d}r|x_t, a) \,.$$

Cependant en pratique, il est souvent impossible d'échantillonner selon la posterior et il est donc impératif d'utiliser une méthode d'approximation. Dans ce chapitre, nous proposons d'utiliser une approche d'Inférence Variationnelle Gaussienne pour approximer la posterior. Pour résoudre le problème d'optimisation d'Inférence Variationnelle, nous proposons d'utiliser l'algorithme de descente de gradient Riemannien (voir la section 1.2.2 du chapitre 1 pour plus de détails sur l'algorithme). Pour rappel, la distribution Variationnelle obtenue après k étapes de descente de gradient Riemannien est $\tilde{\mathrm{q}}_{t,k} = \mathcal{N}(\tilde{\mu}_{t,k}, \tilde{\Sigma}_{t,k})$, où les paramètres $\tilde{\mu}_k$ et $\tilde{\Sigma}_k$ sont mis à jour de la manière suivante:

$$\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - \gamma \nabla U_t(\tilde{\theta}_{t,k}), \tag{2.11}$$
$$\tilde{\Sigma}_{t,k+1} = \tilde{A}_{t,k} \tilde{\Sigma}_{t,k} \tilde{A}_{t,k} \,,$$
$$\tilde{A}_{t,k} = \mathrm{I}_d - \gamma_t (\nabla^2 U_t(\tilde{\theta}_{t,k}) - \tilde{\Sigma}_{t,k}^{-1}) \,, \tag{2.12}$$
$$\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k} \,,$$

où h_t est le pas de discrétisation de l'algorithme et $U_t(\theta) \propto -\log(\hat{\mathrm{p}}_t(\theta))$ est la fonction de potentiel. Cet algorithme présenté dans [Lambert et al., 2022b] permet de résoudre le problème d'optimisation Variationnelle, cependant pour chaque itération k il est nécéssaire d'échantillonner selon une Gaussienne de dimension d, ce qui peut être coûteux en haute dimension.

Première contribution: Nous avons premièrement proposé une version améliorée de l'algorithme de [Lambert et al., 2022b]. Au lieu de considérer la matrice de covariance $\{\tilde{\Sigma}_{t,k}\}$, nous regardons uniquement une de ses racines carrée qui est définie par:

$$B_{t,k+1} = \left\{\mathrm{I}_d - h_t \nabla^2 U_t(\tilde{\theta}_{t,k})\right\} B_{t,k} + h_t (B_{t,k}^{-1})^\top \,. \tag{2.13}$$

Il est alors beaucoup plus efficace d'échantillonner selon la posterior Variationnelle en utilisant la formule suivante:

$$\tilde{\theta}_{t,k} = \tilde{\mu}_{t,k} + B_{t,k} \epsilon_{t,k}, \quad \epsilon_{t,k} \sim \mathcal{N}(0, \mathrm{I}_d) \,.$$

On obtient alors notre premier algorithme appelé `VITS − I` qui à chaque itération t, choisit une action de la manière suivante:

$$a_t = \underset{a\in\mathcal{A}(x_t)}{\operatorname{argmax}} \int r \mathrm{R}_{\tilde{\theta}_{t,K_t}}(\mathrm{d}r|x_t,a) \,. \tag{2.14}$$

Puis on recalcule la distribution Variationnelle en utilisant les equations (2.11) et (2.13) pendant K_{t+1} itérations. Cet algorithme est bien plus efficace que la version de base de [Lambert et al., 2022b], cependant il est toujours coûteux en haute dimension car à chaque itération nous devons inverser la matrice $B_{t,k}$.

Deuxième contribution: Nous avons proposé une nouvelle version de notre algorithme, qui utilise une approximation de Taylor (pour h_t assez petit) qui permet d'approximer l'inverse $B_{t,k}^{-1}$ par $C_{t,k}$ qui est définie recursivement par

$$C_{t,k+1} = C_{t,k}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - \nabla^2 U_t(\tilde{\theta}_{t,k}))\} \,, \quad B_{t,k+1} = (\mathrm{I}_d - h_t\nabla^2 U_t(\tilde{\theta}_{t,k}))B_{t,k} + h_t C_{t,k}^\top \,.$$

Cet algorithme que nous appelons `VITS − II` est encore plus efficace que `VITS − I` et permet choisir une action a_t de la même manière que dans (2.14).

Troisième contribution: Nous avons proposé une dernière version de notre algorithme qui permet d'éviter l'ultime étape coûteuse qui est le calcule de la Hessienne de U_t. Pour cela, nous utilisons la propriété Gaussienne suivante:

$$\int \nabla^2 U_t \mathrm{d}\mathcal{N}(\mu,\Sigma) = \int \Sigma^{-1}(\mathrm{I}_d - \mu)\nabla U_t^\top \mathrm{d}\mathcal{N}(\mu,\Sigma) \,.$$

Le terme de Hessienne $\nabla^2 U_t(\tilde{\theta}_{t,k})$ dans (2.12) est alors remplacé par $C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top$. On obtient alors l'algorithme `VITS − II Hessian-free` qui se base sur les récurrences suivantes:

$$\begin{aligned}
\tilde{\mu}_{t,k+1} &= \tilde{\mu}_{t,k} - h_t\nabla U_t(\tilde{\theta}_{t,k}) \,, \\
C_{t,k+1} &= C_{t,k}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top)\} \,, \\
B_{t,k+1} &= (\mathrm{I}_d - h_t C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top)B_{t,k} + h_t C_{t,k}^\top \,,
\end{aligned}$$

où la distribution Variationnelle est $\tilde{\mathrm{q}}_{t,k} = \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$ et le paramètre $\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k}$. Cet algorithme possède un très faible coût computationnel à la fois théoriquement et empiriquement.

Quatrième contribution: Enfin, nous avons démontré que lorsqu'on considère un Bandit linéaire, notre premier algorithme `VITS − I` possède une garantie de regret de l'ordre de $\tilde{O}(d\sqrt{dT})$. Cette contribution est résumée dans le Théorème suivant:

Theorem 10. *(Informel). Sous certaines hypothèses décrites dans le chapitre 8, si on considère l'algorithme `VITS − I` pour un Bandit linéaire, alors avec probabilité* $1-\delta$ *on a*

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) \leq C_1 d\sqrt{dT}\log\big(3T^3\big)\log\left((1+\frac{C_2 T}{d})/\delta\right) \,,$$

où C_1 *et* C_2 *sont des constantes (voir Chapitre 8 pour plus de détails).*

Ce résultat théorique est la première borne de regret pour une approximation Variationnelle de la posterior. De plus cette borne est optimale pour le problème de Bandit linéaire utilisant une approche basée sur Thompson Sampling classique.

2.2 Chapitre 9: Feel-Good Thompson Sampling avec Langevin Monte Carlo.

Dans ce chapitre, nous allons continuer à étudier les algorithmes de Thompson Sampling pour résoudre le problème de Bandit contextuel. Bien que ces algorithmes soient très performants en pratique et faciles à implémenter, les bornes théoriques de regret obtenues par Thompson Sampling sont généralement moins bonnes que celles des approches fréquentistes. Par exemple, pour le Bandit Linéaire, la meilleure borne de regret pour Thompson Sampling est $\mathrm{CREG} = \tilde{O}(d^{3/2}\sqrt{T})$ ([Agrawal and Goyal, 2012]), tandis que pour l'approche Linear UCB, on obtient une borne $\mathrm{CREG} = \tilde{O}(d\sqrt{T})$ (Dani et al. [2008], Abbasi-Yadkori et al. [2011b]). Pour contourner ce problème, [Zhang, 2022a] a proposé de modifier la fonction de vraisemblance dans TS en ajoutant un terme de pénalité pour favoriser une exploration. Plus précisément, on rappelle que la vraisemblance utilisée dans Thompson Sampling est de la forme suivante:

$$\mathrm{L}_t^{(\mathrm{TS})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\eta(g(\theta,x_s,a_s)-r_s)^2\right),$$

où $g(\theta,x,a)$ est la sortie du modèle ayant comme paramètre θ, associé au contexte x et à l'action a. L'idée de [Zhang, 2022a] est de remplacer cette vraisemblance par une nouvelle vraisemblance qui est définie de la manière suivante:

$$\mathrm{L}_t^{(\mathrm{FG})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\eta(g(\theta,x,a)-r)^2 - \lambda\min(b, g_\star(\theta,x))\right),$$

où b est un paramètre fixé, λ contrôle l'importance de la pénalité, et $g_\star = \max_{a\in\mathcal{A}(x)} g(\theta,x,a)$. Il utilise ensuite exactement la même approche que TS mais avec cette nouvelle fonction de vraisemblance. Cet algorithme est appelé Feel-Good Thompson Sampling (FG-TS). Il démontre ensuite que cette nouvelle version de TS permet d'obtenir des bornes de regret de l'ordre de $\mathrm{CREG} = \tilde{O}(d\sqrt{T})$ qui correspond à la borne optimale de regret minimax. Cependant, cet algorithme n'est pas utilisable en pratique. En effet la distribution à posteriori est encore plus complexe que pour Thompson Sampling classique. Elle n'est même plus Gaussienne dans le cas linéaire. Il est donc obligatoire d'utiliser une méthode d'approximation pour échantillonner selon cette nouvelle posterior.

Dans ce chapitre, nous allons étudier l'algorithme de FG-TS, mais nous y ajoutons une méthode de MCMC pour échantillonner la posterior.

Première contribution Dans un premier temps, nous avons proposé une nouvelle version de la vraisemblance. Celle-ci est une version lissée (smooth) de la vraisemblance de FG-TS. Cette nouvelle version est appelée smooth-FG (sFG) et est définie de la manière suivante:

$$\mathrm{L}_t^{(\mathrm{sFG})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\eta(g(\theta,x,a)-r)^2 - \lambda\left[b-\phi_\varsigma(b-g_\star(\theta,x))\right]\right),$$

où $\phi_\varsigma(u) = \log(1+\exp(\varsigma u))/\varsigma$ et $\varsigma > 0$ est un paramètre qui contrôle la régularité de la vraisemblance. Régulariser la vraisemblance permet d'avoir une posterior lisse (smooth), ce qui améliore largement les performances des méthodes de MCMC. En particulier, les méthodes MCMC basées sur les informations de gradient [Durmus et al., 2018]. Nous avons alors proposé un nouvel algorithme, similaire à FG-TS, mais qui utilise cette nouvelle vraisemblance et qui utilise une approximation de la vrai posterior, cette approximation est notée $\tilde{\mathrm{q}}_t^{(\mathrm{sFG})}$. Utilisant cet algorithme, on obtient le théorème suivant sur la contrôle du regret:

Theorem 11. *(Informel). Sous certaines hypothèses décrites dans le chapitre 9, on a l'inégalité suivante:*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \frac{\lambda}{\eta\epsilon}KT + C_1\lambda T - \frac{Z_T}{\lambda} + (C_2 + \frac{C_3}{\lambda})\sum_{t=0}^{T}\mathbb{E}_{\nu_0}^T[\delta_t],$$

où Z_T est un terme qui dépend de la distribution à priori, $\epsilon \in (0,1)$ est un paramètre fixé et C_1, C_2, C_3 sont des constantes indépendantes du problème. Notons que nous n'avons pas choisit de méthode d'approximation spécifique à ce moment là, le terme $\mathbb{E}_{\nu_0}^T[\delta_t]$ est le terme d'erreur d'approxmation, où $\delta_t = ||\tilde{\mathrm{q}}_t^{(\mathrm{sFG})} - \mu_t^{(\mathrm{sFG})}||_{\mathrm{TV}}$ est la Variation Totale entre la posterior et son approximation.

Deuxième contribution Nous avons ensuite proposé d'utiliser une méthode de MCMC pour échantillonner selon la posterior. Plus précisément, nous avons proposé d'utiliser l'algorithme de Langevin Monte Carlo, qui permet de définir de manière itérative un paramètre de la manière suivante:

$$\theta^{\mathrm{L}}_{t,k+1} = \theta^{\mathrm{L}}_{t,k} + \gamma_t \nabla \log \mu_t^{(\mathrm{sFG})}(\theta^{\mathrm{L}}_{t,k}) + \sqrt{2\gamma_t} Z_{t,k} \,,$$

où $Z_{t,k} \sim \mathcal{N}(0, \mathrm{I}_d)$ est un bruit Gaussien. Vous pouvez trouver plus de détails sur cet algorithme dans la Chapitre1 ou dans le Chapitre 9. Nous avons aussi proposé d'utiliser une version métropolisée de l'algorithme de Langevin Monte Carlo, appelé Metropolized Langevin Algorithm (MALA). Cet algorithme est défini de la manière suivante:

$$\theta^{\mathrm{M}}_{t,k+1} = \begin{cases} \theta^{\mathrm{M}}_{t,k} + \gamma_t \nabla \log \mu_t^{(\mathrm{sFG})}(\theta^{\mathrm{M}}_{t,k}) + \sqrt{2\gamma_t} Z_{t,k} & \text{with probability } 1/\alpha_t^M, \\ \theta^{\mathrm{M}}_{t,k} & \text{sinon,} \end{cases}$$

avec α_t^M la probabilité d'acceptation Metropolos-Hasting, défini dans le Chapitre 9. En utilisant ces algorithmes, nous avons obtenu le résultat suivant:

Corollary 12. *(Informel). Sous certaines hypothèsess décrites dans le chapitre 9, on a l'inégalité suivante:*

$$\begin{aligned} \mathrm{CREG}(\mathbb{Q}^{(\mathrm{sFG})}_{1:T}) \leq \frac{C_4}{\epsilon} \sqrt{\omega d K T \log(dT)} + \big(4\xi + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)\big)T \\ + \; C_5 \sqrt{\frac{\omega K T}{d \log(dT)}} \left(-\log \mathrm{p}_0(\theta_*) + L_g + \xi T + \xi^2 T\right) \\ + \; C_6 \left(1 + \sqrt{\frac{\omega K T}{d \log(dT)}}\right) \sum_{t=0}^{T} \mathbb{E}^T_{\nu_0}[\delta_t] + 4 L_g. \end{aligned}$$

Vous pouvez trouver tous les détails des paramètres utilisés dans le Chapitre 9.

Nous ne sommes pas très loin de pouvoir obtenir une borne de regret à cette étape. Cependant, il nous reste encore à fixer le modèle utilisé, la distribution à priori et aussi à contrôler le terme d'erreur d'approximation.

Troisième contribution Enfin, nous avons utilisé nos résultats précédents sur le Bandit linéaire. C'est-à-dire que le modèle utilisé est $g(\theta, x, a) = \langle \varphi(x, a), \theta \rangle$ où ϕ est la fonction de features. Nous avons aussi considéré une distribution à priori Gaussienne $\mathcal{N}(0, \mathtt{m}_0^{-1} \mathrm{I}_d)$, avec $\mathtt{m}_0 > 0$. Nous avons ainsi obtenu le théorème suivant:

Theorem 13. *(Informel). Sous certaines hypothèsess décrites dans le chapitre 9, on a l'inégalité suivante:*

$$\mathrm{CREG}(\mathbb{Q}^{(\mathrm{sFG})}_{1:T}) \leq \; C_7 \sqrt{\omega_{\mathsf{LG}} T \log^3(dT)} \left(d(\epsilon \wedge \mathtt{m}_0)^{-1} + \sqrt{M} \mathtt{m}_0 \left\| \theta_* \right\|^2 \right) \,,$$

Nous avons donc obtenu une borne de regret de l'ordre de $\mathrm{CREG} = \tilde{O}(d\sqrt{T})$ pour le Bandit linéaire. Cette borne est optimale et correspond à la borne minimax de regret pour ce problème.

CHAPTER 3

SUMMARY OF THE CONTRIBUTIONS

In this chapter, we summarize the main contributions of this thesis, which is organized in two main parts: the theoretical guarantees for Variational Inference and the Thompson Sampling algorithms for Bandit problems.

1 Part I: Theoretical guarantees for Variational Inference

In the first three chapters (Chapter 4-5-6) of this thesis, we study theoretically overparametrized Bayesian Neural Networks trained with Gaussian Variational Inference. The last chapter (Chapter 7) of this first part focuses on a larger Variational family, namely Gaussian mixtures.

1.1 Chapter 4: Variational Inference of overparametrized Bayesian Neural Networks: a theoretical and empirical study of tempering

In this work let's consider a one hidden layer Bayesian Neural Network $f_{\bar{w}} : \mathbb{R}^{\mathsf{d_X}} \to \mathbb{R}^{\mathsf{d_Y}}$, defined by:

$$f_{\bar{w}}(x) = \frac{1}{N}\sum_{j=1}^{N} s(w_j, x)\ , \quad s(w_j, x) = a_j h(b_j, x)\ ,$$

where $h(\cdot,\cdot)$ is the activation function and $\bar{w}$ denotes the model's parameters. The model has a total of N different neurons, each with a parameter $w_j = (a_j, b_j) \in \mathbb{R}^{\mathsf{d_X}} \times \mathbb{R}^{\mathsf{d_Y}}$. The full model parameters is denoted by $\bar{w} = \{w_j\}_{j \leq N}$. Then, let's apply the Bayes-By-Backprop [Blundell et al., 2015] algorithm (see Section 1.2.2 of Chapter 1 for more details) to this NN. Recall that this method uses a Bayesian Variational Inference approach on the NN parameters. Let's define the likelihood function $\mathrm{L}(\bar{w}|\mathrm{D}_p) \propto \prod_{i=1}^{p} \exp(-\ell(f_{\bar{w}}(x_i), y_i))$, the prior distribution $\mathrm{p}_0(\bar{w})$ and consequently it gives us the posterior distribution $\hat{\mathrm{p}}(\bar{w}|\mathrm{D}_p) \propto \mathrm{L}(\bar{w}|\mathrm{D}_p) \times \mathrm{p}_0(\bar{w})$. The objective of such algorithm is to solve the optimization task:

$$\tilde{\mathrm{q}} = \operatorname*{argmin}_{\mathrm{q} \in \mathcal{G}} \mathrm{KL}(\mathrm{q}|\hat{\mathrm{p}})\ , \tag{3.1}$$

where $\mathcal{G} = \left\{\mathrm{q}_{(\mu,\sigma)} = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))\right\}$ is the Mean-Field Gaussian Variational family. The last reminder regarding the Bayes-By-Backprop algorithm is that the optimization of equation (3.1) is equivalent to the optimization of the

ELBO defined by

$$\tilde{\mathrm{q}} = \mathrm{q}_{\boldsymbol{\theta}^{\star,N}}$$
$$\boldsymbol{\theta}^{\star,N} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\,\mathrm{ELBO}^N(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} -\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})|\mathrm{p}_0(\bar{w})) + \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})}[\sum_{i=1}^{p}\log \mathrm{L}(y_i|x_i,\bar{w})]\,. \tag{3.2}$$

Note that the function ELBO^N can be decomposed into two terms: the first one is a regularity term which penalizes the Variational distribution when it is far from the prior, and the second one measures the quality of the Variational distribution with respect to the data. In this work, we study the overparametrized regime, ie, when the number of neurons tends to infinity.

First contribution We showed that in the overparameterized regime, if we consider the ELBO defined in Equation (3.2), the regularization term ($\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})|\mathrm{p}_0(\bar{w}))$) becomes dominant compared to the data-related term. In other words, we will only optimize the first term of the ELBO, causing the variational distribution to match the prior distribution. This contribution is summarized in the following proposition:

Proposition 14. *(Informal). Assume that $\mathcal{G}_\Theta$ is the family of Gaussian distributions with a diagonal covariance, that the prior distribution $\mathrm{p}_0 \in \mathcal{G}_\Theta$, and that X is compact. If ℓ is the squared loss or the cross-entropy loss and the activation function h is Lipschitz. Then, $\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}(\bar{w})|\mathrm{p}_0(\bar{w})) \to 0$ as $N \to \infty$.*

Second contribution We proposed a tempered version of the ELBO that fixes the problem of balancing the two terms of the ELBO. This tempered version is defined by:

$$\mathrm{ELBO}^N_\eta(\boldsymbol{\theta}) = -\eta\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}\,|\,\mathrm{p}_0) + \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w})}[\sum_{i=1}^{p}\log \mathrm{L}(y_i|x_i,\bar{w})]\,. \tag{3.3}$$

The parameter η allows to balance the two terms of the ELBO. The case $\eta = 1$ corresponds to the classical Bayesian approach, $\eta > 1$ is called the "warm posterior" regime, and finally, the case $\eta < 1$ is called the "cold posterior" regime. By choosing $\eta = \tau p/N$ with τ a fixed parameter, we showed that the issue of the variational distribution collapsing onto the prior distribution no longer arises. This contribution is summarized in the following proposition:

Proposition 15. . *(Informal). Assume that $\eta = \tau p/N$ and that the loss function ℓ is the squared loss, then $\limsup_{N\to\infty} \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}, \mathrm{p}_0) > 0$ as $N \to \infty$.*

Third contribution Finally, we showed that this tempered version of the ELBO converges to a well-defined functional when the number of neurons and data tend to infinity. This approach is inspired by recent papers that study the convergence of gradient descent algorithms for one hidden layer neural networks in the overparameterized regime [Chizat and Bach, 2018, Rotskoff et al., 2019, Mei et al., 2018, Tzen and Raginsky, 2019, De Bortoli et al., 2020a]. This contribution is summarized in the following theorem:

Theorem 16. *(Informal). Under certain assumptions described in Chapter 4, if we consider the tempered version of the ELBO defined in* (3.3)*, we obtain the following bound:*

$$\left|\frac{1}{p}\mathrm{ELBO}^N_\eta(\boldsymbol{\theta}) - \mathrm{R}_\tau(\nu^{\boldsymbol{\theta}}_N)\right| \leq \frac{C}{N} + M_G\sqrt{\log(\delta/2)/(2p)}\,,$$

where $\nu^{\boldsymbol{\theta}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ is the empirical distribution of the parameters, C and M_G are constants independent of the problem, and the risk functional is defined by:

$$\mathrm{R}_\tau(\nu) = -\int \ell\left(y, \iint \phi(\theta,z,x)\mathrm{d}\nu(\theta)\mathrm{d}\gamma(z)\right)\mathrm{d}\pi(x,y) - \tau\int \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta|\mathrm{p}_0{}^1)\mathrm{d}\nu(\theta)\,,$$

This result shows that in the overparameterized regime ($N \to \infty$) and the large data regime ($p \to \infty$), the tempered version of the ELBO converges to the risk functional R_τ.

1.2 Chapter 5: Law of Large Numbers for Bayesian two-layer Neural Network trained with Variational Inference.

In this chapter, we consider a setting similar to the one in the previous chapter (Chapter 4). However, some elements of this setting have been simplified to facilitate the theoretical study of the network. Firstly, the neural network considered has only input weights:

$$f_{\bar{w}}(x) = \frac{1}{N}\sum_{i=1}^{N} h(w_i, x)\,, \tag{3.4}$$

recall that h is the activation function. Consequently, the output of the NN is one-dimensional. Finally, we only consider regression problems, so the loss function studied is the quadratic loss, defined by:

$$l(y_1, y_2) = \frac{1}{2}|y_1 - y_2|^2\,.$$

Now, we will recall the "reparameterization trick" from [Blundell et al., 2015]. Let $\tilde{\mathrm{q}}^1_{\theta_i}(w_i) = \mathcal{N}(w_i|\mu_i, \sigma_i)$ be the variational distribution associated with neuron i. Then sampling w_i according to $\tilde{\mathrm{q}}^1_{\theta_i}(w_i)$ is equivalent to sampling z according to $\gamma = \mathcal{N}(0,1)$ and applying the following transformation: $w_i = \mu_i + \sigma_i \odot z$. Therefore let's define the function $\phi : (\theta, z, x) \to h(\mu + \sigma \odot z, x)$ which is the activation function combined with the "reparameterization trick". We have now defined all the necessary notation to derive the algorithms which are studied in this chapter.

The first algorithm we consider is the version of Bayes-By-Backprop that does not use Monte Carlo approximation (more details on this approach can be found in Chapter 1). This algorithm is referred to as **Idealized SGD** in the following work. The principle of Bayes-By-Backprop is to use directly a gradient descent on the ELBO. Let's denote by θ^i_k the variational weights associated with the i^{th} neuron after k steps of gradient descent. We then obtain the following recursion:

$$\begin{aligned}\theta^i_{k+1} = \theta^i_k &- \frac{\eta}{N^2}\sum_{j=1, j\neq i}^{N}\Big(\langle\phi(\theta^j_k, \cdot, x_k), \gamma\rangle - y_k\Big)\langle\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\rangle \\ &- \frac{\eta}{N^2}\Big\langle(\phi(\theta^i_k, \cdot, x_k) - y_k)\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\Big\rangle - \frac{\eta}{N}\nabla_\theta \mathrm{KL}(q^1_{\theta^i_k}|P^1_0)\end{aligned} \tag{3.5}$$

However, this version is unusable in practice because integrals of the form: $\langle U, \gamma\rangle = \int U(z)\mathrm{d}\gamma(z)$ are intractable. Consequently, let's apply the Monte Carlo approximation, by taking B Monte Carlo samples, ie, $\langle U, \gamma\rangle \approx \frac{1}{B}\sum_{l=1}^{B} U(\mathsf{z}^l)$, with z^l sampled according to γ. We then obtain the following algorithm:

$$\theta^i_{k+1} = \theta^i_k - \frac{\eta}{N^2 B}\sum_{j=1}^{N}\sum_{\ell=1}^{B}\big(\phi(\theta^j_k, \mathsf{Z}^{j,\ell}_k, x_k) - y_k\big)\nabla_\theta\phi(\theta^i_k, \mathsf{Z}^{i,\ell}_k, x_k) - \frac{\eta}{N}\nabla_\theta \mathrm{KL}(q^1_{\theta^i_k}|P^1_0) \tag{3.6}$$

Let $(Z^{j,\ell}_k, 1 \leq j \leq N, 1 \leq \ell \leq B, k \geq 0)$ be a sequence of random variables sampled i.i.d. according to γ. This algorithm is called **Bayes-By-Backprop SGD** in the following work.

Let's define the empirical distribution of the network parameters after $\lfloor Nt \rfloor$ steps of gradient descent. This distribution is defined by:

$$\mu^N_t = \frac{1}{N}\sum_{i=1}^{N}\delta_{\theta^N_{\lfloor Nt\rfloor}}$$

Finally, let's define the trajectory of the parameters $\mu^N = \{\mu^N_t, t \in [T]\}$ for T a fixed parameter.

First contribution: We showed that the sequence of trajectories μ^N follows the same law of large numbers for **Idealized SGD** and **Bayes-By-Backprop SGD** algorithms. This contribution is summarized in the following theorem:

Theorem 17. *(Informal). Under certain assumptions described in Chapter 5, if the sequence $\{\mu^N\}_{N\geq 1}$ is described by **Idealized SGD** or by **Bayes-By-Backprop SGD**, then it converges in probability to a unique deterministic solution $\bar{\mu}$ characterized by the following equation: for any test function $f \in \mathcal{C}^\infty(\Theta)$ and any $t \in [T]$ we have:*

$$\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle = -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s$$
$$- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s\big\rangle \mathrm{d}s. \tag{3.7}$$

Second contribution: We analyzed the structure of the limiting equation given by Equation (3.7), and we noted that it can be rewritten in a simpler way. This insight led us to derive a new algorithm, called **Minimal-VI SGD**, that achieves the same limiting equation. The update rule for this algorithm is described by the following equation:

$$\theta_{k+1}^i = \theta_k^i - \frac{\eta}{N^2}\sum_{j=1}^N \big(\phi(\theta_k^j, \mathsf{Z}_k^1, x_k) - y_k\big)\nabla_\theta \phi(\theta_k^i, \mathsf{Z}_k^2, x_k) - \frac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \tag{3.8}$$

Note that at each step, only two Gaussian random variables (Z_k^1, Z_k^2) are sampled for this algorithm, whereas each step of the previous algorithms requires the simulation of $O(N)$ Gaussian random variables. Therefore, this **Minimal-VI SGD** version is much less computationally expensive both theoretically and in practice.

Third contribution: We showed that the sequence of trajectories μ^N follows the same law of large numbers for the **Minimal-VI SGD** algorithm. This contribution is summarized in the following theorem:

Theorem 18. *(Informal). Under certain assumptions described in Chapter 5, if the sequence $(\mu^N)_{N\geq 1}$ is described by **Minimal-VI SGD**, then it satisfies all the statements of Theorem 17.*

This algorithm is therefore much less computationally expensive, and moreover, it has the same limiting behavior.

1.3 Chapter 6: Central Limit Theorem for Bayesian Neural Network trained with Variational Inference.

In this chapter, we follow the work done previously. We consider the same network configuration defined in (3.4), the same Variational family, and the three same algorithms **Idealized SGD** (3.5), **Bayes-By-Backprop** (3.6) and **Minimal VI** (3.8). In the previous chapter, we showed that these three different algorithms follow the same law of large numbers and that the distribution of their parameters all converge to the same deterministic limit distribution $\bar{\mu}_t$. In this chapter, we study the rate of convergence of these distributions by deriving a Central Limit Theorem. Firstly, let's define the fluctuation process:

$$\eta^N : t \in \mathbb{R}_+ \mapsto \sqrt{N}(\mu_t^N - \bar{\mu}_t)\,,$$

First contribution We show a Central Limit Theorem for the **Idealized SGD** algorithm 3.5. This contribution is summarized in the following theorem:

Theorem 19. *(Informal). Under certain assumptions described in Chapter 6, if we consider the **Idealized SGD** algorithm, the sequence $(\eta^N)_{N\geq 1}$ converges in law to a process $\eta^\star$. This process is the unique solution of a certain equation, which is entirely characterized by a G-process and more precisely by its covariance structure. You can find more details about this equation and the definition of the G-process in Chapter 6. For this algorithm, the covariance structure is given by:*

$$Cov(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f](x,y,\bar{\mu}_v), \mathscr{Q}[g](x,y,\bar{\mu}_v))\mathrm{d}v,$$

where $\mathscr{Q}[f](x,y,\bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$.

Second contribution We derived a Central Limit Theorem for the **Bayes-By-Backprop SGD** algorithm (2.6). More precisely, we showed that this algorithm follows exactly the same Central Limit Theorem as the **Idealized SGD** algorithm and therefore satisfies Theorem (6). This further justifies the use of the Monte Carlo approximation, as we see that there is no difference in convergence or in convergence rate between the Monte Carlo approximation and the idealized version.

Third contribution Finally, we derived a Central Limit Theorem for the last algorithm **Minimal-VI SGD** (2.8). This contribution is summarized in the following theorem:

Theorem 20. *(Informal). Under certain assumptions described in Chapter 6, if we consider the **Minimal-VI SGD** algorithm, the sequence $(\eta^N)_{N\geq 1}$ converges in law to a process $\eta^\star$. This process is characterized by a G-process, whose covariance structure is given by*

$$Cov(\mathscr{G}_t[f],\mathscr{G}_s[g]) = \eta^2\int_0^s Cov(\mathscr{C}[f](x,y,z^1,z^2,\bar{\mu}_v),\mathscr{C}[g](x,y,z^1,z^2,\bar{\mu}_v))\mathrm{d}v,$$

where $\mathscr{C}[f](x,y,z^1,z^2,\bar{\mu}_v) = \langle\phi(\cdot,z^1,x)-y,\bar{\mu}_v\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,z^2,x),\bar{\mu}_v\rangle$.

It can be seen that for the **Minimal-VI SGD** algorithm, the G-process has a different covariance structure compared to the other algorithms studied in this chapter. For instance, if we consider a scalar test function, it can be shown that the variance of the G-process associated with **Minimal-VI SGD** is larger than those associated with Bayes-By-Backprop and its Monte Carlo approximation. However, it is important to recall that the Minimal-VI SGD algorithm is much more computationally efficient

1.4 Chapter 7: Theoretical Gaurantees for Variational Inference with Fixed-Variance Mixture of Gaussians.

In the three first chapters of this thesis, we studied the Gaussian Variational Family. However, when the distribution we want to approximate is multimodal, a simple Gaussian does not provide a sufficient approximation. Using a larger variational family, such as Gaussian Mixtures, allows us to better approximate multimodal distributions. Moreover, the family of Gaussian Mixtures is dense in the space of probability distributions with bounded p-th order moments in the Wasserstein-p metric [Delon and Desolneux, 2020]. Therefore, this family of distributions is extremely relevant for the Variational Inference problem. In this study, we propose to consider a simplified setting where the Gaussian components have equal weights and share the same diagonal covariance. This regime breaks down the complexity of the problem, and is still theoretically challenging, but remains a practically relevant scenario. The Variational Family is then defined by:

$$\mathcal{G}_n = \left\{\frac{1}{n}\sum_{i=1}^n \mathcal{N}(x^i,\epsilon^2\mathrm{I}_d),\ x^i\in\mathbb{R}^d\right\} = \left\{k_\epsilon\star\mu,\ \mu=\frac{1}{n}\sum_{i=1}^n\delta_{x^i},\ x^i\in\mathbb{R}^d\right\}.$$

where k_ϵ is the normalized Gaussian kernel, ie, $k_\epsilon\propto\exp\big(-\|x\|^2/(2\epsilon^2)\big)$.

Recall that our objective is to approximate the target distribution $\mu^\star\propto\exp(-V)$, more precisely the optimization problem of Variational Inference is written as follows:

$$\hat{\nu} = \operatorname*{argmin}_{\nu\in\mathcal{G}_n}\mathrm{KL}(\nu,\mu^\star) \tag{3.9}$$

We can now define the objective function:

$$\begin{aligned}\mathcal{F}_\epsilon(\mu) &= \mathrm{KL}(k_\epsilon\star\mu,\mu^\star)\\ &= \int V d(k_\epsilon\star\mu) + \int\log(k_\epsilon\star\mu)d(k_\epsilon\star\mu)\end{aligned}$$

Since the variances of our Gaussians are fixed, Variational Inference aims to optimize the locations of the means $\{x^i\}_{i=1}^n$ of the Gaussian mixture to approximate the target distribution. Therefore, we can view Gaussian mixtures as particle systems that we can evolve along a flow that decreases the objective function $\mathcal{F}_\epsilon$. To accomplish this, we

will use a Wasserstein gradient descent algorithm (see Chapter 7 for more details). Let $\mu_l = \frac{1}{n}\sum_{i=1}^n \delta_{x_l^i}$ be the empirical distribution of the particles after l steps of Wasserstein gradient descent. This empirical distribution is recursively updated according to the following recursion:

$$\mu_{l+1} = (\mathrm{I}_d - \gamma \nabla \mathcal{F}_\epsilon'(\mu_l))_{\#}\mu_l \tag{3.10}$$

where $\gamma > 0$ is the discretization step of the algorithm and $\nabla \mathcal{F}_\epsilon'(\mu_l)$ is the Wasserstein gradient flow of $\mathcal{F}_\epsilon$. Using Equation (3.10), the particles of our system evolve according to the following recursion:

$$x_{l+1}^j = x_l^i - \gamma \left(\int_{\mathbb{R}^d} \nabla V(y) k_\epsilon(y - x^j) \mathrm{d}y + \int_{\mathbb{R}^d} \frac{\sum_{i=1}^n \nabla k_\epsilon(y - x^i)}{\sum_{i=1}^n k_\epsilon(y - x^i)} k_\epsilon(y - x^j) \mathrm{d}y \right),$$

First contribution: We studied the optimization error, i.e. the error obtained when we optimize (3.9). We showed the regularity of our objective function $\mathcal{F}_\epsilon$ and used this result to prove a descent lemma, which is summarized in the following proposition:

Proposition 21. *(Informal) Under certain assumptions described in Chapter 7, the following inequality holds:*

$$\mathcal{F}_\epsilon(\mu_{l+1}) - \mathcal{F}_\epsilon(\mu_l) \leq -\gamma \left(1 - \frac{\gamma}{2} M\right) \|\nabla \mathcal{F}_\epsilon'(\mu_l)\|^2_{L^2(\mu_l)}.$$

where M is a constant.

Consequently, for a sufficiently small γ, this proposition shows that the objective function $\mathcal{F}_\epsilon$ decreases at each iteration l.

Second contribution: We studied the approximation error, which quantifies how far $\hat{\nu}$ is from $\mu^\star$. More precisely, for each $n \in \mathbb{N}$ we assume that we have found the optimizer of (3.9) which is denoted by $\hat{\nu}_n = k_\epsilon \star \mu_n$ with $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i}$. We showed that if the target distribution is a, potentially infinite, mixture of Gaussians then the approximation error is bounded by a function that decreases in $\frac{\log(n)}{n}$. This contribution is summarized in the following theorem:

Theorem 22. *(Informal) Under certain assumptions described in Chapter 7, the following inequality holds:*

$$\mathrm{KL}(\hat{\nu}_n, \mu^\star) \leq C_{\mu^\star}^2 \frac{\log(n) + 1}{n}$$

where $C_{\mu^\star}$ is a constant that depends on the target distribution $\mu^\star$.

2 Part II: Thompson Sampling for Multi-Armed Bandit problems.

In this part of the thesis, we study contextual bandit problems using a Bayesian approach. More specifically, we consider Thompson Sampling algorithm. A complete description of the problem studied, as well as the Thompson Sampling algorithm, is given in Section 2 of Chapter 1.

Classical Thompson Sampling algorithms have two major limitations. First, sampling from the posterior is often difficult or even impossible, requiring the use of approximation methods. Second, these algorithms have weaker theoretical guarantees than those obtained by frequentist approaches, such as UCB. In this part of the thesis, we propose solutions to address these limitations. Specifically, we study the contextual bandit problem using a Gaussian Variational Inference approach in Chapter 8. Finally, in Chapter 9, we studied theoretically an algorithm called Feel-Good Thompson Sampling derived in Zhang [2022a], that improves the theoretical guarantees of classical Thompson Sampling by slightly modifying the posterior.

2.1 Chapter 8: Variational Inference Thompson Sampling for contextual bandits.

Recall the contextual bandit problem: at each iteration $t \in [T]$, an agent observes a context $x_t \in \mathrm{X}$ representing the state of the environment, then chooses an action $a_t = \pi_t(x_t)$, where π is called a policy, and finally receives a reward $r_t \sim \mathrm{R}(\cdot|x_t, a_t)$. The agent's objective is to find a policy that minimizes the cumulative regret defined as:

$$\mathrm{CREG}(\mathbb{Q}_{1:T}) = \sum_{t \leq T} f_\star(x_t) - f(x_t, \pi_t(x_t)) \,,$$

where $f(x,a) = \int r \mathrm{R}(\mathrm{d}r|x,a)$ and $f_\star(x) = \max_{a \in \mathcal{A}(x)} f(x,a)$.

To solve this problem, we use the Thompson Sampling algorithm. First, consider a parametric model R_θ for the reward distribution R. In this chapter, we focus on the exponential family as a model (see Chapter 8 for more details). Next, we apply the Bayesian approach and define the posterior distribution $\hat{\mathrm{p}}_t$. The principle of Thompson Sampling is to sample for each iteration t, a parameter θ_t according to the posterior, and to choose the action as follows:

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}(x_t)} \int r \mathrm{R}_{\theta_t}(\mathrm{d}r|x_t, a) \,.$$

However, in practice, it is often impossible to sample from the posterior. Indeed, this posterior is usually intractable and approximate inference methods have to be used to obtain samples with distributions "close" to the posterior. In this chapter, we focus on Gaussian Variational Inference approach to approximate the posterior. In contextual bandits, the data points progressively accumulate over time. According to Bernstein-Von Mises theorem [Van der Vaart, 2000], the Gaussian approximation becomes increasingly suitable for representing the posterior in this particular setting. The method used to solve the Variational Inference problem is the Riemannian gradient descent algorithm (see Section 1.2.2 of Chapter 1 for more details about the algorithm). As a reminder, the Variational distribution obtained after k steps of Riemannian gradient descent is $\tilde{\mathrm{q}}_{t,k} = \mathcal{N}(\tilde{\mu}_{t,k}, \tilde{\Sigma}_{t,k})$, where the parameters $\tilde{\mu}_k$ and $\tilde{\Sigma}_k$ are updated as follows:

$$\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - \gamma \nabla U_t(\tilde{\theta}_{t,k}) \,, \tag{3.11}$$
$$\tilde{\Sigma}_{t,k+1} = \tilde{A}_{t,k} \tilde{\Sigma}_{t,k} \tilde{A}_{t,k} \,,$$
$$\tilde{A}_{t,k} = \mathrm{I}_d - \gamma_t (\nabla^2 U_t(\tilde{\theta}_{t,k}) - \tilde{\Sigma}_{t,k}^{-1}) \,, \tag{3.12}$$
$$\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k} \,,$$

where h_t is the discretization step of the algorithm and $U_t(\theta) \propto -\log(\hat{\mathrm{p}}_t(\theta))$ is the potential function. The algorithm described by the previous recursion and presented in [Lambert et al., 2022b] allows us to solve the Variational Inference optimization problem. However, at each iteration k, it requires sampling from a d-dimensional Gaussian, which can be computationally expensive in high dimension.

First contribution: We proposed an improved version of the algorithm in [Lambert et al., 2022b]. Instead of considering the covariance matrix $\tilde{\Sigma}_{t,k}$, we only look at one of its square roots, which is defined by:

$$B_{t,k+1} = \left\{\mathrm{I}_d - h_t \nabla^2 U_t(\tilde{\theta}_{t,k})\right\} B_{t,k} + h_t (B_{t,k}^{-1})^\top \,. \tag{3.13}$$

It is then much more efficient to sample from the Variational posterior using the following formula:

$$\tilde{\theta}_{t,k} = \tilde{\mu}_{t,k} + B_{t,k} \epsilon_{t,k}, \quad \epsilon_{t,k} \sim \mathcal{N}(0, \mathrm{I}_d) \,.$$

We obtain our first algorithm called `VITS − I`, which at each iteration t, chooses an action as follows:

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}(x_t)} \int r \mathrm{R}_{\tilde{\theta}_{t,K_t}}(\mathrm{d}r|x_t, a) \,. \tag{3.14}$$

Next, the Variational distribution is recalculated using equations (3.11) and (3.13) for K_{t+1} iterations. This algorithm is much more efficient than the basic version presented in [Lambert et al., 2022b]. However, it is still computationally expensive in high dimension because at each iteration, we must invert the matrix $B_{t,k}$.

Second contribution: We proposed a new version of our algorithm, which uses a Taylor approximation (for h_t small enough) to approximate the inverse $B_{t,k}^{-1}$ by $C_{t,k}$ which is recursively defined by:

$$C_{t,k+1} = C_{t,k}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - \nabla^2 U_t(\tilde{\theta}_{t,k}))\}\,, \quad B_{t,k+1} = (\mathrm{I}_d - h_t\nabla^2 U_t(\tilde{\theta}_{t,k}))B_{t,k} + h_t C_{t,k}^\top\,.$$

This algorithm, which we refert to as VITS − II, is even more efficient than VITS − I and allows choosing an action a_t in the same way as in (3.14). In numerical experiments, this algorithm is much more efficient than VITS − I and obtain similar cumulative regret.

Third contribution: We proposed a final version of our algorithm that avoids the computationally expensive step of calculating the Hessian of U_t. To do this, we leverage the following Gaussian property:

$$\int \nabla^2 U_t \mathrm{d}\mathcal{N}(\mu,\Sigma) = \int \Sigma^{-1}(\mathrm{I}_d - \mu)\nabla U_t^\top \mathrm{d}\mathcal{N}(\mu,\Sigma)\,.$$

Therefore, the hessian term $\nabla^2 U_t(\tilde{\theta}_{t,k})$ in (3.12) is replaced by $C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top$. The algorthm obtained is called VITS − II Hessian-free and it is based on the following recursions:

$$\begin{aligned}
\tilde{\mu}_{t,k+1} &= \tilde{\mu}_{t,k} - h_t\nabla U_t(\tilde{\theta}_{t,k})\,,\\
C_{t,k+1} &= C_{t,k}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top)\}\,,\\
B_{t,k+1} &= (\mathrm{I}_d - h_t C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top)B_{t,k} + h_t C_{t,k}^\top\,,
\end{aligned}$$

where the Variational distribution is $\tilde{\mathrm{q}}_{t,k} = \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$ and the parameter $\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k}$. This algorithm has a very low computational cost both theoretically and empirically.

Fourth contribution: Finally, we showed that in the linear bandit setting, VITS − I obtains a regret guarantee of order $\tilde{O}(d\sqrt{dT})$. This contribution is summarized in the following Theorem:

Theorem 23. *(Informal). Under certain assumptions described in Chapter 8, if we consider VITS − I algorithm for a linear bandit, then with probability at least $1-\delta$ we have:*

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) \leq C_1 d\sqrt{dT}\log\big(3T^3\big)\log\left((1+\frac{C_2 T}{d})/\delta\right),$$

where C_1 and C_2 are constants (see Chapter 8 for more details).

To the best of our knowledge, this is the first regret bound derived for VI in the context of sequential learning. Moreover, It is in the same order as the state-of-the-art cumulative regret obtained in [Agarwal et al., 2012] for the Linear Bandit setting.

2.2 Chapter 9: Feel-Good Thompson Sampling via Langevin Monte Carlo.

In this chapter, we continue to study Thompson Sampling algorithms for solving the contextual bandit problem. While these algorithms are efficient in practice and easy to implement, their theoretical regret bounds are generally worse than those of frequentist approaches. For example, in the case of the Linear Bandit problem, the best regret bound for Thompson Sampling is $\mathrm{CREG} = \tilde{O}(d^{3/2}\sqrt{T})$ [Agrawal and Goyal, 2012], while for the Linear UCB approach, we obtain a bound $\mathrm{CREG} = \tilde{O}(d\sqrt{T})$ [Dani et al., 2008, Abbasi-Yadkori et al., 2011b]. To circumvent this issue, [Zhang, 2022a] proposed to modify the likelihood function in TS by adding a penalty term to enforce more optimistic exploration. More precisely, recall the likelihood used in the Thompson Sampling algorithm:

$$\mathrm{L}_t^{(\mathrm{TS})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\eta(g(\theta,x_s,a_s) - r_s)^2\right),$$

Here, $g(\theta, x, a)$ denotes the output of the model with parameter θ, associated with context x and action a. The idea of [Zhang, 2022a] is to replace this likelihood with a new likelihood defined as follows:

$$\mathrm{L}_t^{(\mathrm{FG})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\eta(g(\theta,x,a)-r)^2 - \lambda\min(b, g_\star(\theta,x))\right),$$

where b is a fixed parameter, λ controls the importance of the penalty, and $g_\star = \max_{a\in\mathcal{A}(x)} g(\theta,x,a)$. The algorithm then uses the same approach as Thompson Sampling, but with the new likelihood function defined above. This algorithm is called Feel-Good Thompson Sampling (FG-TS). The author shows that this new version of TS achieves regret bounds of order $\mathrm{CREG} = \tilde{O}(d\sqrt{T})$, which corresponds to the optimal minimax regret bound. However, this algorithm is unusable in practice. Indeed, the posterior distribution is even more complex than for classical Thompson Sampling. For instance it is no longer Gaussian in the linear case. Therefore, it is necessary to use an approximation method to sample from this new posterior distribution.

In this chapter, we study the FG-TS algorithm. However, to address the challenge of sampling from the complex posterior distribution, we will incorporate a Markov Chain Monte Carlo (MCMC) method.

First contribution We proposed a new version of the likelihood, which is a smoothed version of the FG-TS likelihood. We call this new version smooth-FG (sFG), and it is defined as follows:

$$\mathrm{L}_t^{(\mathrm{sFG})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\eta(g(\theta,x,a)-r)^2 - \lambda\left[b - \phi_\varsigma(b - g_\star(\theta,x))\right]\right),$$

where $\phi_\varsigma(u) = \log(1+\exp(\varsigma u))/\varsigma$ and $\varsigma > 0$ is a parameter that controls the smoothness of the likelihood.

By regularizing the likelihood, we obtain a smooth posterior distribution, which greatly improves the performance of MCMC methods, particularly gradient-based MCMC methods [Durmus et al., 2018]. We proposed a new algorithm that is similar to FG-TS, but uses this new likelihood and an approximation of the true posterior distribution, denoted $\tilde{\mathrm{q}}_t^{(\mathrm{sFG})}$. With this algorithm, we obtain the following theorem on regret control:

Theorem 24. *(Informal). Under certain assumptions described in Chapter 9, we have the following inequality:*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \frac{\lambda}{\eta\epsilon}KT + C_1\lambda T - \frac{Z_T}{\lambda} + (C_2 + \frac{C_3}{\lambda})\sum_{t=0}^{T}\mathbb{E}_{\nu_0}^{T}[\delta_t],$$

where Z_T is a term depending on the prior distribution, $\epsilon \in (0,1)$ is a fixed parameter, and C_1, C_2, C_3 are constants independent of the problem. Note that at this point, we have not chosen a specific approximation method. The term $\mathbb{E}_{\nu_0}^{T}[\delta_t]$ represents the approximation error, where $\delta_t = ||\tilde{\mathrm{q}}_t^{(\mathrm{sFG})} - \mu_t^{(\mathrm{sFG})}||_{\mathrm{TV}}$ is the Total Variation between the posterior distribution and its approximation.

Second contribution We proposed to use an MCMC method to sample from the posterior distribution. More specifically, we proposed to use the Langevin Monte Carlo algorithm, which allows to iteratively define a parameter as follows:

$$\theta_{t,k+1}^{\mathrm{L}} = \theta_{t,k}^{\mathrm{L}} + \gamma_t\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_{t,k}^{\mathrm{L}}) + \sqrt{2\gamma_t}Z_{t,k},$$

where $Z_{t,k} \sim \mathcal{N}(0, \mathrm{I}_d)$ is a Gaussian noise. Find more details about this algorithm in Chapter 1 or in Chapter 9. We also proposed to use a Metropolized version of the Langevin Monte Carlo algorithm, called Metropolized Langevin Algorithm (MALA). This algorithm is defined as follows:

$$\theta_{t,k+1}^{\mathrm{M}} = \begin{cases} \theta_{t,k}^{\mathrm{M}} + \gamma_t\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_{t,k}^{\mathrm{M}}) + \sqrt{2\gamma_t}Z_{t,k} & \text{with probability } 1/\alpha_t^M, \\ \theta_{t,k}^{\mathrm{M}} & \text{otherwise}, \end{cases}$$

where α_t^M is the Metropolis-Hastings acceptance probability and is definied in Chapter 9. Using these algorithms, we obtain the following result:

Corollary 25. *(Informal). Under certain assumptions described in Chapter 9, the following inequality holds:*

$$\begin{aligned}\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \frac{C_4}{\epsilon}\sqrt{\omega dKT\log(dT)} + \big(4\xi + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)\big)T \\ + \ C_5\sqrt{\frac{\omega KT}{d\log(dT)}}\left(-\log \mathrm{p}_0(\theta_*) + L_g + \xi T + \xi^2 T\right) \\ + \ C_6\left(1 + \sqrt{\frac{\omega KT}{d\log(dT)}}\right)\sum_{t=0}^{T}\mathbb{E}_{\nu_0}^{T}[\delta_t] + 4L_g.\end{aligned}$$

Find more details about the parameters used in Chapter 9.

We are almost at the point of obtaining a regret bound. However, there are still some steps to be completed. Specifically, we need to fix the model used, the prior distribution, and also control the approximation error term.

Third contribution Finally, we applied our previous results on the Linear Bandit. Specifically, we used the model $g(\theta, x, a) = \langle \varphi(x, a), \theta\rangle$, where ϕ is the feature function. We also considered a Gaussian prior distribution $\mathcal{N}(0, \mathtt{m}_0^{-1}\mathrm{I}_d)$, with $\mathtt{m}_0 > 0$. We thus obtained the following theorem:

Theorem 26. *(Informal). Under certain assumptions described in Chapter 9, the following inequality holds:*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \ C_7\sqrt{\omega_{\mathsf{LG}} T\log^3(dT)}\left(d(\epsilon \wedge \mathtt{m}_0)^{-1} + \sqrt{M}\mathtt{m}_0\left\|\theta_*\right\|^2\right) ,$$

Consequently, we have obtained a regret bound of order $\mathrm{CREG} = \tilde{O}(d\sqrt{T})$ for the Linear Bandit. This bound is optimal and corresponds to the minimax regret bound for this problem.

Part I

Theoretical guarantees for Variational Inference

CHAPTER 4

Variational Inference of Overparameterized Bayesian Neural Networks: a Theoretical and Empirical Study of Tempering

Chapter abstract: *This chapter studies the Variational Inference (VI) used for training Bayesian Neural Networks (BNN) in the overparameterized regime, i.e, , when the number of neurons tends to infinity. More specifically, we consider overparameterized two-layer BNN trained with VI and point out a critical issue in the mean-field regime. This problem arises from the decomposition of the lower bound on the evidence (ELBO) into two terms: one corresponding to the likelihood function of the model, promoting data-fitting, and the second to the Kullback-Leibler (KL) divergence between the prior distribution and the variational posterior, acting as a regularizer. In particular, we show both theoretically and empirically that there is a reasonable trade-off between these two terms in the overparameterized regime only when the KL term is appropriately re-scaled with respect to the ratio between the the number of observations and neurons. We also illustrate our theoretical results with numerical experiments that highlight the critical choice of this ratio.*

1 Introduction

Bayesian neural networks (BNN) have gained popularity in the field of machine learning because they promise to combine the powerful approximation and discrimination properties of (deep) neural networks (NN) with the decision-theoretic approach of Bayesian inference. Among the advantages of BNN is their ability to provide uncertainty quantification [Arbel et al., 2023], which is a must in many fields - e.g., autonomous driving [Michelmore et al., 2020, McAllister et al., 2017], computer vision [Kendall and Gal, 2017], health [Filos et al., 2019, Abdullah et al., 2022] and many other tasks in artificial intelligence [Papamarkou et al., 2024]. Second, the inclusion of prior information in some cases leads to better generalization error and calibration in classification tasks; see [Jospin et al., 2020, Izmailov et al., 2021] and references therein.

NN can be used to build complex probabilistic models for regression and classification tasks. Given $\bar{w}$ corresponding to the weights and bias of an NN, the network output can be used to define a (conditional) likelihood $\mathrm{L}(\{(x_i, y_i)\}_{i=1}^p | \bar{w})$ of some observed labels $\{y_i\}_{i=1}^p$, $y_i \in \mathsf{Y}$ associated with feature vectors $\{x_i\}_{i=1}^p$, $x_i \in \mathsf{X}$. Specifying a prior distribution for $\bar{w}$ and applying Bayes' rule yields the posterior distribution of weights. In the Bayesian approach, the goal is to find the predictive distribution from new feature vectors defined as an integral with respect to the posterior. One possible approach is to use Markov-Chain Monte Carlo methods - such as Hamiltonian Monte Carlo - for inference in Bayesian neural networks; [Neal, 2011, Hoffman et al., 2014, Betancourt, 2017].

However, the challenge of scaling HMC for applications involving high-dimensional parameter space and large datasets limits its broad application; [Cobb and Jalaian, 2021]. Computationally cheaper MCMC methods have been proposed, see [Welling and Teh, 2011, Chen et al., 2014, Brosse et al., 2018]; but these methods yield biased estimates of posterior expectation, see [Izmailov et al., 2021]. A much simpler alternative from a computational standpoint is to use Variational Inference (VI) [Blundell et al., 2015, Gal and Ghahramani, 2016, Louizos and Welling, 2017, Khan et al., 2018], which approximates the posterior with a parametric distribution. Nevertheless, little is known about the validity or limitations of the latter approach, including the choice of prior, variational family, and their interplay.

A number of recent papers have investigated the limiting behavior of gradient descent type algorithms for one or two hidden layers in the overparameterized regime, [Chizat and Bach, 2018, Rotskoff et al., 2019, Mei et al., 2018, Tzen and Raginsky, 2019, De Bortoli et al., 2020a], i.e, , the number of hidden neurons goes to infinity. More specifically, it was found that the gradient descent applied to (true) risk minimization can be viewed as a temporal and spatial discretization of the Wasserstein gradient flow of a limiting functional, which is defined on the space of probability distributions over the parameters by

$$\mathrm{R}_\mu(\mu) = \int \ell(y, \int s(\bar{w}, x)\mathrm{d}\mu(\bar{w})\mathrm{d}\pi(x,y)) + \mathrm{P}(\mu)\,, \tag{4.1}$$

where π is the data distribution over $\mathsf{X} \times \mathsf{Y}$, $s(\bar{w}, x)$ is the output prediction of the NN with parameter weights $\bar{w}$ and P plays the role of a penalty function. Roughly speaking, identifying this functional consists in noting that the risk R_w over the weights $\bar{w}$ of a NN coincides with R_μ on the set of empirical measures, i.e, for any $\bar{w} = (w_1, \ldots, w_N)$ - where N is the number of neurons-, $\mathrm{R}_w(\bar{w}) = \mathrm{R}_\mu(\mu_N)$ with $\mu_N = N^{-1}\sum_{i=1}^N \delta_{w_i}$. This result emphasizes that in the overparameterized regime, the weights of a NN act as particle discretization of probability measures and the final prediction of a NN has a form of continuous mixture.

We are interested here in performing a similar analysis but for Variational Inference (VI) of two-layer Bayesian Neural Networks (BNN). In this setting, the weights of the NN are no longer fixed, but are sampled from a variational posterior, and the prediction of the NN is the empirical average of the prediction of each sample. The variational posterior is obtained by maximizing an objective function, the Evidence Lower Bound (ELBO) over a parameter space Ξ^N. It was empirically found that the maximization of the "vanilla" ELBO function can lead to very poor inference. To address this problem, a modification of this objective function is often considered, resulting from a decomposition into two terms of this function: one corresponding to the Kullback-Leibler (KL) divergence to the prior and the other to a marginal likelihood term. Based on this decomposition, the modified version of ELBO, called partially tempered ELBO, consists in multiplying the KL term by a temperature parameter. Although this change has been justified intuitively or by purely statistical considerations, to our knowledge no formal results have been derived.

This section is organized as follows. Subsection 2 introduces the background of VI on BNN. Subsection 3 characterizes the inadequacy of these models in the limiting case of the mean field, when the data or prior variance do not scale, and identifies the well-posed regime. Subsection 4 discusses connections to related work and alternative choices of scaling for infinite-witdh NN. In Subsection 5, some numerical experiments are presented to illustrate our claims.

2 Variational inference for BNN objective

Consider a supervised setting where we have access to i.i.d. samples $\{(x_i, y_i)\}_{i=1}^p$, from a distribution π on $\mathsf{X} \times \mathsf{Y} \subset \mathbb{R}^{\mathrm{d}_\mathsf{X}} \times \mathbb{R}^{\mathrm{d}_\mathsf{Y}}$, and aim at predicting y given a new observation x. In this paper, we focus on a fully connected NN with one hidden layer and N neurons, and activation function $h : \mathbb{R}^{\mathrm{d}_\mathsf{X}} \times \mathsf{X} \to \mathbb{R}$. A common example is

$$h(b_j, x) = \sigma(\langle b, x\rangle)\,, \tag{4.2}$$

for $b \in \mathbb{R}^{\mathrm{d}_\mathsf{X}}$ and $x \in \mathsf{X}$, where $\sigma : \mathbb{R} \to \mathbb{R}$ can be the Rectified Linear Unit: $\sigma(t) = \max(0, t)$, or the Sigmoid function: $\sigma(t) = \mathrm{e}^t/(1+\mathrm{e}^t)$, for $t \in \mathbb{R}$. In addition, for each neuron $j \in \{1, \ldots, N\}$, denote by $b_j \in \mathbb{R}^{\mathrm{d}_\mathsf{X}}$ and $a_j \in \mathbb{R}^{\mathrm{d}_\mathsf{Y}}$ the j-th weights of the hidden and output layers respectively, and set $w_j = (b_j, a_j) \in \mathbb{R}^d$, $d = \mathrm{d}_\mathsf{X} + \mathrm{d}_\mathsf{Y}$, and $\bar{w} = (w_j)_{j=1}^N$ all the weights of the NN under consideration. With this notation, for each input $x \in \mathsf{X}$, the output

prediction $f_{\bar{w}} : \mathsf{X} \to \mathbb{R}^{\mathrm{d}_\mathsf{Y}}$ of the neural network can be written as:

$$f_{\bar{w}}(x) = \frac{1}{N}\sum_{j=1}^{N} s(w_j, x), \qquad s(w_j, x) = a_j h(b_j, x). \tag{4.3}$$

Given a loss function $\ell : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}_+$, we use the prediction function $f_{\bar{w}}$ to define the conditional likelihood

$$\mathrm{L}(y|x, \bar{w}) \propto \exp(-\ell(f_{\bar{w}}(x), y)) \,, \tag{4.4}$$

with respect to the Lebesgue measure on $\mathbb{R}^{d\times N}$ denoted by $\mathrm{Leb}_{d\times N}$. Then, choosing a prior pdf p_0 on $\bar{w}$, the posterior pdf $\hat{\mathrm{p}}$ of the weights is proportional to $\bar{w} \mapsto \mathrm{p}_0(\bar{w}) \prod_{i=1}^{p} \mathrm{L}(y_i|x_i, \bar{w})$. We perform Bayesian inference using VI [Khan and Rue, 2023, Blei et al., 2017, Blundell et al., 2015, Graves, 2011, Khan et al., 2018]. The general procedure is to consider a variational family of pdfs $\mathcal{G}_\Theta = \{\tilde{\mathrm{q}}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, for $\Theta \subset \mathbb{R}^{d_\theta}$ and to maximize the Evidence Lower Bound (ELBO) defined for any $\boldsymbol{\theta} \in \Theta$ by:

$$\mathrm{ELBO}^N(\boldsymbol{\theta}) = -\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}} \,|\, \mathrm{p}_0) + \sum_{i=1}^{p} \int_{\mathbb{R}^{N\times d}} \log \mathrm{L}(y_i|x_i, \bar{w}) \tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) \,. \tag{4.5}$$

It is known that maximizing ELBO^N is equivalent to minimizing $\boldsymbol{\theta} \mapsto \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}} \,|\, \hat{\mathrm{p}})$. For this reason, VI consists in approximating the posterior distribution $\hat{\mathrm{p}}$ by $\tilde{\mathrm{q}}_{\boldsymbol{\theta}^\star}$ with $\boldsymbol{\theta}^\star \in \mathrm{argmax}\, \mathrm{ELBO}^N$. The first term in (4.5) acts as a penalty term to control the deviation of $q_{\boldsymbol{\theta}^*}$ from the prior p_0, while the second term plays the role of empirical risk and promotes data-fitting.

In practice, however, it has been shown that the choice of the prior and the variational approximation ELBO^N is crucial for good performance. It was proposed by [Zhang et al., 2018a, Khan et al., 2018, Osawa et al., 2019, Ashukha et al., 2020] to weaken the regularization term KL and consider a partially tempered version of ELBO^N, which for a cooling parameter $\eta > 0$ is given by

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = -\eta \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}} \,|\, \mathrm{p}_0) + \sum_{i=1}^{p} \int_{\mathbb{R}^{N\times d}} \log \mathrm{L}(y_i|x_i, \bar{w}) q_{\boldsymbol{\theta}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) \,. \tag{4.6}$$

It has been shown in [Wenzel et al., 2020, Wilson and Izmailov, 2020] that ELBO_η^N is the same as ELBO^N but considering instead of the true posterior $\hat{\mathrm{p}}$, a partially tempered posterior $\hat{\mathrm{p}}_T \propto \mathrm{L}^{1/T}\mathrm{p}_0$, where the likelihood function is tempered for some temperature $T \geq 0$. The parameter η (or equivalently the temperature T) controls the tradeoff of the likelihood term with respect to the prior. Setting $\eta < 1$ corresponds to a *cold posterior*, where the likelihood term is strengthened so that the posterior is concentrated in regions of high likelihood. The case $\eta = 1$ corresponds to "plain" Bayesian inference, while $\eta > 1$ corresponds to *warm posterior* where the prior has a stronger influence on the posterior.

In a series of paper, [Grünwald, 2012, Grünwald and Van Ommen, 2017, Bhattacharya et al., 2019, Heide et al., 2020, Grunwald et al., 2021] have shown, significantly extending earlier results of [Barron and Cover, 1991, Zhang, 2006], that partially tempered posteriors may have better statistical properties under model misspecification than the "plain" posterior as the number of data points goes to infinity (expressed in terms posterior contraction around the best approximation of the truth). These results have been derived for Generalized Linear Models and it is not clear how these results extend to BNN.

[Wilson and Izmailov, 2020] more informally argues that tempering is not inconsistent with Bayesian principles and that it may be particularly relevant in a parametric setting (where the model is defined by parameters), as opposed to Bayesian Nonparametric approaches - e.g. , Gaussian processes. Namely, while in nonparametric approaches the model capacity is automatically scaled with the available data, this is not the case in parametric approaches, where the model capacity (which is determined by the number of neurons and the neural network architecture) is chosen by the user. Model misspecification is the rule in such case, as we show in Subsection 3 for neural networks with a hidden layer. Other works have questioned the role of data augmentation to justify tempering. In [Aitchison, 2021], the author argues that the choice of likelihood does not reflect properly the data: "curated" datasets such as CIFAR10, where many labelers agree on the label of a data point, are in favor of cold posteriors, while adding noise to the labels reverses this effect. In [Nabarro et al., 2022], the authors investigate the role of data augmentation further and

conclude that the right model should include some data augmentation but tempering still demands an explanation in that setting. In the present work, we consider the role of overparametrization of the models. A priori, both effects - dataset curation and overparametrization of BNN - seem to encourage tempered posteriors through different but complementary aspects. Still, to the best of our knowledge, the choice of temperature with respect to the number of data points and network parameters has not been investigated theoretically, in particular in the context of BNN.

Other studies, e.g. , [Farquhar et al., 2019], noted that a potential cause of the predominance of the KL term in (4.5) stems from the choice of the prior. Indeed, it has been noticed that the role of p_0 is important since it leads to very different inferences, see [Fortuin, 2022]. In particular, using priors on $\bar{w}$ which factorize over the weights, i.e, ,

$$\mathrm{p}_0(\bar{w}) = \prod_{j=1}^{N} \mathrm{p}_0^1(w_j) \,, \tag{4.7}$$

do not yield optimal performance and as a result [Tran et al., 2022, Fortuin et al., 2022, Ober and Aitchison, 2021, Sun et al., 2019] have proposed the design of new priors which introduce correlation amongst the weights and/or heavier tails than Gaussian ones.

In the present work, we take a novel approach to justify the use of ELBO_η^N based on the so-called overparameterized regime and study the impact of the choice of the cooling parameter η. We assume that the prior and posterior factorizes over the neurons, i.e, , the prior takes the form (4.7) and for each $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N) \in \Xi^N$, $\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w}) = \prod_{i=1}^N \tilde{\mathrm{q}}_{\theta_j}^1(w_j)$, where $\mathrm{p}_0{}^1$ and $\{\tilde{\mathrm{q}}_{\theta_j}^1\}_{j=1}^N$ are distributions over $\Xi \subset \mathbb{R}^d$. In this case, the variational parameter space $\Theta = \Xi^N$ and the prior distribution for each neuron $\mathrm{p}_0{}^1$ is the same. Further, we assume that for any $\theta \in \Xi$, the variational distribution writes $\tilde{\mathrm{q}}_\theta^1 = \mathscr{T}_{\theta\#}\gamma$, i.e. as the pushforward of a reference probability measure with density γ by $\mathscr{T}_\theta$ where $\{\mathscr{T}_\theta : \theta \in \Xi\}$ is a family of C^1-diffeomorphisms on $\mathbb{R}^d$.

A common choice for $\mathscr{T}_\theta$ is, setting $\theta = (\mu, \sigma) \in \mathbb{R}^d \times (\mathbb{R}_+^*)^d$,

$$\mathscr{T}_\theta : z \mapsto \mu + \sigma \odot z \,, \tag{4.8}$$

where $\odot$ is the component wise product; but of course much more sophisticated choices are possible. Then, by (4.3)-(4.4) and a change of variable, the ELBO can be expressed as

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = -\eta \sum_{j=1}^{N} \mathrm{KL}(\tilde{\mathrm{q}}_{\theta_j}^1 | \mathrm{p}_0{}^1) - \sum_{i=1}^{p} \mathrm{G}^N(\boldsymbol{\theta}; (x_i, y_i)) \tag{4.9}$$

with denoting the output of a neuron parametrized by $\theta \in \mathbb{R}^d$ for an input x_i by

$$\phi(\theta, z, x_i) = s(\mathscr{T}_\theta(z), x_i) \,, \tag{4.10}$$

and $\boldsymbol{z} = (z_1, \ldots, z_N) \in \mathbb{R}^{d \times N}$,

$$\mathrm{G}^N(\boldsymbol{\theta}; (x, y)) = \int \ell \left(y, \sum_{j=1}^{N} \frac{\phi(\theta_j, z_j, x)}{N} \right) \gamma^{\otimes N}(\mathrm{d}\boldsymbol{z}) \,. \tag{4.11}$$

The decomposition of the KL term in (4.9) as N terms results from the choice of considering priors and posteriors that factorize over neurons. Although the VI framework we are considering may seem overly simplistic in light of the above, it is the one most commonly used in practice, and therefore it is still very important to obtain useful guidelines for implementation in order to optimize its performance. Moreover, it is a first step before considering other VI methods with more complex priors and/or variational families. The expression of ELBO_η^N shows that the parameter η must be chosen to balance the two terms in (4.6)-(4.9) to obtain a well-posed objective functional as $N, p \to +\infty$ and a variational posterior $\tilde{\mathrm{q}}_{\boldsymbol{\theta}^\star}$ different from the prior. Without this parameter, optimizing the ELBO^N (4.5) leads to the collapse of the variational posterior to the prior, as shown in the following proposition.

Proposition 27. *Assume that $\mathcal{G}_\Theta$ is a family of Gaussians with diagonal covariance matrices, that $\mathrm{p}_0 \in \mathcal{G}_\Theta$ and that X is compact. Let $\boldsymbol{\theta}^{*,N} = \mathrm{argmax}_{\boldsymbol{\theta} \in \Theta} \mathrm{ELBO}^N(\boldsymbol{\theta})$. Assume also that ℓ is the square loss or cross-entropy, and that σ is Lipschitz. Then, $\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}, \mathrm{p}_0) \to 0$ as $N \to \infty$.*

This result and its proof, that can be found in Subsection 7, are inspired from [Coker et al., 2022, Theorem 1,2] who show that the moments of the predictive posterior collapse to the ones of the prior and that the KL converges to 0 as $N \to \infty$, when ℓ is the square loss or logistic loss and σ is odd. Proposition 27 states an analog result, but which holds for additional losses (i.e., also cross-entropy) and more general activation functions (e.g., non odd ones as ReLU). This is partly due to our different scaling of the output of the neural network in N^{-1}, see (4.3), that differs from theirs, in $N^{-1/2}$. To obtain their result, [Coker et al., 2022] fundamentally rely on a kind of central limit theorem, which explains their scaling. This is the main reason why they must assume that the activation function is an odd function. Note that this latter condition is not satisfied for ReLU. In contrast, by considering a mean-field regime with a scaling in N^{-1}, we can get rid off this condition by relying on a law of large numbers and encompass a larger set of losses and activation functions, including the ReLU activation function. The result of Proposition 27 highlights that optimizing ELBO^N becomes ill-posed as $N \to \infty$. This suggests that the optimal variational posterior tends to ignore the data fitting term in (4.9), and that η must be chosen to rebalance ELBO^N. In the next Subsection, we provide a theoretical framework supporting tempering and then present our main results regarding the choice of η.

3 Identifying well-posed regimes for the ELBO with product priors

We follow the approach outlined in [Chizat and Bach, 2018, Rotskoff et al., 2019, Mei et al., 2018] for ERM. We first generalize the definition of ELBO_η^N defined in (4.9) over Ξ^N, to probability measures ν on Ξ. Indeed, the following result states that ELBO_η^N can be expressed as a functional of the empirical measure over the weights $\nu_N^{\boldsymbol{\theta}}$ defined for each variational parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N) \in \Xi^N$ by

$$\nu_N^{\boldsymbol{\theta}} = N^{-1} \sum_{i=1}^N \delta_{\theta_i} \,, \tag{4.12}$$

where δ_θ is the Dirac mass at $\theta \in \Xi$. Define $\mathcal{P}_N(\Xi)$ the subset of $\mathcal{P}(\Xi)$ which can be written as (4.12) for some $\boldsymbol{\theta} \in \Xi^N$ (i.e. discrete measures supported on N parameters).

Proposition 28. *For any $N \in \mathbb{N}$, there exists a function F_η^N defined over $\mathcal{P}_N(\Xi)$ and valued in $\mathbb{R} \cup \{+\infty\}$ such that $\mathrm{F}_\eta^N(\nu_N^{\boldsymbol{\theta}}) = \mathrm{ELBO}_\eta^N(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Xi^N$.*

Proof. Denote by $\mathscr{S}_N$ the set of permutations over $\{1, \dots, N\}$ and for any $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N) \in \Theta$, $\tau \in \mathscr{S}_N$, $\boldsymbol{\theta}^\tau = (\theta_{\tau(1)}, \dots, \theta_{\tau(N)})$. Note that for any $\tau \in \mathscr{S}_N$, $\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}^\tau)$. The proof is then completed upon using that $\boldsymbol{\theta} \mapsto \nu_N^{\boldsymbol{\theta}}$ is a bijection from $\Xi^N / \sim$ to $\mathcal{P}_N(\Xi)$, where $\sim$ is the equivalence relation defined by $\boldsymbol{\theta} \sim \boldsymbol{\theta}'$ if $\exists \tau \in \mathscr{S}_N$ s.t. $\boldsymbol{\theta}' = \boldsymbol{\theta}^\tau$. □

Proposition 28 is a first step towards identifying an objective functional defined on $\mathcal{P}(\Xi)$, since F_η^N is a reparametrization of the ELBO (i.e., it has the same value) but is defined on empirical measures supported on N atoms. The main caveat is that F_η^N cannot be non-trivially extended to a functional defined for a general probability measure on Ξ, because it depends on N through the integration of the loss function with respect to the $N \times d$ dimensional Gaussian noise in (4.11). However, in our next result, we show that, when restricted to empirical probabilities, as $N \to +\infty$, F_η^N is a perturbation of the functional $\tilde{\mathrm{F}}_\eta^N$ defined *over all probabilities in* $\mathcal{P}(\Xi)$ by

$$\tilde{\mathrm{F}}_\eta^N(\nu) = -\sum_{i=1}^p \tilde{\mathrm{G}}(\nu; (x_i, y_i)) - \eta N \int \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta | \mathrm{p_0}^1) \mathrm{d}\nu(\theta) \,, \tag{4.13}$$

where

$$\tilde{\mathrm{G}}(\nu; (x, y)) = \ell\left(y, \iint \phi(\theta, z, x) \mathrm{d}\nu(\theta) \mathrm{d}\gamma(z)\right) \,, \tag{4.14}$$

and ϕ is given by (4.10). The main difference between F_η^N and $\tilde{\mathrm{F}}_\eta^N$ is the place where the integration occurs with respect to the Gaussian measure. For F_η^N, the integration is on the loss function in (4.11), while in $\tilde{\mathrm{F}}_\eta^N$, the integration is only on the second argument of the loss in (4.14). We now define for any $\theta \in \Xi$ and $x \in \mathsf{X}$, $\tilde{\phi}(\theta, x) = \int \phi(\theta, z, x) \mathrm{d}\gamma(z)$. Consider the following assumption:

Assumption 1.

(i) There exists $\mathtt{L}_\ell > 0$ *such that for any* $y \in \mathsf{Y}$*, the function* $\tilde{y} \mapsto \ell(y,\tilde{y})$ *is* $\mathtt{L}_\ell$*-smooth: for any* $\tilde{y}_1, \tilde{y}_2 \in \mathsf{Y}$*,*

$$\|\nabla_{\tilde{y}} \ell(y,\tilde{y}_1) - \nabla_{\tilde{y}} \ell(y,\tilde{y}_2)\| \leq \mathtt{L}_\ell \|\tilde{y}_1 - \tilde{y}_2\| \ . \tag{4.15}$$

(ii) There exists $C_\phi \geq 0$*, such that for any* $\theta \in \Xi$*,* $x \in \mathsf{X}$*,*

$$\int \left\|\phi(\theta,z,x) - \tilde{\phi}(\theta,x)\right\|^2 \mathrm{d}\gamma(z) \leq C_\phi \ . \tag{4.16}$$

Note that Assumption 1-(i) is satisfied for the quadratic or logistic loss if Y is bounded. We give practical conditions on the activation function σ, the prior p_0[1] and the set Ξ to ensure that Assumption 1-(ii) holds in the case where $\mathscr{T}_\theta$ is supposed to be of the form (4.8) for any $\theta \in \Xi$, later in this Subsection after stating our general results.

Theorem 29. *Assume Assumption 1. Then, there exists* $C \geq 0$ *such that for any* $N, p \in \mathbb{N}$*,* $\{(x_i,y_i)\}_{i=1}^p \in (\mathsf{X} \times \mathsf{Y})^p$*,* $\boldsymbol{\theta} \in \Xi^N$ *and* $\eta > 0$*,*

$$\left|\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}})\right| \leq Cp/N \ , \tag{4.17}$$

where $\nu_N^{\boldsymbol{\theta}}$ *is defined in* (4.12)*.*

Proof. Using that for any $y \in \mathsf{Y}$, the function $\tilde{y} \mapsto \ell(y,\tilde{y})$ is $\mathtt{L}_\ell$-smooth, we get by [Nesterov, 2004, Lemma 1.2.3], Proposition 28 and the definitions (4.9)-(4.11)-(4.13)-(4.14),

$$\left|\mathrm{F}_\eta^N(\nu_N^{\boldsymbol{\theta}}) - \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}})\right| \leq \frac{\mathtt{L}_\ell}{2N^2} \sum_{i=1}^p \int \left\| \sum_{j=1}^N \phi(\theta_j,z_j,x_i) - \tilde{\phi}(\theta_j,x_i) \right\|^2 \mathrm{d}\gamma^{\otimes N}(\boldsymbol{z}) \tag{4.18}$$

$$\leq \frac{\mathtt{L}_\ell}{2N^2} \sum_{i=1}^p \sum_{j=1}^N \int \left\|\phi(\theta_j,z_j,x_i) - \tilde{\phi}(\theta_j,x_i)\right\|^2 \mathrm{d}\gamma(z) \ . \tag{4.19}$$

The proof follows from Assumption 1-(ii). □

We also show in the following theorem that the minimization of F_η^N over $\mathcal{P}_N(\Xi^N)$ provides a good approximation for the minimization problem corresponding to $\tilde{\mathrm{F}}_\eta^N$ for sufficiently large N.

Theorem 30. *Assume Assumption 1 and that there exists* $\nu_\star \in \mathcal{P}(\Xi)$ *such that* $\nu_\star \in \mathrm{argmax}_{\mathcal{P}(\Xi)} \tilde{\mathrm{F}}_\eta^N$*. Suppose in addition that there exists* $C_\phi^{\nu_\star} \geq 0$ *such that for any* $x \in \mathsf{X}$*,*

$$\int \left\| \tilde{\phi}(\theta,x) - \int \tilde{\phi}(\theta',x) \mathrm{d}\nu_\star(\theta') \right\|^2 \mathrm{d}\nu_\star(\theta) \leq C_\phi^{\nu_\star} \ . \tag{4.20}$$

Then, there exists $C \geq 0$ *such that for any* $N, p \in \mathbb{N}$*,* $\{(x_i,y_i)\}_{i=1}^p \in (\mathsf{X} \times \mathsf{Y})^p$ *and* $\eta > 0$*,*

$$\left| \sup_{\boldsymbol{\theta} \in \Xi^N} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \sup_{\nu \in \mathcal{P}(\Xi)} \tilde{\mathrm{F}}_\eta^N(\nu) \right| \leq Cp/N \ . \tag{4.21}$$

Proof. Using Theorem 29, we easily get that for any $\boldsymbol{\theta} \in \Xi^N$,

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) \leq \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}}) + Cp/N \leq \sup_\nu \tilde{\mathrm{F}}_\eta^N(\nu) + Cp/N \ , \tag{4.22}$$

for some $C \geq 0$ independent of $\{(x_i,y_i)\}_{i=1}^p \in (\mathsf{X} \times \mathsf{Y})^p$ and $\eta > 0$. On the other hand, we have using that $\nu_\star$ is a maximizer of $\tilde{\mathrm{F}}_\eta^N$,

$$\sup_{\boldsymbol{\theta} \in \Xi^N} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) \geq \sup_\nu \tilde{\mathrm{F}}_\eta^N(\nu) - \int \left|\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}})\right| \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) - \int \left|\tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}}) - \tilde{\mathrm{F}}_\eta^N(\nu_\star)\right| \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \ . \tag{4.23}$$

Using Assumption 1, for any $y \in \mathsf{Y}$, $\tilde{y} \mapsto \ell(y,\tilde{y})$ is L_ℓ-smooth, we get by [Nesterov, 2004, Lemma 1.2.3], setting $\tilde{\phi}_N = (1/N)\sum_{j=1}^N \tilde{\phi}(\theta_j, x_i)$,

$$\int \left|\tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}}) - \tilde{\mathrm{F}}_\eta^N(\nu_\star)\right| \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \leq \mathrm{L}_\ell \sum_{i=1}^p \int \left\| \tilde{\phi}_N - \int \tilde{\phi}(\theta', x_i)\mathrm{d}\nu_\star(\theta') \right\|^2 \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \\ \leq \mathrm{L}_\ell p C_{\tilde{\phi}}^{\nu_\star}/N \,. \tag{4.24}$$

Combining (4.23), (4.24) and Theorem 29 concludes the proof. □

The bounds of Theorems 29 and 30, concentrate for $p/N \to 0$, which corresponds to the current practical overparametrized regimes for NN, where the number of parameters of the network is larger than the number of data points. We now set the cooling parameter as $\eta = \tau p/N$ with $\tau > 0$. As stated in the next proposition, whose proof can be found in Subsection 7, this tempering prevents the collapse of the variational posterior onto the prior when optimizing the tempered ELBO, in contrast with Proposition 27.

Proposition 31. *Let $\boldsymbol{\theta}^{*,N} = \mathrm{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta})$. Assume $\eta = \tau p/N$ and that ℓ is the square loss. Then, $\limsup_{N\to\infty} \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}, \mathrm{p}_0) > 0$ as $N \to \infty$.*

With the particular choice of tempering $\eta = \tau p/N$, the functional $\tilde{\mathrm{F}}_\eta^N$ depends only on the number of observations p but no longer on the number of neurons N. We denote, for that particular choice of η,

$$\mathrm{F}_\tau^p(\nu) = p^{-1}\tilde{\mathrm{F}}_\eta^N(\nu) = -\frac{1}{p}\sum_{i=1}^p \tilde{\mathrm{G}}(\nu;(x_i,y_i)) - \tau \int \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta | \mathrm{p}_0{}^1)\mathrm{d}\nu(\theta) \,. \tag{4.25}$$

In our next result, we show that with high probability, $\mathrm{F}_\tau^p(\nu)$ provides a good approximation as $p \to \infty$ of the function

$$\mathrm{R}_\tau(\nu) = -\int \tilde{\mathrm{G}}(\nu;(x,y))\mathrm{d}\pi(x,y) - \tau \int \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta | \mathrm{p}_0{}^1)\mathrm{d}\nu(\theta) \,, \tag{4.26}$$

where $\tilde{\mathrm{G}}$ is defined by (4.14).

Proposition 32. *Assume Assumption 1 and that there exists $M_G > 0$, such that for any $\nu \in \mathcal{P}(\Xi)$, $0 \leq \tilde{\mathrm{G}}(\nu;(x,y)) \leq M_G$, for π-almost all $(x,y) \in \mathsf{X}\times\mathsf{Y}$. Suppose in addition that $\{(x_i,y_i)\}_{i=1}^p$ are i.i.d. with distribution π. Then, for any $\nu \in \mathcal{P}(\Xi)$ and $\delta > 0$, with probability $1-\delta$ at least, it holds*

$$|\mathrm{F}_\tau^p(\nu) - \mathrm{R}_\tau(\nu)| \leq M_G\sqrt{\log(\delta/2)/(2p)} \,. \tag{4.27}$$

The proof follows from applying Hoeffding's inequality on the bounded i.i.d. variables $\tilde{\mathrm{G}}(\nu;(x_i,y_i))$ for $i = 1,\ldots,p$.

It is worth noting that the limiting risk R_τ is similar to the one obtained in the analysis of the limiting behavior of gradient descent type algorithms for two-layer NN in the overparameterized regime, by [Chizat and Bach, 2018, Rotskoff et al., 2019, Mei et al., 2018, Tzen and Raginsky, 2019, De Bortoli et al., 2020a] - see (4.1). Moreover, the maximization of the ELBO using gradient descent can be viewed as a temporal and spatial discretization of the Wasserstein gradient flow of the limiting function (4.26).

We conclude this Subsection by illustrating when the previous results hold, i.e. when the assumptions are satisfied for a mean-field variational family associated with the family of C^1-diffeomorphisms $\{\mathscr{T}_\theta : \theta \in \Xi\}$ given in (4.8). Consider the following assumption:

Assumption 2. *(i) The subset Ξ is a compact set of $\mathbb{R}^d \times (\mathbb{R}_+^*)^d$, and X,Y are compact sets of $\mathbb{R}^{\mathrm{d}_\mathsf{X}}, \mathbb{R}^{\mathrm{d}_\mathsf{Y}}$.*

(ii) The probability measure γ satisfies $\int \|z\|^4 \mathrm{d}\gamma(z) < +\infty$.

(iii) For any $x \in \mathsf{X}$, there exists $\mathrm{L}_h \geq 0$ such that the function $b \mapsto h(b,x)$ is L_h-Lipschitz on $\mathbb{R}^{\mathrm{d}_\mathsf{X}}$ and $\sup_{x\in\mathsf{X}, b\in\mathbb{R}^{\mathrm{d}_\mathsf{X}}} |h(b,x)|/(1+\|b\|) < +\infty$.

(iv) The prior density $\mathrm{p}_0{}^1$ is positive on $\mathbb{R}^d$ and satisfies $\theta \mapsto \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta | \mathrm{p}_0{}^1)$ is continuous on Ξ.

Note that the condition that for any $x \in \mathsf{X}$, the condition $b \mapsto h(b,x)$ is $\mathtt{L}_h$-Lipschitz is automatically satisfied for h of the form (4.2) with σ the RELU or sigmoid function if X is bounded. Also, we verify in the next proposition, whose proof can be found in Subsection 7, that $\theta \mapsto \mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta | \mathrm{p}_0{}^1)$ is continuous if $\mathrm{p}_0{}^1$ and γ are non-degenerate Gaussian distributions.

Proposition 33. *Assume Assumption 1-(i) and Assumption 2. Then Assumption 1-(ii) and the conditions of Theorem 30 hold.*

Following the submission of our results, several papers have studied the theoretical properties of the method derived in this work for the overparameterized regime. In [Descours et al., 2023a], the authors derive a Law of Large Numbers for three different training schemes, one of which (Bayes-by-Backprop SGD) is exactly the training method described in this paper. Similarly, in [Descours et al., 2024], they derive a Central Limit Theorem for the same three training schemes presented in [Descours et al., 2023a]. Note that both paper use the re-scaling of the ELBO advocated by our results.

4 Discussion on the Lazy versus Mean Field regime for BNN

In (4.3), we chose to scale the output of the Bayesian Neural Network by a factor $1/N$, where N is the number of neurons. In the Bayesian Neural Network literature, it is also common to consider scaling the output of the neural network by $1/\sqrt{N}$, i.e.

$$f_{\bar{w}}(x) = N^{-1/2} \sum_{j=1}^{N} s(w_j, x), \tag{4.28}$$

Regarding standard (non Bayesian) neural networks, both the scalings $1/\sqrt{N}$ and $1/N$ have been considered to study overparametrized (infinite-width) neural networks, and are referred to respectively as the Neural Tangent Kernel (NTK) regime [Jacot et al., 2018] and the Mean Field (MF) regime [Chizat and Bach, 2018, Mei et al., 2018, Sirignano and Spiliopoulos, 2020a]. These choices are briefly discussed and compared in [Chizat et al., 2019]. The scaling $1/\sqrt{N}$ results in a so-called lazy training regime where the output of the neural network hardly varies with respect to its initialization. In contrast, the scaling $1/N$ allows to converge as $N \to \infty$ to a non degenerate dynamic described by a partial differential equation. We now discuss these choices of possible scalings for Bayesian neural networks.

When the output of a BNN is scaled as (4.28), it is well-known that the output prediction function $f_{\bar{w}} : \mathsf{X} \to \mathbb{R}^{\mathrm{d_Y}}$ under the prior converges weakly to a (prior) Gaussian Process as $N \to \infty$ [Neal, 2012, Lee et al., 2019, Matthews et al., 2018, Garriga-Alonso et al., 2019, Novak et al., 2019, Hron et al., 2020a]; and under the posterior to a (posterior) Gaussian process [Hron et al., 2020b] (assuming the likelihood is a bounded continuous function of the NN output).

We establish in the supplement, Subsection 7, that even when choosing the scaling $1/\sqrt{N}$, our conclusions and in particular Theorem 30 still hold. Indeed, we show that the problem of optimizing the ELBO using (4.28) can be reformulated as the one corresponding to a $1/N$ scaling. However the choice of the temperature in the resulting ELBO is more subtle.

Here we consider the Mean Field scaling (4.3) and we obtain results on tempering that are easy to interpret, since the whole (KL) regularization term is reweighted with respect to the number of observations and neurons. We advocate more generally that the MF regime for Bayesian Neural networks deserves to be analyzed since it provides a simple model for overparametrized NN which comes with many interesting consequences that we derived previously; which are first steps towards the full understanding of NN training and performance. We now turn to a practical evaluation of our study.

Remark 34. *One may argue that with the $1/N$ scaling* (4.3)*, the variance of the prior distribution collapses to zero as $N \to \infty$. However, the distribution which really matters is the (variational) posterior for which the variance is expected not to vanish with the tempering we propose, with non odd activation functions. We confirm this statement using a numerical experiment provided in Subsection 7.*

5 Experiments

In this section we illustrate our findings and their practical implications for image classification on standard datasets (MNIST, CIFAR-10). The reader may refer to Subsection 7 for additional experiments, including on regression tasks, that highlight the importance of rescaling the ELBO. In this Subsection, we illustrate the influence of the parameter τ through different metrics.

Evaluation. Let $\mathcal{D} = (x_i, y_i)_{i=1}^p$ be a dataset, where $y_i = c \in \{1, \dots, n_l\}$ is a discrete class label. For an input $x \in \mathsf{X}$, the predictive probability of a class c by a neural network with weights $\bar{w}$ is defined by $\Psi_c(f_{\bar{w}}(x))$, where $\Psi_c(f_{\bar{w}}(x))$ denotes the c-th component of the softmax function applied to the output $f_{\bar{w}}(x) \in \mathbb{R}^{n_l}$ of the neural network. The cross entropy loss writes $\ell_{\mathrm{CE}}(y, f_{\bar{w}}(x)) = -\sum_{c=1}^{n_l} \tilde{y}_c \log(\Psi_c(f_{\bar{w}}(x)))$, where $\tilde{y}_c$ denotes the c-th coordinate of a one-hot representation of the label y and the Negative Log Likelihood (NLL) $\sum_{i=1}^p \int_{\mathbb{R}^{N\times d}} \ell_{\mathrm{CE}}(y_i, f_{\bar{w}}(x_i)) q_{\boldsymbol{\theta}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w})$. The calibration performance of the model can be estimated by the Expected Calibration Error (ECE) [Naeini et al., 2015], see also Subsection 7. We recall that model is calibrated if the predictive posterior is the true probability for each class $c \in \{1, \dots, n_l\}$. However, since these probabilities are unknown, they have to be estimated, e.g. through ECE. As the NLL, ECE penalizes low probabilities assigned to correct predictions and high probabilities assigned to wrong ones; but these evalutation metrics are not strictly equivalent.

To make our prediction, for $x \in \mathsf{X}$, we use the posterior predictive distribution defined for a class c as $\int \Psi_c(f_{\bar{w}}(x)) \tilde{\mathrm{q}}_{\boldsymbol{\theta}^\star}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w})$ with $\boldsymbol{\theta}^\star$ obtained by minimization of ELBO_η^N by Bayes by Backprop. This integral is estimated by an empirical version

$$\int \Psi_c(f_{\bar{w}}(x)) \tilde{\mathrm{q}}_{\boldsymbol{\theta}^\star}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) \approx \frac{1}{m} \sum_{l=1}^m \Psi_c(f_{\bar{w}_l}(x)) \,, \tag{4.29}$$

where for $l = 1, \dots, m$, $\bar{w}_l$ are i.i.d. samples from $\tilde{\mathrm{q}}_{\boldsymbol{\theta}^\star}$. All the evaluation metrics mentioned above (NLL, ECE), as well as the accuracy are estimated using the same procedure. We will present our results on the MNIST dataset (where $p = 6.10^4$) and the CIFAR-10 dataset (where $p = 5.10^4$) [Krizhevsky et al., 2009].

Setup. We use a Linear BNN on MNIST, and ResNet20 architecture [He et al., 2016] on CIFAR-10 [Simonyan and Zisserman, 2015]. For CIFAR-10, we use the standard data augmentation techniques, see [Khan et al., 2018]. For each neuron, we use a centered Gaussian prior with variance $1/5$, following [Osawa et al., 2019]. We train each BNN by Bayes by Backprop [Blundell et al., 2015] with the reparametrization trick (see Subsection 7) and using batch normalization [Ioffe and Szegedy, 2015].

Results Figures 4.1 and 4.2 illustrate the performance of the different models and data sets for different values of τ. We evaluate the models on the test set in terms of their accuracy, NLL, ECE, and average confidence over the test set. In all experiments, we take $m = 50$ to approximate a BNN prediction and average our results over 5 experiments for each τ. It is worth noting that for a large τ, the accuracy decreases while the NLL increases. This is hardly a surprise, since the KL regularization forces the VI posterior to stay close to the prior distribution, resulting in underfitting. At the same time, the ECE value is low because of the poor confidence in the model, which is reflected in the accuracy. For small values of τ, the data fitting term is privileged, so the accuracy of the model is high, while the NLL is low. At the same time, the confidence in the model is very high, resulting in a low ECE. For intermediate values of τ, the accuracy of the models starts to decrease, but slower than the confidence in the model, which explains an increase in ECE. We also illustrate the different regimes for the parameter τ with additional experiments in Subsection 7, including analysis of the weights distribution and out-of-distribution detection.

Figure 4.3 displays the NLL and accuracy of 7 networks with an increasing number of neurons, and trained on the same dataset (MNIST) by optimizing the classical ELBO. The ratio p/N evolves here as the dataset is fixed and only the network size changes. Figure 4.3 highlights that the performance decreases as N increases and suggests the critical role of the ratio p/N.

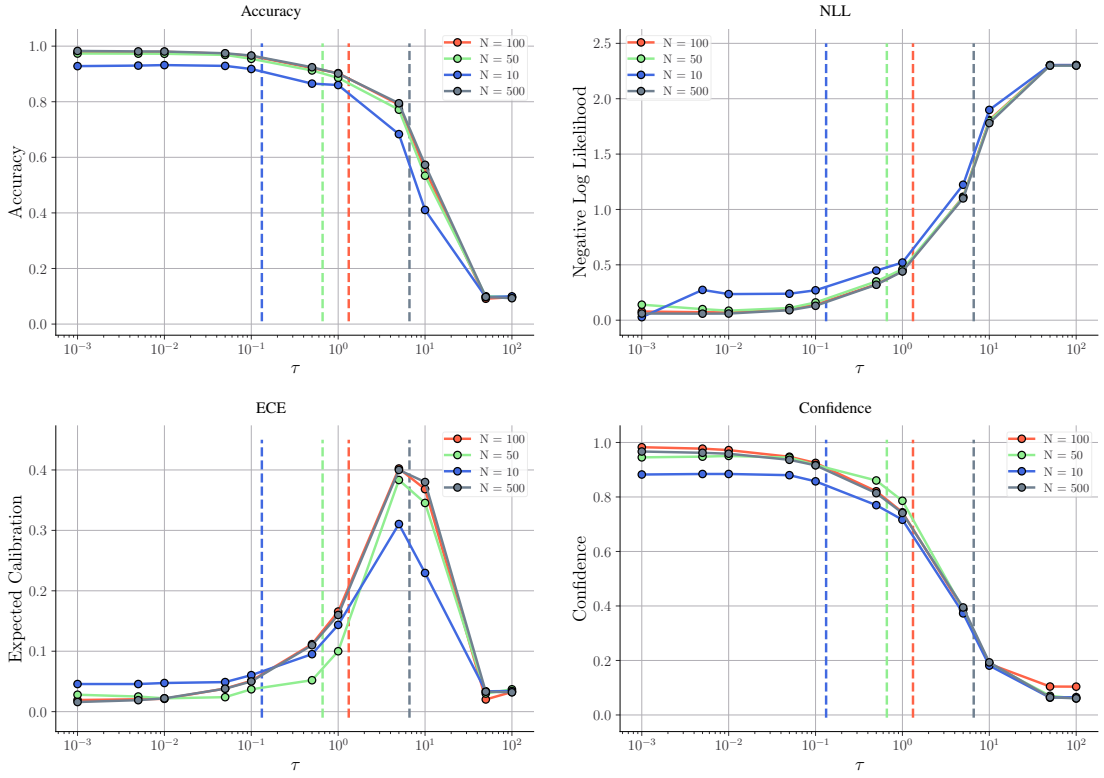


Fig. 4.1 Effect of the temperature for a Linear BNN trained on MNIST. No cooling $\eta = 1$ is indicated by a red line.

6 Conclusion

In this work, we studied BNN trained with mean-field VI in the overparameterized regime. We have highlighted both theoretically and numerically that the partially tempered ELBO_η^N advocated for VI for BNN effectively addresses the potential imbalance between the data fitting and KL terms. For mean-field VI and product prior distributions, we found that the cooling parameter must be chosen proportional to the ratio between the number of observations and neurons to achieve a balance between the data fitting and KL regularizer. With this choice, ELBO_η^N converges to a limiting functional that has the same structure as the one given by [Chizat and Bach, 2018, Rotskoff et al., 2019, Mei et al., 2018, Tzen and Raginsky, 2019, De Bortoli et al., 2020a] for empirical risk minimization. We also explained why, in the absence of cooling, the KL term can dominate the data fitting term, typically leading to underfitting of the model, which in practice translates into poor results on all metrics considered. Our work therefore provides a well-grounded theoretical justification for the importance of using a partial tempering in the overparameterized framework, which completes the justifications given by [Wenzel et al., 2020, Izmailov et al., 2021, Nabarro et al., 2022, Noci et al., 2021, Laves et al., 2021]. While our theoretical results apply to a neural network with a single hidden layer, we have shown numerically that similar conclusions can be drawn for more general NN architectures. We emphasize that the introduction of a cooling factor into the Mean-Field VI for BNN is not without implications for the validity of Bayesian inference, and that the conclusions that can be drawn in this framework-in particular, Bayesian uncertainty quantification-must therefore be used with care (even though the accuracy, NLL, and ECE metrics obtained with Mean-Field VI compare favorably to their "classical" ERM learning counterparts).

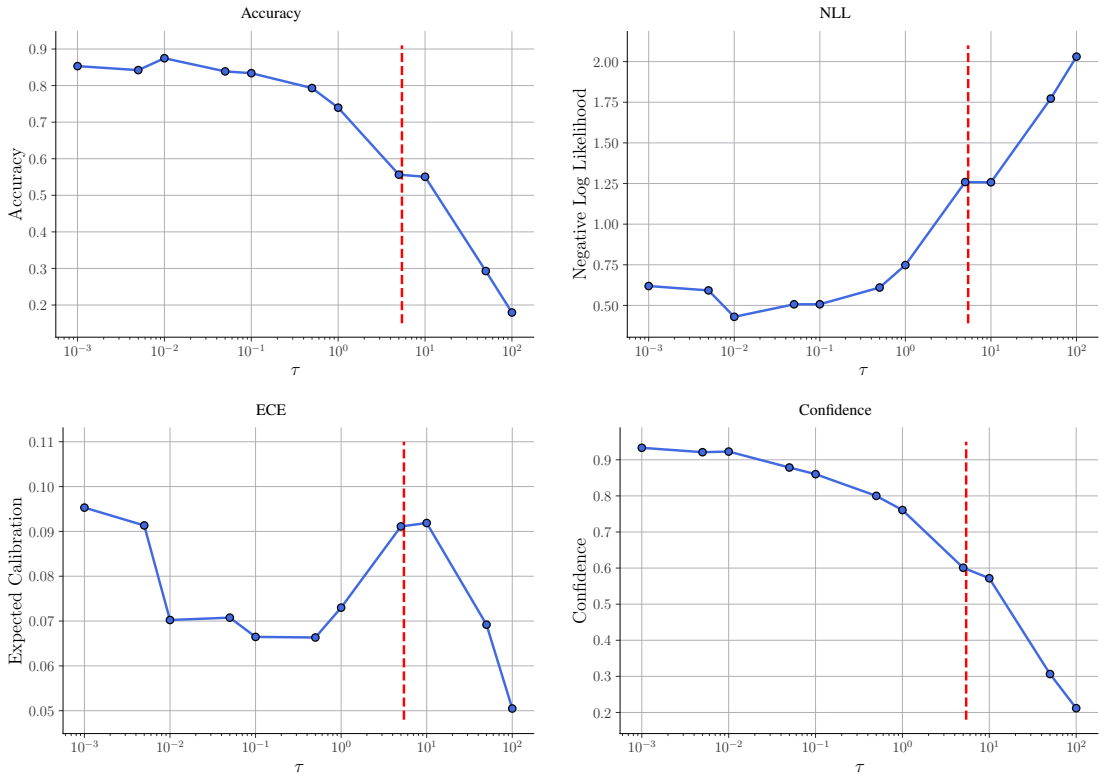


Fig. 4.2 Effect of the temperature for a Resnet20 trained on CIFAR-10. No cooling $\eta = 1$ is indicated by a red line.

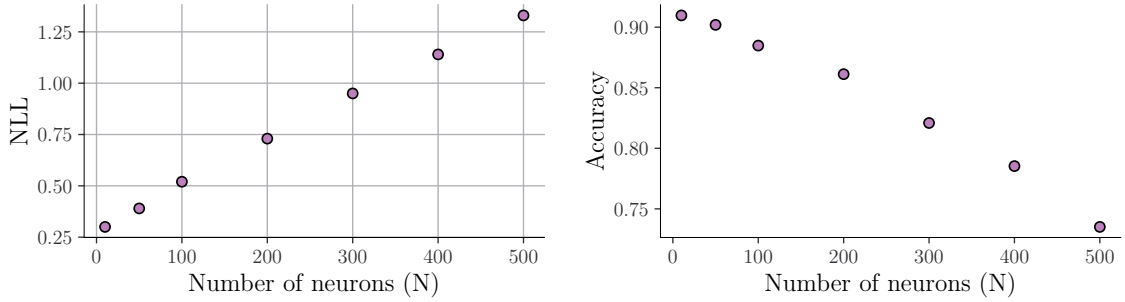


Fig. 4.3 NLL and accuracy after training using the classical ELBO on MNIST with an increasing number of neurons.

7 Appendix

Proof of Proposition 27

We have assumed that $\mathcal{G}_\Theta$ is a family of Gaussians with diagonal covariance matrices, and that $\mathrm{p}_0 \in \mathcal{G}_\Theta$ hence there exists $\boldsymbol{\theta}_0 = (\mu, \sigma) \in \mathbb{R}^{N(\mathrm{d_X}+\mathrm{d_Y})} \times \mathbb{R}^{N(\mathrm{d_X}+\mathrm{d_Y})}$ such that $\mathrm{p}_0 = q_{\boldsymbol{\theta}_0}$. For ease of notations, we work with p_0 standard Gaussian:

$$\mathrm{p}_0(\bar{w}) = \prod_{j=1}^{N}\prod_{l=1}^{\mathrm{d_Y}} \mathcal{N}(a_{i,j}; 0, 1) \times \prod_{j=1}^{N}\prod_{l=1}^{\mathrm{d_X}} \mathcal{N}(b_{j,l}; 0, 1) \tag{4.30}$$

Our results hold for more general parameters for p_0 but we fix these ones for convenience of notations. The posterior $\tilde{\mathrm{q}}_{\boldsymbol{\theta}} \in \mathcal{G}_\Theta$ is:

$$\tilde{\mathrm{q}}_{\boldsymbol{\theta}}(\bar{w}) = \prod_{j=1}^{N}\prod_{l=1}^{\mathsf{d_Y}} \mathcal{N}(a_{j,l}; \mu_{a_{j,l}}, \sigma^2_{a_{j,l}}) \times \prod_{j=1}^{N}\prod_{l=1}^{\mathsf{d_X}} \mathcal{N}(b_{j,l}; \mu_{b_{j,l}}, \sigma^2_{b_{j,l}}) \tag{4.31}$$

We define $a_j = (a_{j,1}, \ldots, a_{j,\mathsf{d_Y}}) \in \mathbb{R}^{\mathsf{d_Y}}$ and $b_j = (b_{1,j}, \ldots, b_{\mathsf{d_X},j}) \in \mathbb{R}^{\mathsf{d_X}}$ respectively the j^{th} row of the first layer weight matrix and the j^{th} column of the second layer weight matrix. We denote $\mu_{a_j} = (\mu_{a_{j,1}}, \ldots, \mu_{a_{j,\mathsf{d_Y}}}) \in \mathbb{R}^{\mathsf{d_Y}}$, $\mu_a = (\mu_{a_1}, \ldots, \mu_{a_N}) \in \mathbb{R}^{N\mathsf{d_Y}}$.

Recall that

$$\mathrm{ELBO}^N(\boldsymbol{\theta}) = -\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}) - \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}} | q_{\boldsymbol{\theta}_0}), \quad \text{with } \mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}) = -\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}}\left[\sum_{i=1}^{p} \log(\mathrm{L}(y_i|x_i,\bar{w}))\right], \tag{4.32}$$

where $\mathrm{L}(y|x,\bar{w}) \propto \exp(-\ell(f_{\bar{w}}(x), y))$ is defined by (4.4).

By the optimality of $\boldsymbol{\theta}^\star$, we have:

$$\mathrm{ELBO}^N(\boldsymbol{\theta}^\star) \geq \mathrm{ELBO}^N(\boldsymbol{\theta}_0), \tag{4.33}$$

Hence,

$$\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*} | q_{\boldsymbol{\theta}_0}) \leq \mathcal{L}(q_{\boldsymbol{\theta}_0}) - \mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*}). \tag{4.34}$$

We now deal separately with the square loss (Case 1) and cross-entropy loss (Case 2). Throughout, we will often use the notation $\sigma_j = \sigma(< b_j, x >)$ for any $j = 1, \ldots, N$ and a generic point $x \in \mathsf{X}$. Since we have assumed that σ is L-Lipschitz, for any $y \in \mathbb{R}$, $|\sigma(y)| \leq |\sigma(0)| + L|y|$. Also, to explicit the dependence of $\boldsymbol{\theta}^*, \boldsymbol{\theta}_0$ in N we will write their associated distributions $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}$ and $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}$ respectively.

Case of the square loss The idea of the proof is to show that the right hand side term of (4.34) converges to zero by showing that the two negative log likelihoods converge to the same finite limit, and hence their difference to zero as N goes to infinity. When l is the square loss, for any $q^N_{\boldsymbol{\theta}} \in \mathcal{G}_\Theta$, by (4.32) we have

$$\mathcal{L}(q^N_{\boldsymbol{\theta}}) = \sum_{i=1}^{p} \mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}\left[\|y_i\|^2 + \|f_{\bar{w}}(x_i)\|^2 - 2\langle y_i, f_{\bar{w}}(x_i)\rangle + \log(Z)\right], \tag{4.35}$$

where Z is the normalization constant of the model defined by (4.4). We will show that for both the prior $q^N_{\boldsymbol{\theta}_0}$ and optimal posterior $q^N_{\boldsymbol{\theta}^*}$, the first and second moment of the predictive distribution converge to zero as N goes to infinity.

Under the prior distribution (4.30), for any $x \in \mathsf{X}$, the first of the predictive distribution can be written:

$$\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}}[f_{\bar{w}}(x)] = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\sigma_j a_j\right] = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\sigma_j]\mathbb{E}[a_j] = 0\,,$$

and second moments,

$$\begin{aligned}
\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}}[\|f_{\bar{w}}(x)\|^2] &= \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}}\left[\frac{1}{N^2}\left(\sum_{j=1}^{N}\sigma_j^2\|a_j\|^2 + 2\sum_{j=1}^{N}\sum_{k<j}\sigma_j\sigma_k\langle a_j, a_k\rangle\right)\right] \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}}[\sigma_j^2] \leq \frac{1}{N^2}\left(\sum_{j=1}^{N}|\sigma(0)|^2 + L^2\|x\|^2\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}}[\|b_j\|^2]\right) \\
&\xrightarrow[N\to\infty]{} 0\,,
\end{aligned}$$

Hence we first obtain:

$$\lim_{N\to\infty} \mathcal{L}(q^N_{\boldsymbol{\theta}_0}) = \sum_{i=1}^{p} \|y_i\|^2 + \log Z. \tag{4.36}$$

We now turn to showing that $\mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*})$ has the same limit. First notice that since $\mathcal{L}$ is a positive function, by (4.34) we have: $\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0}) \leq \mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})$. Since the right-hand term is a converging sequence, it means that $\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0})$ is bounded by a constant C_{KL} independent of N.

By applying Lemmas 35 and 36, we have:

$$\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}}[< y_i, f_{\bar{w}}(x_i) >] \quad \leq \|y_i\| \|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}}[f_{\bar{w}}(x_i)]\| \leq \frac{\phi(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}} \leq \frac{\phi(C_{\mathrm{KL}}, \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}} \tag{4.37}$$

$$\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}}[\|f_{\bar{w}}(x)\|^2] \leq \frac{\psi(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}} \leq \frac{\psi(C_{\mathrm{KL}}, \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}} \tag{4.38}$$

where the most right hand side inequalities come from the fact that $\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0})$ is bounded by a constant C_{KL} independent of N; and $\phi(C_{\mathrm{KL}}, \mathsf{X}, \mathrm{d}_{\mathsf{Y}}), \psi(C_{\mathrm{KL}}, \mathsf{X}, \mathrm{d}_{\mathsf{Y}})$ are constants that only depend on the data points $(x_i, y_i)_{i=1}^p$, the spaces X, Y and parameters of the prior distribution (through C_{KL}). Hence, the first and second moments of the predictive under the posterior $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}$ converge to 0. Hence, we obtain:

$$\lim_{N\to\infty} \mathcal{L}(q^N_{\boldsymbol{\theta}^*}) = \sum_{i=1}^{p} \|y_i\|^2 + \log Z. \tag{4.39}$$

From (4.34), (4.36) and (4.39) we finally that

$$\lim_{N\to\infty} \mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) = 0. \tag{4.40}$$

Case of the cross-entropy Similarly to the square loss case, the idea of the proof is to show that $\mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*})$ have the same limit. We will make use of Lemma 37 which specify that limit under a null moment assumption.

Under the prior distribution $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}$,

$$\|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}}[\frac{1}{N}\sum_{j=1}^{N}\sigma_j a_j]\| = \frac{1}{N}\|\sum_{j=1}^{N}\mathbb{E}[\sigma_j]\mathbb{E}[a_j]\| = 0, \tag{4.41}$$

hence by Lemma 37:

$$\lim_{N\to\infty} \mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) = p(\log(\mathrm{d}_{\mathsf{Y}}) + \log Z).$$

We now turn to the predictive distribution under the posterior $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}$. Recall that since $\mathcal{L}$ is a positive function, using the optimality of the posterior we have: $\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) \leq \mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})$. Since the right-hand term is a converging sequence, it means that $\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})$ is bounded by a constant C_{KL} independent of N.

By Lemma 35, we can bound the first moment of the predictive distribution as:

$$\|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}}[f_{\bar{w}}(x)]\| \leq \frac{\phi(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}} \leq \frac{\phi(C_{\mathrm{KL}}, \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}}, \tag{4.42}$$

where the last inequality comes from the fact that the KL term is bounded by a constant C_{KL} independent of N for the optimal variational parameter $\boldsymbol{\theta}^{*,N}$. Moreover, by using similar argument than in the proof of Lemma 35, we can show that each coordinate μ, σ of $\boldsymbol{\theta}^{*,N}$ is bounded as:

- $\mu \leq \sqrt{2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})} \leq \sqrt{2C_{\mathrm{KL}}}$

- $\sigma \leq 2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) + 1 \leq 2C_{\mathrm{KL}} + 1$

It means that each neuron weight has bounded mean and variance. We can thus apply Lemma 37, which yields:

$$\lim_{N\to\infty} \mathcal{L}(q^N_{\boldsymbol{\theta}}) = p(\log(\mathrm{d}_{\mathsf{Y}}) - \log Z).$$

As $0 \leq \mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) \leq \mathcal{L}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) - \mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*})$ we obtain:

$$\lim_{N\to\infty} \mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) = 0. \tag{4.43}$$

Lemma 35. *Assume the conditions of Proposition 27 hold. Then there exists a function ϕ, increasing in its first variable, such that*

$$\|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| \leq \frac{\phi(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}}.$$

Proof. By Cauchy-Schwartz inequality, the first moment of the predictive distribution under the variational posterior can be upper bounded as:

$$\|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| = \frac{1}{N}\|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}}[\sum_{j=1}^N \sigma(\langle b_j, x\rangle)a_j]\| \leq \frac{1}{N}\sum_{j=1}^N |\mathbb{E}[\sigma(\langle b_j, x\rangle)]|\|\mu_{a_j}\|.$$

Since σ is Lipschitz, $|\sigma(x)| \leq C_0 + L|x|$ where $C_0 = |\sigma(0)|$. Hence,

$$|\mathbb{E}[\sigma(\langle b_j, x\rangle)]| \leq |C_0 + L\mathbb{E}[|\langle b_j, x\rangle|]| \leq C_0 + L\sum_{l=1}^{\mathrm{d}_{\mathsf{X}}} \mathbb{E}[|b_{j,l}||x_l|]$$

Let's start by finding an upper bound for $\mathbb{E}[|b_{j,l}|]$. If $b_{j,l} \sim \mathcal{N}(\mu_{b_{j,l}}, \sigma^2_{b_{j,l}})$, then $|b_{j,l}|$ has an absolute Gaussian distribution and denoting Φ the CDF of a standard Gaussian, we have

$$\mathbb{E}[|b_{j,l}|] = \sigma_{b_{j,l}}\sqrt{\frac{2}{\pi}}\exp\left(\frac{-\mu^2_{b_{j,l}}}{2\sigma^2_{b_{j,l}}}\right) + \mu_{b_{j,l}}\left[1 - 2\Phi\left(-\frac{\mu_{b_{j,l}}}{\sigma_{b_{j,l}}}\right)\right] \leq \sigma_{b_{j,l}}\sqrt{\frac{2}{\pi}} + |\mu_{b_{j,l}}|.$$

Recall that the KL between the posterior $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}$ and prior $\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}$ can be written:

$$\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) = \frac{1}{2}\sum_{j=1}^N \left[\sum_{l=1}^{\mathrm{d}_{\mathsf{X}}}(\mu^2_{b_{j,l}} + \sigma^2_{b_j,l} - \log\left(\sigma^2_{b_j,l}\right) - 1) + \sum_{l=1}^{\mathrm{d}_{\mathsf{Y}}}(\mu^2_{a_{j,l}} + \sigma^2_{a_j,l} - \log\left(\sigma^2_{a_j,l}\right) - 1)\right]$$

Hence, for any $j = 1, \ldots, N$ and $l = 1, \ldots, \mathrm{d}_{\mathsf{X}}$:

$$|\mu_{b_{j,l}}| \leq \|\mu\|_2 \leq \sqrt{2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})}, \tag{4.44}$$

$$\sigma_{b_{j,l}} \leq \|\sigma\|_2 \leq |\sigma_{b_{j,l}} + 1 - 1| \leq |\sigma^2_{b_{j,l}} - \log\left(\sigma^2_{b_{j,l}}\right) - 1| + 1 \leq 2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) + 1, \tag{4.45}$$

and

$$\mathbb{E}[|b_{j,l}|] \leq \sqrt{\frac{2}{\pi}}(2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) + 1) + \sqrt{2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})} := D(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}))$$

Where D is increasing. Hence, since X is compact, there exists C_{X} such that $\|x\|_1 \leq C_{\mathsf{X}}$ and:

$$|\mathbb{E}[\sigma(\langle b_j, x\rangle)]| \leq C_0 + LC_{\mathsf{X}}D(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})) := E(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}} | \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}),$$

Where E is increasing in its first variable. Finally, since

$$N^{-1}\sum_{j=1}^N \|\mu_{a_j}\|_2 \leq N^{-1}\|\mu_a\|_1 \leq N^{-1}\sqrt{N\mathrm{d}_{\mathsf{Y}}}\|\mu_a\|_2 \leq N^{-\frac{1}{2}}\sqrt{\mathrm{d}_{\mathsf{Y}}}\sqrt{2\mathrm{KL}(q^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})}, \tag{4.46}$$

the first moment of the predictive distribution can be upper bounded as:

$$\|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| \leq \frac{E(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), X)\sqrt{\mathrm{d}_{\mathsf{Y}}}\sqrt{2\mathrm{KL}(q^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})}}{\sqrt{N}} := \frac{\phi(\mathrm{KL}(q^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{\sqrt{N}},$$

where ϕ is increasing in its first variable. □

Lemma 36. *Assume the conditions of Proposition 27 hold. Then there exists a function ψ depending only on $\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})$, X, and d_{Y} such that G, increasing in its first variable, such that:*

$$\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}}[\|f_{\bar{w}}(x)\|^2] \leq \frac{\psi(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}), \mathsf{X}, \mathrm{d}_{\mathsf{Y}})}{N}.$$

Proof. For a posterior of the form (4.31), we can write the second moment of the predictive distribution as:

$$\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}}[\|f_w(x)\|^2] = \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}[\sigma_j^2]\mathbb{E}[\|a_j\|^2] + \frac{2}{N^2}\sum_{j=1}^N\sum_{k<j}^N \mathbb{E}[\sigma_j]\mathbb{E}[\sigma_k]\mathbb{E}[\langle a_j, a_k\rangle].$$

We start with the second term on the right hand side of (4.47). Using $\mathbb{E}[\langle a_j, a_k\rangle] = \langle \mu_{a_j}, \mu_{a_k}\rangle \leq 1/2(\|\mu_{a_j}\|^2 + \|\mu_{a_k}\|^2)$, along with (4.44) and (7), we have

$$\frac{1}{N^2}\sum_{j=1}^N\sum_{k=1}^N \mathbb{E}[\sigma_j]\mathbb{E}[\sigma_k]\langle \mu_{a_j}, \mu_{a_k}\rangle \leq \frac{E^2(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}))2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})}{N^2}. \tag{4.47}$$

We now turn to the first term on the right hand side of (4.47). We first have for any $j = 1, \ldots, N$, using (7) that:

$$\mathbb{E}[\|a_j\|^2] = \sum_{l=1}^{\mathrm{d}_{\mathsf{Y}}} \mathbb{E}[a_{j,l}^2] = \sum_{l=1}^{\mathrm{d}_{\mathsf{Y}}} (\sigma^2_{a_{j,l}} + \mu^2_{a_{j,l}}) \leq 2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) + \mathrm{d}_{\mathsf{Y}}(2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) + 1)^2 := F(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})).$$

Then, using that σ is L-Lipschitz, Cauchy-Schwartz inequality and that since X is compact there exists c_{X} such that $\|x\| \leq c_{\mathsf{X}}$, we have:

$$\mathbb{E}[\sigma_j^2] \leq \mathbb{E}[(C_0 + L| < b_j, x > |)^2] = C_0^2 + 2C_0 c_{\mathsf{X}} L\mathbb{E}[\|b_j\|] + L^2 c^2_{\mathsf{X}}\mathbb{E}[\|b_j\|^2],$$

where, using (7) and (7),

$$\mathbb{E}[\|b_j\|] \leq \sum_{l=1}^{\mathrm{d}_{\mathsf{X}}} \mathbb{E}[|b_{j,l}|] \leq \mathrm{d}_{\mathsf{X}} D(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})),$$
$$\mathbb{E}[\|b_j\|^2] = \sum_{l=1}^{\mathrm{d}_{\mathsf{X}}} \mathbb{E}[b_{j,l}^2] = \sum_{l=1}^{\mathrm{d}_{\mathsf{X}}} (\sigma^2_{b_{j,l}} + \mu^2_{b_{j,l}}) \leq \mathrm{d}_{\mathsf{X}}(2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) + 1)^2 + 2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}).$$

Hence,

$$\mathbb{E}[\sigma_j^2] \leq C_0^2 + 2C_0 c_{\mathsf{X}} L\mathrm{d}_{\mathsf{X}} D(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})) + L^2 c^2_{\mathsf{X}} 2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}}|\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}) := G(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})),$$

with R increasing. Hence, the first term on the right hand side of (4.47) can be bounded as:

$$\frac{1}{N^2}\sum_{j=1}^N \mathbb{E}[\sigma_j^2]\mathbb{E}[\|a_j\|^2] \leq \frac{G(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}))F(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}))}{N}.$$

Finally, we obtain the desired result with:

$$\psi(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})) := G(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}))F(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})) + E^2(\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0}))\sqrt{2\mathrm{KL}(\tilde{\mathrm{q}}^N_{\boldsymbol{\theta}^*}, \tilde{\mathrm{q}}^N_{\boldsymbol{\theta}_0})}.$$

□

Lemma 37. *Let l be the cross-entropy loss, and $q_{\boldsymbol{\theta}}^N \in \mathcal{G}_\Theta$ where $\mathcal{G}_\Theta$ is a family of Gaussians with diagonal covariance matrices, i.e. for any $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta} = (\mu, \sigma) \in \mathbb{R}^{N\mathsf{d_X}} \times R^{N\mathsf{d_Y}}$. Assume that each coordinate of $\boldsymbol{\theta}$ is bounded by a constant (independent of N) and that $\lim_{N\to\infty} \|\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| = 0$ for any $x \in \mathsf{X}$. Then,*

$$\lim_{N\to\infty} \mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N) = p(\log(\mathsf{d_Y}) + \log(Z)).$$

Proof. For any $i = 1, \ldots, p$, denote

$$\begin{array}{rccc} l_{y_i}: & \mathbb{R}^{\mathsf{d_Y}} & \longrightarrow & \mathbb{R} \\ & (z_1, \ldots, z_{\mathsf{d_Y}}) & \longmapsto & -\log\left(\frac{e^{z y_i}}{\sum_{j=1}^{\mathsf{d_Y}} e^{z_j}}\right) \end{array}, \tag{4.48}$$

so that $\forall z = (z_1, \ldots z_{\mathsf{d_Y}}) \in \mathbb{R}^{\mathsf{d_Y}}$,

$$|l_{y_i}(z)| = \left| -\log(\exp(z_{y_i})) + \log\left(\sum_{k=1}^{\mathsf{d_Y}} \exp(z_k)\right)\right|. \tag{4.49}$$

By the definition of $\mathcal{L}$ and plugging $-\log(\mathsf{d_Y}) - \log(Z)$ in (4.49), we have:

$$|\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N) - p(\log(\mathsf{d_Y}) + \log(Z))| \leq \sum_{i=1}^{p} |\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}[f_{\bar{w}}(x, y_i)]| + |\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}[\log \frac{1}{\mathsf{d_Y}} \sum_{k=1}^{\mathsf{d_Y}} e^{f_{\bar{w}}(x,k)}]|,$$

where $f_{\bar{w}}(x, k)$ denotes the k-th coordinate of $f_{\bar{w}}(x) \in \mathbb{R}^{\mathsf{d_Y}}$ for $l = 1, \ldots, \mathsf{d_Y}$. The first term on the right hand side of the previous inequality converges to 0 as N goes to infinity by assumption. Hence, we can focus on the second term. For any $k = 1, \ldots, \mathsf{d_Y}$, since σ is L-Lipschitz,

$$f_{\bar{w}}(x, k) = \frac{1}{N}\sum_{j=1}^{N} \sigma(< b_j, x >) a_{j,k} \leq \frac{1}{N}\sum_{j=1}^{N} C_0 a_{j,k} + \frac{L}{N}\sum_{j=1}^{N}\sum_{l=1}^{\mathsf{d_X}} |b_{j,l}||x_l| a_{j,k}.$$

Using the previous inequality along with Jensen's inequality, we have

$$\left|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}\left[\log\left(\frac{1}{\mathsf{d_Y}}\sum_{k=1}^{\mathsf{d_Y}} e^{\frac{L}{N}\sum_{j=1}^{N}\sum_{l=1}^{\mathsf{d_X}} |b_{j,l}||x_l| a_{j,k}}\right)\right]\right| \leq \left|\log \frac{1}{\mathsf{d_Y}}\sum_{k=1}^{\mathsf{d_Y}}\prod_{j=1}^{N}\prod_{l=1}^{\mathsf{d_X}} \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}\left[e^{\frac{L|b_{j,l}||x_l| a_{j,k}}{N}}\right]\right|.$$

Since the posterior is of the form (4.31) we have for any index (j, k, l):

$$\begin{aligned} \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}\left[e^{L\frac{|b_{j,l}||x_l| a_{j,k}}{N}}\right] &\leq \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}\left[e^{L\frac{|b_{j,l}||x_l| a_{j,k}|}{N}}\right] \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,1); v\sim\mathcal{N}(0,1)}\left[e^{\frac{L|\sigma_{b_{j,l}} u + \mu_{b_{j,l}}||x||\sigma_{a_{j,k}} v + \mu_{a_{j,k}}|}{N}}\right] \\ &\leq \mathbb{E}_{u\sim\mathcal{N}(0,1); v\sim\mathcal{N}(0,1)}\left[e^{\frac{|Cb+C||x||Ca+C|}{N}}\right], \end{aligned}$$

for some constant $C > 0$ since by assumption each coordinate of the variational parameter is bounded. By the dominated convergence theorem, when N goes to infinity we have:

$$\mathbb{E}_{u\sim\mathcal{N}(0,1); v\sim\mathcal{N}(0,1)}\left[e^{\frac{L|\sigma_{b_{j,l}} u + \mu_{b_{j,l}}||x|(\sigma_{a_{j,k}} v + \mu_{a_{j,k}})}{N}}\right] = 1 + o\left(\frac{1}{N}\right).$$

Hence,

$$\left|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}\left[\log\frac{1}{\mathrm{d}_{\mathsf{Y}}}\sum_{k=1}^{\mathrm{d}_{\mathsf{Y}}}e^{\frac{L}{N}\sum_{j=1}^{N}\sum_{l=1}^{\mathrm{d}_{\mathsf{X}}}|b_{j,l}||x_l|a_{j,k}}\right]\right| \leq N\mathrm{d}_{\mathsf{X}}\log\left(1+o\left(\frac{1}{N}\right)\right)$$

Similarly, we can prove that:

$$\lim_{N\to\infty}\left|\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N}\left[\log\frac{1}{\mathrm{d}_{\mathsf{Y}}}\sum_{k=1}^{\mathrm{d}_{\mathsf{Y}}}e^{\frac{\sigma(0)}{N}\sum_{j=1}^{N}a_{j,k}}\right]\right| = 0$$

Finally, we have:

$$\lim_{N\to\infty}|\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^N) - p(\log(\mathrm{d}_{\mathsf{Y}})+\log(Z))| \leq \lim_{N\to\infty}\left|N\mathrm{d}_{\mathsf{X}}\log\left(1+o\left(\frac{1}{N}\right)\right)\right| = 0.$$

□

Proof of Proposition 33

We will first need the following technical result.

Lemma 38. *Assume Assumption 1-(i) and Assumption 2. Then for any $x\in\mathsf{X}$, the function $\theta\mapsto\tilde{\phi}(\theta,x)$ is continuous. In addition, there exists $C\geq 0$ such that for any $x\in\mathsf{X}$ and $\theta\in\Xi$, $\|\tilde{\phi}(\theta,x)\|\leq C$.*

Proof. Since $\phi(\theta,z,x)=ah(b,x)$ and since by Assumption 2, $b\mapsto h(b,x)$ is continuous for any $x\in\mathsf{X}$, it follows that for any $x\in\mathsf{X}$, $z\in\mathbb{R}^d$, $\theta\mapsto\phi(\theta,z,x)$ is continuous on Ξ. Using (4.52) and the condition that Ξ is compact, an application of the Lebesgue dominated convergence theorem implies that for any $x\in\mathsf{X}$, the function $\theta\mapsto\tilde{\phi}(\theta,x)$ is continuous. Finally, Eq. (4.52) and the condition that Ξ is compact shows that there exists $C\geq 0$ such that for any $x\in\mathsf{X}$ and $\theta\in\Xi$, $\|\tilde{\phi}(\theta,x)\|\leq C$. □

We now prove Proposition 33. We first prove Assumption 1-(ii). Recall, that for $\theta,z,x\in\Xi\times\mathbb{R}^d\times\mathsf{X}$, $\phi(\theta,z,x)=s(\mathscr{T}_\theta(z),x)$ where $\mathscr{T}_\theta(z)=\mu+\sigma\odot z$. Therefore, by (4.3), decomposing each weight as $w=(a,b)$ where a is the output weight and b is the hidden weight, $\phi(\theta,z,x)=ah(b,x)$, with $a=\mu_a+\sigma_a\odot z_a$ and $b=\mu_b+\sigma_b\odot z_b$, $\theta=(\theta_a,\theta_b)$, $\theta_a=(\mu_a,\sigma_a)\in\mathbb{R}^{\mathrm{d}_{\mathsf{Y}}}\times(\mathbb{R}_+^*)^{\mathrm{d}_{\mathsf{Y}}}$ and $\theta_b=(\mu_b,\sigma_b)\in\mathbb{R}^{\mathrm{d}_{\mathsf{X}}}\times(\mathbb{R}_+^*)^{\mathrm{d}_{\mathsf{X}}}$. Hence,

$$\|a\|^2\leq 2\|\mu_a\|^2+2\|\sigma_a\|^2\|z_a\|^2\leq 2\|\theta\|^2(1+\|z_a\|^2)\,, \tag{4.50}$$

$$\|b\|^2\leq 2\|\theta\|^2(1+\|z_b\|^2)\,. \tag{4.51}$$

Also, by Assumption 2, there exist $C_0,C_1\geq 0$ such that for any x,b, $|h(b,x)|\leq C_0+C_1\|b\|$. Hence, we have for any $\theta\in\Xi$, $z\in\mathbb{R}^d$ and $x\in\mathsf{X}$,

$$\begin{aligned}\|\phi(\theta,z,x)\|^2 &\leq \|a\|^2(C_0+C_1\|b\|)^2\\ &\leq 2\|\theta\|^2(1+\|z_a\|^2)[C_0+2C_1\|\theta\|(1+\|z_b\|)^{\frac{1}{2}}]^2\\ &\leq C_3(1+\|z\|^4)(1+\|\theta\|^2)\,,\end{aligned} \tag{4.52}$$

for some constant $C_3>0$. As Ξ is compact and $\int\|z\|^4\mathrm{d}\gamma(z)<+\infty$, it follows that Assumption 1-(ii) holds. We now show that $\mathrm{argmax}_{\mathcal{P}(\Xi)}\tilde{\mathrm{F}}_\eta^N\neq\emptyset$. By Proposition 38 in the supplement, $\tilde{\phi}$ is bounded and for any $x\in\mathsf{X}$, $\theta\mapsto\tilde{\phi}(\theta,x)$ is continuous. Using that under Assumption 2 for any $y\in\mathsf{Y}$, $\tilde{y}\mapsto\ell(y,\tilde{y})$ is continuous, it follows that $\nu\mapsto\tilde{\mathrm{G}}(\nu;(x,y))$ is continuous for the weak topology on $\mathcal{P}(\Xi)$ for any $(x,y)\in\mathsf{X}\times\mathsf{Y}$. In addition, since $\theta\mapsto\mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta|\mathrm{p}_0)$ is continuous, we get since Ξ is compact that $\nu\mapsto\int\mathrm{KL}(\tilde{\mathrm{q}}^1{}_\theta|\mathrm{p}_0{}^1)\mathrm{d}\nu(\theta)$ is continuous for the weak topology. It follows that $\nu\mapsto\tilde{\mathrm{F}}_\eta^N(\nu)$ is continuous for the weak topology. Using Ξ is compact, $\mathcal{P}(\Xi)$ is compact for the weak topology by [Ambrosio et al., 2008, Theorem 5.1.3], and it follows that $\mathrm{argmax}_{\mathcal{P}(\Xi)}\tilde{\mathrm{F}}_\eta^N\neq\emptyset$. The last condition (4.20) of Theorem 30 easily follows from Proposition 38.

Proof of Proposition 31

We will prove the result by contradiction. Let $\boldsymbol{\theta}^{*,N} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta})$. Assume that $\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}, \mathrm{p}_0) \to 0$ as $N \to \infty$.

Recall that the prior writes as (4.30). Since the posterior $\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}} \in \mathcal{G}_\Theta$ writes:

$$\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}(\bar{w}) = \prod_{j=1}^{N}\prod_{l=1}^{\mathrm{d_Y}} \mathcal{N}(a_{j,l}; \mu^*_{a_{j,l}}, \sigma^{*2}_{a_{j,l}}) \times \prod_{j=1}^{N}\prod_{r=1}^{\mathrm{d_X}} \mathcal{N}(b_{j,k}; \mu^*_{b_{j,r}}, \sigma^{*2}_{b_{j,r}}), \tag{4.53}$$

assuming $\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}, \mathrm{p}_0) \to 0$ as $N \to \infty$ is equivalent to assuming that for all $j = 1, \dots, N$, $l = 1, \dots, \mathrm{d_Y}$, $r = 1, \dots \mathrm{d_X}$,

$$\lim_{N\to+\infty} \mu^*_{a_{j,l}} = 0, \ , \lim_{N\to+\infty} \mu_{b^*_{j,r}} = 0, \ \lim_{N\to+\infty} \sigma^*_{a_{j,l}} = 1, \ \lim_{N\to+\infty} \sigma^*_{b_{j,r}} = 1. \tag{4.54}$$

We consider a small perturbation of the optimal variational distribution. More precisely, we consider the parameter $\tilde{\boldsymbol{\theta}}^{*,N}$ defined by:

- for all $j = 1, \dots, N$, $r = 1, \dots, \mathrm{d_X}$, $\tilde{\mu}_{b_{j,r}} = \mu^\star_{b_{j,r}}$
- for all $j = 1, \dots, N$, $r = 1, \dots, \mathrm{d_X}$, $\tilde{\sigma}_{b_{j,r}} = \sigma^\star_{b_{j,r}}$
- for all $j = 1, \dots, N$, $l = 1, \dots, \mathrm{d_Y}$, if $(j,l) \neq (k,m)$, $\tilde{\mu}_{a_{j,l}} = \mu^\star_{a_{j,l}}$
- for all $j = 1, \dots, N$, $l = 1, \dots, \mathrm{d_Y}$, $\sigma_{a_{j,l}} = \sigma^\star_{a_{j,l}}$
- $\tilde{\mu}_{a_{k,m}} = \mu^\star_{a_{k,m}} + \epsilon$

for some $\epsilon > 0$ and $k \in \{1, \dots, N\}, m \in \{1, \dots, \mathrm{d_Y}\}$.

Recall that

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = -\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}) - \eta \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}} | q_{\boldsymbol{\theta}_0}), \quad \text{with } \mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}) = -\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}}}\left[\sum_{i=1}^{p} \log(\mathrm{L}(y_i|x_i,\bar{w}))\right], \tag{4.55}$$

where $\mathrm{L}(y|x,\bar{w}) \propto \exp(-\ell(f_{\bar{w}}(x), y))$ is defined by (4.4). By the optimality of $\boldsymbol{\theta}^{*,N}$, we have $\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}^{*,N}) \geq \mathrm{ELBO}_\eta^N(\tilde{\boldsymbol{\theta}}^{*,N})$, i.e.

$$\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}) + \eta \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}} | \mathrm{p}_0) \leq \mathcal{L}(\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}) + \eta \mathrm{KL}(\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}} | \mathrm{p}_0) \tag{4.56}$$

which results in

$$\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}) - \mathcal{L}(\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}) \leq \eta\left(\mathrm{KL}(\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}} | \mathrm{p}_0) - \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}} | \mathrm{p}_0)\right) = \frac{\eta}{2}\left(\epsilon \mu^*_{a_{k,m}} + \frac{\epsilon^2}{2}\right) \tag{4.57}$$

where the equality on the right-hand side follows from the construction of $\tilde{\boldsymbol{\theta}}^{*,N}$ w.r.t. $\boldsymbol{\theta}^{*,N}$ and the formula of KL between Gaussians (7).

Now, for the square loss, by (4.32) we have, denoting $\sigma_j = \sigma(\langle b_j, x_i \rangle)$ (we mask the dependence in the data index i for lighter notations) for any $j = 1, \dots, N$:

$$\begin{aligned}
\mathcal{L}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}) - \mathcal{L}(\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}) &= \sum_{i=1}^{p} \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[\|f_{\bar{w}}(x_i)\|^2\right] - \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\|f_{\bar{w}}(x_i)\|^2\right] \\
&\quad - 2\left\langle y_i, \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[f_{\bar{w}}(x_i)\right] - \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[f_{\bar{w}}(x_i)\right]\right\rangle \\
&= \frac{1}{2N}\sum_{i=1}^{p}\sum_{l=1}^{d_Y}\left(\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[\sum_{j=1}^{N}\sigma_j^2 a_{j,l}^2\right] - \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\sum_{j=1}^{N}\sigma_j^2 a_{j,l}^2\right]\right) \\
&+ \frac{1}{2N^2}\sum_{i=1}^{p}\sum_{l=1}^{d_Y}\sum_{j=1}^{N}\sum_{s=1,s\neq i}^{N}\left(\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_j a_{j,l}\sigma_s a_{s,l}\right] - \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\sigma_j a_{j,l}\sigma_s a_{s,l}\right]\right) \\
&+ 2\sum_{i=1}^{p}\left\langle y_i, \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}[\sigma_j]\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}[a_j] - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[\sigma_j]\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[a_j]\right\rangle \\
&:= A_N + B_N + C_N,
\end{aligned}$$

First, since the difference is null for $l \neq m$ and $j \neq k$, we have:

$$A_N = \frac{1}{2N}\sum_{i=1}^{p}\left(\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_k^2 a_{k,m}^2\right] - \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\sigma_k^2 a_{k,m}^2\right]\right) \tag{4.58}$$

$$= \frac{1}{2N}\sum_{i=1}^{p}\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_k^2\right]\left(\mu_{a_{k,m}}^{\star 2} + \sigma_{a_{k,m}}^{\star 2} - (\mu_{a_{k,m}}^{\star} + \epsilon)^2 - \sigma_{a_{k,m}}^{\star 2}\right) \tag{4.59}$$

$$= \frac{1}{2N}\sum_{i=1}^{p}\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_k^2\right](-2\mu_{a_{k,m}}^{\star}\epsilon - \epsilon^2). \tag{4.60}$$

Then, let's define $\Delta_{j,s,l} = \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[\sigma_j a_{j,l}\sigma_s a_{s,l}] - \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\tilde{\boldsymbol{\theta}}^{*,N}}}[\sigma_j a_{j,l}\sigma_s a_{s,l}])$. Firstly, if $l \neq m$, then $\Delta_{j,s,l} = 0$ for any $j, s \in \{1, \dots, N\}$. Now, if $l = m$, there are 3 different combinations for the indexes j and s:

- If $j \neq k$ and $s \neq k$, then $\Delta_{j,s,m}$ is also null.
- If $j \neq k$ and $s = k$ then $\Delta_{j,k,m} = \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[\sigma_j a_{j,m}\sigma_k](\mu_{a_{k,m}}^{\star} - (\mu_{a_{k,m}}^{\star} + \epsilon))$
- if $j = k$ and $s \neq k$ then $\Delta_{k,s,m} = \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[\sigma_k a_{s,m}\sigma_s](\mu_{a_{k,m}}^{\star} - (\mu_{a_{k,m}}^{\star} + \epsilon))$,

Hence

$$B_N = -\frac{1}{N^2}\sum_{i=1}^{p}\sum_{s=1,s\neq k}^{N}\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[\sigma_k a_{s,m}\sigma_s]\epsilon. \tag{4.61}$$

And for the last term, we have:

$$C_N = \frac{2}{N}\sum_{i=1}^{p} y_{i,m}\mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{*,N}}}[\sigma_k]\epsilon. \tag{4.62}$$

Since from the previous computations,

$$\frac{1}{2}\left(\epsilon\mu_{a_{k,m}}^{*} + \frac{\epsilon^2}{2}\right) \geq \frac{1}{\eta}\left(A_N + B_N + C_N\right), \tag{4.63}$$

we now study the limit of the right-hand side of the previous inequality under the tempering $\eta = \tau p/N$.

Recall that by assuming the collapse of the posterior onto the prior (4.54), since $\sigma_s = \sigma(\langle b_s, x_i \rangle)$, we have for any $s = 1, \dots, N$

$$m_i^{(1)} := \lim_{N\to\infty} \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*,N}}[\sigma_s] = \mathbb{E}_{b\sim\mathcal{N}(0,I_{\mathrm{d}_\mathsf{X}})}[\sigma(\langle b, x_i\rangle)]\,,$$

and,

$$m_i^{(2)} := \lim_{N\to\infty} \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*,N}}[\sigma_s^2] = \mathbb{E}_{b\sim\mathcal{N}(0,I_{\mathrm{d}_\mathsf{X}})}[\sigma^2(\langle b, x_i\rangle)]. \tag{4.64}$$

Hence, we have by assumption (4.54), successively:

$$\lim_{N\to\infty} \frac{N}{\tau p} A_N = \lim_{N\to\infty} \frac{1}{2\tau p} \sum_{i=1}^{p} \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*,N}}\left[\sigma_k^2\right](-2\mu^\star_{a_{k,m}}\epsilon - \epsilon^2) = -\frac{\epsilon^2}{2\tau p}\sum_{i=1}^{p} m_i^{(2)}, \tag{4.65}$$

Then, we obtain, using again assumption (4.54):

$$\lim_{N\to\infty} \frac{N}{\tau p} B_N = -\lim_{N\to\infty} \frac{1}{N\tau p} \sum_{i=1}^{p}\sum_{s=1,s\neq k}^{N} \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*,N}}[\sigma_k a_{s,m}\sigma_s] = -\lim_{N\to\infty} \frac{1}{N\tau p} \sum_{i=1}^{p}\sum_{s=1,s\neq k}^{N} (m_i^{(1)})^2 \mu^*_{a_{s,m}} = 0, \tag{4.66}$$

and finally:

$$\lim_{N\to\infty} \frac{N}{\tau p} C_N = \frac{2\epsilon}{\tau p}\sum_{i=1}^{p} y_{i,m} \lim_{N\to\infty} \mathbb{E}_{\bar{w}\sim\tilde{\mathrm{q}}_{\boldsymbol{\theta}^*,N}}[\sigma_k] = \frac{2\epsilon}{\tau p}\sum_{i=1}^{p} y_{i,m} m_i^{(1)}. \tag{4.67}$$

Hence, considering (4.63) when $N\to\infty$, we have

$$\lim_{N\to\infty}\left(\epsilon\mu^*_{a_{k,m}} + \frac{\epsilon^2}{2}\right) = \frac{\epsilon^2}{2} \geq -\frac{\epsilon^2}{\tau p}\sum_{i=1}^{p} m_i^{(2)} + \frac{4\epsilon}{\tau p}\sum_{i=1}^{p} y_{i,m} m_i^{(1)}, \tag{4.68}$$

reordering:

$$\epsilon^2\left(1 + \frac{1}{\tau p}\sum_{i=1}^{p} m_i^{(2)}\right) \geq \frac{4\epsilon}{\tau p}\sum_{i=1}^{p} y_{i,m} m_i^{(1)}. \tag{4.69}$$

Now, let $\epsilon = sign(\sum_{i=1}^{p} y_{i,m} m_i^{(1)})\xi$ for some $\xi\in\mathbb{R}^{+*}$. The previous inequality writes:

$$\xi a := \xi\left(1 + \frac{1}{\tau p}\sum_{i=1}^{p} m_i^{(2)}\right) \geq \frac{4}{\tau p}\left|\sum_{i=1}^{p} y_{i,m} m_i^{(1)}\right| := b. \tag{4.70}$$

Choosing $\xi < b/a$, we see that the previous inequality is false, contradicting the assumption (4.54).

Tempering the ELBO objective when scaling the output of a neural network by $1/\sqrt{N}$

We now rewrite the ELBO (4.6) when the scaling in $1/\sqrt{N}$ is adopted; by rewriting both the KL term and likelihood term independently. We denote $\boldsymbol{\theta}^N = (\theta_1^N, \ldots, \theta_N^N)$ the variational parameters of a neural network scaled by $1/N$ and $\boldsymbol{\theta}^{\sqrt{N}} = (\theta_1^{\sqrt{N}}, \ldots, \theta_N^{\sqrt{N}})$ if scaled by $1/\sqrt{N}$. Denoting $a'_j = a_j/\sqrt{N}, b'_j = b_j$ for any $j = 1, \ldots, N$, the output of the neural network under (4.3) can be written

$$N^{-1}\sum_{j=1}^{N} a_j\sigma(\langle b_j, x\rangle) := N^{-1/2}\sum_{j=1}^{N} a'_j\sigma(\langle b'_j, x\rangle). \tag{4.71}$$

Hence, the mean and variance of input and output weights can be written respectively

$$(\mu_{b'_j}, \sigma_{b'_j}) = (\mu_{b_j}, \sigma_{b_j}),\ (\mu_{a'_j}, \sigma_{a'_j}) = (\mu_{a_j}N^{-\frac{1}{2}}, \sigma_{a_j}N^{-\frac{1}{2}}). \tag{4.72}$$

Recall as well that for any $j = 1, \dots, N$, $\theta_j^N = (\mu_j, \sigma_j) \in \mathbb{R}^{2d}$, $\mu_j = (\mu_{b_j}, \mu_{a_j}) \in \mathbb{R}^{\mathrm{d}_X} \times \mathbb{R}^{\mathrm{d}_Y}$, $\sigma_j = (\sigma_{b_j}, \sigma_{a_j}) \in \mathbb{R}^{\mathrm{d}_X} \times \mathbb{R}^{\mathrm{d}_Y}$. Similarly, $\theta_j^{\sqrt{N}}$ correspond to the parameters b'_j, a'_j for $j = 1, \dots, N$. Hence, using the formula for KL divergence between Gaussians, denoting $\tilde{\mathrm{q}}_{\boldsymbol{\theta}}^{\sqrt{N}}$ the distribution over weights (a', b') we have

$$\mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}^{\sqrt{N}}} | \mathrm{p}_0) = \frac{1}{2} \sum_{j=1}^{N} \Bigg[\sum_{l=1}^{\mathrm{d}_X} (\mu_{b_{j,l}}^2 + \sigma_{b_j,l}^2 - \log\Big(\sigma_{b_j,l}^2\Big)) + \sum_{l=1}^{\mathrm{d}_Y} \left(\frac{\mu_{a_{j,l}}^2}{N} + \frac{\sigma_{a_j,l}^2}{N} - \log\left(\frac{\sigma_{a_j,l}^2}{N}\right)\right) - \mathrm{d}_X - \mathrm{d}_Y \Bigg] := \alpha_N + \beta_N + C_N, \quad (4.73)$$

where C_N is a constant (i.e. does not depend on the variational parameters) depending on N. We can notice than in contrast with the $1/N$ scaling, now the two terms (α_N, β_N) in the KL term, corresponding to the input and output weights, are unbalanced. We propose to rescale the latter as $\eta_N^1 \alpha_N + \beta_N$. We now turn to the data fitting term of the ELBO. Notice first that

$$\phi(\theta_j^{\sqrt{N}}, z, x_i) = a'_j h(b'_j, x) = \frac{a_j}{\sqrt{N}} h(b_j, x) = \frac{\phi(\theta_j^N, z, x_i)}{\sqrt{N}}. \quad (4.74)$$

Hence we have for the likelihood term:

$$\sum_{i=1}^{p} \int_{\mathbb{R}^{N \times d}} \log \mathrm{L}(y_i | x_i, \bar{w}) q_{\boldsymbol{\theta}^{\sqrt{N}}}(\bar{w}) \mathrm{dLeb}_{d \times N}(\bar{w}) = \sum_{i=1}^{p} \int \ell\left(y, \sum_{j=1}^{N} \frac{\phi(\theta_j^{\sqrt{N}}, z_j, x)}{N}\right) \gamma^{\otimes N}(\mathrm{d}\boldsymbol{z}) \,, \quad (4.75)$$

$$= \sum_{i=1}^{p} \int \ell\left(y, \sum_{j=1}^{N} \frac{\phi(\theta_j^{N}, z_j, x)}{N}\right) \gamma^{\otimes N}(\mathrm{d}\boldsymbol{z}) \quad (4.76)$$

$$= \sum_{i=1}^{p} \mathrm{G}_\Theta^N(\boldsymbol{\theta}^N; (x, y)). \quad (4.77)$$

In the end, we define:

$$\mathrm{ELBO}_{\eta_N^1}^N(\boldsymbol{\theta}^{\sqrt{N}}) = -\eta_N^1 \alpha_N - \beta_N - \sum_{i=1}^{p} \mathrm{G}_\Theta^N(\boldsymbol{\theta}^{\sqrt{N}}; (x_i, y_i)), \quad (4.78)$$

which is balanced by choosing $\eta_N^1 = \frac{\tau p}{N}$.

Additional Experiments

About the posterior variance collapse in the mean field regime Here we discuss how the variance of the variational posterior behaves when optimizing the balanced ELBO as proposed in Subsection 3. For the square loss and since the distribution of each neuron is independent, the data-fitting term writes:

$$\mathrm{G}_\Theta^N(\boldsymbol{\theta}; (x, y)) = C - \frac{2y}{N} \mathbb{E}_{z \sim \gamma^{\otimes N}} \Big[\sum_{j=1}^{N} a_j \sigma(\langle b_j, x \rangle) \Big] + \frac{1}{N^2} \mathbb{E}_{z \sim \gamma^{\otimes N}} \Big[\sum_{j=1}^{N} \|a_j\|^2 \sigma^2(\langle b_j, x \rangle) \Big]$$

$$= C - \frac{2y}{N} \mathbb{E} \Big[\sum_{j=1}^{N} (\mu_{a_j} + \sigma_{a_j} \odot z_{a_j}) \sigma(\Big\langle \mu_{b_j} + \sigma_{b_j} \odot z_{b_j}, x \Big\rangle) \Big]$$

$$+ \frac{1}{N^2} \mathbb{E} \Big[\sum_{j=1}^{N} \|\mu_{a_j} + \sigma_{a_j} \odot z_{a_j}\|^2 \sigma^2(\Big\langle \mu_{b_j} + \sigma_{b_j} \odot z_{b_j}, x \Big\rangle) \Big]$$

where the z_{a_j}, z_{b_j} are independent for any $j = 1, \dots, N$. Consider σ odd, e.g. σ is the identity function. Optimizing over the variance for the variational posterior σ, we have that the σ minimizing $\mathrm{G}_\Theta^N(\boldsymbol{\theta}; (x, y))$ minimizes the last

term above, i.e. $N^{-1}\sigma^2$. Hence, adding the KL to a standard Gaussian prior, the variational posterior variance σ minimizing the (rescaled) ELBO minimizes:

$$\frac{\sigma^2}{N} + (\sigma^2 - \log(\sigma^2)). \tag{4.79}$$

As $N \to \infty$, the first term becomes negligible, and the optimal variance σ collapses to the one of the prior. Moreover, we always have that as $N \to \infty$, the variance of the prior collapses to 0. Indeed, when choosing standard Gaussians as priors, the prior variance is equal to $\frac{1}{N^2}\sum_{j=1}^{N} 1 = 1/N \to 0$. However, when σ is non odd (e.g. RelU), the latter phenomenon - the posterior variance collapse onto zero - does not happen, as shown in the following experiment.

We consider a toy regression $\{(x_i, y_i)\}_{i=1}^{p}$ where $x_i \sim \mathrm{U}[-1,1]$, $y_i = x_i^3 + \epsilon$ and $\epsilon \sim \mathcal{N}(0, 0.001)$. The following dataset is represented in Figure 4.4. The models considered are one hidden layer Bayesian Neural Network with N number of neurons and a non odd activation (ReLU) trained by variational inference with a tempered ELBO ($\eta_N = \tau p/N$). We consider 9 possible values for N ($10^1, 2\times10^1, 10^2, 2\times10^2, 10^3, 2\times10^3, 10^4, 2\times10^4, 10^5$). And we study the dynamic, with respect to N, of the predictive distribution variance, ie, $\mathbb{V}_{w\sim q_\theta^N}[f_w(x)]$. Figure 4.4 represents the expected predictive distribution variance, ie, $\mathbb{E}_{x\sim U[-1,1]}[\mathbb{V}_{w\sim q_\theta}[f_w(x)]]$ approximated with Monte Carlo after the training. We cannot see any pattern demonstrating that the variance of the BNN's output tends to zero as N grows. Hence, when the ELBO is well tempered and the activation is non odd then the predictive distribution variance over the prior collapse to zero, but not necessary the one over the variational posterior .

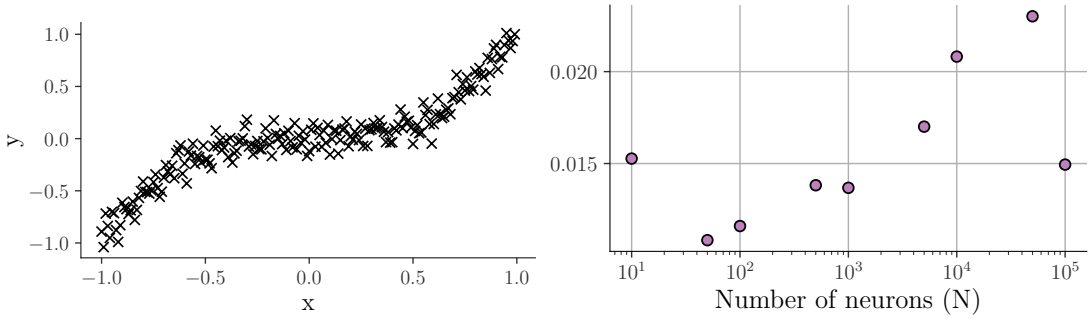


Fig. 4.4 Expected standard deviation of the prediction for different sizes of model

Balanced ELBO with cooling We first support with a very simple experiment the theoretical results of Subsection 3 and the relevance of the form of the parameter $\eta = \tau p/N$ we find. This experiment does not require training, since the goal here is to illustrate how introducing this parameter allows to balance the contributions of the two terms in the decomposition of ELBO_η^N in (4.6). We choose the architecture of a one hidden layer neural network with RelU activation functions, to which we will refer to as *Linear BNN*. We consider a regression task on the Boston dataset and a classification task on MNIST. We choose a zero-mean Gaussian prior with variance $1/5$ for each neuron. Also, we initialize the variational parameters $\theta = (\mu, \sigma)$ where μ is close to zero and $\sigma = 10^{-3}$. Figures 4.5 and 4.6 illustrate the ratio between the likelihood and KL terms in ELBO_η^N when the number of weights grows, for $\eta = 1$ (no cooling), $\eta = \tau p/N$ and different values of the hyperparameter τ, on the MNIST and BOSTON datasets respectively. They confirm that when the number of data points p is fixed and ELBO_η^N is not rescaled, one of the two terms become dominant contrary to the case where we set $\eta = \tau p/N$.

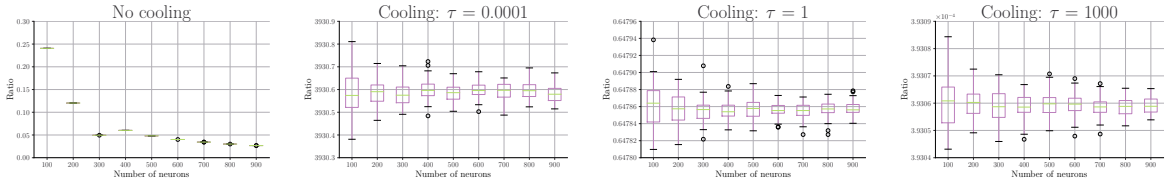


Fig. 4.5 Ratio of the two ELBO^N terms, for a Linear BNN (non trained) on MNIST.

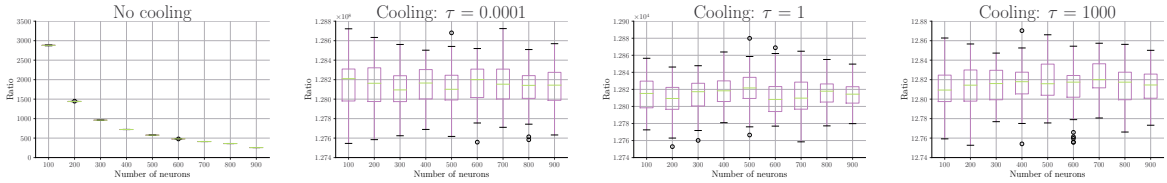


Fig. 4.6 Ratio of the two ELBO^N terms, for a Linear BNN (non trained) on BOSTON.

ECE definition For any input x, define $\mathrm{conf}(x) = \max_{c\in\{1,\dots,n_l\}} \Psi_c(f_{\bar{w}}(x))$, i.e., the maximal predicted probability of the network. This quantity can be viewed as a prediction confidence for the input x. ECE discretizes the interval $[0,1]$ into a given number of bins B and groups predictions based on the confidence score: $S_b = \{i \in \{1,\dots,p\}, \mathrm{conf}(x_i) \in [b/B, (1+b)/B[\}$. The calibration error is the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence).

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{p} |\mathrm{acc}(S_b) - \mathrm{conf}(S_b)| \,, \tag{4.80}$$

where p is the total number of data points, and $|S_b|$, $\mathrm{acc}(S_b)$ and $\mathrm{conf}(S_b)$ are the number of predictions, the accuracy and confidence of bin S_b respectively.

Cooling effect on the distribution of the variational parameters Figure 4.7 illustrates the distribution of the variational parameters after training a linear BNN (i.e., single hidden layer with RelU) on MNIST. For a large τ, the distribution of the variational parameters is close to the prior (a centered Gaussian with standard deviation 0.2). For a small τ, we can see that the network has learnt values of σ that are very different from the prior (e.g., close to zero). Intermediate values of τ interpolate between the two previous regimes

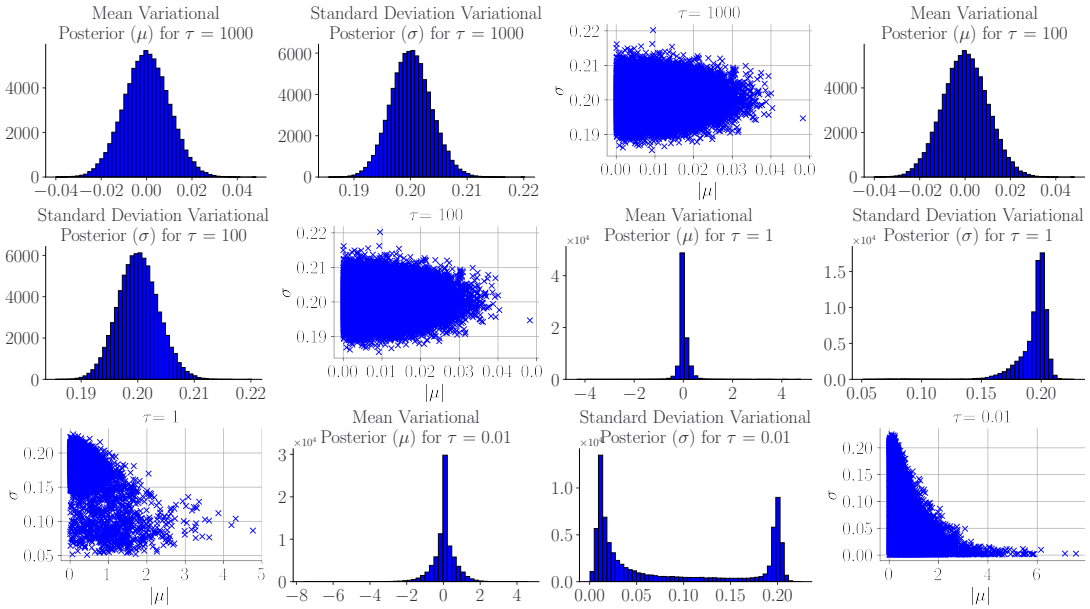


Fig. 4.7 Histograms of the variational parameters $\theta = (\mu, \sigma)$ for a Linear BNN trained on MNIST. From left to right: histogram of variational means, standard deviations, and standard deviation as a function of the norm of the mean.

OOD detection We also compare the performance on out-of-distribution of a Resnet20 trained on CIFAR-10 with Bayes by Backprop. We compute the the histogram of predictive entropies for 5000 in-distribution samples

and out-of-distribution samples. Recall that the negative entropy is defined for a vector of class probabilities $[p(y=c|x,\mathcal{D})]_{c\in\{1,\dots,n_l\}}$ as $-\sum_{c=1}^{n_l} p(y=c|x,\mathcal{D})\log(p(y=c|x,\mathcal{D}))$. The first ones correspond to samples from the test set of CIFAR-10; while the out-of-distribution samples are chosen from another image dataset, namely SVHN [Netzer et al., 2011]. Our results are to be found in Figure4.8 and illustrate again the importance of the parameter τ. When τ is very small, the model is highly confident for in-distribution samples, and has diffuse predictive entropies for out-distribution samples. As τ increases, the model starts to be less confident, resulting in higher entropies on both in-distribution and out-distribution samples, especially for the out-distribution samples. Finally if τ is too large, as the model sticks to the prior distribution, it is not confident neither on the in-distribution nor out-distribution, resulting on a spiky distribution of predictive entropies at high values.

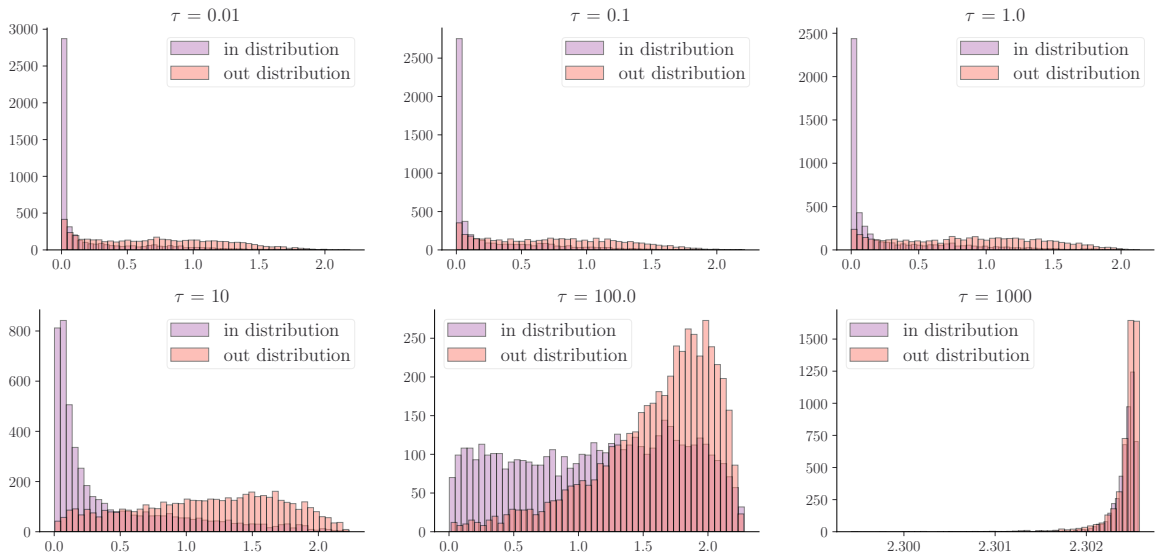


Fig. 4.8 Histogram of the predictive entropies for a Resnet20 trained on CIFAR-10, on 5000 in-distribution (from the test set of CIFAR-10 dataset) and out-of-distribution (from SVHN dataset) samples

Bayes by Backprop

Several methods have been proposed to optimize ELBO^N. A first and straightforward approach is to apply stochastic gradient descent (SGD), using samples from $q_{\boldsymbol{\theta}_c}$ where $\boldsymbol{\theta}_c$ is the current point, to obtain stochastic estimates for $\nabla_\theta \mathrm{ELBO}^N$. However, the resulting estimation of the gradient suffers from high variance. Alternative algorithms have been proposed to mitigate this effect, such as Probabilistic Backpropagation [Hernández-Lobato and Adams, 2015] or Bayes by Backprop [Blundell et al., 2015]. Given a fixed distribution $\bar{\gamma}$ and a parameterized function $g(\boldsymbol{\theta},\cdot)$, the network parameter $\bar{w}$ is obtained as $\bar{w}=g(\boldsymbol{\theta},\boldsymbol{z})$, where $\boldsymbol{z}$ is sampled from $\bar{\gamma}$, e.g., from a standard normal distribution. While a new $\boldsymbol{z}$ is sampled at each iteration, its distribution is constant, unlike that of the network parameters $\bar{w}$. As soon as $g(\boldsymbol{\theta},\cdot)$ is invertible and $\boldsymbol{\gamma}, q(\cdot|\boldsymbol{\theta})$ are non-degenerated probability distributions, we have $q(\bar{w}|\boldsymbol{\theta})d\bar{w}=\bar{\gamma}(\boldsymbol{z})d\boldsymbol{z}$ (see [Jospin et al., 2020, Appendix A]), and for any differentiable function f:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}_{\bar{w}\sim q(.|\boldsymbol{\theta})}[f(\bar{w},\boldsymbol{\theta})]=\mathbb{E}_{\boldsymbol{z}\sim\bar{\gamma}}\left[\frac{\partial f(\bar{w},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+\frac{\partial\bar{w}}{\partial\boldsymbol{\theta}}\frac{\partial f(\bar{w},\boldsymbol{\theta})}{\partial\bar{w}}\right]. \tag{4.81}$$

Algorithm 2 Bayes by Backprop

Input: step-size $\delta > 0$, number of iterations m_{iter}, number of samples $M_{samples}$.
for each m_{iter} iterations **do**
for each $m = 1, \dots, M_{samples}$ **do**
1. Sample $\boldsymbol{z} \sim \gamma^{\otimes N}$
2. Let $\bar{w} = \boldsymbol{\mu} + \log(1 + \exp(\boldsymbol{\rho})) \circ \boldsymbol{z}$.

end for
3. Compute

$$g(\bar{w}, \boldsymbol{\theta}) \approx \frac{1}{M_{samples}} \sum_{m=1}^{M_{batch}} \log q(\bar{w}_i|\boldsymbol{\theta}) - \log \mathrm{p}_0(\bar{w}_i) P(\mathcal{D}|\bar{w}_i) \tag{4.82}$$

5. Calculate the gradient with respect to the mean and standard deviation parameter ρ

$$\Delta_\mu = \frac{\partial g(w,\theta)}{\partial w} + \frac{\partial g(w,\theta)}{\partial \mu} \tag{4.83}$$

$$\Delta_\rho = \frac{\partial g(w,\theta)}{\partial w} \frac{\epsilon}{1 + \exp(\rho)} + \frac{\partial g(w,\theta)}{\partial \rho} \tag{4.84}$$

6. Update the variational parameters:

$$\mu \longleftarrow \mu - \delta \Delta_\mu \tag{4.85}$$

$$\rho \longleftarrow \rho - \delta \Delta_\rho \tag{4.86}$$

end for

Bayes by Backprop uses the previous equality to estimate the gradient of F, because $F = \mathbb{E}_{\bar{w} \sim q(\cdot|\boldsymbol{\theta})}[f(\bar{w}, \boldsymbol{\theta})]$ with $f(\bar{w}, \boldsymbol{\theta}) = \log q(\bar{w}|\boldsymbol{\theta}) - \log \mathrm{p}_0(\bar{w}) - \log P(\mathrm{d}|\bar{w})$. More specifically, it performs a stochastic gradient descent for F using a new sample $\boldsymbol{z}$ at each time step to estimate the gradient of F as the parameter $\boldsymbol{\theta}$ is updated. When the step size in this algorithm goes to zero, the Bayes by Backprop dynamics corresponds to a Wasserstein gradient flow of a particular functional defined on the space of probability distributions over $\boldsymbol{\theta}$, which we introduce in the next section.

As in [Blundell et al., 2015], we will use a variance reparameterization; $\sigma = \log(1 + \exp(\rho)) \in \mathbb{R}^+$ for $\rho \in \mathbb{R}$. Consequently, the variational parameter is given by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times 2d}$ with $\theta_j = (\mu_j, \rho_j) \in \mathbb{R}^{2d}$. We denote by $g : \mathbb{R}^{2d} \times \mathbb{R}^d \to \mathbb{R}, (\theta, z) \mapsto \mu + \log(1 + \exp(\rho)) \odot z$, where $\odot$ denotes the entry-wise multiplication and γ denotes the standard normal distribution over $\mathbb{R}^d$. The Bayes-by-backprop algorithm in this setting is summarized in Algorithm 2.

This algorithm is well suited for minibatch optimisation, when the dataset $\mathcal{D}$ is split into a partition of L subsets (minibatches) $\mathcal{D}_1, \dots, \mathcal{D}_L$. In this case [Graves, 2011] proposes to minimise a rescaled NELBO^N for each minibatch $\mathcal{D}_l$, $l = 1, \dots, L$ as

$$\mathrm{NELBO}_l^N = \frac{1}{L} \mathrm{KL}(\tilde{\mathrm{q}}_{\boldsymbol{\theta}}|\mathrm{p}_0) - \mathbb{E}_{\bar{w} \sim \tilde{\mathrm{q}}_{\boldsymbol{\theta}}}[\log P(\mathcal{D}_l|\bar{w})]. \tag{4.87}$$

CHAPTER 5

Law of Large Numbers for Bayesian two-layer Neural Network trained with Variational Inference

Chapter abstract: *We provide a rigorous analysis of training by variational inference (VI) of Bayesian neural networks in the two-layer and infinite-width case. We consider a regression problem with a regularized evidence lower bound (ELBO) which is decomposed into the expected log-likelihood of the data and the Kullback-Leibler (KL) divergence between the a priori distribution and the variational posterior. With an appropriate weighting of the KL, we prove a law of large numbers for three different training schemes: (i) the idealized case with exact estimation of a multiple Gaussian integral from the reparametrization trick, (ii) a minibatch scheme using Monte Carlo sampling, commonly known as Bayes by Backprop, and (iii) a new and computationally cheaper algorithm which we introduce as Minimal VI. An important result is that all methods converge to the same mean-field limit. Finally, we illustrate our results numerically and discuss the need for the derivation of a central limit theorem.*

1 Introduction

Deep Learning has led to a revolution in machine learning with impressive successes. However, some limitations of DL have been identified and, despite, many attempts, our understanding of DL is still limited. A long-standing problem is the assessment of predictive uncertainty: DL tends to be overconfident in its predictions [Abdar et al., 2021], which is a problem in applications such as autonomous driving [McAllister et al., 2017, Michelmore et al., 2020], medical diagnosis [Kendall and Gal, 2017, Filos et al., 2019], or finance; cf [Krzywinski and Altman, 2013, Ghahramani, 2015]. Therefore, on the one hand, analytical efforts are being made to thoroughly investigate the performance of DL; and on the other hand, many approaches have been proposed to alleviate its shortcomings. The Bayesian paradigm is an attractive way to tackle predictive uncertainty, as it provides a framework for training uncertainty-aware neural networks (NNs) (e.g. [Ghahramani, 2015, Blundell et al., 2015, Gal and Ghahramani, 2016]).

Thanks to a fully probabilistic approach, Bayesian Neural Networks (BNN) combine the impressive neural-network expressivity with the decision-theoretic approach of Bayesian inference, making them capable of providing predictive uncertainty; see [Blundell et al., 2015, Michelmore et al., 2020, McAllister et al., 2017, Filos et al., 2019]. However, Bayesian inference requires deriving the posterior distribution of the NN weights. This posterior distribution is typically not tractable. A classical approach is to sample the posterior distribution using Markov Chain Monte Carlo methods (such as Hamilton-Monte-Carlo methods). There are however long-standing difficulties, such as the proper choice of the prior and fine-tuning of the sampler. Such difficulties often become prohibitive in large-dimensional cases,[Cobb and Jalaian, 2021]. An alternative is to use variational inference, which has a long

history [Hinton and Camp, 1993, MacKay, 1995, MacKay et al., 1995]. Simpler methods that do not require exact computation of integrals over the variational posterior were then developed, e.g. first by [Graves, 2011] thanks to some approximation and then by [Blundell et al., 2015] with the *Bayes by Backprop* approach. In the latter, the posterior distribution is approximated by a parametric distribution and a generalisation of the reparametrization trick used by [Kingma and Welling, 2014] leads to an unbiased estimator of the gradient of the ELBO; see also [Gal and Ghahramani, 2016, Louizos and Welling, 2017, Khan et al., 2018]. Despite the successful application of this approach, little is known about the overparameterized limit and appropriate weighting that must be assumed to obtain a nontrivial Bayesian posterior, see [Izmailov et al., 2021]. Recently, [Huix et al., 2022] outlined the importance of balancing in ELBO the integrated log-likelihood term and the KL regularizer, to avoid both overfitting and dominance of the prior. However, a suitable limiting theory has yet to be established, as well as guarantees for the practical implementation of the stochastic gradient descent (SGD) used to estimate the parameters of the variational distribution.

Motivated by the need to provide a solid theoretical framework, asymptotic analysis of NN has gained much interest recently. The main focus has been on the gradient descent algorithm and its variants [Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018, Mei et al., 2018, Sirignano and Spiliopoulos, 2020b, Descours et al., 2022a]. In much of these works, a mean-field analysis is performed to characterize the limiting nonlinear evolution of the weights of a two-layer NN, allowing the derivation of a law of large numbers and a central limit theorem for the empirical distribution of neuron weights. A long-term goal of these works is to demonstrate convergence toward a global minimum of these limits for the mean field. Despite some progress in this direction, this is still an open and highly challenging problem; cf [Chizat and Bach, 2018, Chizat, 2022, Chizat et al., 2022]. Nevertheless, this asymptotic analysis is also of interest in its own right, as we show here in the case of variational inference for Bayesian neural networks. Indeed, based on this asymptotic analysis, we develop an efficient and new variant of the stochastic gradient descent (SGD) algorithm for variational inference in BNN that computes only the information necessary to recover the limit behavior.

Our goal, then, is to work at the intersection of analytical efforts to gain theoretical guarantees and insights and of practical methods for a workable variational inference procedure. By adapting the framework developed by [Descours et al., 2022a], we produce a rigorous asymptotic analysis of BNN trained in a variational setting for a regression task. From the limit equation analysis, we first find that a proper regularisation of the Kullback-Leibler divergence term in relation with the integrated loss leads to their right asymptotic balance. Second, we prove the asymptotic equivalence of the idealized and Bayes-by-Backprop SGD schemes, as both preserve the same core contributions to the limit. Finally, we introduce a computationally more favourable scheme, directly stemming from the effective asymptotic contributions. This scheme is the true mean-field algorithmic approach, as only deriving from non-interacting terms.

This Section is organized as follows: Subsection 2 introduces the variational inference in BNN, as well as the SGD schemes commonly considered, namely the *idealized* and *Bayes-by-backprop* variants. Then, in Subsection 3 we establish our initial result, the LLN for the *idealized SGD*. In Subsection 4 we prove the LLN for the *Bayes-by-backprop* SGD and its variants. We show that both SGD schemes have the same limit behavior. Based on an analysis of the obtained limit equation, we present in Subsection 5 the new *minimal- VI*. Finally, in Subsection 6 we illustrate our findings using numerical experiments. The proofs of the mean-field limits, which are original and quite technically demanding, are gathered in the appendix paper.

Related works. Law of Large Numbers (LLN) for mean-field interacting particle systems, have attracted a lot of attentions; see for example [Hitsuda and Mitoma, 1986, Sznitman, 1991, Fernandez and Méléard, 1997, Jourdain and Méléard, 1998, Delarue et al., 2019, Del Moral and Guionnet, 1999, Kurtz and Xiong, 2004] and references therein. The use of mean-field particle systems to analyse two-layer neural networks with random initialization have been considered in [Mei et al., 2018, 2019], which establish a LLN on the empirical measure of the weights at fixed times - we consider in this paper the trajectory convergence, i.e. the whole empirical measure process (time indexed) converges uniformly w.r.t. Skorohod topology. It enables not only to use the limiting PDE, for example to study the convergence of the weights towards the infimum of the loss function (see [Chizat and Bach, 2018] for preliminary results), but is is also crucial to establish the central limit theorem, see for example [Descours et al., 2022a]. [Rotskoff and Vanden-Eijnden, 2018] give conditions for global convergence of GD for exact mean-square loss and online stochastic gradient descent (SGD) with mini-batches increasing in size with the number of weights N. A LLN for the entire trajectory of the empirical measure is also given in [Sirignano and Spiliopoulos, 2020b]

for a standard SGD. [De Bortoli et al., 2020b] establish the propagation of chaos for SGD with different step size schemes. Compared to the existing literature dealing with the SGD empirical risk minimization in two-layer neural networks, [Descours et al., 2022a] provide the first rigorous proof of the existence of the limit PDE, and in particular its uniqueness, in the LLN.

We are interested here in deriving a LLN but for Variational Inference (VI) of two-layer Bayesian Neural Networks (BNN), where we consider a regularized version of the Evidence Lower Bound (ELBO).

2 Variational inference in BNN:Notations and common SGD schemes

2.1 Variational inference and Evidence Lower Bound

Setting. Let X and Y be subsets of $\mathbb{R}^n$ ($n \geq 1$) and $\mathbb{R}$ respectively. For $N \geq 1$ and $\boldsymbol{w} = (w^1, \dots, w^N) \in \left(\mathbb{R}^d\right)^N$, let $f^N_{\boldsymbol{w}} : \mathsf{X} \to \mathbb{R}$ be the following two-layer neural network: for $x \in \mathsf{X}$,

$$f^N_{\boldsymbol{w}}(x) := \frac{1}{N} \sum_{i=1}^{N} s(w_i, x) \in \mathbb{R},$$

where $s : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}$ is the activation function. We work in a Bayesian setting, in which we seek a distribution of the latent variable $\boldsymbol{w}$ which represents the weights of the neural network. The standard problem in Bayesian inference over complex models is that the posterior distribution is hard to sample. To tackle this problem, we consider Variational Inference, in which we consider a family of distribution $\mathcal{G}^N = \{q^N_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Xi^N\}$ (where Ξ is some parameter space) easy to sample. The objective is to find the best $q^N_{\boldsymbol{\theta}} \in \mathcal{G}^N$, the one closest in KL divergence (denoted $\mathscr{D}_{\mathrm{KL}}$) to the exact posterior. Because we cannot compute the KL, we optimize the evidence lower bound (ELBO), which is equivalent to the KL up to an additive constant.

Denoting by $\mathfrak{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ the negative log-likelihood (by an abuse of language, we call this quantity the *loss*), the ELBO (see [Blei et al., 2017]) is defined, for $\boldsymbol{\theta} \in \Xi^N$, $(x, y) \in \mathsf{X} \times \mathsf{Y}$, by

$$\mathrm{E}_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) := -\int_{\left(\mathbb{R}^d\right)^N} \mathfrak{L}(y, f^N_{\boldsymbol{w}}(x)) q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0),$$

where P^N_0 is some prior on the weights of the NN. The ELBO is decomposed into two terms: one corresponding to the Kullback-Leibler (KL) divergence between the variational density and the prior and the other to a marginal likelihood term. It was empirically found that the maximization of the ELBO function is prone to yield very poor inferences [Coker et al., 2022]. It is argued in [Coker et al., 2022] and [Huix et al., 2022] that optimizing the ELBO leads as $N \to \infty$ to the collapse of the variational posterior to the prior. [Huix et al., 2022] proposed to consider a regularized version of the ELBO, which consists in multiplying the KL term by a parameter which is scaled by the inverse of the number of neurons:

$$\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) := -\int_{\left(\mathbb{R}^d\right)^N} \mathfrak{L}(y, f^N_{\boldsymbol{w}}(x)) q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0), \tag{5.1}$$

A first objective of this paper is to show that the proposed regularization leads to a stable asymptotic behavior and the effect of both the integrated loss and Kullback-Leibler terms on the limiting behavior are balanced in the limit $N \to \infty$. The maximization of $\mathrm{E}^N_{\mathrm{lbo}}$ is carried out using SGD.

The variational family $\mathcal{G}^N$ we consider is a Gaussian family of distributions. More precisely, we assume that for any $\boldsymbol{\theta} = (\theta^1, \dots, \theta^N) \in \Xi^N$, the variational distribution $q^N_{\boldsymbol{\theta}}$ factorizes over the neurons: for all $\boldsymbol{w} = (w^1, \dots, w^N) \in \left(\mathbb{R}^d\right)^N$, $q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) = \prod_{i=1}^N q^1_{\theta^i}(w^i)$, where $\theta = (m, \rho) \in \Xi := \mathbb{R}^d \times \mathbb{R}$ and q^1_θ is the probability density function (pdf) of $\mathcal{N}(m, g(\rho)^2 I_d)$, with $g(\rho) = \log(1 + e^\rho),\ \rho \in \mathbb{R}$.

In the following, we simply write $\mathbb{R}^{d+1}$ for $\mathbb{R}^d \times \mathbb{R}$. In addition, following the reparameterisation trick of [Blundell et al., 2015], $q^1_\theta(w)\mathrm{d}w$ is the pushforward of a reference probability measure with density γ by Ψ_θ (see more precisely Assumption **A1**). In practice, γ is the pdf of $\mathcal{N}(0, \mathrm{I}_d)$ and $\Psi_\theta(z) = m + g(\rho)z$. With these notations, (5.1) writes

$$\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) = -\int_{\left(\mathbb{R}^d\right)^N} \mathfrak{L}\Big(y, \frac{1}{N} \sum_{i=1}^{N} s(\Psi_{\theta^i}(z^i), x)\Big) \gamma(z^1) \dots \gamma(z^N) \mathrm{d}z_1 \dots \mathrm{d}z_N - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0).$$

Loss function and prior distribution. In this work, we focus on the regression problem, i.e. $\mathfrak{L}$ is the Mean Square Loss: for $y_1, y_2 \in \mathbb{R}$, $\mathfrak{L}(y_1, y_2) = \frac{1}{2}|y_1 - y_2|^2$. We also introduce the function $\phi : (\theta, z, x) \in \mathbb{R}^{d+1} \times \mathbb{R}^d \times \mathsf{X} \mapsto s(\Psi_\theta(z), x)$. On the other hand, we assume that the prior distribution P_0^N write, for all$\boldsymbol{w} \in (\mathbb{R}^d)^N$, $P_0^N(\boldsymbol{w}) = \prod_{i=1}^N P_0^1(w^j)$, where $P_0^1 : \mathbb{R}^d \to \mathbb{R}_+$ is the pdf of $\mathcal{N}(m_0, \sigma_0^2 I_d)$, and $\sigma_0 > 0$. Therefore $\mathscr{D}_{\mathrm{KL}}(q_{\boldsymbol{\theta}}^N | P_0^N) = \sum_{i=1}^N \mathscr{D}_{\mathrm{KL}}(q_{\theta^i} | P_0^1)$ and, for $\theta = (m, \rho) \in \mathbb{R}^{d+1}$,

$$\mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) = \int_{\mathbb{R}^d} q_\theta^1(x) \log\big(q_\theta^1(x)/P_0^1(x)\big) \mathrm{d}x = \frac{\|m - m_0\|_2^2}{2\sigma_0^2} + \frac{d}{2}\Big(\frac{g(\rho)^2}{\sigma_0^2} - 1\Big) + \frac{d}{2}\log\Big(\frac{\sigma_0^2}{g(\rho)^2}\Big).$$

Note that $\mathscr{D}_{\mathrm{KL}}$ has at most a quadratic growth in m and ρ.

Note that we assume here a Gaussian prior to get an explicit expression of the Kullback-Leibler divergence. Most arguments extend to sufficiently regular densities and are essentially the same for exponential families, using conjugate families for the variational approximation.

2.2 Common SGD schemes in backpropagation in a variational setting

Idealized SGD. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Consider a data set $\{(x_k, y_k)\}_{k \geq 0}$ i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$, the space of probability measures over $\mathsf{X} \times \mathsf{Y}$. For $N \geq 1$ and given a learning rate $\eta > 0$, the maximization of $\theta \in \mathbb{R}^{d+1} \mapsto \mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}, x, y)$ with a SGD algorithm writes as follows: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\left\{ \begin{array}{l} \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} \mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}_k, x_k, y_k) \\ \boldsymbol{\theta}_0 \sim \mu_0^{\otimes N}, \end{array} \right. \tag{5.2}$$

where $\mu_0 \in \mathcal{P}(\mathbb{R}^{d+1})$ and $\boldsymbol{\theta}_k = (\theta_k^1, \dots, \theta_k^N)$. We now compute $\nabla_{\boldsymbol{\theta}} \mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}, x, y)$.

First, under regularity assumptions on the function ϕ (which will be formulated later, see **A1** and **A3** below) and by assumption on $\mathfrak{L}$, we have for all $i \in \{1, \dots, N\}$ and all $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

$$\begin{aligned} &\int_{(\mathbb{R}^d)^N} \nabla_{\theta^i} \mathfrak{L}\Big(y, \frac{1}{N}\sum_{j=1}^N \phi(\theta^j, z^j, x)\Big) \gamma(z^1) \dots \gamma(z^N) \mathrm{d}z^1 \dots \mathrm{d}z^N \\ &= -\frac{1}{N^2} \sum_{j=1}^N \int_{(\mathbb{R}^d)^N} (y - \phi(\theta^j, z^j, x)) \nabla_\theta \phi(\theta^i, z^i, x) \gamma(z^1) \dots \gamma(z^N) \mathrm{d}z^1 \dots \mathrm{d}z^N \\ &= -\frac{1}{N^2} \Big[\sum_{j=1, j \neq i}^N (y - \langle \phi(\theta^j, \cdot, x), \gamma \rangle) \langle \nabla_\theta \phi(\theta^i, \cdot, x), \gamma \rangle + \langle (y - \phi(\theta^i, \cdot, x)) \nabla_\theta \phi(\theta^i, \cdot, x), \gamma \rangle \Big], \end{aligned} \tag{5.3}$$

where we have used the notation $\langle U, \nu \rangle = \int_{\mathbb{R}^q} U(z) \nu(\mathrm{d}z)$ for any integrable function $U : \mathbb{R}^q \to \mathbb{R}$ w.r.t. a measure ν (with a slight abuse of notation, we denote by γ the measure $\gamma(z)\mathrm{d}z$). Second, for $\theta \in \mathbb{R}^{d+1}$, we have

$$\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) = \begin{pmatrix} \nabla_m \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) \\ \partial_\rho \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_0^2}(m - m_0) \\ \frac{d}{\sigma_0^2} g'(\rho) g(\rho) - d \frac{g'(\rho)}{g(\rho)} \end{pmatrix}. \tag{5.4}$$

In conclusion, the SGD (5.2) writes: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\left\{ \begin{array}{l} \theta_{k+1}^i = \theta_k^i \frac{\eta}{N^2} \sum_{j=1, j \neq i}^N \Big(\langle \phi(\theta_k^j, \cdot, x_k), \gamma \rangle - y_k \Big) \langle \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \rangle \\ \qquad \frac{\eta}{N^2} \Big\langle (\phi(\theta_k^i, \cdot, x_k) - y_k) \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \Big\rangle - \frac{\eta}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \\ \theta_0^i \sim \mu_0. \end{array} \right. \tag{5.5}$$

We shall call this algorithm *idealised* SGD because it contains an intractable term given by the integral w.r.t. γ. This has motivated the development of methods where this integral is replaced by an unbiased Monte Carlo estimator (see [Blundell et al., 2015]) as detailed below.

***Bayes-by-Backprop* SGD**. The second SGD algorithm we study is based on an approximation, for $i \in \{1,\dots,N\}$, of $\int_{(\mathbb{R}^d)^N}(y-\phi(\theta^j,z^j,x))\nabla_\theta\phi(\theta^i,z^i,x)\gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N$ (see (5.3)) by

$$\frac{1}{B}\sum_{\ell=1}^{B}\big(y-\phi(\theta^j,\mathsf{Z}^{j,\ell},x)\big)\nabla_\theta\phi(\theta^i,\mathsf{Z}^{i,\ell},x) \tag{5.6}$$

where $B \in \mathbf{N}^*$ is a fixed integer and $(\mathsf{Z}^{q,\ell}, q \in \{i,j\}, 1 \le \ell \le B)$ is a i.i.d finite sequence of random variables distributed according to $\gamma(z)\mathrm{d}z$. In this case, for $N \ge 1$, given a dataset $(x_k,y_k)_{k\ge 0}$, the maximization of $\theta \in \mathbb{R}^{d+1} \mapsto \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta},x,y)$ with a SGD algorithm is the following: for $k \ge 0$ and $i \in \{1,\dots,N\}$,

$$\begin{cases} \theta^i_{k+1} = \theta^i_k - \frac{\eta}{N^2B}\sum_{j=1}^N\sum_{\ell=1}^B\big(\phi(\theta^j_k,\mathsf{Z}^{j,\ell}_k,x_k)-y_k\big)\nabla_\theta\phi(\theta^i_k,\mathsf{Z}^{i,\ell}_k,x_k) - \frac{\eta}{N}\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k}|P^1_0) \\ \theta^i_0 = (m^i_0,\rho^i_0) \sim \mu_0, \end{cases} \tag{5.7}$$

where $\eta > 0$ and $(\mathsf{Z}^{j,\ell}_k, 1 \le j \le N, 1 \le \ell \le B, k \ge 0)$ is a i.i.d sequence of random variables distributed according to γ.

3 Law of large numbers for the idealized SGD

Assumptions and notations. When E is a metric space and $\mathscr{F} = \mathbb{R}_+$ or $\mathscr{F} = [0,T]$ ($T \ge 0$), we denote by $\mathcal{D}(\mathscr{F},E)$ the Skorohod space of càdlàg functions on $\mathscr{F}$ taking values in E and $\mathcal{C}(\mathscr{F},E)$ the space of continuous functions on $\mathscr{F}$ taking values in E. The evolution of the parameters $(\{\theta^i_k, i=1,\dots,N\})_{k\ge 1}$ defined by (5.5) is tracked through their empirical distribution ν^N_k (for $k \ge 0$) and its scaled version μ^N_t (for $t \in \mathbb{R}_+$), which are defined as follows:

$$\nu^N_k := \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_k} \ \text{ and } \ \mu^N_t := \nu^N_{\lfloor Nt\rfloor}, \text{ where the } \theta^i_k\text{'s are defined (5.5).} \tag{5.8}$$

Fix $T > 0$. For all $N \ge 1$, $\mu^N := \{\mu^N_t, t \in [0,T]\}$ is a random element of $\mathcal{D}([0,T],\mathcal{P}(\mathbb{R}^{d+1}))$, where $\mathcal{P}(\mathbb{R}^{d+1})$ is endowed with the weak convergence topology. For $N \ge 1$ and $k \ge 1$, we introduce the following $\boldsymbol{\sigma}$-algebras:

$$\mathcal{F}^N_0 = \boldsymbol{\sigma}(\theta^i_0, 1 \le i \le N) \ \text{ and } \ \mathcal{F}^N_k = \boldsymbol{\sigma}(\theta^i_0,(x_q,y_q),1 \le i \le N, 0 \le q \le k-1). \tag{5.9}$$

Recall $q^1_\theta : \mathbb{R}^d \to \mathbb{R}_+$ be the pdf of $\mathcal{N}(m,g(\rho)^2\mathrm{I}_d)$ ($\theta = (m,\rho) \in \mathbb{R}^{d+1}$). In this work, we assume the following.

A1. There exists a pdf $\gamma : \mathbb{R}^d \to \mathbb{R}_+$ such that for all $\theta \in \mathbb{R}^{d+1}$, $q^1_\theta \mathrm{d}x = \Psi_\theta \# \gamma \mathrm{d}x$, where $\{\Psi_\theta, \theta \in \mathbb{R}^{d+1}\}$ is a family of $\mathcal{C}^1$-diffeomorphisms over $\mathbb{R}^d$ such that for all $z \in \mathbb{R}^d$, $\theta \in \mathbb{R}^{d+1} \mapsto \Psi_\theta(z)$ is of class $\mathcal{C}^\infty$. Finally, there exists $\mathfrak{b} : \mathbb{R}^d \to \mathbb{R}_+$ such that for all multi-index $\alpha \in \mathbf{N}^{d+1}$ with $|\alpha| \ge 1$, there exists $C_\alpha > 0$, for all $z \in \mathbb{R}^d$ and $\theta = (\theta_1,\dots,\theta_{d+1}) \in \mathbb{R}^{d+1}$,

$$\big|\partial_\alpha \Psi_\theta(z)\big| \le C_\alpha \mathfrak{b}(z) \quad \text{with for all } q \ge 1,\ \langle \mathfrak{b}^q, \gamma\rangle < +\infty, \tag{5.10}$$

where $\partial_\alpha = \partial^{\alpha_1}_{\theta_1}\dots\partial^{\alpha_{d+1}}_{\theta_{d+1}}$ and $\partial^{\alpha_j}_{\theta_j}$ is the partial derivatives of order α_j w.r.t. to θ_j.

A2. The sequence $\{(x_k,y_k)\}_{k\ge 0}$ is i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$. The set $\mathsf{X}\times\mathsf{Y} \subset \mathbb{R}^d \times \mathbb{R}$ is compact. For all $k \ge 0$, $(x_k,y_k) \perp\!\!\!\perp \mathcal{F}^N_k$, where $\mathcal{F}^N_k$ is defined in (5.9).

A3. The activation function $s : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}$ belongs to $\mathcal{C}^\infty_b(\mathbb{R}^d \times \mathsf{X})$ (the space of smooth functions over $\mathbb{R}^d \times \mathsf{X}$ whose derivatives of all order are bounded).

A4. The initial parameters $(\theta^i_0)_{i=1}^N$ are i.i.d. w.r.t. $\mu_0 \in \mathcal{P}(\mathbb{R}^{d+1})$ which has compact support.

Note that **A1** is satisfied when γ is the pdf of $\mathcal{N}(0,\mathrm{I}_d)$ and $\Psi_\theta(z) = m + g(\rho)z$, with $\mathfrak{b}(z) = 1+|z|$. With these assumptions, for every fixed $T > 0$, the sequence $(\{\theta^i_k, i=1,\dots,N\})_{k=0,\dots,\lfloor NT\rfloor}$ defined by (5.5) is a.s. bounded:

Lemma 39 (Uniform bound on the parameters)**.** *Assume **A1**→**A4**. Then, there exists $C > 0$ such that a.s. for all $T > 0$, $N \geq 1$, $i \in \{1, \dots, N\}$, and $0 \leq k \leq \lfloor NT \rfloor$, $|\theta_k^i| \leq Ce^{[C(2+T)]T}$.*

Lemma 39 implies that a.s. for all $T > 0$ and $N \geq 1$, $\mu^N \in \mathcal{D}([0,T], \mathcal{P}(\Theta_T))$, where

$$\Theta_T = \{\theta \in \mathbb{R}^{d+1}, |\theta| \leq Ce^{[C(2+T)]T}\}.$$

The first main result of this work is the following Theorem which derives the law of large numbers for $(\mu^N)_{N\geq 1}$ (defined in (5.8)).

Theorem 40. *Assume **A1**→**A4**. Let $T > 0$. Then, the sequence $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,T], \mathcal{P}(\Theta_T))$ defined in* (5.8) *converges in probability to the unique deterministic solution $\bar{\mu} \in \mathcal{C}([0,T], \mathcal{P}(\Theta_T))$ to the following measure-valued evolution equation: $\forall f \in \mathcal{C}^\infty(\Theta_T)$ and $\forall t \in [0,T]$,*

$$\begin{aligned}\langle f, \bar{\mu}_t \rangle - \langle f, \mu_0 \rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma \big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s \big\rangle \mathrm{d}s. \end{aligned} \tag{5.11}$$

The proof of Theorem 40 is given in Appendix 8. We stress here the most important steps and used techniques. In a first step, we derive an identity satisfied by $(\mu^N)_{N\geq 1}$, namely the pre-limit equation (5.28); see Sec. 8.1. Then we show in Sec. 8.2.2 that $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,T], \mathcal{P}(\Theta_T))$. To do so, we check that the sequence $(\mu^N)_{N\geq 1}$ satisfies all the required assumptions of [Jakubowski, 1986, Theorem 3.1] when $E = \mathcal{P}(\Theta_T)$ there. In Sec. 8.2.3 we prove that every limit point of $(\mu^N)_{N\geq 1}$ satisfies the limit equation (5.11). Then, in Section 8.2.4, we prove that there is a unique solution of the measure-valued equation (5.11). To prove the uniqueness of the solution of (5.11), we use techniques developed in [Piccoli et al., 2015] which are based on a representation formula for solution to measure-valued equations [Villani, 2021, Theorem 5.34] together with estimates in Wasserstein distances between two solutions of (5.11) derived in [Piccoli and Rossi, 2016]. In Section 8.2.4, we also conclude the proof of Theorem 40. Compared to [Descours et al., 2022a, Theorem 1], the fact that $(\{\theta_k^i, i = 1, \dots, N\})_{k=0,\dots,\lfloor NT \rfloor}$ defined by (5.5) are a.s. bounded allows to use different and more straightforward arguments to prove (i) the relative compactness in $\mathcal{D}([0,T], \mathcal{P}(\Theta_T))$ of $(\mu^N)_{N\geq 1}$ (defined in (5.8)) (ii) the continuity property of the operator $\mathsf{m} \mapsto \Lambda_t[f](\mathsf{m})$ defined in (5.35) w.r.t. the topology of $\mathcal{D}([0,T], \mathcal{P}(\Theta_T))$ and (iii) $(\mu^N)_{N\geq 1}$ has limit points in $\mathcal{C}([0,T], \mathcal{P}(\Theta_T))$. Step (ii) is necessary in order to pass to the limit $N \to +\infty$ in the pre-limit equation and Step (iii) is crucial since we prove that there is at most one solution of (5.11) in $\mathcal{C}([0,T], \mathcal{P}(\Theta_T))$. It is worthwhile to emphasize that, as $N \to \infty$, the effects of the integrated loss and of the KL terms are balanced, as conjectured in [Huix et al., 2022].

To avoid further technicalities, we have chosen what may seem restrictive assumptions on the data or the activation function. Note however that it readily extends to unbounded set X, and also unbounded Y assuming that π as polynomial moments of sufficiently high order. Also, RELU (or more easily leaky RELU) may be considered by using weak derivatives (to consider the singularity at 0), and a priori moment bounds on the weights.

4 LLN for the *Bayes-by-Backprop* SGD

The sequence $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ defined recursively by the algorithm (5.7) is in general not bounded, since $\nabla_\theta \phi(\theta, \mathsf{Z}, x)$ is not necessarily bounded if $\mathsf{Z} \sim \gamma(s)\mathrm{d}z$. Therefore, we cannot expect Lemma 39 to hold for $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ set by (5.7). Thus, the sequence $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ is considered on the whole space $\mathbb{R}^{d+1}$.

Wasserstein spaces and results. For $N \geq 1$, and $k \geq 1$, we set

$$\mathcal{F}_k^N = \boldsymbol{\sigma}\Big(\theta_0^i, \mathsf{Z}_q^{j,\ell}, (x_q, y_q), 1 \leq i, j \leq N, 1 \leq \ell \leq B, 0 \leq q \leq k-1\Big). \tag{5.12}$$

In addition to **A1**→**A4** (where in **A2**, when $k \geq 1$, $\mathcal{F}_k^N$ is now the one defined in (5.12)), we assume:

A5. The sequences $(\mathsf{Z}_k^{j,\ell}, 1 \leq j \leq N, 1 \leq \ell \leq B, k \geq 0)$ and $((x_k, y_k), k \geq 0)$ are independent. In addition, for $k \geq 0$, $\big((x_k, y_k), \mathsf{Z}_k^{j,\ell}, 1 \leq j \leq N, 1 \leq \ell \leq B\big) \perp\!\!\!\perp \mathcal{F}_k^N$.

Note that the last statement of **A5** implies the last statement of **A2**. We introduce the scaled empirical distribution of the parameters of the algorithm (5.7), i.e. for $k \geq 0$ and $t \geq 0$:

$$\nu_k^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_k^i} \text{ and } \mu_t^N := \nu_{\lfloor Nt \rfloor}^N, \text{ where the } \theta_k^i\text{'s are defined (5.7).} \tag{5.13}$$

One can no longer rely on the existence of a compact subset $\Theta_T \subset \mathbb{R}^{d+1}$ such that a.s. $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,T], \mathcal{P}(\Theta_T))$, where $\mu^N = \{t \geq 0 \mapsto \mu_t^N\}$ is defined in (5.13). For this reason, we will work in Wasserstein spaces $\mathcal{P}_q(\mathbb{R}^{d+1})$, $q \geq 0$, which, we recall, are defined by

$$\mathcal{P}_q(\mathbb{R}^{d+1}) = \Big\{\nu \in \mathcal{P}(\mathbb{R}^{d+1}), \int_{\mathbb{R}^{d+1}} |\theta|^q \nu(\mathrm{d}\theta) < +\infty\Big\}. \tag{5.14}$$

These spaces are endowed with the Wasserstein metric W_q, see e.g. [Santambrogio, 2015, Chapter 5] for more materials on Wasserstein spaces. For all $q \geq 0$, $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbb{R}_+, \mathcal{P}_q(\mathbb{R}^{d+1}))$. The second main results of this work is a LLN for $(\mu^N)_{N\geq 1}$ defined in (5.13).

Theorem 41. *Assume **A1**→**A5**. Let $\gamma_0 > 1 + \frac{d+1}{2}$. Then, the sequence $(\mu^N)_{N\geq 1}$ defined in* (5.13) *converges in probability in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ to a deterministic element $\bar{\mu} \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$, where $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ is the unique solution in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ to the following measure-valued evolution equation:$\forall f \in \mathcal{C}_b^\infty(\mathbb{R}^{d+1})$ and $\forall t \in \mathbb{R}_+$,*

$$\begin{aligned}\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s\big\rangle \mathrm{d}s.\end{aligned} \tag{5.15}$$

Theorem 41 is proved in the appendix 9. Since $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ defined by (5.7) is not bounded in general, we work in the space $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. The proof of Theorem 41 is more involved than that of Theorem 40, and generalizes the latter to the case where the parameters of the SGD algorithm are unbounded. We prove that $(\mu^N)_{N\geq 1}$ (defined in (5.13)) is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. To this end we now use [Jakubowski, 1986, Theorem 4.6]. The compact containment, which is the purpose of Lemma 58, is not straightforward since $\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1})$ is not compact contrary to Theorem 40 where we used the compactness of $\mathcal{P}(\Theta_T)$. More precisely, the compact containment of $(\mu^N)_{N\geq 1}$ relies on a characterization of the compact subsets of $\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1})$ (see Proposition 56) and moment estimates on $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ (see Lemma 55). We also mention that contrary to what is done in the proof of Theorem 40, we do not show that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ is continuous in time but we still manage to prove that they all satisfy (5.15). Then, using the duality formula for the W_1-distance together with rough estimates on the jumps of $t \mapsto \langle f, \mu_t^N\rangle$ (for f uniformly Lipschitz over $\mathbb{R}^{d+1}$), we then show that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ belongs a.s. to $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$. Again this is important since we have uniqueness of (5.15) in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$.

We conclude this section with the following important uniqueness result.

Proposition 42. *Under the assumptions of Theorems 40 and 41, the solution to* (5.11) *is independent of T and is equal to the solution to* (5.15)*.*

This uniqueness result states that both idealized and *Bayes-by-backprop* SGD have the same limiting behavior. It is also noteworthy that the mini-batch B is held fixed B. The effect of batch size can be seen at the level of the central limit theorem, which we leave for future work.

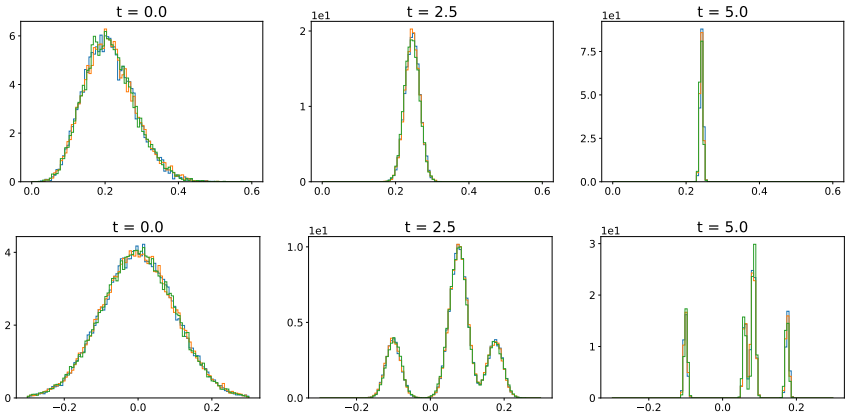


Fig. 5.1 Histograms of $\{F(\theta^i_{\lfloor NT \rfloor}), i = 1, \dots, N\}$, at different times (initialization ($t = 0$), half ($t = 2.5$) and end of training ($T = 5$)), when $N = 10000$. First line: $F(\theta) = \|m\|_2$, where $\theta = (m, \rho) \in \mathbb{R}^d \times \mathbb{R}$. Second line: $F(\theta) = m \in \mathbb{R}^d$. Idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green).

5 The *Minimal-VI* SGD algorithm

The idea behing the *Bayes-by-Backprop* SGD stems from the fact that there are integrals wrt γ in the loss function that cannot be computed in practice and it is quite natural up to a reparameterization trick, to replace these integrals by a Monte Carlo approximation (with i.i.d. gaussian random variables). To devise a new cheaper algorithm based on the only terms impacting the asymptotic limit, we directly analyse the limit equation (5.11) and remark that it can be rewritten as, $\forall f \in \mathcal{C}^\infty(\Theta_T)$ and $\forall t \in [0, T]$,

$$\begin{aligned}
&\langle f, \bar{\mu}_t \rangle - \langle f, \mu_0 \rangle \\
&= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}\times(\mathbf{R}^d)^2} \big\langle \phi(\cdot, z_1, x) - y, \bar{\mu}_s \big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z_2, x), \bar{\mu}_s \big\rangle \gamma^{\otimes 2}(\mathrm{d}z_1 \mathrm{d}z_2) \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\
&\quad - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \bar{\mu}_s \big\rangle \mathrm{d}s.
\end{aligned}$$

Thus, the integration over $\gamma^{\otimes 2}$ can be considered as that over π, i.e., we can consider them as two more data variables that only need to be sampled at each new step. In this case, the SGD (5.7) becomes: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \theta^i_{k+1} = \theta^i_k - \frac{\eta}{N^2} \sum_{j=1}^N \big(\phi(\theta^j_k, \mathsf{Z}^1_k, x_k) - y_k\big) \nabla_\theta \phi(\theta^i_k, \mathsf{Z}^2_k, x_k) - \frac{\eta}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k} | P^1_0) \\ \theta^i_0 = (m^i_0, \rho^i_0) \sim \mu_0, \end{cases} \tag{5.16}$$

where $\eta > 0$ and $(\mathsf{Z}^p_k, p \in \{1, 2\}, k \geq 0)$ is a i.i.d sequence of random variables distributed according to $\gamma^{\otimes 2}$. We call this backpropagation scheme *minimal- VI SGD* which is much cheaper in terms of computational complexity, with the same limiting behavior as we now discuss.

We introduce the $\boldsymbol{\sigma}$-algebra for $N, k \geq 1$:

$$\mathcal{F}^N_k = \boldsymbol{\sigma}\Big(\theta^i_0, \mathsf{Z}^p_q, (x_q, y_q), 1 \leq i \leq N, p \in \{1, 2\}, 0 \leq q \leq k-1\}\Big). \tag{5.17}$$

In addition to **A1**→**A4** (where in **A2**, $\mathcal{F}^N_k$ is now the one defined above in (5.17) when $k \geq 1$), the following assumption

A6. The sequences $(\mathsf{Z}^p_k, p \in \{1, 2\}, k \geq 0)$ and $((x_k, y_k), k \geq 0)$ are independent. In addition, for $k \geq 0$, $\big((x_k, y_k), \mathsf{Z}^p_k, p \in \{1, 2\}\big) \perp\!\!\!\perp \mathcal{F}^N_k$, where $\mathcal{F}^N_k$ is defined in (5.17).

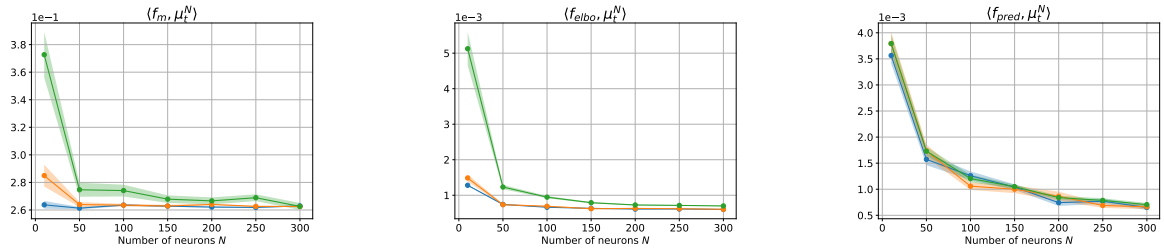


Fig. 5.2 Convergence of $\langle f, \mu_T^N \rangle$ to $\langle f, \bar{\mu}_T \rangle$, for the idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green) SGD algorithms over 50 realizations.

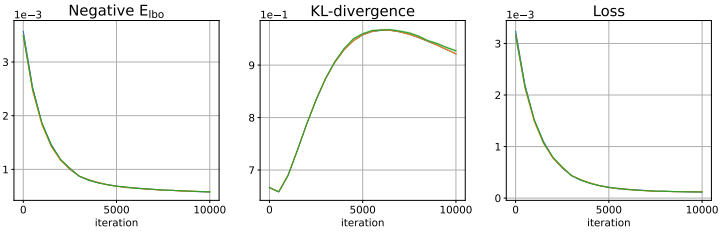


Fig. 5.3 Decay of the negative ELBO (left) and its two components (KL (middle), loss (right)) during the training process done by the idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green) SGD algorithms, for $N = 10000$.

Set for $k \geq 0$ and $t \geq 0$, $\nu_k^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_k^i}$ and $\mu_t^N := \nu_{\lfloor Nt \rfloor}^N$, where the θ_k^i's are defined in (5.16). The last main result of this work states that the sequence $(\mu^N)_{N\geq 1}$ satisfies the same law of large numbers when $N \to +\infty$ as the one satisfied by (5.13), whose proof will be omitted as it is the same as the one made for Theorem 41.

Theorem 43. *Assume **A1**→**A4** and **A6**. Then, the sequence of $(\mu^N)_{N\geq 1}$ satisfies all the statements of Theorem 41.*

6 Numerical experiments

In this section we illustrate the theorems 40, 41, and 43 using the following toy model. We set $d = 5$. Given $\theta^* \in \mathbb{R}^d$ (drawn from a normal distribution and scaled to the unit norm), we draw i.i.d observations as follows: Given $x \sim \mathcal{U}([-1,1]^d)$, we draw $y = \tanh(x^\top \theta^*) + \epsilon$, where ϵ is zero mean with variance 10^{-4}. The initial distribution of parameters is centered around the prior: $\boldsymbol{\theta}_0 \sim (\mathcal{N}(m_0, 0.01\mathrm{I}_d) \times \mathcal{N}(g^{-1}(\sigma_0), 0.01))^{\otimes N}$, with $m_0 = 0$ and $\sigma_0 = 0.2$. Since the idealized algorithm cannot be implemented exactly, a mini-batch of size 100 is used as a proxy for the following comparisons of the different algorithms. For the algorithm (5.7) SGD we set $B = 1$.

6.1 Evolution and limit of the distribution

Fig. 5.1 displays the histograms of $\{F(\theta_{\lfloor Nt \rfloor}^i), i = 1, \ldots, N\}$ ($F(\theta) = \|m\|_2, g(\rho)$ or m, where $\theta = (m, \rho) \in \mathbb{R}^d \times \mathbb{R}$), for $N = 10000$, at initialization, halfway through training, and at the end of training. The empirical distributions illustrated by these histograms are very similar over the course of training. It can be seen that for $N = 10000$ the limit of the mean field is reached.

6.2 Convergence with respect to the numbers of neurons.

We investigate here the speed of convergence of μ_t^N to $\bar{\mu}_t$ (as $N \to +\infty$), when tested against test functions f. More precisely, we fix a time T (end of training) and Figure 5.2 represents the empirical mean of $\langle f, \mu_T^N \rangle$ over 50 realizations. The test functions used for this experiment are $f_m(\theta) = \|m\|_2$, $f_{\mathrm{Elbo}(\theta)} = -\hat{\mathrm{E}}_{\mathrm{lbo}}(\theta)^N$ where $\hat{\mathrm{E}}_{\mathrm{lbo}}$ is the empirical $\mathrm{E}_{\mathrm{lbo}}^N$ (see (5.1)) computed with 100 samples of (x, y) and $(z^1, \ldots, z^N)$. Finally, $f_{pred}(\theta) = \hat{\mathbb{E}}_x \left[\hat{\mathbb{V}}_{\boldsymbol{w} \sim q_{\boldsymbol{\theta}}^N} [f_{\boldsymbol{w}}^N(x)]^{1/2} \right]$

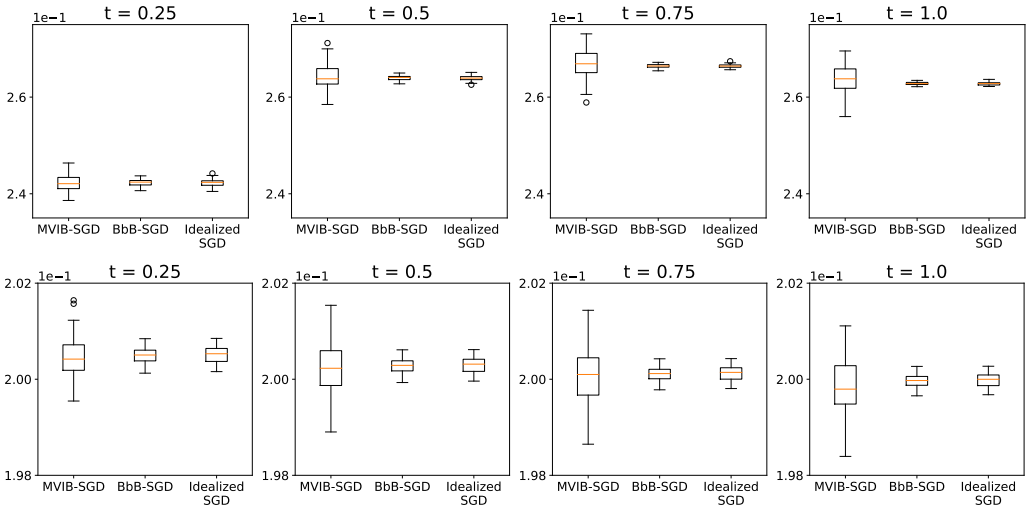


Fig. 5.4 Boxplots for 50 runs of $\langle f, \mu_t^N \rangle$ for the three SGD schemes for $f(\theta) = \|m\|_2$ on the first line and $f(\theta) = g(\rho)$ on the second line. MVIB-SGD:*Minimal-VI* SGD. BbB-SGD:*Bayes-by-Backprop* SGD.

where $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ denote respectively the empirical mean and the empirical variance over 100 samples. All algorithms are converging to the same limit and are performing similarly even with a limited number of neurons ($N = 300$ in this example).

6.3 Convergence with respect to time.

This section illustrates the training process of a BNN with a given number of neurons $N = 10000$. In Figure 5.3, we plot the negative ELBO on a test set and its two components, the loss and the KL-divergence terms. Figure 5.3 shows that the BNN is able to learn on this specific task and all algorithms exhibit a similar performance. It illustrates the trajectorial convergence of $\{\mu_t^N, t \in [0,T]\}_{N \geq 1}$ to $\{\bar{\mu}_t, t \in [0,T]\}$ as $N \to +\infty$.

6.4 Behavior around the limit $\bar{\mu}$.

On Figure 5.4, we plot the boxplots of $\langle f, \mu_t^N \rangle$ for 50 realizations and $N = 10000$, at different times of the training. *Minimal-VI* scheme (which is computationally cheaper as explained in 5) exhibit a larger variance than the other algorithms.

7 Conclusion

By establishing the limit behavior of the idealized SGD for the variational inference of BNN with the weighting suggested by [Huix et al., 2022], we have rigorously shown that the most-commonly used in practice *Bayes-by-Backprop* scheme indeed exhibits the same limit behavior. Furthermore, the analysis of the limit equation led us to validate the correct scaling of the KL divergence term in with respect to the loss. Notably, the mean-field limit dynamics has also helped us to devise a far less costly new SGD algorithm, the *Minimal-VI*. This scheme shares the same limit behavior, but only stems from the non-vanishing asymptotic contributions, hence the reduction of the computational cost. Aside from confirming the analytical results, the first simulations presented here show that the three algorithms, while having the same limit, may differ in terms of variance. Thus, deriving a CLT result and discussing the right trade-off between computational complexity and variance will be done in future work. Also, on a more general level regarding uncertainty quantification, an interesting question is to analyse the impact of the correct scaling of the KL divergence term on the error calibration and how to apply the same analysis in the context of deep ensembles.

8 Proof of Theorem 40

For simplicity, we prove Theorem 40 when $T = 1$, and we denote Θ_1 simply by Θ. In this section we assume **A1**–**A4**.

8.1 Pre-limit equation (5.28) and error terms in (5.28)

8.1.1 Derivation of the pre-limit equation

The aim of this section is to establish the so-called pre-limit equation (5.28), which will be our starting point to derive Equation (5.11). Let $N \geq 1$, $k \in \{0, \dots, N\}$, and $f \in \mathcal{C}^\infty(\Theta)$. Recall that by Lemma 39 and since $0 \leq k \leq N$, a.s. $\theta_k^i \in \Theta$, and thus a.s. $f(\theta_k^i)$ is well-defined. The Taylor-Lagrange formula yields

$$\begin{aligned}\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle &= \frac{1}{N} \sum_{i=1}^N f(\theta_{k+1}^i) - f(\theta_k^i) \\ &= \frac{1}{N} \sum_{i=1}^N \nabla_\theta f(\theta_k^i) \cdot (\theta_{k+1}^i - \theta_k^i) + \frac{1}{2N} \sum_{i=1}^N (\theta_{k+1}^i - \theta_k^i)^T \nabla^2 f(\widehat{\theta_k^i})(\theta_{k+1}^i - \theta_k^i),\end{aligned}$$

where, for all $i \in \{1, \dots, N\}$, $\widehat{\theta_k^i} \in (\theta_k^i, \theta_{k+1}^i) \subset \Theta$. Using (5.5), we then obtain

$$\begin{aligned}\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle = &- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \big(\langle \phi(\theta_k^j, \cdot, x_k), \gamma \rangle - y_k \big) \langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \rangle \\ &- \frac{\eta}{N^2} \big\langle \big(\phi(\cdot, \cdot, x_k) - y_k \big) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x_k), \nu_k^N \otimes \gamma \big\rangle \\ &- \frac{\eta}{N} \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \nu_k^N \big\rangle + \mathbb{R}_k^N[f], \end{aligned} \tag{5.18}$$

where

$$\mathbb{R}_k^N[f] := \frac{1}{2N} \sum_{i=1}^N (\theta_{k+1}^i - \theta_k^i)^T \nabla^2 f(\widehat{\theta_k^i})(\theta_{k+1}^i - \theta_k^i). \tag{5.19}$$

Let us define

$$\begin{aligned}\mathbf{D}_k^N[f] := &\mathbf{E}\Big[- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \big(\langle \phi(\theta_k^j, \cdot, x_k), \gamma \rangle - y_k \big) \langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \rangle \Big| \mathcal{F}_k^N \Big] \\ &- \mathbf{E}\Big[\frac{\eta}{N^2} \big\langle (\phi(\cdot, \cdot, x_k) - y_k) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x_k), \nu_k^N \otimes \gamma \big\rangle \Big| \mathcal{F}_k^N \Big]. \end{aligned} \tag{5.20}$$

Note that using (5.45) and (5.47) together with the fact that $|\nabla_\theta f(\theta_k^i)| \leq \sup_{\theta \in \Theta} |\nabla_\theta f(\theta)|$, the integrant in (5.20) is integrable and thus $\mathbf{D}_k^N[f]$ is well defined. Using the fact that $(x_k, y_k) \perp\!\!\!\perp \mathcal{F}_k^N$ by **A2** and that $\{\theta_k^i, i = 1, \dots, N\}$ is $\mathcal{F}_k^N$-measurable by (5.5), we have:

$$\begin{aligned}\mathbf{D}_k^N[f] = &- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\langle \phi(\theta_k^j, \cdot, x), \gamma \rangle - y \big) \langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) \\ &- \frac{\eta}{N^2} \int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot, \cdot, x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \nu_k^N \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y). \end{aligned} \tag{5.21}$$

Introduce also

$$\begin{aligned}\mathbf{M}_k^N[f] := &- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N (\langle \phi(\theta_k^j, \cdot, x_k), \gamma \rangle - y_k) \langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \rangle \\ &- \frac{\eta}{N^2} \langle (\phi(\cdot, \cdot, x_k) - y_k) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x_k), \nu_k^N \otimes \gamma \rangle - \mathbf{D}_k^N[f]. \end{aligned}$$

Note that $\mathbf{E}\big[\mathbf{M}_k^N[f]|\mathcal{F}_k^N\big]=0$. Equation (5.18) then writes

$$\langle f,\nu_{k+1}^N\rangle-\langle f,\nu_k^N\rangle=\mathbf{D}_k^N[f]+\mathbf{M}_k^N[f]-\frac{\eta}{N}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\nu_k^N\big\rangle+\mathbb{R}_k^N[f]. \tag{5.22}$$

Notice also that

$$\begin{aligned}
\mathbf{D}_k^N[f]=&-\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1}^N\int_{\mathsf{X}\times\mathsf{Y}}(\langle\phi(\theta_k^j,\cdot,x),\gamma\rangle-y)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&+\frac{\eta}{N^3}\sum_{i=1}^N\int_{\mathsf{X}\times\mathsf{Y}}(\langle\phi(\theta_k^i,\cdot,x),\gamma\rangle-y)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&-\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
=&-\frac{\eta}{N}\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\nu_k^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&+\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\langle\phi(\cdot,\cdot,x),\gamma\rangle-y)\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\nu_k^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&-\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y).
\end{aligned}\tag{5.23}$$

Now, we define for $t\in[0,1]$:

$$\mathbf{D}_t^N[f]:=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{D}_k^N[f],\ \ \mathbb{R}_t^N[f]:=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbb{R}_k^N[f],\ \text{ and }\ \mathbf{M}_t^N[f]:=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]. \tag{5.24}$$

We can rewrite $\mathbf{D}_t^N[f]$ has follows:

$$\mathbf{D}_t^N[f]=\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}N\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s=N\int_0^t\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s-N\int_{\frac{\lfloor Nt\rfloor}{N}}^t\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s.$$

Since $\nu_{\lfloor Ns\rfloor}^N=\mu_s^N$ (by definition, see (5.8)), we have, using also (5.23) with $k=\lfloor Ns\rfloor$,

$$\begin{aligned}
\mathbf{D}_t^N[f]=&-\eta\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&+\frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&-\frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s-\mathbf{V}_t^N[f],
\end{aligned}\tag{5.25}$$

where

$$\begin{aligned}
\mathbf{V}_t^N[f]:=&-\eta\int_{\frac{\lfloor Nt\rfloor}{N}}^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&+\frac{\eta}{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&-\frac{\eta}{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

On the other hand, we also have for $t\in[0,1]$,

$$\sum_{k=0}^{\lfloor Nt\rfloor-1}-\frac{\eta}{N}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\nu_k^N\big\rangle=-\eta\int_0^{\frac{\lfloor Nt\rfloor}{N}}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\mu_s^N\big\rangle\mathrm{d}s. \tag{5.26}$$

We finally set:

$$\mathbf{W}_t^N[f] := -\mathbf{V}_t^N[f] + \eta \int_{\frac{\lfloor Nt \rfloor}{N}}^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s. \tag{5.27}$$

Since $\langle f, \mu_t^N \rangle - \langle f, \mu_0^N \rangle = \sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle$, we deduce from (5.22), (5.24), (5.25), (5.26) and (5.27), the so-called pre-limit equation satisfied by μ^N: for $N \geq 1$, $t \in [0,1]$, and $f \in \mathcal{C}^\infty(\Theta)$,

$$\begin{aligned}
\langle f, \mu_t^N \rangle - \langle f, \mu_0^N \rangle = & -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\
& - \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s \\
& + \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\
& - \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\
& + \mathbf{M}_t^N[f] + \mathbf{W}_t^N[f] + \mathbb{R}_t^N[f].
\end{aligned} \tag{5.28}$$

8.1.2 The last five terms in (5.28) are error terms

The purpose of this section is to show that the last five terms appearing in the r.h.s. of (5.28) are error terms when $N \to +\infty$. For $J \in \mathbf{N}^*$ and $f \in \mathcal{C}^J(\Theta)$, set $\|f\|_{\mathcal{C}^J(\Theta)} := \sum_{|k| \leq J} \|\partial_k f\|_{\infty,\Theta}$, where $\|g\|_{\infty,\Theta} = \sup_{\theta\in\Theta} |g(\theta)|$ for $g : \Theta \to \mathbb{R}^m$.

Lemma 44 (Error terms)**.** *Assume **A1**→**A4**. Then, there exists $C > 0$ such that a.s. for all $f \in \mathcal{C}^\infty(\Theta)$ and $N \geq 1$,*

1. $\frac{\eta}{N} \int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \leq C \|f\|_{\mathcal{C}^1(\Theta)} / N.$
2. $\frac{\eta}{N} \int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \leq C \|f\|_{\mathcal{C}^1(\Theta)} / N.$
3. $\sup_{t\in[0,1]} |\mathbf{W}_t^N[f]| + \sup_{t\in[0,1]} |\mathbb{R}_t^N[f]| \leq C \|f\|_{\mathcal{C}^2(\Theta)} / N.$

Finally, $\sup_{t\in[0,1]} \mathbf{E}\big[|\mathbf{M}_t^N[f]|\big] \leq C \|f\|_{\mathcal{C}^1(\Theta)} / \sqrt{N}$.

Proof. All along the proof, $C > 0$ denotes a positive constant independent of $N \geq 1, k \in \{0, \ldots, N-1\}, (s,t) \in [0,1]^2, (x,y) \in X \times Y, \theta \in \Theta, z \in \mathbb{R}^d$, and $f \in \mathcal{C}^\infty(\Theta)$ which can change from one occurrence to another. Using (5.47), the Cauchy-Schwarz inequality, and the fact that $\nabla_\theta f$ is bounded over Θ imply:

$$|\langle \nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle| \leq \langle |\nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta,\cdot,x)|, \gamma \rangle \leq C \|f\|_{\mathcal{C}^1(\Theta)}. \tag{5.29}$$

Combining (5.45) and (5.29), we obtain:

$$\int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \leq C \|f\|_{\mathcal{C}^1(\Theta)}$$

and

$$\int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_m f \cdot \nabla_m \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \leq C \|f\|_{\mathcal{C}^1(\Theta)},$$

which proves Items 1 and 2.

Let us now prove Item 3. By (5.45) and (5.29), $\sup_{t\in[0,1]} |\mathbf{V}_t^N[f]| \leq C \|f\|_{\mathcal{C}^1(\Theta)} / N$. On the other hand, because $f \in \mathcal{C}^\infty(\Theta)$ and $\theta \mapsto \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)$ is continuous (see (5.4)) over Θ which is compact, it holds, $\|\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)\|_{\infty,\Theta} < +\infty$. Hence, it holds:

$$\sup_{t\in[0,1]} \Big| \int_{\frac{\lfloor Nt \rfloor}{N}}^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s \Big| \leq C \|f\|_{\mathcal{C}^1(\Theta)} / N.$$

Using (5.27), it then holds $\sup_{t\in[0,1]}|\mathbf{W}_t^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N$. Since $f\in\mathcal{C}^\infty(\Theta)$, we have, by (5.19), for $N\ge 1$ and $0\le k\le N-1$, $|\mathbb{R}_k^N[f]| \le \|f\|_{\mathcal{C}^2(\Theta)}\frac{C}{N}\sum_{i=1}^N|\theta_{k+1}^i-\theta_k^i|^2$. By (5.48) and Lemma 39, $|\theta_{k+1}^i-\theta_k^i|^2\le C/N^2$ and consequently, one has:

$$|\mathbb{R}_k^N[f]| \le C\|f\|_{\mathcal{C}^2(\Theta)}/N^2. \tag{5.30}$$

Hence, for all $t\in[0,1]$, $|\mathbb{R}_t^N[f]|\le C\|f\|_{\mathcal{C}^2(\Theta)}/N$. This proves Item 3.

Let us now prove the last item in Lemma 44. Let $t\in[0,1]$. We have, by (5.24),

$$|\mathbf{M}_t^N[f]|^2 = \sum_{k=0}^{\lfloor Nt\rfloor-1}\left|\mathbf{M}_k^N[f]\right|^2 + 2\sum_{k<j}\mathbf{M}_k^N[f]\mathbf{M}_j^N[f].$$

For all $0\le k<j<\lfloor Nt\rfloor$, $\mathbf{M}_k^N[f]$ is $\mathcal{F}_j^N$-measurable (see (5.9)), and since $\mathbf{E}\big[\mathbf{M}_j^N[f]|\mathcal{F}_j^N\big]=0$, one deduces that $\mathbf{E}\big[\mathbf{M}_k^N[f]\mathbf{M}_j^N[f]\big] = \mathbf{E}\big[\mathbf{M}_k^N[f]\mathbf{E}\big[\mathbf{M}_j^N[f]|\mathcal{F}_j^N\big]\big]=0$. Hence, $\mathbf{E}[|\mathbf{M}_t^N[f]|^2]=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[|\mathbf{M}_k^N[f]|^2]$. By (5.45) and (5.29), one has a.s. for all $0\le k\le N-1$,

$$|\mathbf{M}_k^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N. \tag{5.31}$$

Hence, $\mathbf{E}[|\mathbf{M}_t^N[f]|^2]\le C\|f\|_{\mathcal{C}^1(\Theta)}/N$, which proves the last inequality in Lemma 44. □

8.2 Convergence to the limit equation as $N\to+\infty$

In this section we prove the relative compactness of $(\mu^N)_{N\ge1}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$. We then show that any of its limit points satisfies the limit equation (5.11).

8.2.1 Wasserstein spaces and duality formula

In this section we recall some basic results which will be used throughout this work on the space $\mathcal{P}(\mathcal{S})$ when $(\mathcal{S},\,d)$ is a Polish space. First when endowed with the weak convergence topology, $\mathcal{P}(\mathcal{S})$ is a Polish space [Billingsley, 1999, Theorem 6.8]. In addition, $\mathcal{P}_q(\mathcal{S})=\{\nu\in\ P(\mathcal{S}),\int_{\mathcal{S}}\,d(w_0,w)^q\nu(\mathrm{d}w)<+\infty\}$, where $w_0\in\ S$ is arbitrary (note that this space was defined previously in (5.14) when $\mathcal{S}=\mathbb{R}^{d+1}$) when endowed with the W_q metric is also a Polish space [Villani, 2009, Theorem 6.18]. Recall also the duality formula for the W_1-distance on $\mathcal{P}_1(\mathcal{S})$ (see e.g [Villani, 2009, Remark 6.5]):

$$\mathsf{W}_1(\mu,\nu)=\sup\Big\{\Big|\int_{\mathcal{S}}f(w)\mathrm{d}\mu(w)-\int_{\mathcal{S}}f(w)\nu(\mathrm{d}w)\Big|,\ \|f\|_{\text{Lip}}\le1\Big\}. \tag{5.32}$$

Finally, when $\mathcal{K}\subset\mathbb{R}^{d+1}$ is compact, the convergence in W_q-distance is equivalent to the usual weak convergence on $\mathcal{P}(\mathcal{K})$ (see e.g. [Villani, 2009, Corollary 6.13]).

8.2.2 Relative compactness

The main result of this section is to prove that $(\mu^N)_{N\ge1}$ is relatively compact in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$, which is the purpose of Proposition 46 below. To this end, we need to prove that for all $f\in\mathcal{C}^\infty(\Theta)$, every sequence $(\langle f,\mu_t^N\rangle)_{N\ge1}$ satisfies some regularity conditions, which is the purpose of the next result.

Lemma 45 (Regularity condition)**.** *Assume **A1**→**A4**. Then there exists $C>0$ such that a.s. for all $f\in\mathcal{C}^\infty(\Theta)$, $0\le r<t\le1$, and $N\ge1$:*

$$|\langle f,\mu_t^N\rangle-\langle f,\mu_r^N\rangle| \le C\|f\|_{\mathcal{C}^2(\Theta)}\Big[|t-r|+\frac{|t-r|}{N}+\frac{1}{N}\Big]. \tag{5.33}$$

Proof. Let $f \in \mathcal{C}^\infty(\Theta)$ and let $N \geq 1$ and $0 \leq r < t \leq 1$. In the following $C > 0$ is a positive constant independent of $f \in \mathcal{C}^\infty(\Theta)$, $N \geq 1$, and $0 \leq r < t \leq 1$, which can change from one occurrence to another. From (5.28), we have

$$\begin{aligned}\langle f, \mu_t^N \rangle - \langle f, \mu_r^N \rangle = \mathbf{A}_{r,t}^N[f] - \eta \int_r^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s \\ + \mathbf{M}_t^N[f] - \mathbf{M}_r^N[f] + \mathbf{W}_t^N[f] - \mathbf{W}_r^N[f] + \mathbb{R}_t^N[f] - \mathbb{R}_r^N[f], \end{aligned} \tag{5.34}$$

where

$$\begin{aligned}\mathbf{A}_{r,t}^N[f] = -\eta \int_r^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) \\ + \frac{\eta}{N} \int_r^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\ - \frac{\eta}{N} \int_r^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y).\end{aligned}$$

By (5.45) and (5.29), $|\mathbf{A}_{r,t}^N[f]| \leq C\|f\|_{\mathcal{C}^1(\Theta)} \big[|t-r| + \frac{|t-r|}{N}\big]$. In addition, since $\theta \mapsto \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)$ is bounded over Θ (since it is smooth and Θ is compact),

$$\Big| \int_r^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s \Big| \leq C\|f\|_{\mathcal{C}^1(\Theta)} |t-r|.$$

Furthermore, using (5.31),

$$|\mathbf{M}_t^N[f] - \mathbf{M}_r^N[f]| = \Big| \sum_{k=\lfloor Nr \rfloor}^{\lfloor Nt \rfloor - 1} \mathbf{M}_k^N[f] \Big| \leq (\lfloor Nt \rfloor - \lfloor Nr \rfloor) C \|f\|_{\mathcal{C}^1(\Theta)} / N.$$

Next, we have, by Item 3 in Lemma 44, $|\mathbf{W}_t^N[f] - \mathbf{W}_r^N[f]| \leq |\mathbf{W}_t^N[f]| + |\mathbf{W}_r^N[f]| \leq C\|f\|_{\mathcal{C}^2(\Theta)}/N$. Finally, by (5.30),

$$|\mathbb{R}_t^N[f] - \mathbf{R}_r^N[f]| = \Big| \sum_{k=\lfloor Nr \rfloor}^{\lfloor Nt \rfloor - 1} \mathbb{R}_k^N[f] \Big| \leq (\lfloor Nt \rfloor - \lfloor Nr \rfloor) C \|f\|_{\mathcal{C}^2(\Theta)} / N^2.$$

The proof of Proposition 45 is complete plugging all the previous estimates in (5.34). □

Proposition 46 (Relative compactness)**.** *Assume **A1**→**A4**. Then, the sequence $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,1], \mathcal{P}(\Theta))$.*

Proof. The proof consists in applying [Jakubowski, 1986, Theorem 3.1] with $E = \mathcal{P}(\Theta)$ endowed with the weak convergence topology. Set $\mathbb{F} = \{\mathfrak{L}_f, f \in \mathcal{C}^\infty(\Theta)\}$ where

$$\mathsf{L}_f : \nu \in \ P(\Theta) \mapsto \langle f, \nu \rangle.$$

The class of continuous functions $\mathbb{F}$ on $\mathcal{P}(\Theta)$ satisfies Conditions [Jakubowski, 1986, (3.1) and (3.2) in Theorem 3.1].

On the other hand, the condition [Jakubowski, 1986, (3.3) in Theorem 3.1] is satisfied since $\mathcal{P}(\Theta)$ is compact because Θ is compact (see e.g. [Panaretos and Zemel, 2020, Corollary 2.2.5] together with [Villani, 2009, Corollary 6.13]).

It remains to verify Condition (3.4) of [Jakubowski, 1986, Theorem 3.1], i.e. that for all $f \in \mathcal{C}^\infty(\Theta)$, $(\langle f, \mu^N \rangle)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,1], \mathbb{R})$. To this end, we apply [Billingsley, 1999, Theorem 13.2]. Condition (i) in [Billingsley, 1999, Theorem 13.2] is satisfied because $|\langle f, \mu_t^N \rangle| \leq \|f\|_{\infty,\Theta}$ for all $t \in [0,1]$ and $N \geq 1$. Let us now show that Condition (ii) in [Billingsley, 1999, Theorem 13.2] holds. For this purpose, we use Lemma 45. For $\delta, \beta > 0$ sufficiently small, it is possible to construct a subdivision $\{t_i\}_{i=0}^v$ of $[0,1]$ such that $t_0 = 0$, $t_v = 1$, $t_{i+1} - t_i = \delta + \beta$ for $i \in \{0, \ldots, v-2\}$ and $\delta + \beta \leq t_v - t_{v-1} \leq 2(\delta + \beta)$. According to the terminology introduced

in [Billingsley, 1999, Section 12], $\{t_i\}_{i=0}^v$ is δ-sparse. Then, by Lemma 45, there exists $C>0$ such that a.s. for all $\delta,\beta>0$, all such subdivision $\{t_i\}_{i=0}^v$, $i\in\{0,\dots,v-1\}$, and $N\geq 1$,

$$\sup_{t,r\in[t_i,t_{i+1}]}|\langle f,\mu_t^N\rangle-\langle f,\mu_r^N\rangle|\leq C\Big(|t_{i+1}-t_i|+\frac{|t_{i+1}-t_i|}{N}+\frac{1}{N}\Big)\leq C\Big(2(\delta+\beta)+\frac{2(\delta+\beta)}{N}+\frac{1}{N}\Big).$$

Thus, one has:

$$\inf_{\beta>0}\max_i\sup_{t,r\in[t_i,t_{i+1}]}|\langle f,\mu_t^N\rangle-\langle f,\mu_r^N\rangle|\leq C\Big(2\delta+\frac{2\delta}{N}+\frac{1}{N}\Big).$$

Consequently, there exists $C>0$ such that a.s. for all $\delta>0$ small enough and $N\geq 1$,

$$w'_{\langle f,\mu^N\rangle}(\delta):=\inf_{\substack{\{t_i\}\\ \delta\text{-sparse}}}\max_i\sup_{t,r\in[t_i,t_{i+1}]}|\langle f,\mu_t^N\rangle-\langle f,\mu_r^N\rangle|\leq C\Big(2\delta+\frac{2\delta}{N}+\frac{1}{N}\Big).$$

This implies $\lim_{\delta\to 0}\limsup_{N\to+\infty}\mathbf{E}[w'_{\langle f,\mu^N\rangle}(\delta)]=0$. By Markov's inequality, this proves Condition (ii) of [Billingsley, 1999, Theorem 13.2]. Therefore, for all $f\in\mathcal{C}^\infty(\Theta)$, using also Prokhorov theorem, the sequence $(\langle f,\mu^N\rangle)_{N\geq 1}\subset\mathcal{D}([0,1],\mathbb{R})$ is relatively compact. In conclusion, according to [Jakubowski, 1986, Theorem 3.1], $(\mu^N)_{N\geq 1}\subset\mathcal{D}([0,1],\mathcal{P}(\Theta))$ is tight. □

8.2.3 Limit points satisfy the limit equation (5.11)

In this section we prove that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ satisfies (5.11).

Lemma 47. *Let $\mathsf{m},(\mathsf{m}^N)_{N\geq 1}\subset\mathcal{D}([0,1],\mathcal{P}(\Theta))$ be such that $\mathsf{m}^N\to m$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, for all Lipschitz continuous function $f:\Theta\to\mathbb{R}$, we have $\langle f,\mathsf{m}^N\rangle\to\langle f,\mathsf{m}\rangle$ in $\mathcal{D}([0,1],\mathbb{R})$.*

Proof. Let f be such a function. By [Billingsley, 1999, p.124], $\mathsf{m}^N\to m$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ iff there exist functions $\lambda_N:[0,1]\to[0,1]$ continuous, increasing onto itself such that $\sup_{t\in[0,1]}|\lambda_N(t)-t|\to_{N\to\infty}0$ and $\sup_{t\in[0,1]}\mathsf{W}_1(\mathsf{m}^N_{\lambda_N(t)},\mathsf{m}_t)\to_{N\to\infty}0$. Then $\langle f,\mathsf{m}^N\rangle\to\langle f,\mathsf{m}\rangle$ in $\mathcal{D}([0,1],\mathbb{R})$ since by (5.32), $\sup_{t\in[0,1]}|\langle f,\mathsf{m}^N_{\lambda_N(t)}\rangle-\langle f,\mathsf{m}_t\rangle|\leq\|f\|_{\mathrm{Lip}}\sup_{t\in[0,1]}\mathsf{W}_1(\mathsf{m}^N_{\lambda_N(t)},\mathsf{m}_t)\to_{N\to\infty}0$. □

Proposition 48 (Continuity of the limit points of $\langle f,\mu^N\rangle$)**.** *Let $f\in\mathcal{C}^\infty(\Theta)$. Then, any limit point of $(\langle f,\mu^N\rangle)_{N\geq 1}\subset\mathcal{D}([0,1],\mathbb{R})$ belong a.s. to $\mathcal{C}([0,1],\mathbb{R})$.*

Proof. Fix $t\in(0,1]$.Letting $r\to t$ in (5.33), we obtain $|\langle f,\mu_t^N\rangle-\langle f,\mu_{t^-}^N\rangle|\leq C/N$. Therefore $\sup_{t\in(0,1]}|\langle f,\mu_t^N\rangle-\langle f,\mu_{t^-}^N\rangle|\xrightarrow{\mathscr{D}}0$ as $N\to+\infty$.The result follows from [Billingsley, 1999, Theorem 13.4]. □

Proposition 49 (Continuity of the limit points of μ^N)**.** *Let $\mu^*\in\mathcal{D}([0,1],\ P(\Theta))$ be a limit point of $(\mu^N)_{N\geq 1}\subset\mathcal{D}([0,1],\ P(\Theta))$. Then, a.s. $\mu^*\in\mathcal{C}([0,1],\ P(\Theta))$.*

Proof. Up to extracting a subsequence, we assume that $\mu^N\xrightarrow{\mathscr{D}}\mu^*$. By Skorohod representation theorem, there exists another probability space $(\hat\Omega,\hat{\mathcal{F}},\hat{\mathbf{P}})$ on which are defined random elements $(\hat\mu^N)_{N\geq 1}$ and $\hat\mu^*$, where,

$$\hat\mu^*\overset{\mathscr{D}}{=}\mu^*,\ \text{ and for all } N\geq 1,\ \hat\mu^N\overset{\mathscr{D}}{=}\mu^N,$$

and such that $\hat{\mathbf{P}}$-a.s., $\hat\mu^N\to\hat\mu^*$ in $\mathcal{D}([0,1],\ P(\Theta))$ as $N\to+\infty$. Fix $f\in\mathcal{C}^\infty(\Theta)$. We have, by Lemma 47,

$$\hat{\mathbf{P}}-\text{a.s.},\ \langle f,\hat\mu^N\rangle\to_{N\to+\infty}\langle f,\hat\mu^*\rangle\ \ \mathrm{in}mathcalD([0,1],\mathbb{R}).$$

In particular, $\langle f,\hat\mu^N\rangle\to_{N\to+\infty}\langle f,\hat\mu^*\rangle$ in distribution. By Proposition 48, there exists $\hat\Omega_f\subset\hat\Omega$ of $\hat{\mathbf{P}}$-mass 1 such that for all $\omega\in\hat\Omega_f$, $\langle f,\hat\mu^*(\omega)\rangle\in\mathcal{C}([0,1],\mathbb{R})$. Denote by $\mathscr{F}$ the class polynomial functions with rational coefficients. Since this class is countable, the set $\hat\Omega_{\mathscr{F}}:=\cap_{f\in\mathscr{F}}\hat\Omega_f$ is of $\hat{\mathbf{P}}$-mass 1.

Consider now an arbitrary $f \in \mathcal{C}(\Theta)$ and let us show that for all $\omega \in \hat{\Omega}_{\mathscr{F}}$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1],\mathbb{R})$. By the Stone-Weierstrass theorem, there exist $(f_n)_{n\geq 1} \subset \mathscr{F}$ such that $\|f_n - f\|_{\infty,\Theta} \to_{n\to+\infty} 0$. On $\hat{\Omega}_{\mathscr{F}}$, for all n, $t \in [0,1] \mapsto \langle f_n, \hat{\mu}_t^*\rangle$ is continuous and converges uniformly to $t \in [0,1] \mapsto \langle f, \hat{\mu}_t^*\rangle$.

Hence, for all $\omega \in \hat{\Omega}_{\mathscr{F}}$ and $f \in \mathcal{C}(\Theta)$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1],\mathbb{R})$, i.e. for all $\omega \in \hat{\Omega}_{\mathscr{F}}$, $\hat{\mu}^*(\omega) \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$. This concludes the proof. □

Now, we introduce, for $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, the function $\Lambda_t[f] : \mathcal{D}([0,1],\mathcal{P}(\Theta)) \to \mathbb{R}_+$ defined by:

$$\begin{aligned}\Lambda_t[f] : \mathsf{m} \mapsto \Big| &\langle f, \mathsf{m}_t\rangle - \langle f, \mu_0\rangle \\ &+ \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &+ \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s\rangle \mathrm{d}s \Big|. \end{aligned} \tag{5.35}$$

We now study the continuity of $\Lambda_t[f]$.

Lemma 50. *Let $(\mathsf{m}^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$ converge to $\mathsf{m} \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, for all continuity point $t \in [0,1]$ of m and all $f \in \mathcal{C}^\infty(\Theta)$, we have $\Lambda_t[f](\mathsf{m}^N) \to \Lambda_t[f](\mathsf{m})$.*

Proof. Let $f \in \mathcal{C}^\infty(\Theta)$ and denote by $\mathcal{C}(\mathsf{m}) \subset [0,1]$ the set of continuity points of m. Let $t \in \mathcal{C}(\mathsf{m})$. From [Billingsley, 1999, p. 124], we have, for all $s \in \mathcal{C}(\mathsf{m})$,

$$\mathsf{m}_s^N \to \mathsf{m}_s \text{ in } \mathcal{P}(\Theta). \tag{5.36}$$

Thus, $\langle f, \mathsf{m}_t^N\rangle \to_{N\to\infty} \langle f, \mathsf{m}_t\rangle$. For all $z \in \mathbb{R}^d$ and $(x,y) \in \mathsf{X}\times\mathsf{Y}$, **A1** and **A3** ensure that the functions $\theta \in \Theta \mapsto \phi(\theta,z,x) - y$ and $\theta \in \Theta \mapsto \nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta,z,x)$ are continuous and also bounded because Θ is compact. Hence, for all $s \in [0,t] \cap \mathcal{C}(\mathsf{m})$, using (5.36),

$$\langle \phi(\cdot,z,x) - y, \mathsf{m}_s^N\rangle \to \langle \phi(\cdot,z,x) - y, \mathsf{m}_s\rangle \text{ and } \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z,x), \mathsf{m}_s^N\rangle \to \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z,x), \mathsf{m}_s\rangle$$

Since $[0,1]\backslash\mathcal{C}(\mathsf{m})$ is at most countable (see [Billingsley, 1999, p. 124]) we have that for a.e. $(s,z',z,x,y) \in [0,t]\times\mathbb{R}^d\times\mathbb{R}^d\times\mathsf{X}\times\mathsf{Y}$,

$$\langle \phi(\cdot,z',x) - y, \mathsf{m}_s^N\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z,x), \mathsf{m}_s^N\rangle \to \langle \phi(\cdot,z',x) - y, \mathsf{m}_s\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z,x), \mathsf{m}_s\rangle.$$

Since $\phi(\theta,z',x) - y$ is bounded and by (5.46), there exists $C > 0$ such that for all $(s,z',z,x,y) \in [0,t]\times\mathbb{R}^d\times\mathbb{R}^d\times\mathsf{X}\times\mathsf{Y}$, $\langle |\phi(\cdot,z',x) - y|, \mathsf{m}_s^N\rangle \langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z,x)|, \mathsf{m}_s^N\rangle \leq C\|\nabla_\theta f\|_{\infty,\Theta} \mathfrak{b}(z)$. By the dominated convergence theorem, we then have:

$$\begin{aligned}&\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ \underset{N\to+\infty}{\longrightarrow} &\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.\end{aligned}$$

With the same arguments as above, one shows that $\int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mathsf{m}_s^N\rangle \mathrm{d}s \to \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mathsf{m}_s\rangle \mathrm{d}s$. The proof of the lemma is complete. □

Proposition 51 (Convergence to the limit equation)**.** *Let $\mu^* \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$ be a limit point of $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, a.s. μ^* satisfies* (5.11).

Proof. Up to extracting a subsequence, we can assume that $\mu^N \xrightarrow{\mathscr{D}} \mu^*$ as $N \to +\infty$. Let $f \in \mathcal{C}^\infty(\Theta)$. The pre-limit equation (5.28) and Lemma 44 imply that a.s. for all $N \geq 1$ and $t \in [0,1]$, $\Lambda_t[f](\mu^N) \leq C/N + \mathbf{M}_t^N[f]$. Hence, using the last statement in Lemma 44, it holds for all $t \in [0,1]$,

$$\lim_{N\to\infty} \mathbf{E}[\Lambda_t[f](\mu^N)] = 0.$$

In particular, $\Lambda_t[f](\mu^N) \xrightarrow{\mathscr{D}} 0$. Let us now show that $\Lambda_t[f](\mu^N) \xrightarrow{\mathscr{D}} \Lambda_t[f](\mu^*)$. Denoting by $\mathsf{D}(\Lambda_t[f])$ the set of discontinuity points of $\Lambda_t[f]$, we have, from Proposition 49 and Lemma 50, for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$,

$$\mathbf{P}(\mu^* \in D(\Lambda_t[f])) = 0.$$

By the continuous mapping theorem, $\Lambda_t[f](\mu^N) \xrightarrow{\mathscr{D}} \Lambda_t[f](\mu^*)$. By uniqueness of the limit in distribution, we have that for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, a.s. $\Lambda_t[f](\mu^*) = 0$. Let us now prove that a.s. for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, $\Lambda_t[f](\mu^*) = 0$.

On the one hand, for all $f \in \mathcal{C}^\infty(\Theta)$ and $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$, the function $t \mapsto \Lambda_t[f](\mathsf{m})$ is right-continuous. Since $[0,1]$ is separable, we have that for all $f \in \mathcal{C}^\infty(\Theta)$, a.s. for all $t \in [0,1]$, $\Lambda_t[f](\mu^*) = 0$.

One the other hand $\mathcal{C}^\infty(\Theta)$ is separable (when endowed with the norm $\|f\|_{\mathcal{C}^\infty(\Theta)} = \sum_{k\geq 0} 2^{-k} \min(1, \sum_{|j|=k} \|\partial_j f\|_{\infty,\Theta})$) and the function $f \in \mathcal{C}^\infty(\Theta) \mapsto \Lambda_t[f](\mathsf{m})$ is continuous (for fixed $t \in [0,1]$ and $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$) relatively to the topology induced by $\|f\|_{\mathcal{C}^\infty(\Theta)}$.

Hence, we obtain that a.s. for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, $\Lambda_t[f](\mu^*) = 0$. The proof of the proposition is thus complete. □

8.2.4 Uniqueness and end of the proof of Theorem 40

Proposition 52. *There exists a unique solution to* (5.11) *in* $\mathcal{C}([0,1], \mathcal{P}(\Theta))$.

Proof. First of all, the fact that there is a solution to (5.11) is provided by Propositions 46, 49 and 51. The proof of the fact that there is a unique solution to (5.11) relies on the same arguments as those used in the proof of [Descours et al., 2022a, Proposition 2.14].

For $\mu \in \mathcal{P}(\mathbb{R}^{d+1})$, we introduce $v[\mu] : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ defined, for $\theta = (m, \rho) \in \mathbb{R}^{d+1}$, by

$$v[\mu](\theta) = -\eta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu \otimes \gamma \rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) - \eta \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0). \tag{5.37}$$

In addition, if $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ is solution to (5.11), it satisfies also (5.11) with test functions $f \in \mathcal{C}^\infty_c(\mathbb{R}^{d+1})$. Then, adopting the terminology of [Santambrogio, 2015, Section 4.1.2], any solution $\bar{\mu}$ to (5.11) is a *weak solution*[1] on $[0,T]$ of the measure-valued equation

$$\begin{cases} \partial_t \bar{\mu}_t = \mathrm{div}(v[\bar{\mu}_t]\bar{\mu}_t) \\ \bar{\mu}_0 = \mu_0. \end{cases} \tag{5.38}$$

Let us now prove that:

1. There exists $C > 0$ such that for all $\mu \in \mathcal{P}(\mathbb{R}^{d+1})$ and $\theta \in \mathbb{R}^{d+1}$,
$$|\mathrm{J}_\theta v[\mu](\theta)| \leq C.$$
2. There exists $C > 0$ such that for all $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ solution to (5.11), $0 \leq s, t \leq 1$, and $\theta \in \mathbb{R}^{d+1}$,
$$|v[\bar{\mu}_t](\theta) - v[\bar{\mu}_s](\theta)| \leq C|t-s|.$$
3. There exists $L' > 0$ such that for all $\mu, \nu \in P_1(\mathbb{R}^{d+1})$,
4. $$\sup_{\theta \in \mathbb{R}^d} |v[\mu](\theta) - v[\nu](\theta)| \leq L' \mathsf{W}_1(\mu, \nu).$$

[1] We mention that according to [Santambrogio, 2015, Proposition 4.2], the two notions of solutions of (5.38) (namely the weak solution and the *distributional* solution) are equivalent.

Before proving the three items above, we quickly conclude the proof of the proposition. Items 1 and 2 above imply that $v(t,\theta) = v[\bar{\mu}_t](\theta)$ is globally Lipschitz continuous over $[0,1]\times\mathbb{R}^{d+1}$ when $\bar{\mu}\in\mathcal{C}([0,1],\mathcal{P}(\Theta))$ is a solution to (5.11). Since $\bar{\mu}\in\mathcal{C}([0,1],\mathcal{P}(\Theta))\subset\mathcal{C}([0,1],\mathcal{P}(\mathbb{R}^{d+1}))$, this allows to use the representation theorem [Villani, 2021, Theorem 5.34] for the solution of (5.38) in $\mathcal{C}([0,1],\mathcal{P}(\mathbb{R}^{d+1}))$, i.e. it holds:

$$\forall t\in[0,1],\ \bar{\mu}_t = \phi_t\#\mu_0, \tag{5.39}$$

where ϕ_t is the flow generated by the vector field $v[\bar{\mu}_t](\theta)$ over $\mathbb{R}^{d+1}$. Equation (5.39) and the fact that $\mathcal{C}([0,1],\mathcal{P}(\Theta))\subset\mathcal{C}([0,1],\mathcal{P}_1(\mathbb{R}^{d+1}))$ together with Item 3 above and the same arguments as those used in the proof of [Descours et al., 2022a, Proposition 2.14] (which we recall is based estimates in Wasserstein distances between two solutions of (5.11) derived in [Piccoli and Rossi, 2016]), one deduces that there is a unique solution to (5.11).

Let us prove Item 1. Recall $g(\rho)=\ln(1+e^{\rho})$. The functions

$$\rho\mapsto g''(\rho)g(\rho),\ \rho\mapsto g'(\rho),\ \rho\mapsto\frac{g'(\rho)}{g(\rho)},\ \text{and}\ \rho\mapsto\frac{g''(\rho)}{g(\rho)}$$

are bounded on $\mathbb{R}$. Thus, in view of (5.4), $\|\mathrm{Hess}_\theta\mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)\|_{\infty,\mathbb{R}^{d+1}}<+\infty$. On the other hand, by **A1** and **A3**, for $x\in\mathsf{X}$, $z\in\mathbb{R}^d$, $\theta\in\Theta\mapsto\phi(\theta,z,x)$ is smooth and there exists $C>0$, for all $x\in\mathsf{X}$, $\theta\in\mathbb{R}^{d+1}$, $z\in\mathbb{R}^d$:

$$|\mathrm{Hess}_\theta\phi(\theta,z,x)|\le C(\mathfrak{b}(z)^2+\mathfrak{b}(z)).$$

This bound allows us to differentiate under the integral signs in (5.37) and proves that $|\mathrm{J}_\theta\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu\otimes\gamma\rangle\langle\nabla_\theta\phi(\theta,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)|\le C$, where $C>0$ is independent of $\mu\in\mathcal{P}(\Theta)$ and $\theta\in\Theta$. The proof of Item 1 is complete.

Let us prove Item 2. Let $\bar{\mu}\in\mathcal{C}([0,1],\mathcal{P}(\Theta))$ be a solution to (5.11), $0\le s\le t\le 1$, and $\theta\in\mathbb{R}^{d+1}$. We have

$$v[\bar{\mu}_t](\theta)-v[\bar{\mu}_s](\theta) = -\eta\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),(\bar{\mu}_t-\bar{\mu}_s)\otimes\gamma\rangle\langle\nabla_\theta\phi(\theta,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y). \tag{5.40}$$

Let $z\in\mathbb{R}^d$ and $x\in\mathsf{X}$. By **A1** and **A3**, $\phi(\cdot,z,x)\in\mathcal{C}^\infty(\Theta)$. Therefore, by (5.11),

$$\begin{aligned}\langle\phi(\cdot,z,x),\bar{\mu}_t-\bar{\mu}_s\rangle &= -\eta\int_s^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x')-y,\bar{\mu}_r\otimes\gamma\rangle\langle\nabla_\theta\phi(\cdot,z,x)\cdot\nabla_\theta\phi(\cdot,\cdot,x'),\bar{\mu}_r\otimes\gamma\rangle\pi(\mathrm{d}x',\mathrm{d}y)\mathrm{d}r\\ &\quad-\eta\int_s^t\langle\nabla_\theta\phi(\cdot,z,x)\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\bar{\mu}_r\rangle\mathrm{d}r\end{aligned}$$

We have $\|\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)\|_{\infty,\Theta}<+\infty$. Using also (5.46) and the fact that $\mathsf{X}\times\mathsf{Y}$ is a compact (see **A2**), it holds:

$$|\langle\phi(\cdot,z,x),\bar{\mu}_t-\bar{\mu}_s\rangle|\le C\mathfrak{b}(z)|t-s|.$$

Hence, for all $x'\in\mathsf{X}$,

$$|\langle\phi(\cdot,\cdot,x'),(\bar{\mu}_t-\bar{\mu}_s)\otimes\gamma\rangle|\le\langle|\langle\phi(\cdot,\cdot,x'),\bar{\mu}_t-\bar{\mu}_s\rangle|,\gamma\rangle\le C|t-s|.$$

Thus, by (5.40) and (5.47), $|v[\bar{\mu}_t](\theta)-v[\bar{\mu}_s](\theta)|\le C|t-s|$. This ends the proof of Item 2.

Let us now prove Item 3. Fix $\mu,\nu\in P_1(\mathbb{R}^{d+1})$ and $\theta\in\mathbb{R}^{d+1}$. We have

$$v[\mu](\theta)-v[\nu](\theta) = -\eta\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),(\mu-\nu)\otimes\gamma\rangle\langle\nabla_\theta\phi(\theta,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y) \tag{5.41}$$

For all $x\in\mathsf{X}$, using (5.32) and (5.46), it holds:

$$\begin{aligned}|\langle\phi(\cdot,\cdot,x),(\mu-\nu)\otimes\gamma\rangle| &\le\int_{\mathbb{R}^d}|\langle\phi(\cdot,z,x),\mu\rangle-\langle\phi(\cdot,z,x),\nu\rangle|\gamma(z)\mathrm{d}z\\ &\le C\int_{\mathbb{R}^d}\mathsf{W}_1(\mu,\nu)\mathfrak{b}(z)\gamma(z)\mathrm{d}z\le C\,W_1(\mu,\nu).\end{aligned}$$

Finally, using in addition (5.47) and (5.41), we deduce Item 3.

This ends the proof of the proposition. □

We are now ready to prove Theorem 40.

Proof. [Proof of Theorem 40] Recall Lemma 39 ensures that a.s. $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$. By Proposition 46, this sequence is relatively compact. Let $\mu^* \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$ be a limit point. Along some subsequence N', it holds:

$$\mu^{N'} \xrightarrow{\mathscr{D}} \mu^*.$$

In addition, a.s. $\mu^* \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$ (by Proposition 49) and μ^* satisfies (5.11) (by Proposition 51). By Proposition 52, (5.11) admits a unique solution $\bar{\mu} \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$. Hence, a.s. $\mu^* = \bar{\mu}$. Therefore,

$$\mu^{N'} \xrightarrow{\mathscr{D}} \bar{\mu}.$$

Since the sequence $(\mu^N)_{N\geq 1}$ admits a unique limit point, the whole sequence converges in distribution to $\bar{\mu}$. The convergence also holds in probability since $\bar{\mu}$ is deterministic. The proof of Theorem 40 is complete. □

8.3 Proof of Lemma 39

In this section we prove Lemma 39. We start with the following simple result.

Lemma 53. *Let $T > 0$, $N \geq 1$, and $c_1 > 0$. Consider a sequence $(u_k)_{0\leq k\leq \lfloor NT\rfloor} \subset \mathbb{R}_+$ for which there exists v_0 such that $u_0 \leq v_0$ and for all $1 \leq k \leq \lfloor NT\rfloor$, $u_k \leq c_1(1 + \frac{1}{N}\sum_{\ell=0}^{k-1} u_\ell)$. Then, for all $0 \leq k \leq \lfloor NT\rfloor$, $u_k \leq v_0 e^{c_1 T}$.*

Proof. Define $v_k = c_1(1 + \frac{1}{N}\sum_{\ell=0}^{k-1} v_\ell)$. For all $0 \leq k \leq \lfloor NT\rfloor$, $u_k \leq v_k$ and $v_k = v_{k-1}(1 + c_1/N)$. Hence $v_k = v_0\big(1 + c_1/N\big)^k \leq v_0\big(1 + c_1/N\big)^{\lfloor NT\rfloor} \leq v^0 e^{c_1 T}$. This ends the proof of the Lemma. □

Proof. [Proof of Lemma 39] Since $\rho \mapsto g'(\rho)$ and $\rho \mapsto g'(\rho)/g(\rho)$ are bounded continuous functions over $\mathbb{R}$, and since $|g(\rho)| \leq C(1 + |\rho|)$, according to (5.4), there exists $c > 0$, for all $\theta \in \mathbb{R}^{d+1}$,

$$|\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \leq c(1 + |\theta|). \tag{5.42}$$

All along the proof, $C > 0$ is a constant independent of $N \geq 1$, $T > 0$, $i \in \{1,\dots,N\}$, $1 \leq k \leq \lfloor NT\rfloor$, $(x,y) \in X \times Y$, $\theta \in \mathbb{R}^{d+1}$, and $z \in \mathbb{R}^d$, which can change from one occurence to another. It holds:

$$|\theta_k^i| \leq |\theta_0^i| + \sum_{\ell=0}^{k-1} |\theta_{\ell+1}^i - \theta_\ell^i|. \tag{5.43}$$

Using (5.5), we have, for $0 \leq \ell \leq k-1$,

$$\begin{aligned} |\theta_{\ell+1}^i - \theta_\ell^i| \leq \frac{\eta}{N^2} \sum_{j=1, j\neq i}^{N} \Big| (\langle \phi(\theta_\ell^j, \cdot, x_\ell), \gamma\rangle - y_\ell) \langle \nabla_\theta \phi(\theta_\ell^i, \cdot, x_\ell), \gamma\rangle \Big| \\ + \frac{\eta}{N^2} \Big| \Big\langle (\phi(\theta_\ell^i, \cdot, x_\ell) - y_\ell) \nabla_\theta \phi(\theta_\ell^i, \cdot, x_\ell), \gamma \Big\rangle \Big| + \frac{\eta}{N} |\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_\ell^i}^1 | P_0^1)|. \end{aligned} \tag{5.44}$$

For all $\theta \in \mathbb{R}^{d+1}$, $z \in \mathbb{R}^d$, $(x,y) \in \mathsf{X} \times \mathsf{Y}$, we have, by **A2** and **A3**, since $\phi(\theta, z, x) = s(\Psi_\theta(z), x)$,

$$|\phi(\theta, z, x) - y| \leq C. \tag{5.45}$$

Moreover, we have $\nabla_\theta \phi(\theta, z, x) = \nabla_1 s(\Psi_\theta(z), x) \mathrm{J}_\theta \Psi_\theta(z)$ (here $\nabla_1 s$ refers to the gradient of s w.r.t. its first variable). By **A3**, $|\nabla_1 s(\Psi_\theta(z), x)| \leq C$ and, hence, denoting by J_θ the Jacobian w.r.t. θ, using (5.10),

$$|\nabla_\theta \phi(\theta, z, x)| \leq C |\mathrm{J}_\theta \Psi_\theta(z)| \leq C \mathfrak{b}(z). \tag{5.46}$$

Therefore, by (5.10),

$$\langle |\nabla_\theta \phi(\theta, \cdot, x)|, \gamma \rangle \leq C. \tag{5.47}$$

Hence, we obtain, using (5.44) and (5.42),

$$|\theta^i_{\ell+1} - \theta^i_\ell| \leq \frac{\eta}{N^2} \sum_{j=1, j\neq i}^{N} C + \frac{\eta}{N^2} C + \frac{c\eta}{N}(1 + |\theta^i_\ell|) \leq \frac{C}{N}(1 + |\theta^i_\ell|). \tag{5.48}$$

Using **A4**, there exists $K_0 > 0$ such that a.s. for all i, $|\theta^i_0| \leq K_0$. Then, from (5.43) and (5.48), for $1 \leq k \leq \lfloor NT \rfloor$, it holds:

$$|\theta^i_k| \leq K_0 + \frac{C}{N} \sum_{\ell=0}^{k-1} (1 + |\theta^i_\ell|) \leq K_0 + CT + \frac{C}{N} \sum_{\ell=0}^{k-1} |\theta^i_\ell| \leq C_{0,T}(1 + \frac{1}{N} \sum_{\ell=0}^{k-1} |\theta^i_\ell|),$$

with $C_{0,T} = \max(K_0 + CT, C) \leq K_0 + C(1+T)$. Then, by Lemma 53 and **A4**, we have that for all $N \geq 1$, $i \in \{1, \ldots, N\}$ and $0 \leq k \leq \lfloor NT \rfloor$, $|\theta^i_k| \leq K_0 e^{[K_0 + C(1+T)]T}$. The proof of Lemma 39 is thus complete. □

9 Proof of Theorem 41

In this section, we assume **A1** → **A5** (where in **A2**, when $k \geq 1$, $\mathcal{F}^N_k$ is now the one defined in (5.12)) and the θ^i_k's (resp. μ^N) are those defined by (5.7) for $i \in \{1, \ldots, N\}$ and $k \geq 0$ (resp. by (5.13) for $N \geq 1$).

9.1 Preliminary analysis and pre-limit equation

9.2 Notation and weighted Sobolev embeddings

For $J \in \mathbf{N}$ and $\beta \geq 0$, let $\mathcal{H}^{J,\beta}(\mathbb{R}^{d+1})$ be the closure of the set $\mathcal{C}^\infty_c(\mathbb{R}^{d+1})$ for the norm

$$\|f\|_{\mathcal{H}^{J,\beta}} := \Big(\sum_{|k| \leq J} \int_{\mathbb{R}^{d+1}} \frac{|\partial_k f(\theta)|^2}{1 + |\theta|^{2\beta}} \mathrm{d}\theta \Big)^{1/2}.$$

The space $\mathcal{H}^{J,\beta}(\mathbb{R}^{d+1})$ is a separable Hilbert space and we denote its dual space by $\mathcal{H}^{-J,\beta}(\mathbb{R}^{d+1})$ (see e.g. [Fernandez and Méléard, 1997, Jourdain and Méléard, 1998]). The associated scalar product on $\mathcal{H}^{J,\beta}(\mathbb{R}^{d+1})$ will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}^{J,\beta}}$. For $\Phi \in \mathcal{H}^{-J,\beta}(\mathbb{R}^{d+1})$, we use the notation

$$\langle f, \Phi \rangle_{J,\beta} = \Phi[f],\ f \in \mathcal{H}^{J,\beta}(\mathbb{R}^{d+1}).$$

For ease of notation, and if no confusion is possible, we simply denote $\langle f, \Phi \rangle_{J,\beta}$ by $\langle f, \Phi \rangle$. The set $\mathcal{C}^{J,\beta}_0(\mathbb{R}^{d+1})$ (resp. $\mathcal{C}^{J,\beta}(\mathbb{R}^{d+1})$) is defined as the space of functions $f : \mathbb{R}^{d+1} \to \mathbb{R}$ with continuous partial derivatives up to order $J \in \mathbf{N}$ such that

$$\text{for all } |k| \leq J,\ \lim_{|\theta| \to \infty} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta} = 0 \ \ (\text{resp. } \sum_{|k| \leq J} \sup_{\theta \in \mathbb{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta} < +\infty).$$

The spaces $\mathcal{C}^{J,\beta}(\mathbb{R}^{d+1})$ and $\mathcal{C}^{J,\beta}_0(\mathbb{R}^{d+1})$ is endowed with the norm

$$\|f\|_{\mathcal{C}^{J,\beta}} := \sum_{|k| \leq J} \sup_{\theta \in \mathbb{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta}.$$

We note that

$$\theta \in \mathbb{R}^{d+1} \mapsto (1 - \chi(\theta))|\theta|^\alpha \in\ H^{J,\beta}(\mathbb{R}^{d+1}) \text{ if } \beta - \alpha > (d+1)/2, \tag{5.49}$$

where $\chi \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})$ equals 1 near 0. We recall that from [Fernandez and Méléard, 1997, Section 2], for $m' > (d+1)/2$ and $\alpha, j \geq 0$, $\mathcal{H}^{m'+j,\alpha}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}_0^{j,\alpha}(\mathbb{R}^{d+1})$. In the following, we consider $\gamma_0, \gamma_1 \in \mathbb{R}$ and $L_0 \in \mathbf{N}$ such that

$$\gamma_1 > \gamma_0 > \frac{d+1}{2} + 1 \text{ and } L_0 > \frac{d+1}{2} + 1.$$

We finally recall the following standard result.

Proposition 54. *Let $q > p \geq 1$ and $C > 0$. The set $\mathscr{K}_C^q := \{\mu \in \mathcal{P}_p(\mathbb{R}^{d+1}), \int_{\mathbb{R}^{d+1}} |x|^q \mu(\mathrm{d}x) \leq C\}$ is compact.*

9.3 Bound on the moments of the θ_k^i's

We have the following uniform bound in $N \geq 1$ on the moments of the sequence $\{\theta_k^i, i \in \{1,\dots,N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ defined by (5.7).

Lemma 55. *Assume **A1**$\to$ **A5**. For all $T > 0$ and $p \geq 1$, there exists $C > 0$ such that for all $N \geq 1$, $i \in \{1,\dots,N\}$ and $0 \leq k \leq \lfloor NT \rfloor$,*

$$\mathbf{E}[|\theta_k^i|^p] \leq C.$$

Proof. Let $p \geq 1$. By **A4**, $\mathbf{E}[|\theta_0^i|^p] \leq C_p$ for all $i \in \{1,\dots,N\}$. Let $T > 0$. In the following $C > 0$ is a constant independent of $N \geq 1$, $i \in \{1,\dots,N\}$, and $1 \leq k \leq \lfloor NT \rfloor$. Using (5.7), the fact that ϕ is bounded, Y is bounded, and (5.46), we have, for $0 \leq n \leq k-1$,

$$\begin{aligned}
|\theta_{n+1}^i - \theta_n^i| &\leq \frac{C}{N^2 B} \sum_{j=1}^N \sum_{\ell=1}^B \mathfrak{b}(\mathsf{Z}_n^{i,\ell}) + \frac{C}{N} |\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta_n^i} | P_0^1)| \\
&\leq \frac{C}{NB} \sum_{\ell=1}^B (1 + \mathfrak{b}(\mathsf{Z}_n^{i,\ell})) + \frac{C}{N}(1 + |\theta_n^i|), \qquad (5.50)
\end{aligned}$$

where we have also used (5.42) for the last inequality. Let us recall the following convexity inequality: for $m, p \geq 1$ and $x_1, \dots, x_p \in \mathbb{R}_+$,

$$\Big(\sum_{n=1}^m x_n \Big)^p \leq m^{p-1} \sum_{n=1}^m x_n^p. \qquad (5.51)$$

Using (5.43), **A1** with $q = p$, and the fact that $1 \leq k \leq \lfloor NT \rfloor$, one has setting $u_k = \mathbf{E}[|\theta_k^i|^p]$, $u_k \leq C(1 + \frac{1}{N}\sum_{n=0}^{k-1} u_n)$. The result then follows from Lemma 53. □

9.4 Pre-limit equation

In this section, we derive the pre-limit equation for μ^N defined by (5.13). For simplicity we will keep the same notations as those introduced in Section 8.1.1, though these objects will now be defined with θ_k^i set by (5.7), and on $\mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$, for all integer $k \geq 0$, and all time $t \geq 0$. Let $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$. Then, set for $k \geq 0$,

$$\begin{aligned}
\mathbf{D}_k^N[f] = &-\frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\langle \phi(\theta_k^j, \cdot, x), \gamma \rangle - y \big) \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&- \frac{\eta}{N^2} \int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \nu_k^N \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y).
\end{aligned}$$

Note that $\mathbf{D}_k^N$ above is the one defined in (5.21) but now on $\mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$ and with θ_k^i defined by (5.7). For $k \geq 0$, we set

$$\mathbf{M}_k^N[f] = -\frac{\eta}{N^3 B} \sum_{i,j=1}^N \sum_{\ell=1}^B (\phi(\theta_k^j, \mathsf{Z}_k^{j,\ell}, x_k) - y_k) \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \mathsf{Z}_k^{i,\ell}, x_k) - \mathbf{D}_k^N[f]. \qquad (5.52)$$

By Lemma 55 together with (5.45) and (5.46), $\mathbf{M}_k^N[f]$ is integrable. Also, using **A5** and the fact that θ_k^j is $\mathcal{F}_k^N$-measurable (see (5.12)),

$$\mathbf{E}[\mathbf{M}_k^N[f]|\mathcal{F}_k^N] = 0.$$

Set $\mathbf{M}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]$, $t \geq 0$. We now extend the definition of $\mathbf{W}_t^N[f]$ and $\mathbf{R}_k^N[f]$ in (5.27) and (5.19) to any time $t \geq 0$, $k \geq 0$, and $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$, and with θ_k^i set by (5.7). We then set

$$\mathbb{R}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbf{R}_k^N[f],\ t \geq 0.$$

With the same algebraic computations as those made in Section 8.1.1, one obtains the following pre-limit equation: for $N \geq 1$, $t \geq 0$, and $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$,

$$\begin{aligned}
\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle &= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mu_s^N\big\rangle \mathrm{d}s \\
&\quad + \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad - \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad + \mathbf{M}_t^N[f] + \mathbf{W}_t^N[f] + \mathbb{R}_t^N[f].
\end{aligned} \tag{5.53}$$

We will now show that the sequence $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$.

9.5 Relative compactness and convergence to the limit equation

9.6 Relative compactness in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$

In this section we prove the following result.

Proposition 56. *Assume **A1**$\to$ **A5**. Recall $\gamma_0 > \frac{d+1}{2} + 1$. Then, the sequence $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$.*

We start with the following lemma.

Lemma 57. *Assume **A1**$\to$ **A5**. Then, $\forall T > 0$ and $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$,*

$$\sup_{N\geq 1} \mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \mu_t^N\rangle^2\Big] < +\infty.$$

Proof. Let $T > 0$. In what follows, $C > 0$ is a constant independent of $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$, $(s,t) \in [0,T]^2$, and $z \in \mathbb{R}^d$ which can change from one occurence to another. We have by **A4**, $\mathbf{E}[\langle f, \mu_0^N\rangle^2] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}$. By (5.53) and (5.45), it holds:

$$\begin{aligned}
\sup_{t\in[0,T]} \langle f, \mu_t^N\rangle^2 &\leq C\Big[\|f\|^2_{\mathcal{C}^{2,\gamma_1}} + \int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \big|\langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma\rangle, \mu_s^N\Big\rangle\Big|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad \int_0^T \big|\big\langle |\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1)|, \mu_s^N\big\rangle\big|^2 \mathrm{d}s \\
&\quad + \frac{1}{N^2} \int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \big|\langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma\rangle, \mu_s^N\Big\rangle\Big|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad + \sup_{t\in[0,T]} |\mathbf{M}_t^N[f]|^2 + \sup_{t\in[0,T]} |\mathbf{W}_t^N[f]|^2 + \sup_{t\in[0,T]} |\mathbb{R}_t^N[f]|^2.\Big].
\end{aligned} \tag{5.54}$$

We have using (5.46), for $s \in [0,T]$ and $z \in \mathbb{R}^d$,

$$|\nabla_\theta f(\theta^i_{\lfloor Ns \rfloor}) \cdot \nabla_\theta \phi(\theta^i_{\lfloor Ns \rfloor}, z, x)| \le C\|f\|_{\mathcal{C}^{1,\gamma_1}} \mathfrak{b}(z)(1 + |\theta^i_{\lfloor Ns \rfloor}|^{\gamma_1}). \tag{5.55}$$

Thus, using Lemma 55,

$$\mathbf{E}\big[\big\langle \langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma \rangle, \mu^N_s \big\rangle^2\big] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}. \tag{5.56}$$

Using (5.42), for $s \in [0,T]$, it holds:

$$\big|\nabla_\theta f(\theta^i_{\lfloor Ns \rfloor}) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_{\lfloor Ns \rfloor}} | P^1_0)\big| \le C\|f\|_{\mathcal{C}^{1,\gamma_1}}(1 + |\theta^i_{\lfloor Ns \rfloor}|^{\gamma_1+1}). \tag{5.57}$$

Thus, using Lemma 55,

$$\mathbf{E}\big[\big|\big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \mu^N_s \big\rangle\big|^2\big] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}. \tag{5.58}$$

On the other hand, we have using (5.51):

$$\sup_{t\in[0,T]} |\mathbf{M}^N_t[f]|^2 \le \lfloor NT \rfloor \sum_{k=0}^{\lfloor NT \rfloor - 1} |\mathbf{M}^N_k[f]|^2. \tag{5.59}$$

Recall (5.52). By (5.21), (5.51), **A1**, and (5.55), it holds:

$$|\mathbf{D}^N_k[f]|^2 \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}\big[\frac{1}{N^4}\sum_{i\neq j=1}^N (1 + |\theta^i_k|^{2\gamma_1}) + \frac{1}{N^4}(1 + \langle |\cdot|^{2\gamma_1}, \nu^N_k \rangle)\big] \le \frac{C}{N^2}\|f\|^2_{\mathcal{C}^{1,\gamma_1}}(1 + |\theta^i_k|^{2\gamma_1})$$

and

$$|\mathbf{M}^N_k[f]|^2 \le \frac{C}{N^4 B}\sum_{i,j=1}^N \sum_{\ell=1}^B \|f\|^2_{\mathcal{C}^{1,\gamma_1}} |\mathfrak{b}(\mathsf{Z}^{i,\ell}_k)|^2 (1 + |\theta^i_{\lfloor Ns \rfloor}|^{2\gamma_1}) + |\mathbf{D}^N_k[f]|^2.$$

By Lemma 55 and **A1**, one deduces that

$$\mathbf{E}[|\mathbf{M}^N_k[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N^2. \tag{5.60}$$

Going back to (5.59), we then have $\mathbf{E}[\sup_{t\in[0,T]} |\mathbf{M}^N_t[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}$. Using the same arguments as those used so far, one also deduces that for $t \in [0,T]$

$$\begin{aligned}
\sup_{t\in[0,T]} |\mathbf{W}^N_t[f]|^2 &\le \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2} \sup_{t\in[0,T]} \big(1 + \langle |\cdot|^{\gamma_1+1}, \nu^N_{\lfloor Nt \rfloor} \rangle\big)^2 \\
&= \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2} \max_{0\le k\le \lfloor NT \rfloor} \big(1 + \langle |\cdot|^{\gamma_1+1}, \nu^N_k \rangle\big)^2 \\
&\le \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2} \sum_{k=0}^{\lfloor NT \rfloor} \big(1 + \langle |\cdot|^{\gamma_1+1}, \nu^N_k \rangle\big)^2.
\end{aligned}$$

and thus

$$\mathbf{E}\Big[\sup_{t\in[0,T]} |\mathbf{W}^N_t[f]|^2\Big] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N. \tag{5.61}$$

Let us finally deal with the term involving $\mathbb{R}^N_t[f]$. One has using (5.51):

$$\sup_{t\in[0,T]} |\mathbb{R}^N_t[f]|^2 \le \lfloor NT \rfloor \sum_{k=0}^{\lfloor NT \rfloor - 1} |\mathbb{R}_k[f]|^2.$$

For $0 \le k \le \lfloor NT \rfloor - 1$, we have, from (5.19),

$$|\mathbb{R}_k^N[f]|^2 \le \frac{C\|f\|_{\mathcal{C}^{2,\gamma_1}}^2}{N} \sum_{i=1}^N |\theta_{k+1}^i - \theta_k^i|^4 (1 + |\hat{\theta}_k^i|^{\gamma_1})^2$$
$$\le \frac{C\|f\|_{\mathcal{C}^{2,\gamma_1}}^2}{N} \sum_{i=1}^N |\theta_{k+1}^i - \theta_k^i|^4 (1 + |\theta_{k+1}^i|^{2\gamma_1} + |\theta_k^i|^{2\gamma_1}).$$

Using (5.50),

$$|\theta_{k+1}^i - \theta_k^i|^4 \le C\Big[\frac{1}{N^4} + \frac{|\theta_k^i|^4}{N^4} + \frac{1}{N^4 B} \sum_{\ell=1}^B |\mathfrak{b}(\mathsf{Z}_k^{i,\ell})|^4\Big].$$

By Lemma 55 and **A1**, it then holds $\mathbf{E}[|\theta_{k+1}^i - \theta_k^i|^4 (1 + |\theta_{k+1}^i|^{2\gamma_1} + |\theta_k^i|^{2\gamma_1})] \le C/N^4$. Hence, one deduces that

$$\mathbf{E}[\sup_{t\in[0,T]} |\mathbb{R}_t^N[f]|^2] \le C\|f\|_{\mathcal{C}^{2,\gamma_1}}^2 / N^2. \tag{5.62}$$

This ends the proof of Lemma 57. □

Lemma 58 (Compact containment for $(\mu^N)_{N\ge1}$)**.** *Assume **A1**→ **A5**. Let $0 < \epsilon < \gamma_1 - \gamma_0$. For every $T > 0$,*

$$\sup_{N\ge1} \mathbf{E}\Big[\sup_{t\in[0,T]} \int_{\mathbb{R}^{d+1}} |x|^{\gamma_0+\epsilon} \mu_t^N(\mathrm{d}x)\Big] < +\infty. \tag{5.63}$$

Proof. Apply Lemma 57 with $f : \theta \mapsto (1-\chi)|\theta|^{\gamma_0+\epsilon} \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$. □

Lemma 59. *Assume **A1**→**A5**. Let $T > 0$ and $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$. Then, there exists $C > 0$ such that for all $\delta > 0$ and $0 \le r < t \le T$ such that $t - r \le \delta$, one has for all $N \ge 1$,*

$$\mathbf{E}\big[|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle|^2\big] \le C(\delta^2 + \delta/N + 1/N).$$

Proof. Using (5.53), Jensen's inequality, (5.45), (5.56), and (5.58), one has for $f \in \mathcal{C}^{2,\gamma_1}(\mathbb{R}^{d+1})$,

$$\mathbf{E}\big[|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle|^2\big] \le C\Big[(t-r)^2(1+1/N^2)\|f\|_{\mathcal{C}^{1,\gamma_1}}^2 + \mathbf{E}\big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]\Big|^2\big]\Big]$$
$$+ \mathbf{E}\big[\Big|\mathbf{W}_t^N[f] - \mathbf{W}_r^N[f]\Big|^2\big] + \mathbf{E}\big[\Big|\mathbb{R}_t^N[f] - \mathbb{R}_r^N[f]\Big|^2\big]. \tag{5.64}$$

We also have with the same arguments as those used just before (5.31)

$$\mathbf{E}\big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]\Big|^2\big] = \sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{E}[|\mathbf{M}_k^N[f]|^2].$$

Using in addition (5.60), one has $\mathbf{E}\big[\big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]\big|^2\big] \le C(N\delta+1)\|f\|_{\mathcal{C}^{1,\gamma_1}}^2/N^2$. Note that with this argument, we also deduce that

$$\mathbf{E}[|\mathbf{M}_t^N[f]|^2] \le C\|f\|_{\mathcal{C}^{1,\gamma_1}}^2/N. \tag{5.65}$$

On the other hand, by (5.61) and (5.62), one has

$$\mathbf{E}\big[\Big|\mathbf{W}_t^N[f] - \mathbf{W}_r^N[f]\Big|^2\big] \le C\|f\|_{\mathcal{C}^{1,\gamma_1}}^2/N \text{ and } \mathbf{E}\big[\Big|\mathbb{R}_t^N[f] - \mathbb{R}_r^N[f]\Big|^2\big] \le C\|f\|_{\mathcal{C}^{2,\gamma_1}}^2/N^2.$$

One then plugs all the previous estimates in (5.64) to deduce the result of Lemma 59. □

We are now in position to prove Proposition 56. **Proof.** [Proof of Proposition 56] The proof consists in applying [Jakubowski, 1986, Theorem 4.6] with $E = \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1})$ and $\mathbb{F} = \{\mathsf{H}_f, f \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})\}$ where

$$\mathsf{H}_f : \nu \in \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}) \mapsto \langle f, \nu \rangle.$$

The set $\mathbb{F}$ on $\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1})$ satisfies Conditions [Jakubowski, 1986, (3.1) and (3.2) in Theorem 3.1]. Condition (4.8) there follows from Proposition 54, Lemma 58, and Markov's inequality. Let us now show [Jakubowski, 1986, Condition (4.9)] is verified, i.e. that for all $f \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})$, the family $(\langle f, \mu^N \rangle)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathbb{R})$. To do this, it suffices to use Lemma 59 and [Descours et al., 2022a, Proposition A.1] (with $\mathcal{H}_1 = \mathcal{H}_2 = \mathbb{R}$ there). In conclusion, according to [Jakubowski, 1986, Theorem 4.6], the sequence $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ is relatively compact. □

9.7 Limit points satisfy the limit equation (5.15)

For $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbb{R}^{d+1})$ and $t \geq 0$, we introduce for $\mathsf{m} \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$,

$$\begin{aligned}\Phi_t[f] : \mathsf{m} \mapsto \Big| &\langle f, \mathsf{m}_t \rangle - \langle f, \mu_0 \rangle \\ &+ \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &+ \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s \big\rangle \mathrm{d}s \Big|. \end{aligned} \tag{5.66}$$

Note that $\Phi_t[f]$ is the function $\Lambda_t[f]$ previously defined in (5.35) for test functions $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbb{R}^{d+1})$ and for $\mathsf{m} \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$.

Lemma 60. *Assume **A1**→ **A5**. Let $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbb{R}^{d+1})$. Then $\Phi_t[f]$ is well defined. In addition, if a sequence $(\mathsf{m}^N)_{N\geq 1}$ converges to m in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$, then, for all continuity point $t \geq 0$ of m, we have $\Phi_t[f](\mathsf{m}^N) \to \Phi_t[f](\mathsf{m})$.*

Proof. Using **A1**, and because Y is bounded and the function ϕ is bounded, $\mathscr{G}_1^{x,y} : \theta \mapsto \langle \phi(\theta,\cdot,x) - y, \gamma \rangle \in \mathcal{C}_b^\infty(\mathbb{R}^{d+1})$. In addition, for all multi-index $\alpha \in \mathbf{N}^{d+1}$, there exists $C > 0$, for all $x, y \in X \times Y$ and all $\theta \in \mathbb{R}^{d+1}$, $|\partial_\alpha \mathscr{G}_1^{x,y}(\theta)| \leq C$. The same holds for the function $\mathscr{G}_2^x : \theta \in \mathbb{R}^{d+1} \mapsto \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle$. Consequently, $\theta \mapsto \nabla_\theta f(\theta) \cdot \mathscr{G}_2^x(\theta) \in \mathcal{C}^{0,\gamma_0-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{0,\gamma_0}(\mathbb{R}^{d+1})$. Then, there exists $C > 0$ independent of $(x, y) \in X \times Y$ and $s \in [0, t]$ such that

$$|\langle \mathscr{G}_1^{x,y}, \mathsf{m}_s \rangle| \leq C,$$

and

$$|\langle \nabla_\theta f \cdot \mathscr{G}_2^x, \mathsf{m}_s \rangle| \leq C \|f\|_{\mathcal{C}^{1,\gamma_0-1}} \langle 1 + |.|^{\gamma_0},\, m_s \rangle.$$

Finally, the function $\theta \mapsto \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)$ is smooth (see (6.3)) and (5.42) extends to all its derivatives, i.e. for all multi-index $\alpha \in \mathbf{N}^{d+1}$, there exists $c > 0$, for all $\theta \in \mathbb{R}^{d+1}$,

$$|\partial_\alpha \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \leq c(1 + |\theta|).$$

Thus, $\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) \in \mathcal{C}^{0,\gamma_0}(\mathbb{R}^{d+1})$ and for some $C > 0$ independent of $s \in [0, t]$

$$|\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s \rangle| \leq C \|f\|_{\mathcal{C}^{1,\gamma_0-1}} \langle 1 + |.|^{\gamma_0},\, m_s \rangle.$$

Since in addition $\sup_{s\in[0,t]} \langle 1 + |.|^{\gamma_0},\, m_s \rangle < +\infty$ (since $s \mapsto \langle 1 + |.|^{\gamma_0},\, m_s \rangle \in \mathcal{D}(\mathbb{R}_+, \mathbb{R})$), $\Phi_t[f]$ is well defined. To prove the continuity property of $\Phi_t[f]$ it then suffices to use the previous upper bounds together similar arguments as those used in the proof of Lemma 50 (see also [Descours et al., 2022a]). □

Proposition 61. *Assume **A1**→**A5**. Let μ^* be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. Then, μ^* satisfies a.s. Equation* (5.15)*.*

Proof. Let us consider $f \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})$ and μ^* be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. Recall that by [Ethier and Kurtz, 2009, lemma 7.7 in Chapter 3], the complementary of the set

$$\mathcal{C}(\mu^*) = \{t \geq 0, \mathbf{P}(\mu^*_{t-} = \mu^*_t) = 1\}$$

is at most countable. Let $t_* \in \mathcal{C}(\mu^*)$. Then, by Lemma 60, one has that $\mathbf{P}(\mu^* \in \mathsf{D}(\Phi_{t_*}[f])) = 0$. Thus, by the continuous mapping theorem, it holds

$$\Phi_{t_*}[f](\mu^N) \xrightarrow{\mathscr{D}} \Phi_{t_*}[f](\mu^*).$$

On the other hand, using (6.25) and the estimates (5.62), (5.61), (5.65), (5.56), and (5.58), it holds

$$\lim_{N\to\infty} \mathbf{E}[\Phi_{t_*}[f](\mu^N)] = 0.$$

Consequently, for all $f \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})$ and $t_* \in \mathcal{C}(\mu^*)$, it holds a.s. $\Phi_{t_*}[f](\mu^*) = 0$. On the other hand, for all $\psi \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})$, $\mathsf{m} \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$, and $s \geq 0$, the mappings

$$t \geq 0 \mapsto \Phi_t[\psi](\mathsf{m})$$

is right continuous, and

$$f \in \mathcal{H}^{L_0,\gamma_0-1}(\mathbb{R}^{d+1}) \mapsto \Phi_s[f](\mathsf{m})$$

is continuous (because $\mathcal{H}^{L_0,\gamma_0-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}_0^{1,\gamma_0-1}(\mathbb{R}^{d+1})$). In addition, $\mathcal{H}^{L_0,\gamma_0-1}(\mathbb{R}^{d+1})$ admits a dense and countable subset of elements in $\mathcal{C}_c^\infty(\mathbb{R}^{d+1})$. Moreover, there exists a countable subset $\mathcal{T}_{\mu^*}$ of $\mathcal{C}(\mu^*)$ such that for all $t \geq 0$ and $\epsilon > 0$, there exists $s \in \mathcal{T}_{\mu^*}$, $s \in [t, t+\epsilon]$. We prove this claim. Since $\mathbb{R}_+$ is a metric space, $\mathcal{C}(\mu^*)$ is separable and thus admits a dense subset $\mathcal{O}_{\mu^*}$. Since $[t+\epsilon/4, t+3\epsilon/4] \cap \mathcal{C}(\mu^*) \neq \emptyset$, there exists $u \in [t+\epsilon/4, t+3\epsilon/4] \cap \mathcal{C}(\mu^*)$. Consider now $s \in \mathcal{O}_{\mu^*}$ such that $|s-u| \leq \epsilon/4$. It then holds $t \leq s \leq t+\epsilon$, proving the claim with $\mathcal{T}_{\mu^*} = \mathcal{O}_{\mu^*}$.

Hence, we have with a classical argument that a.s. for all $f \in \mathcal{H}^{L_0,\gamma_0-1}(\mathbb{R}^{d+1})$ and $t \geq 0$, $\Lambda_t[f](\mu^*) = 0$. Note also that $\mathcal{C}_b^\infty(\mathbb{R}^{d+1}) \subset \mathcal{H}^{L_0,\gamma_0-1}(\mathbb{R}^{d+1})$ since $2\gamma_0 > d+1$. This ends the proof of the proposition. □

9.8 Uniqueness of the limit equation and end of the proof of Theorem 41

In this section, we prove that there is a unique solution to (5.15) in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$. To this end, we first need to prove that every limit points of $(\mu^N)_{N\geq 1}$ a.s. belongs to $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$.

9.9 Limit points belong to $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$

Proposition 62. *Assume **A1**→**A5**. Let $\mu^* \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. Then, a.s. $\mu^* \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$.*

Proof. Note that since $\mathsf{W}_1 \leq W_{\gamma_0}$, $\mu^{N'} \xrightarrow{\mathscr{D}} \mu^*$ also in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$, along some subsequence N'. According to [Jacod and Shiryaev, 1987, Proposition 3.26 in Chapter VI], $\mu^* \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ a.s. if for all $T > 0$, $\lim_{N\to+\infty} \mathbf{E}\big[\sup_{t\in[0,T]} W_1(\mu^N_{t-}, \mu^N_t)\big] = 0$. Using (5.32), this is equivalent to prove that

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\leq 1} |\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle|\Big] = 0. \tag{5.67}$$

Let us consider $T > 0$ and a Lipschitz function $f : \mathbb{R}^{d+1} \to \mathbb{R}$ such that $\|f\|_{\mathrm{Lip}} \leq 1$. We have $\langle f, \mu^N_t\rangle = \langle f, \mu^N_0\rangle + \sum_{k=0}^{\lfloor Nt\rfloor - 1} \langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle$ (with usual convention $\sum_0^{-1} = 0$). Thus the discontinuity points of

$t \in [0,T] \mapsto \langle f, \mu_t^N \rangle$ lies exactly at $\{1/N, 2/N, \dots, \lfloor NT \rfloor / N\}$ and

$$|\langle f, \mu_{t_-}^N \rangle - \langle f, \mu_t^N \rangle| \leq \max_{k=0,\dots,\lfloor NT \rfloor - 1} |\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle|, \ \ \forall t \in [0,T], \, f \text{ Lipschitz}. \tag{5.68}$$

Pick $k = 0, \dots, \lfloor NT \rfloor - 1$. We have by (5.50),

$$|\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle| \leq \frac{1}{N} \sum_{i=1}^N |\theta_{k+1}^i - \theta_k^i| \leq \frac{C}{N} \sum_{i=1}^N \Big[\frac{1}{NB} \sum_{\ell=1}^B (1 + \mathfrak{b}(\mathsf{Z}_k^{i,\ell})) + \frac{1}{N}(1 + |\theta_k^i|) \Big] =: d_k^N \tag{5.69}$$

Hence, it holds:

$$|d_k^N|^2 \leq \frac{C}{N} \sum_{i=1}^N \Big[\frac{1}{N^2 B} \sum_{\ell=1}^B (1 + \mathfrak{b}^2(\mathsf{Z}_k^{i,\ell})) + \frac{1}{N^2}(1 + |\theta_k^i|^2) \Big],$$

where thanks to Lemma 55 and **A1**, for all $k = 0, \dots, \lfloor NT \rfloor - 1$, $\mathbf{E}[|d_k^N|^2] \leq C/N^2$ for some $C > 0$ independent of $N \geq 1$ and $k = 0, \dots, \lfloor NT \rfloor - 1$. Thus, using (5.68) and (5.69),

$$\begin{aligned}
\mathbf{E}\Big[\sup_{t \in [0,T]} \sup_{\|f\|_{\text{Lip}} \leq 1} |\langle f, \mu_{t_-}^N \rangle - \langle f, \mu_t^N \rangle| \Big] &\leq \mathbf{E}\Big[\sup_{\|f\|_{\text{Lip}} \leq 1} \max_{k=0,\dots,\lfloor NT \rfloor - 1} |\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle| \Big] \\
&\leq \mathbf{E}\Big[\max_{k=0,\dots,\lfloor NT \rfloor - 1} d_k^N \Big] \\
&\leq \mathbf{E}\Big[\sqrt{\sum_{k=0}^{\lfloor NT \rfloor - 1} |d_k^N|^2} \Big] \\
&\leq \sqrt{\mathbf{E}\Big[\sum_{k=0}^{\lfloor NT \rfloor - 1} |d_k^N|^2 \Big]} \leq \frac{C}{\sqrt{N}}.
\end{aligned}$$

This concludes the proof of Proposition 62. □

9.10 Uniqueness of the solution to (5.15)

Proposition 63. *There is a unique solution $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ to (5.15).*

Proof. First of all, the existence of a solution is provided by Propositions 56, 62 and 61. Let us now prove that there is a unique solution to (5.15) in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$.

Recall the definition of $v[\mu]$ in (5.37). We claim that for all $T > 0$ and all solution $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ of (5.15), there exists $C > 0$ such that

$$|v[\bar{\mu}_t](\theta) - v[\bar{\mu}_s](\theta)| \leq C|t - s|, \ \text{ for all } 0 \leq s \leq t \leq T \text{ and } \theta \in \mathbb{R}^{d+1}. \tag{5.70}$$

The proof of item (5.70) is the same as the one made for Item 2 in Proposition 52 since it holds using (5.42) and (5.46), for all $0 \leq s \leq t \leq T$ and $z \in \mathbb{R}^d$,

$$\begin{aligned}
\Big| \int_s^t \langle \nabla_\theta \phi(\cdot, z, x) \cdot \nabla_\theta \mathscr{G}_{\text{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_r \rangle \mathrm{d}r \Big| &\leq C\mathfrak{b}(z) \int_s^t \langle (1 + |\cdot|), \bar{\mu}_r \rangle \mathrm{d}r \\
&\leq C\mathfrak{b}(z) \max_{r \in [0,T]} \langle (1 + |\cdot|), \bar{\mu}_r \rangle |t - s|.
\end{aligned}$$

We now conclude the proof of Proposition 63. Item 1 in the proof of Proposition 52 and (5.70) imply that $v(t, \theta) = v[\bar{\mu}_t](\theta)$ is globally Lipschitz on $[0,T] \times \mathbb{R}^{d+1}$, for all $T > 0$, when $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ is a solution of (5.15). Since in addition a solution $\bar{\mu}$ to (5.15) is a weak solution on $\mathbb{R}_+$ to (5.38) in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}(\mathbb{R}^{d+1}))$, it holds by [Villani, 2021, Theorem 5.34]:

$$\forall t \geq 0, \ \bar{\mu}_t = \phi_t \# \mu_0, \tag{5.71}$$

where ϕ_t is the flow generated by the vector field $v[\bar{\mu}_t](\theta)$ over $\mathbb{R}^{d+1}$. Together with Item 3 in the proof of Proposition 52 and using the same arguments as those used in Step 3 of the proof of [Descours et al., 2022a, Proposition 2.14], two solutions agrees on each $[0,T]$ for all $T>0$. One then deduces the uniqueness of the solution to (5.11). The proof of Proposition 63 is complete. □

We are now in position to end the proof of Theorem 41.

Proof. [Proof of Theorem 41] By Proposition 56, $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. Let $\mu^1,\mu^2 \in \mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ be two limit points of this sequence. By Proposition 62, a.s. $\bar{\mu}^1,\bar{\mu}^2 \in \mathcal{C}(\mathbb{R}_+,\mathcal{P}_1(\mathbb{R}^{d+1}))$. In addition, according to Proposition 61, μ^1 and μ^2 are a.s. solutions of (5.15). Denoting by $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ the unique solution to (5.15) (see Proposition 63), we have a.s.

$$\bar{\mu}^1 = \bar{\mu} \text{ and } \bar{\mu}^2 = \bar{\mu} \text{ in } \mathcal{C}(\mathbb{R}_+,\mathcal{P}_1(\mathbb{R}^{d+1})).$$

In particular $\bar{\mu} \in \mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ and $\bar{\mu}^j = \bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$, $j \in \{1,2\}$. As a consequence, $\bar{\mu}$ is the unique limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ and the whole sequence $(\mu^N)_{N\geq 1}$ converges to $\bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$. Since $\bar{\mu}$ is deterministic, the convergence also holds in probability. The proof of Theorem 41 is complete. □

Let us now prove Proposition 42.

Proof. [Proof of Proposition 42] Any solution to (5.11) in $\mathcal{C}([0,T],\mathcal{P}(\Theta_T))$ is a solution to (5.15) in $\mathcal{C}([0,T],\mathcal{P}_1(\mathbb{R}^{d+1}))$. The result follows from Proposition 63. □

CHAPTER 6

Central Limit Theorem for Bayesian Neural Network trained with Variational Inference

Chapter abstract: *In this chapter, we rigorously derive Central Limit Theorems (CLT) for Bayesian two-layer neural networks in the infinite-width limit and trained by variational inference on a regression task. The different networks are trained via different maximization schemes of the regularized evidence lower bound: (i) the idealized case with exact estimation of a multiple Gaussian integral from the reparametrization trick, (ii) a minibatch scheme using Monte Carlo sampling, commonly known as Bayes-by-Backprop, and (iii) a computationally cheaper algorithm named Minimal VI. The latter was recently introduced by leveraging the information obtained at the level of the mean-field limit. Laws of large numbers are already rigorously proven for the three schemes that admits the same asymptotic limit. By deriving CLT, this work shows that the idealized and Bayes-by-Backprop schemes have similar fluctuation behavior, that is different from the Minimal VI one. Numerical experiments then illustrate that the Minimal VI scheme is still more efficient, in spite of bigger variances, thanks to its important gain in computational complexity.*

1 Introduction

Neural networks (NN), especially with a deep learning architecture, are one of the most powerful function approximators, in particular in a regime of abundant data. Their flexibility may however lead to some overfitting issues, which justify the introduction of a regularization term in the loss. Therefore, Bayesian Neural Networks (BNN) are an interesting alternative. Thanks to a full probabilistic approach, they directly model the uncertainty on the learnt weights through the introduction of a prior distribution, which acts as some natural regularization. Thus, BNN combine the expressivity power of NN, while showing more robustness, in particular when dealing with small datasets, and providing predictive uncertainty [Blundell et al., 2015, Michelmore et al., 2020, McAllister et al., 2017, Filos et al., 2019]. During training, the probabilistic modelling however requires to compute integrals over the posterior distribution. This can be computationally demanding, as these integrals are most of the time not tractable. Alternative techniques as Markov-chain Monte Carlo methods and variational inference are most commonly used instead. The convergence time of the former may prove too prohibitively long in large-dimensional cases [Cobb and Jalaian, 2021]. Therefore variational inference [Hinton and Camp, 1993, MacKay, 1995, MacKay et al., 1995] comes often as the most efficient alternative, especially while using the reparametrization trick and the Bayes-by-backprop (BbB) approach. The variational approach relies on an approximation of the posterior distribution by the closest realization of a parametric one, according to a Kullback-Leibler (KL) divergence. Using a generalisation of the

reparametrization trick [Kingma and Welling, 2014], the *Bayes-by-Backprop* approach [Blundell et al., 2015] leads to an unbiased estimator of the gradient of the ELBO, which enables training by stochastic gradient descent (SGD).

There are now many successful applications of this approach, e.g. [Gal and Ghahramani, 2016, Louizos and Welling, 2017, Khan et al., 2018]. This comes in contrast with the lack of analytical understanding of the behavior of BNN trained with variational inference, especially regarding their overparametrized limit. For instance, it was but only recently shown in [Descours et al., 2023b] what is the appropriate balance in the ELBO of the integrated log-likelihood term and of the KL regularizer, in order to avoid a trivial Bayesian posterior [Izmailov et al., 2021]. To achieve such results, a proper limiting theory was rigorously derived in [Descours et al., 2023b]. Such mean-field analysis, as done in [Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018, Mei et al., 2018, Sirignano and Spiliopoulos, 2020b, Descours et al., 2022b], enables the determination of the limiting nonlinear evolution of the weights of the NN, trained by a gradient descent or some variants. It then allows the derivation of a Law of Large Numbers (LLN) and a Central Limit Theorem (CLT). The main practical goal of such asymptotic analysis is to show convergence towards some global minimizer, it however remains an open and highly-challenging question. Nevertheless, such asymptotic analysis can still be of direct and practical relevance. On top of the proper balance in ELBO, it was recently shown in [Descours et al., 2023b] for BNN on a regression task that the mean-field limit can be leveraged to develop a new SGD training scheme, named *Minimal VI* (MiVI). Indeed, in this limit, the microcospic correlations between each pair of neurons can be shown to be equivalent to some averaged effect of the whole system. Therefore, the *Minimal VI* scheme, which backpropagates only these average fields, is proven to follow the same LLN as standard SGD schemes, but only requires a fraction of the previously needed computations to recover the same limit behavior. Furthermore, numerical experiments showed that the convergence to the mean-field limit arises quite fast with the number of neurons ($N = 300$ [Descours et al., 2023b]). The *Minimal VI* scheme would emerge as a genuinely competitive alternative under these conditions. However, unsurprisingly, numerical experiments also showed a larger variance for the *Minimal VI* scheme, compared to others. Therefore the work presented here directly deals with a precise study of the fluctuation behaviors present at finite width N, as done in [Descours et al., 2022b] for a two-layer NN, but here for the different variational training schemes of a BNN. Independently from the question of scheme comparison, the issue of quantifying the deviations of finite-width BNN from their infinite-width limit is of direct and fundamental relevance.

In more details, we push on the analytical effort to further characterize the limiting behaviors of the three schemes and derive CLT. By framing the fluctuation behaviors of the different schemes, this work is thus of practical and direct relevance for a robust and efficient variational inference framework.

The paper is organized as follows: Section 2 presents the BNN setting as well as the different training algorithms, i.e. idealized, BbB and MiVI, as well as recalls the LLN derived in [Descours et al., 2023b], that shows their asymptotic equivalence at first order. Then, in Section 3, we prove for each algorithm a CLT for the rescaled and centered empirical measure with identified covariance based on non trivial extensions of [Descours et al., 2022b]. Whereas, covariances of the $\mathfrak{G}$-process driving the limit SPDE may be compared, the asymptotic variances of the rescaled centered empirical process are not easily comparable. Therefore we produce numerical experiments in Section 4 showing the good performance of MiVI needing few additional neurons to get comparable variances with less complexity. The proofs for CLT can be found in Section 6.

Related works. The derivation of LLN and CLT for mean-field interacting particle systems have garnered significant attention; refer to, for instance, [Hitsuda and Mitoma, 1986, Sznitman, 1991, Fernandez and Méléard, 1997, Jourdain and Méléard, 1998, Delarue et al., 2019, Del Moral and Guionnet, 1999, Kurtz and Xiong, 2004] and references therein. The use of such approaches to study the asymptotic limit of two-layer NN were introduced in [Mei et al., 2018] (see also [Mei et al., 2019]), which establishes a LLN on the empirical measure of the weights at fixed times. Formal arguments in [Rotskoff and Vanden-Eijnden, 2018] led to conditions to achieve a global convergence of Gradient Descent for exact mean-square loss and online SGD with mini-batches. Regarding fluctuation behaviors, they observe with increasing mini-batch size in the SGD the reduction of the variance of the process leading the fluctuations of the empirical measure of the weights (see [Rotskoff and Vanden-Eijnden, 2018] (Arxiv-V2. Sec 3.3)). See also [Chen et al., 2020a] for a dynamical CLT and [De Bortoli et al., 2020b] on propagation of chaos for SGD on a two-layer NN with different step-size schemes, however limited to finite time horizon. In [Descours et al., 2022b], a LLN and CLT for the entire trajectory, and not only at fixed times, of the empirical measure of a two-layer NN are rigorously derived, especially when proving the uniqueness of the limit PDE. These results are obtained for a large class of variants of SGD (minibatches, noise), that extend in addition to rigorize the work done in [Sirignano

and Spiliopoulos, 2020b] and [Sirignano and Spiliopoulos, 2020c]. Regarding the fluctuation behavior, the results in [Descours et al., 2022b] agree with the observations of [Rotskoff and Vanden-Eijnden, 2018] on the minibatch impact and further exhibit a possible particular fluctuation behavior in a large noise regime. Finally, regarding BNN, [Descours et al., 2023b] rigorously prove a LLN for the entire trajectory for a two-layer BNN trained on a regression task with three different schemes (idealized, BbB, MiVI).

We rigorously prove a CLT for the entire trajectory of the empirical measure of the weights of a two-layer BNN trained by three different maximization schemes (idealized, BbB, MiVI) of a regularized version of ELBO. Remark that a trajectorial CLT is necessary to understand the evolution of the variance of the scaled centered covariance.

2 Setting and proven mean-field limit

2.1 Variational Inference and Evidence Lower Bound

In this section, we first recall the setting of Bayesian neural networks as well as the minimization problem in Variational Inference. We then introduce the three maximization algorithms of the ELBO and recall the respective Law of Large Numbers which were derived in [Descours et al., 2023b], which are the starting points of this work.

The Evidence Lower Bound Let X and Y be subsets of $\mathbb{R}^d$ ($d \geq 1$) and $\mathbb{R}$ respectively. For $N \geq 1$ and $\boldsymbol{w} = (w^1, \dots, w^N) \in (\mathbb{R}^d)^N$, we consider the following two-layer neural network $f^N_{\boldsymbol{w}} : \mathsf{X} \to \mathbb{R}$ defined by:

$$f^N_{\boldsymbol{w}}(x) := \frac{1}{N} \sum_{i=1}^{N} s(w_i, x) \in \mathbb{R},$$

where $x \in \mathsf{X}$ and $s : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}$ is the so-called activation function. In a Bayesian setting, one needs to be able to efficiently sample according to the posterior distribution $\mathfrak{p}^N$ of the latent variable $\boldsymbol{w}$ ($\boldsymbol{w}$ are the weights of the neural network). The classical issue in Bayesian inference over complex models is that the posterior distribution $\mathfrak{p}^N$ is quite hard to sample. For that reason, in variational inference, one looks for the closest distribution to $\mathfrak{p}^N$ in a family of distributions $\mathcal{Q}^N = \{q^N_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Xi^N\}$ which are much easier to sample than $\mathfrak{p}^N$. Here, Ξ is the parameter space. To measure the distance between $q \in \mathcal{Q}^N$ and $\mathfrak{p}^N$, one typically considers the KL divergence distance, denoted by $\mathscr{D}_{\mathrm{KL}}$ in the following. In other words, this minimization problem writes:

$$\mathrm{argmin}_{q \in \mathcal{Q}^N} \mathscr{D}_{\mathrm{KL}}(q | \mathfrak{p}^N).$$

This minimization problem is hard to solve since the KL is not easily computable in practice. A routine computation shows that the above minimization problem, which also writes $\mathrm{argmin}_{\boldsymbol{\theta} \in \Xi^N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | \mathfrak{p}^N)$, is equivalent to the maximization of the Evidence Lower Bound over $\theta \in \Xi^N$. In practice $N \gg 1$, and in this regime, it has been shown in [Coker et al., 2022] and [Huix et al., 2022] that optimizing the ELBO leads to the collapse of the variational posterior to the prior. It has been suggested in [Huix et al., 2022] to rather consider a regularized version of the ELBO, which consists in multiplying the KL term by a parameter which is scaled by the inverse of the number of neurons:

$$\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) = -\int_{(\mathbb{R}^d)^N} \mathfrak{L}\big(y, f^N_{\boldsymbol{w}}(x)\big) q^N_{\boldsymbol{\theta}}(\mathrm{d}w) - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0).$$

In conclusion, the maximization problem we will consider in this work is

$$\mathrm{argmax}_{\boldsymbol{\theta} \in \Xi^N} \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y).$$

Loss function and prior distribution The variational family $\mathcal{Q}^N$ we consider is a Gaussian family of distributions. More precisely, it is assumed throughout this work that for any $\boldsymbol{\theta} = (\theta^1, \dots, \theta^N) \in \Xi^N$, the variational distribution $q^N_{\boldsymbol{\theta}}$ factorizes over the neurons: for all $\boldsymbol{w} = (w^1, \dots, w^N) \in (\mathbb{R}^d)^N$, $q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) = \prod_{i=1}^N q^1_{\theta^i}(w^i)$, where $\theta^i = (m^i, \rho^i) \in \Xi := \mathbb{R}^d \times \mathbb{R}$ and $q^1_{\theta^i}$ is the probability density function (pdf) of $\mathcal{N}(m^i, g(\rho^i)^2 \mathrm{I}_d)$, with $g(\rho) = \log(1 + e^\rho),\ \rho \in \mathbb{R}$.

Let us simply write $\mathbb{R}^{d+1}$ for $\mathbb{R}^d \times \mathbb{R}$. Following the reparameterisation trick of [Blundell et al., 2015], $q^1_\theta(w)\mathrm{d}w$ is the pushforward of a reference probability measure with density γ by Ψ_θ (see Assumption **A1**). In practice, γ is the pdf of $\mathcal{N}(0, I_d)$ and $\Psi_\theta(z) = m + g(\rho)z$. In addition, in all this work, we consider the regression problem, i.e. $\mathfrak{L}$ is the Mean Square Loss: for $\mathfrak{a}, \mathfrak{b} \in \mathbb{R}$, $\mathfrak{L}(\mathfrak{a}, \mathfrak{b}) = \frac{1}{2}|\mathfrak{a} - \mathfrak{b}|^2$.

Set $\phi : (\theta, z, x) \in \mathbb{R}^{d+1} \times \mathbb{R}^d \times \mathsf{X} \mapsto s(\Psi_\theta(z), x)$. Throughout this work, we assume that the prior distribution P_0^N is the function defined by:

$$\forall \boldsymbol{w} \in (\mathbb{R}^d)^N,\ P_0^N(\boldsymbol{w}) = \prod_{i=1}^N P_0^1(w^i), \tag{6.1}$$

where $P_0^1 : \mathbb{R}^d \to \mathbb{R}_+$ is the pdf of $\mathcal{N}(m_0, \sigma_0^2 I_d)$, and $\sigma_0 > 0$. With all these assumptions and notations, we have:

$$\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) = -\frac{1}{2}\int \Big| y - \frac{1}{N}\sum_{i=1}^N s(\Psi_{\theta^i}(z^i), x)\Big|^2 \gamma(z^1)\dots\gamma(z^N)\mathrm{d}z_1\dots\mathrm{d}z_N - \frac{1}{N}\sum_{i=1}^N \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i}|P_0^1). \tag{6.2}$$

Remark 64. *We recall that* (6.1) *implies that* $\mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}}|P_0^N)$ *has a rather nice expression, given by:* $\mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}}|P_0^N) = \sum_{i=1}^N \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i}|P_0^1)$ *and, for* $\theta = (m, \rho) \in \mathbb{R}^{d+1}$,

$$\mathscr{D}_{\mathrm{KL}}(q^1_\theta|P_0^1) = \int_{\mathbb{R}^d} q^1_\theta(x)\log\big(q^1_\theta(x)/P_0^1(x)\big)\mathrm{d}x = \frac{\|m - m_0\|_2^2}{2\sigma_0^2} + \frac{d}{2}\Big(\frac{g(\rho)^2}{\sigma_0^2} - 1\Big) + \frac{d}{2}\log\Big(\frac{\sigma_0^2}{g(\rho)^2}\Big).$$

We also note that $\mathscr{D}_{\mathrm{KL}}$ *has at most a quadratic growth in* m *and* ρ. *In addition, for* $\theta \in \mathbb{R}^{d+1}$, *we have*

$$\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\theta|P_0^1) = \begin{pmatrix} \nabla_m \mathscr{D}_{\mathrm{KL}}(q^1_\theta|P_0^1) \\ \partial_\rho \mathscr{D}_{\mathrm{KL}}(q^1_\theta|P_0^1) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_0^2}(m - m_0) \\ \frac{d}{\sigma_0^2} g'(\rho)g(\rho) - d\frac{g'(\rho)}{g(\rho)} \end{pmatrix}. \tag{6.3}$$

We assume here a Gaussian prior to get an explicit expression of the Kullback-Leibler divergence. Most arguments extend to sufficiently regular densities and are essentially the same for exponential families, using conjugate families for the variational approximation.

2.2 Stochastic Gradient Descent and maximization algorithms

In this section, we present the three different maximization algorithms of the ELBO we are going to consider. In what follows, $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space and we write $\langle U, \nu\rangle = \int_{\mathbb{R}^q} U(z)\nu(\mathrm{d}z)$ for any integrable function $U : \mathbb{R}^q \to \mathbb{R}$ w.r.t. a measure ν (with a slight abuse of notation, we denote by γ the measure $\gamma(z)\mathrm{d}z$). Also we define the σ-algebra $\mathcal{F}_0^N = \boldsymbol{\sigma}(\theta_0^i, 1 \le i \le N)$.

Idealized SGD Consider a data set $\{(x_k, y_k)\}_{k\ge 0}$ i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$, the space of probability measures over $\mathsf{X} \times \mathsf{Y}$. For $N \ge 1$ and given a learning rate $\kappa > 0$, the maximization of $\theta \in \mathbb{R}^{d+1} \mapsto \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y)$ with a SGD algorithm writes as follows: for $k \ge 0$,

$$\begin{cases} \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \kappa\nabla_{\boldsymbol{\theta}}\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}_k, x_k, y_k) \\ \boldsymbol{\theta}_0 \sim \mu_0^{\otimes N}, \end{cases} \tag{6.4}$$

where $\mu_0 \in \mathcal{P}(\mathbb{R}^{d+1})$ (the space of probability measures over $\mathbb{R}^{d+1}$) and $\boldsymbol{\theta}_k = (\theta_k^1, \dots, \theta_k^N)$.

Using the computation of $\nabla_{\boldsymbol{\theta}}\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}_k, x_k, y_k)$ performed in [Descours et al., 2023b], (6.4) writes: for $k \ge 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \theta^i_{k+1} = \theta^i_k - \frac{\kappa}{N^2}\sum_{j=1, j\neq i}^N \Big(\langle\phi(\theta^j_k, \cdot, x_k), \gamma\rangle - y_k\Big)\langle\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\rangle \\ \qquad\qquad -\frac{\kappa}{N^2}\Big\langle(\phi(\theta^i_k, \cdot, x_k) - y_k)\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\Big\rangle - \frac{\kappa}{N}\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k}|P_0^1), \\ \theta^i_0 \sim \mu_0. \end{cases} \tag{6.5}$$

We shall call this algorithm *idealised* SGD because it contains an intractable term given by the integral w.r.t. the probability distribution γ. This has motivated the development of methods where this integral is replaced by an unbiased Monte Carlo estimator (see [Blundell et al., 2015]) as detailed below with the BbB SGD scheme. For the Idealized SGD, and for later purposes, we set for $N \geq 1$ and $k \geq 1$:

$$\mathcal{F}_k^N = \boldsymbol{\sigma}(\theta_0^i, (x_q, y_q), 1 \leq i \leq N, 0 \leq q \leq k-1) \tag{6.6}$$

***Bayes-by-Backprop* (BbB) SGD** For $N \geq 1$, given a dataset $(x_k, y_k)_{k\geq 0}$, the maximization of $\theta \in \mathbb{R}^{d+1} \mapsto \mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}, x, y)$ with a BbB SGD algorithm is the following: for $k \geq 0$ and $i \in \{1, \ldots, N\}$,

$$\begin{cases} \theta_{k+1}^i = \theta_k^i - \frac{\kappa}{N^2} \sum_{j=1}^N \big(\phi(\theta_k^j, \mathsf{Z}_k^j, x_k) - y_k\big) \nabla_\theta \phi(\theta_k^i, \mathsf{Z}_k^i, x_k) - \frac{\kappa}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1), \\ \theta_0^i = (m_0^i, \rho_0^i) \sim \mu_0, \end{cases} \tag{6.7}$$

where $(\mathsf{Z}_k^j, 1 \leq j \leq N, k \geq 0)$ is a i.i.d sequence of random variables distributed according to γ. We recall that this algorithm is based on the Monte Carlo approximation, for $i \in \{1, \ldots, N\}$, of the term

$$\int_{(\mathbb{R}^d)^N} (y - \phi(\theta^j, z^j, x)) \nabla_\theta \phi(\theta^i, z^i, x) \gamma(z^1) \ldots \gamma(z^N) \mathrm{d}z^1 \ldots \mathrm{d}z^N$$

which is the gradient w.r.t. to θ^i of the integral term in the left-hand-side of (6.2). We mention that we consider here in (6.7) the BbB SGD with a batch size of 1, corresponding to $|B| = 1$ in [Descours et al., 2023b].

For the BbB SGD, we set for $N \geq 1$ and $k \geq 1$:

$$\mathcal{F}_k^N = \boldsymbol{\sigma}\Big(\theta_0^i, \mathsf{Z}_q^j, (x_q, y_q), 1 \leq i, j \leq N, 0 \leq q \leq k-1\big\}\Big). \tag{6.8}$$

***Minimal VI* (MiVI) SGD** The last algorithm studied, denoted MiVI SGD, was proposed in [Descours et al., 2023b] as an efficient alternative to the first two algorithm above. It is the following: for $k \geq 0$ and $i \in \{1, \ldots, N\}$,

$$\begin{cases} \theta_{k+1}^i = \theta_k^i - \frac{\kappa}{N^2} \sum_{j=1}^N \big(\phi(\theta_k^j, \mathsf{Z}_k^1, x_k) - y_k\big) \nabla_\theta \phi(\theta_k^i, \mathsf{Z}_k^2, x_k) - \frac{\kappa}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \\ \theta_0^i = (m_0^i, \rho_0^i) \sim \mu_0, \end{cases} \tag{6.9}$$

where $(\mathsf{Z}_k^p, p \in \{1,2\}, k \geq 0)$ is a i.i.d sequence of random variables distributed according to $\gamma^{\otimes 2}$. Thus, the MiVI descent backpropagates through two common Gaussian variables $(\mathsf{Z}_k^1, \mathsf{Z}_k^2)$ to all neurons, instead of a different Gaussian random variable Z_k^i for each neuron.

We finally set for $N, k \geq 1$:

$$\mathcal{F}_k^N = \boldsymbol{\sigma}\Big(\theta_0^i, \mathsf{Z}_q^p, (x_q, y_q), i \in [1, N], p \in \{1,2\}, q \in [0, k-1]\Big). \tag{6.10}$$

2.3 Mean-field limit and Law of Large Numbers

Empirical distributions and assumptions We introduce the empirical distribution ν_k^N of the parameters $\{\theta_k^i, i \in \{1, \ldots, N\}\}$ at iteration $k \geq 0$ (where the θ_k^i's are generated either by the algorithm (6.5), (6.7), or by (6.9)) as well as its scaled version μ_t^N, which are defined by:

$$\nu_k^N := \frac{1}{N} \sum_{i=1}^N \delta_{\theta_k^i} \text{ and } \mu_t^N := \nu_{\lfloor Nt \rfloor}^N. \tag{6.11}$$

Note that for all $N \geq 1$, $\mu^N := \{\mu_t^N, t \geq 0\}$ is a random element of the Skorokhod space $\mathcal{D}(\mathbb{R}_+, \mathcal{P}(\mathbb{R}^{d+1}))$, when $\mathcal{P}(\mathbb{R}^{d+1})$ is endowed with the weak convergence topology. Let us recall that for $q \geq 0$, the Wasserstein spaces $\mathcal{P}_q(\mathbb{R}^{d+1})$ are defined by $\mathcal{P}_q(\mathbb{R}^{d+1}) = \{\mu \in \mathcal{P}(\mathbf{R}^{d+1}), \int_{\mathbb{R}^{d+1}} |\theta|^q \mu(\mathrm{d}\theta) < +\infty\}$. The space $\mathcal{P}_q(\mathbb{R}^{d+1})$ is endowed with the standard Wasserstein metric W_q. Note that for all $q \geq 0$, $(\mu^N)_{N\geq 1}$ is also a random sequence of elements in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_q(\mathbb{R}^{d+1}))$. We denote by $\mathcal{C}_b^\infty(\mathbb{R}^d \times \mathsf{X})$ the space of smooth functions over $\mathbb{R}^d \times \mathsf{X}$ whose derivatives of all order are bounded.

We now introduce the assumptions [Descours et al., 2023b] we will work with in this work:

A1. There exists a pdf $\gamma : \mathbb{R}^d \to \mathbb{R}_+$ such that for all $\theta \in \mathbb{R}^{d+1}$, $q_\theta^1 \mathrm{d}x = \Psi_\theta \# \gamma \mathrm{d}x$, where $\{\Psi_\theta, \theta \in \mathbb{R}^{d+1}\}$ is a family of $\mathcal{C}^1$-diffeomorphisms over $\mathbb{R}^d$ such that for all $z \in \mathbb{R}^d$, $\theta \in \mathbb{R}^{d+1} \mapsto \Psi_\theta(z)$ is of class $\mathcal{C}^\infty$. Finally, there exists $\mathfrak{p}_0 \in \mathbf{N}^*$ such that for all multi-index $\alpha \in \mathbf{N}^{d+1}$ with $|\alpha| \geq 1$, there exists $C_\alpha > 0$, for all $z \in \mathbb{R}^d$ and $\theta = (\theta_1, \ldots, \theta_{d+1}) \in \mathbb{R}^{d+1}$,

$$\left|\partial_\alpha \Psi_\theta(z)\right| \leq C_\alpha \mathfrak{b}(z) \quad \text{with } \forall q \geq 1,\ \langle \mathfrak{b}^q, \gamma \rangle < +\infty, \tag{6.12}$$

where $\partial_\alpha = \partial_{\theta_1}^{\alpha_1} \ldots \partial_{\theta_{d+1}}^{\alpha_{d+1}}$ and $\partial_{\theta_j}^{\alpha_j}$ is the partial derivatives of order α_j w.r.t. to θ_j, and $\mathfrak{b}(z) = 1 + |z|^{\mathfrak{p}_0}$.

A2. The sequence $\{(x_k, y_k)\}_{k \geq 0}$ is i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$. The set $\mathsf{X} \times \mathsf{Y} \subset \mathbb{R}^d \times \mathbb{R}$ is compact. For all $k \geq 0$, $(x_k, y_k) \perp\!\!\!\perp \mathcal{F}_k^N$ (where, depending on the considered algorithms, $\mathcal{F}_k^N$ is defined by (6.6), (6.8), or (6.10)).

A3. The (activation) function $s : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}$ belongs to $\mathcal{C}_b^\infty(\mathbb{R}^d \times \mathsf{X})$.

A4. The initial parameters $(\theta_0^i)_{i=1}^N$ are i.i.d. w.r.t. $\mu_0 \in \mathcal{P}(\mathbb{R}^{d+1})$. Furthermore, μ_0 has compact support.

We moreover assume when considering the BbB algorithm (6.7) (resp. the MiVI algorithm (6.9)):

A5. The sequences $(\mathsf{Z}_k^j, 1 \leq j \leq N, k \geq 0)$ (resp. $(\mathsf{Z}_k^p, p \in \{1,2\}, k \geq 0)$) and $((x_k, y_k), k \geq 0)$ are independent. For $k \geq 0$, $\big((x_k, y_k), \mathsf{Z}_k^j, 1 \leq j \leq N\big) \perp\!\!\!\perp \mathcal{F}_k^N$, see (6.8) (resp. $\big((x_k, y_k), \mathsf{Z}_k^p, p \in \{1,2\}\big) \perp\!\!\!\perp \mathcal{F}_k^N$, see (6.10)).

In the following we simply denote all the above assumptions by **A**. Let us remark that **A3** may seem restrictive, see however Remark 4 in [Descours et al., 2022b] to consider a more general setting.

Law of Large Numbers for the sequence of rescaled empirical distribution As already explained, the starting points to derive Central Limit Theorems for the sequence $(\mu^N)_{N \geq 1}$ defined in (6.11) for the three algorithms introduced above are the Law of Large Numbers obtained in [Descours et al., 2023b] (see more precisely Theorems 1, 2, and 3 there), that we now recall.

Theorem 65 ([Descours et al., 2023b]). *Let $\gamma_0 > 1 + \frac{d+1}{2}$. Assume **A**. Let the $\{\theta_k^i, k \geq 0, i \in \{1, \ldots, N\}\}$'s be generated either by the algorithm* (6.5)*,* (6.7)*, or* (6.9)*. Then, $(\mu^N)_{N \geq 1}$ (see* (6.11)*) converges in $\mathbf{P}$-probability in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1}))$ to a deterministic element $\bar{\mu} \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. In addition, $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ and it is the unique solution in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ to the following measure-valued evolution equation: $\forall f \in \mathcal{C}_b^\infty(\mathbb{R}^{d+1})$ and $\forall t \in \mathbb{R}_+$:*

$$\begin{aligned}\langle f, \bar{\mu}_t \rangle - \langle f, \mu_0 \rangle = &- \kappa \int_0^t \int_{\mathsf{X} \times \mathsf{Y}} \big\langle \phi(\cdot, \cdot, x) - y, \bar{\mu}_s \otimes \gamma \big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \bar{\mu}_s \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\ &- \kappa \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{G}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s \big\rangle \mathrm{d}s. \end{aligned} \tag{6.13}$$

Let us mention that the statement of Theorem 65 differs slightly from the one of Th. 2 in [Descours et al., 2023b] when the Idealized SGD (6.5) is concerned. Since this was possible, we have decided here to work in $\mathcal{P}(\Theta)$ (with $\Theta \subset \mathbb{R}^{d+1}$ compact) instead of $\mathcal{P}_{\gamma_0}(\mathbb{R}^{d+1})$. Nevertheless, Theorem 3 in [Descours et al., 2023b], by following its proof, also holds for the scaled empirical measure μ^N of the parameters θ_k^i's generated by the Idealized SGD (6.5).

3 Main results: Central Limit Theorems

For $\mathfrak{J} \in \mathbf{N}$ and $\mathfrak{j} \geq 0$, let $\mathcal{H}^{\mathfrak{J},\mathfrak{j}}(\mathbb{R}^{d+1})$ be the closure of the set $\mathcal{C}_c^\infty(\mathbb{R}^{d+1})$ for the norm $\|f\|_{\mathcal{H}^{\mathfrak{J},\mathfrak{j}}}$ defined by

$$\|f\|_{\mathcal{H}^{\mathfrak{J},\mathfrak{j}}}^2 = \sum_{|k| \leq \mathfrak{J}} \int_{\mathbb{R}^{d+1}} \frac{|\partial_k f(\theta)|^2}{1 + |\theta|^{2\mathfrak{j}}} \mathrm{d}\theta.$$

The space $\mathcal{H}^{\mathfrak{J},\mathfrak{j}}(\mathbb{R}^{d+1})$ was introduced e.g. in [Fernandez and Méléard, 1997, Jourdain and Méléard, 1998]. It is a separable Hilbert space. Its dual space is denoted by $\mathcal{H}^{-\mathfrak{J},\mathfrak{j}}(\mathbb{R}^{d+1})$. The associated scalar product on $\mathcal{H}^{\mathfrak{J},\mathfrak{j}}(\mathbb{R}^{d+1})$

will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}^{\mathrm{J},\mathrm{j}}}$. For $\Phi \in \mathcal{H}^{-\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1})$, we use the notation $\langle f, \Phi \rangle_{\mathrm{J},\mathrm{j}} = \Phi[f],\ f \in \mathcal{H}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1})$. We will simply denote $\langle f, \Phi \rangle_{J,\beta}$ by $\langle f, \Phi \rangle$ when no confusion is possible. The set $\mathcal{C}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1})$ is defined as the space of functions $f : \mathbb{R}^{d+1} \to \mathbf{R}$ which have continuous partial derivatives up to the order $\mathrm{J} \in \mathbf{N}$ and satisfy, for all $|k| \leq \mathrm{J}$, $\frac{|\partial_k f(\theta)|}{1+|\theta|^{\mathrm{j}}} \to 0$ as $|\theta| \to +\infty$. It is endowed with the norm $\|f\|_{\mathcal{C}^{\mathrm{J},\mathrm{j}}} := \sum_{|k| \leq \mathrm{J}} \sup_{\theta \in \mathbb{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1+|\theta|^{\mathrm{j}}} < +\infty$.

We denote by $x \mapsto \lceil x \rceil$ the ceiling function and we finally set:

$$\mathrm{j}_3 = \lceil \frac{d+1}{2} \rceil + 1 \text{ and } \mathrm{J}_3 = 4 \lceil \frac{d+1}{2} \rceil + 8.$$

The fluctuation process is defined by

$$\eta^N : t \in \mathbb{R}_+ \mapsto \sqrt{N}(\mu_t^N - \bar{\mu}_t)\,, \tag{6.14}$$

where μ^N is defined in (6.11) and $\bar{\mu}_t$ is its limiting process, see Theorem 65. We will show below that the three fluctuation processes converge in law to a limiting process which is the unique (weak) solution an equation (namely Equation **(EqL)** below). The equation **(EqL)** is fully characterizes by the covariance structure of a so-called $\mathfrak{G}$-*process*, a process we introduce now.

Definition 66. *We say that a $\mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$-valued process $\mathscr{G}$ is a $\mathfrak{G}$-process if for all $k \geq 1$ and all $f_1, \ldots, f_k \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1})$, $\{t \in \mathbb{R}_+ \mapsto (\mathscr{G}_t[f_1], \ldots, \mathscr{G}_t[f_k])^T\}$ is a $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^k)$-valued process with zero-mean, independent Gaussian increments (and thus a martingale) and with covariance structure prescribed by $Cov(\mathscr{G}_t[f_i], \mathscr{G}_s[f_j])$, for $0 \leq s \leq t$.*

We mention that two $\mathfrak{G}$-processes are equal in law if and only if they have the same covariance structure (see [Descours et al., 2022b]). For a $\mathfrak{G}$-process $\mathscr{G} \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$, we say that a $\mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$-valued process η is a solution of **(EqL)** if it satisfies a.s. the equation:

$$\begin{aligned}
&\forall f \in \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1}), \forall t \in \mathbb{R}_+, \\
&\langle f, \eta_t \rangle - \langle f, \eta_0 \rangle = -\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&\qquad\qquad - \kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \qquad \textbf{(EqL)} \\
&\qquad\qquad - \kappa \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \eta_s \rangle \mathrm{d}s + \mathscr{G}_t[f].
\end{aligned}$$

We now define, as in the classical theory of stochastic differential equations (see [Kallenberg, 2002]), the notion of weak solution of **(EqL)**.

Definition 67. *Let ν be a $\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$-valued random variable. We say that weak existence holds for **(EqL)** with initial distribution ν if: there exist a probability space $\mathscr{P}$, a process $\eta \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$ and a $\mathfrak{G}$-process $\mathscr{G}$ on $\mathscr{P}$ satisfying **(EqL)** with in addition $\eta_0 = \nu$ in law. In this case, we will simply say that η is a weak solution of **(EqL)**. In addition, we say that weak uniqueness holds if for any two weak solutions η° and $\eta^\star$ of **(EqL)** with the same initial distributions, it holds $\eta^\circ = \eta^\star$ in law.*

We are now in position to state the main theoretical result of this work: Central Limit Theorems for the trajectory of the scaled empirical measures μ^N of the $\{\theta_k^i, i \in \{1, \ldots, N\}\}$'s generated either by the algorithm (6.5), (6.7), or by (6.9).

Theorem 68. *Assume **A**. Then,*

1. *The sequence $(\eta^N)_{N \geq 1}$ converges in distribution in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$ to a $\mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$-valued process $\eta^\star$.*
2. *The process $\eta^\star$ is the unique weak solution of **(EqL)** with initial distribution ν_0, where ν_0 is the unique (in distribution) $\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$-valued random variable such that for all $k \geq 1$ and $f_1, \ldots, f_k \in \mathcal{H}^{J_3-1,j_3}(\mathbb{R}^{d+1})$, $(\langle f_1, \nu_0 \rangle, \ldots, \langle f_k, \nu_0 \rangle)^T \sim \mathcal{N}(0, \mathfrak{C}(f_1, \ldots, f_k))$, where $\mathfrak{C}(f_1, \ldots, f_k)$ is the covariance matrix of $(f_1(\theta_0^1), \ldots, f_k(\theta_0^1))^T$. Moreover, the $\mathfrak{G}$-process $\mathscr{G}$ has covariance structure given by, for all $f, g \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1})$ and all $0 \leq s \leq t$:*

- *When the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the idealized algorithm* (6.5) *or by the BbB algorithm* (6.7)*,*

$$Cov(\mathcal{G}_t[f],\mathcal{G}_s[g]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f](x,y,\bar{\mu}_v),\mathscr{Q}[g](x,y,\bar{\mu}_v))\mathrm{d}v,$$

where $\mathscr{Q}[f](x,y,\bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$.

- *When the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the MiVI algorithm* (6.9)*,*

$$Cov(\mathcal{G}_t[f],\mathcal{G}_s[g]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f](x,y,z^1,z^2,\bar{\mu}_v),\mathscr{Q}[g](x,y,z^1,z^2,\bar{\mu}_v))\mathrm{d}v,$$

where $\mathscr{Q}[f](x,y,z^1,z^2,\bar{\mu}_v) = \langle \phi(\cdot,z^1,x) - y, \bar{\mu}_v\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z^2,x), \bar{\mu}_v\rangle$.

Let us begin by the following remark: when $f = g$ it follows directly from Jensen's inequality that the variance of the $\mathfrak{G}$-process leading the limiting SPDE of the CLT of the *Minimal VI* algorithm is greater than the corresponding variance of the *BbB* algorithm. It is however not clear if this hierarchy is conserved through the SPDE. However numerical experiments presented in Section 4 tend to this conclusion.

The strategy of the proof of Theorem 68 is the same whenever one considers that the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by (6.5), (6.7) or (6.9), except for the convergence of the martingale sequence $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$ towards a $\mathfrak{G}$-process which requires more inlvolved analysis (see more precisely Section 6.3). Appendix 6 below is dedicated to the detailed proof of the Central Limit Theorem when the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by (6.7). The other two cases are treated very similarly except, as already mentioned, the convergence of the martingale term towards a $\mathfrak{G}$-process, which is therefore proved for each of the three algorithms in Section 6.3. The proof of Theorem 68 is inspired by the one made for Th. 2 in [Descours et al., 2022b]. Nonetheless, two difficulties arise in the proof of Theorem 68 compared to [Descours et al., 2022b]. The first one comes from the fact that the term $\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)$, appearing in all of the three algorithms, is not bounded in θ (see indeed (6.3)). The second difficulty deals with the convergence of the martingale sequence $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$, defined in (6.24), when the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by (6.7). In this case, we have to introduce and study the convergence of the empirical distribution of both the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s and the Z^i's (see (6.74) and Lemma 83).

4 Numerical simulations

In this section, we begin by illustrating Theorem 68 of this paper, followed by a comparative analysis between MiVI SGD algorithm and its two counterparts, idealized (I-SGD) and BbB SGD.

For our experimental setup, we draw uniformly the input data $x \sim \mathcal{U}([-1,1]^{d_{in}})$. Then, the output data is given by $y = \tanh(\langle x, w_{in}^\star\rangle) \cdot w_{out}^\star + \beta \cdot \epsilon$. Here, $\beta \in \mathbf{R}$ represents the noise level and $\epsilon \sim \mathcal{N}(0, \mathrm{I}_{d_{in}})$ is the Gaussian noise. Therefore, we are trying to learn the noisy prediction of a two-layer Neural Network with an hyperbolic tangent activation function. The true parameters of this network are defined by $w_{in}^\star \in \mathbf{R}^{d_{in}}$ and $w_{out}^\star \in \mathbf{R}^{d_{out}}$. These true parameters are initialized randomly, sampled from a standard Gaussian distribution.

We consider two distinct settings in our evaluation. The first is a noiseless and low-dimensional scenario with parameters set to $\beta = 0$, $d_{in} = 10$, and $d_{out} = 1$. In contrast, the second setting is more complex, involving noise with $\beta = 1$, and higher dimensions with $d_{in} = 50$ and $d_{out} = 10$.

For all algorithms (MiVI-SGD, BbB-SGD, and I-SGD), the prior distribution is $P_0^N = \mathcal{N}(0, \mathrm{I}_{N\times(d_{in}+d_{out})})$. The variational parameters θ are randomly initialized, centered around the prior distribution. Since the Idealized-SGD cannot be implemented due to intractable integral calculation, we approximate it using Monte Carlo with a mini-batch of 100. For the algorithm BbB-SGD, we set the number of Monte Carlo samples to 1. The number of gradient descent steps used by all algorithms is set to $\lfloor t \cdot N \rfloor$, where $t = 10$ for the simple setting. However, due to computational limitations, we set $t = 3$ for the complex setting. For all experiments, we consider three different test functions. If $\theta = (m, \rho)$, we define $f_{mean}(\theta) = \|m\|_2$, $f_{std}(\theta) = |g(\rho)|$, and $f_{pred}(\theta) = \hat{\mathbb{E}}_x\left[\hat{\mathbb{V}}_{w\sim q_\theta^1}[s(w,x)]^{\frac{1}{2}}\right]$. Here, $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ represent the empirical mean and variance over 100 samples, respectively. These functions are used to compute $\langle f, \mu_t^N\rangle$ and $\langle f, \eta_t^N\rangle$.

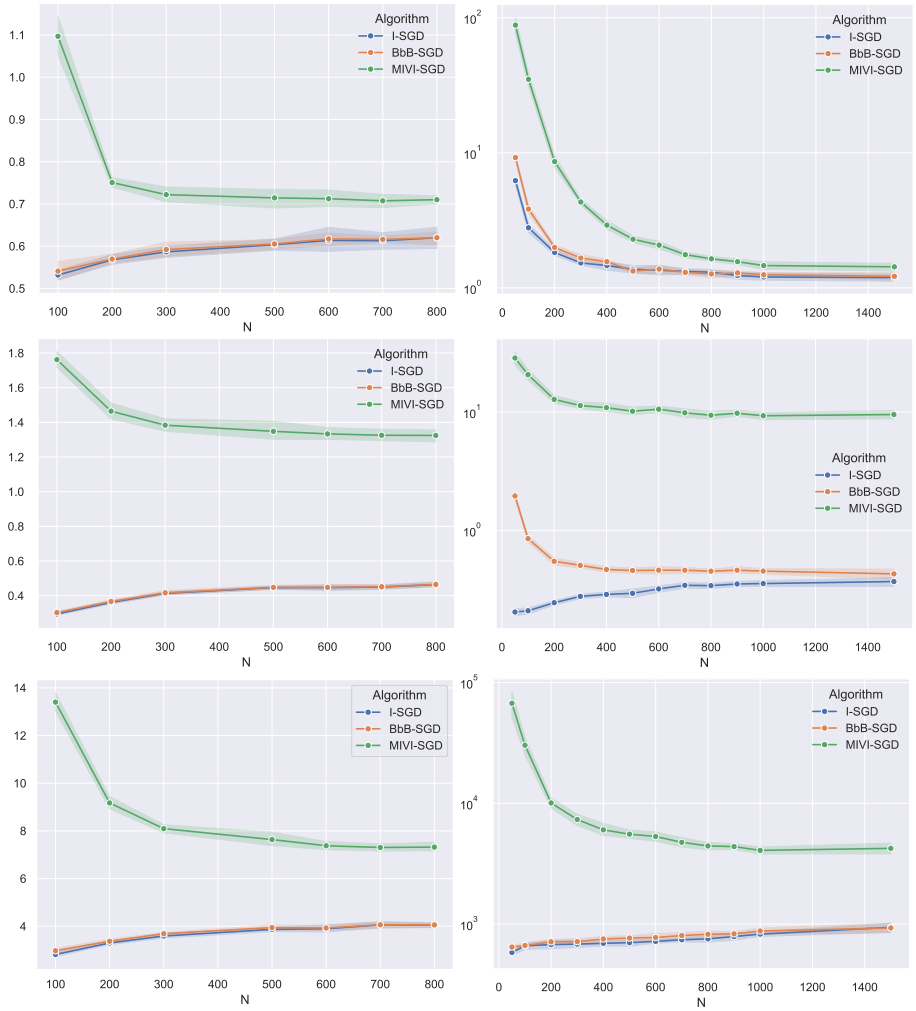


Fig. 6.1 Convergence of $\mathbb{V}[\langle f, \eta_t^N \rangle]$ in the simple (left column) and complex (right column) setting, for f_{mean} (1^{st} line), f_{std} (2^{nd} line) and f_{pred} (3^{rd} line).

Illustration of Theorem 68: Using the definition of η_t^N in equation 6.14, and that $\bar{\mu_t}$ is deterministic, then we deduce that $\mathbb{V}[\langle f,\, \eta_t^N \rangle] = N \cdot \mathbb{V}[\langle f,\, \mu_t^N \rangle]$. Figure 6.1 displays the convergence of $N \cdot \mathbb{V}[\langle f,\, \mu_t^N \rangle]$ in the simple and complex setting. The variance is estimated using its empirical version with 300 samples, and the 95% confidence interval is calculated based on 10 samples. These plots clearly show that the $\mathfrak{G}$-process associated with the limiting fluctuation process η_t derived from BbB-SGD shares the same covariance as the one derived from I-SGD, but differs from the covariance derived from MiVI-SGD, which exhibit larger values. These plots clearly illustrates the main result of Theorem 68 and the following remark.

Comparison MiVI-SGD, BbB-SGD and I-SGD: The objective of this paragraph is to compare, at a fixed number of neurons N, the performances of algorithms MiVI-SGD, BbB-SGD and I-SGD. Recall that algorithm BbB-SGD randomly samples N Gaussian vectors of dimension $d_{in} + d_{out}$ at each training step. Consequently, during the full training, this algorithm samples $\lfloor t \cdot N \rfloor N$ Gaussian vectors. In contrast, MiVI-SGD samples only 2 Gaussian vectors per training step, resulting in a total of $2\lfloor t \cdot N \rfloor$ sampled Gaussian vectors. Therefore, algorithm MiVI-SGD becomes more suitable (in terms of the number of Gaussian vectors sampled) for $N \geq 2$. Figure 6.2 show the variance of $\langle f,\, \mu_t^N \rangle$ with respect to N, in the simple and complex setting. Similarly to the previous paragraph, the variance is estimated using 300 samples, and the 95% confidence interval is computed based on 10 samples.

This figure shows that, in the simple setting MiVI-SGD with $N = 700$ obtains the same performance (in term of $\mathbb{V}[\langle f,\, \mu_t^N \rangle]$), than BbB-SGD and I-SGD with $N = 620$ for f_{mean}, $N = 210$ for f_{std} and $N = 380$. Similarly, in the

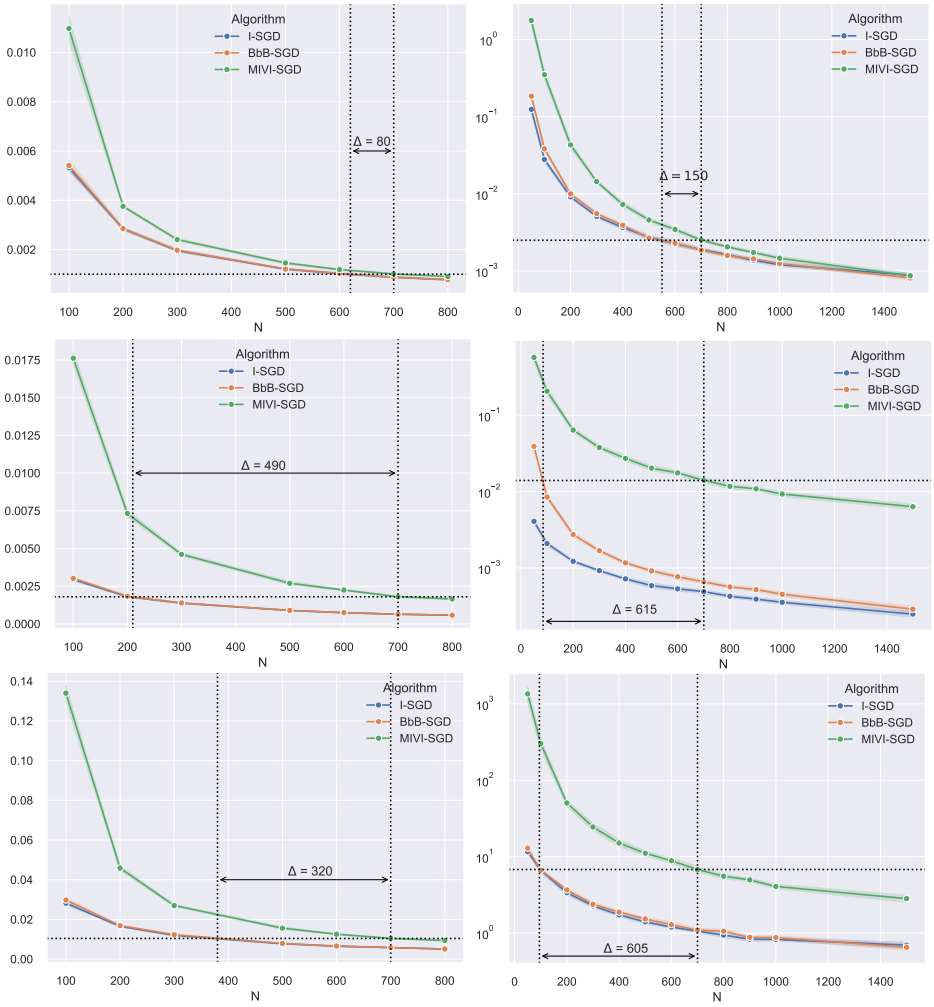


Fig. 6.2 $\mathbb{V}[\langle f, \mu_t^N \rangle]$ with respect to N, in the simple (left column) and complex (right column) setting, for f_{mean} (1^{st} line), f_{std} (2^{nd} line) and f_{pred} (3^{rd} line).

complex setting MiVI-SGD with $N = 700$ obtains the same performance (in term of $\mathbb{V}[\langle f,\ \mu_t^N \rangle]$), than BbB-SGD and I-SGD with $N = 550$ for f_{mean}, $N = 75$ for f_{std} and $N = 95$.

Consequently, in both settings, algorithm MiVI-SGD appears to be more efficient (in terms of the number of sampled vectors) than other algorithms for achieving the same value of $\mathbb{V}[\langle f, ; \mu_t^N \rangle]$.

5 Conclusion

In this work, we have rigorously shown CLT for a two-layer BNN trained by variational inference with different SGD schemes. It appears that the idealized SGD and the most-commonly used *Bayes-by-Backprop* SGD schemes have the same fluctuation behaviors, i.e. driven by a SPDE with a $\mathfrak{G}$-process having the same covariance structure, in addition to admitting the same mean-field limit. Introduced in [Descours et al., 2023b], the less costly *Minimal VI* SGD scheme exhibits a different fluctuation behavior, with a $\mathfrak{G}$-process of different covariance structure, which can be argued to lead to larger variances. Though, numerical experiments show that the trade-off between computational complexity and variance is still vastly in favour of the *Minimal VI* scheme. This opens the interesting perspective of exploring whether additional practical improvements can be derived from the asymptotic results at the mean-field level. This becomes even more intriguing and a justified approach given that neural networks appear to reach such limits rapidly.

6 Central Limit Theorem: proof of Theorem 68

In all this section, the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the algorithm (6.7), except in Section 6.3 which, we recall, is dedicated to the study of the convergence of the sequences of martingale $(\sqrt{N}\mathbf{M}^N)_N$ (see (6.24)). Recall the definition of the σ-algebra $\mathcal{F}_k^N$ in (6.8). We also recall the following paramount result which aims at giving uniform bounds, see Lemma 17 in [Descours et al., 2023b] on the moments of the parameters $\{\theta_k^i, i \in \{1,\dots,N\}\}$ up to iteration $\lfloor NT \rfloor$, for a fixed $T > 0$.

Lemma 69. *Assume **A**. Then, for all $T > 0$ and all $p \geq 1$, there exists $C > 0$ such that for all $N \geq 1$, $i \in \{1,\dots,N\}$ and $0 \leq k \leq \lfloor NT \rfloor$, $\mathbf{E}[|\theta_k^i|^p] \leq C$.*

Let us now recall some Sobolev embeddings which will be also used in the proof of Theorem 68. For $\mathfrak{N}, \mathfrak{j} > (d+1)/2$ and $\mathfrak{k}, \mathfrak{J} \geq 0$, $\mathcal{H}^{\mathfrak{N}+\mathfrak{J},\mathfrak{k}}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{\mathfrak{J},\mathfrak{k}}(\mathbb{R}^{d+1})$ and $\mathcal{H}^{\mathfrak{N}+\mathfrak{J},\mathfrak{k}}(\mathbb{R}^{d+1}) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{\mathfrak{J},\mathfrak{k}+\mathfrak{j}}(\mathbb{R}^{d+1})$ (see Section 2 in [Fernandez and Méléard, 1997]). Recall $\mathfrak{J}_3 = 4\lceil \frac{d+1}{2} \rceil + 8$ and $\mathfrak{j}_3 = \lceil \frac{d+1}{2} \rceil + 1$. Set $\mathfrak{J}_0 = \lceil \frac{d+1}{2} \rceil + 3$, $\mathfrak{J}_1 = 2\lceil \frac{d+1}{2} \rceil + 4$, $\mathfrak{J}_2 = 3\lceil \frac{d+1}{2} \rceil + 6$, and $\mathfrak{j}_2 = 2\lceil \frac{d+1}{2} \rceil + 2$, $\mathfrak{j}_1 = 3\lceil \frac{d+1}{2} \rceil + 4$, $\mathfrak{j}_0 = 4\lceil \frac{d+1}{2} \rceil + 5$. Hence, the following Hilbert-Schmidt embeddings hold: $\mathcal{H}^{\mathfrak{J}_3-1,\mathfrak{j}_3}(\mathbb{R}^{d+1}) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{\mathfrak{J}_2,\mathfrak{j}_2}(\mathbb{R}^{d+1})$, $\mathcal{H}^{\mathfrak{J}_2,\mathfrak{j}_2}(\mathbb{R}^{d+1}) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{\mathfrak{J}_1+1,\mathfrak{j}_1-1}(\mathbb{R}^{d+1})$, $\mathcal{H}^{\mathfrak{J}_1,\mathfrak{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{\mathfrak{J}_0,\mathfrak{j}_0}(\mathbb{R}^{d+1})$. One also has the following continuous embeddings: $\mathcal{H}^{\mathfrak{J}_0,\mathfrak{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{2,\mathfrak{j}_0}(\mathbb{R}^{d+1})$ and $\mathcal{H}^{\mathfrak{N},\mathfrak{j}}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{H}^{\mathfrak{N},\mathfrak{j}+\mathfrak{k}}(\mathbb{R}^{d+1})$, where $\mathfrak{N}, \mathfrak{j}, \mathfrak{k} \geq 0$.

We finally recall some useful inequality which will be used throughout this work (see the proof of Lemma 1 in [Descours et al., 2023b]) and which are direct consequences of **A**: for all $\theta \in \mathbb{R}^{d+1}$, $z \in \mathbb{R}^d$, and $(x,y) \in \mathsf{X} \times \mathsf{Y}$, it holds:

I. $|\phi(\theta,z,x) - y| \leq C$ and $|\nabla_\theta \phi(\theta,z,x)| \leq C|\mathrm{J}_\theta \Psi_\theta(z)| \leq C\mathfrak{b}(z)$ (where J_θ denotes the Jacobian operator w.r.t. θ).

In addition,

II. For all $x \in \mathsf{X}$,

$$\mathfrak{H}(\cdot,x) : \theta \mapsto \int_{\mathbb{R}^d} \phi(\theta,z,x)\gamma(z)\mathrm{d}z = \langle \phi(\theta,\cdot,x), \gamma \rangle \tag{6.15}$$

is smooth and all its derivatives of non negative order are uniformly bounded over $\mathbb{R}^{d+1}$ w.r.t $x \in \mathsf{X}$.

Moreover, for any multi-index $\alpha \in \mathbf{N}^{d+1}$, (see Remark 64), it holds for some $C > 0$ and all $\theta \in \mathbb{R}^{d+1}$:

$$|\partial_\alpha \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \leq C(1 + |\theta|) \text{ if } |\alpha| = 1 \text{ and } |\partial_\alpha \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \leq C \text{ for } |\alpha| \geq 2. \tag{6.16}$$

6.1 Relative compactness of the fluctuation sequence $(\eta^N)_{N \geq 1}$

Recall that the fluctuation process is defined by $\eta^N : t \in \mathbb{R}_+ \mapsto \sqrt{N}(\mu_t^N - \bar{\mu}_t)$, $N \geq 1$. The aim of this section is to prove the following relative compactness result on the sequence $(\eta^N)_{N \geq 1}$.

Proposition 70. *Assume **A**. Then, $(\eta^N)_{N \geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathfrak{J}_3+1,\mathfrak{j}_3}(\mathbb{R}^{d+1}))$.*

We mention that Proposition 70 also holds when the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the two other algorithms (6.5) and (6.9). Before starting the proof of Proposition 70, we need to introduce an auxiliary system of particles, this is the purpose of the next lemma.

For any $\mu \in \mathcal{P}(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^{d+1}))$, we consider $\mathscr{P}_\mu \in \mathcal{P}(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^{d+1}))$ defined as the law of the process $(X_t)_{t \geq 0}$ solution to

$$(\mathbf{E}_\mu) \begin{cases} \mathrm{d}X_t = -\kappa \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_t \otimes \gamma \rangle \langle \nabla_\theta \phi(X_t,\cdot,x), \gamma \rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}t - \kappa \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{X_t}^1 | P_0^1)\mathrm{d}t, \\ X_0 \sim \mu_0. \end{cases}$$

We then denote by $\mathscr{F}(\mu)$ the function $t \in \mathbb{R}_+ \mapsto (\mathscr{P}_\mu)_t = \mathscr{P}_\mu \circ \pi_t^{-1}$ the law $(X_s)_{s \geq 0}$ at time t, where π_t is the natural projection from $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^{d+1})$ to $\mathbb{R}$ define by $\pi_t(f) = f(t)$.

Lemma 71. *Assume **A**. Then, $\bar{\mu} = \mathscr{F}(\bar{\mu})$ (where $\bar{\mu}$ is given by Theorem 65), i.e for the solution $(\bar{X}_t)_{t\geq 0}$ of $(\mathbf{E}_{\bar{\mu}})$, it holds $\bar{X}_t \sim \bar{\mu}_t$ for all $t \geq 0$.*

Proof. We claim that $\mathscr{F}(\mu) \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$, for all $\mu \in \mathcal{P}(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^{d+1}))$. Let us prove this claim. Let $(X_t)_{t\geq 0}$ be the solution of $(\mathbf{E}_\mu)$. Then, by **I**, **II**, and **A**, together with (6.16), there exists $c_0 > 0$ such that a.s. for all $t \geq 0$,

$$|X_t| \leq c_0(1+t) + c_0 \int_0^t |X_s| \mathrm{d}s.$$

Therefore, a.s., for all $T > 0$ and $0 \leq t \leq T$, by Gronwall lemma, one has $|X_t| \leq c_0(1+T)e^{c_0 T}$. With this bound, one deduces that there exists $c_1 > 0$ such that a.s. for all $0 \leq s \leq t \leq T$, $|X_t - X_s| \leq c_1(1+T)e^{c_1 T}(t-s)$, which proves the claim.

Let $\mu \in \mathcal{P}(\mathbb{R}^{d+1})$. Define $\mathscr{V}[\mu] : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ by:

$$\mathscr{V}[\mu](\theta) = -\kappa \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu \otimes \gamma \rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) - \kappa \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0). \tag{6.17}$$

By the analysis carried out in Section B.3.2 in [Descours et al., 2023b] (based on Th. 5.34 in [Villani, 2021]), $\bar{\mu}$ is the unique *weak solution*[1] in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$ of the measure-valued equation

$$\begin{cases} \partial_t \mu^*_t = \mathrm{div}(\mathscr{V}[\bar{\mu}_t]\mu^*_t) \\ \mu^*_0 = \mu_0. \end{cases} \tag{6.18}$$

On the other hand, using the equality $g(X_t) - g(X_0) = \int_0^t \nabla g(X_u) \cdot \frac{\mathrm{d}}{\mathrm{d}t} X_u \mathrm{d}u$ valid for any $\mathcal{C}^1$ function g with compact support, together with $(\mathbf{E}_\mu)$, we deduce that $\mathscr{F}(\bar{\mu})$ is a weak solution of (6.18). By uniqueness, $\bar{\mu} = \mathscr{F}(\bar{\mu})$. The proof is complete. □

Let us now introduce N independent processes $\bar{X}^i$, $i \in \{1, \dots, N\}$, solution to $(\mathbf{E}_{\bar{\mu}})$. It then holds thanks to Lemma 71, for all $i \in \{1, \dots, N\}$ and $t \geq 0$:

$$(\mathbf{S}) \begin{cases} \mathrm{d}\bar{X}^i_t = -\kappa \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_t \otimes \gamma \rangle \langle \nabla_\theta \phi(\bar{X}^i_t,\cdot,x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}t - \kappa \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\bar{X}^i_t} | P^1_0)\mathrm{d}t, \\ \bar{X}^i_0 \sim \mu_0,\ \bar{X}^i_t \sim \bar{\mu}_t. \end{cases}$$

Their empirical distribution is denoted by $\bar{\mu}^N_t = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{X}^i_t}$, for $N \geq 1$ and $t \in \mathbb{R}_+$. Recall that from the proof of Lemma 71, there exists $c_1 > 0$ such that a.s. for all $0 \leq s \leq t \leq T$ and all $i \in \{1, \dots, N\}$:

$$|\bar{X}^i_t| \leq c_1(1+T)e^{c_1 T} \text{and } |\bar{X}^i_t - \bar{X}^i_s| \leq c_1(1+T)e^{c_1 T}(t-s). \tag{6.19}$$

We now decompose η^N using the following two processes:

$$\Upsilon^N := \sqrt{N}(\mu^N - \bar{\mu}^N) \text{ and } \Theta^N := \sqrt{N}(\bar{\mu}^N - \bar{\mu}). \tag{6.20}$$

We denote by $\mathcal{C}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1})^*$ the dual space of $\mathcal{C}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1})$ ($\mathrm{J}, \mathrm{j} \geq 0$). One the one hand, $\bar{\mu}^N \in \mathcal{C}(\mathbb{R}_+, \mathcal{C}^{1,\mathrm{j}}(\mathbb{R}^{d+1})^*)$, $\mathrm{j} \geq 0$. This is indeed a direct consequence of (6.19). On the other hand, for any $\mathrm{j} \geq 0$, $\mu^N \in \mathcal{D}(\mathbb{R}_+, \mathcal{C}^{0,\mathrm{j}}(\mathbb{R}^{d+1})^*)$. Hence, it holds for all $\mathrm{j} \geq 0$ a.s.

$$\Upsilon^N \in \mathcal{D}(\mathbb{R}_+, \mathcal{C}^{1,\mathrm{j}}(\mathbb{R}^{d+1})^*). \tag{6.21}$$

Concerning Θ^N, we have the following result.

Lemma 72. *Assume **A**. Then, for any $\mathrm{J} > 1 + (d+1)/2$ and $\mathrm{k} \geq 0$, $\bar{\mu}^N, \bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J},\mathrm{k}}(\mathbb{R}^{d+1}))$. Therefore, a.s. $\Theta^N \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J},\mathrm{k}}(\mathbb{R}^{d+1}))$. Finally, (6.13) also holds for any test function $f \in \mathcal{H}^{\mathrm{J},\mathrm{k}}(\mathbb{R}^{d+1})$ ($\mathrm{J} > 1 + (d+1)/2$ and $\mathrm{k} \geq 0$).*

[1] See Section 4.1.2 in [Santambrogio, 2015] for the definition.

Proof. Let $\mathtt{J} > 1+(d+1)/2$ and $\mathtt{k} \geq 0$. It then holds $\mathcal{H}^{\mathtt{J},\mathtt{k}}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\mathtt{k}}(\mathbb{R}^{d+1})$. This implies that $\mathcal{C}^{1,\mathtt{k}}(\mathbb{R}^{d+1})^* \hookrightarrow \mathcal{H}^{-\mathtt{J},\mathtt{k}}(\mathbb{R}^{d+1})$, and consequently, $\bar{\mu}^N \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathtt{J},\mathtt{k}}(\mathbb{R}^{d+1}))$.

Let us now prove that $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathtt{J},\mathtt{k}}(\mathbb{R}^{d+1}))$ for $\mathtt{k} \geq 0$. Set $\mathtt{j} = \mathtt{k}+1$. Recall that one can choose any $\gamma_0 > 1 + \frac{d+1}{2}$ in Theorem 65. Pick thus such a γ_0 such that $\mathtt{j} \leq \gamma_0$. We then have $\mathcal{H}^{\mathtt{J},\mathtt{j}-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\mathtt{j}-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\gamma_0-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{0,\gamma_0-1}(\mathbb{R}^{d+1})$. Since μ_0 has compact support, $\mu_0 \in \mathcal{C}^{0,\gamma_0-1}(\mathbb{R}^d)^* \hookrightarrow \mathcal{H}^{-\mathtt{J},\mathtt{j}-1}(\mathbb{R}^d)$. Let $f \in \mathcal{C}^\infty_c(\mathbb{R}^{d+1})$ and $0 \leq s \leq t \leq T$. Thanks to (6.16) and Assumption **A**, we deduce that:

$$\begin{aligned}
|\langle f, \bar{\mu}_t\rangle - \langle f, \bar{\mu}_s\rangle| &\leq C|t-s|(\|f\|_{\mathcal{C}^{1,\gamma_0}} + \|f\|_{\mathcal{C}^{1,\gamma_0-1}}) \sup_{u\in[0,T]} |\langle 1+|\cdot|^{\gamma_0}, \bar{\mu}_u\rangle| \\
&\leq C|t-s|\|f\|_{\mathcal{C}^{1,\gamma_0-1}} \sup_{u\in[0,T]} |\langle 1+|\cdot|^{\gamma_0}, \bar{\mu}_u\rangle| \\
&\leq C|t-s|\|f\|_{\mathcal{H}^{\mathtt{J},\mathtt{j}-1}} \sup_{u\in[0,T]} |\langle 1+|\cdot|^{\gamma_0}, \bar{\mu}_u\rangle|.
\end{aligned}$$

We have that $\sup_{u\in[0,T]} |\langle 1+|\cdot|^{\gamma_0}, \bar{\mu}_u\rangle| < +\infty$ since $u \geq 0 \mapsto \langle 1+|\cdot|^{\gamma_0}, \bar{\mu}_u\rangle \in \mathcal{D}(\mathbb{R}_+, \mathbb{R})$ (this follows from the fact that $\bar{\mu} \in \mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\gamma_0}(\mathbb{R}^d))$ together with Th. 6.9 in [Villani, 2009]). We have thus proved that $\bar{\mu}_t \in \mathcal{H}^{-\mathtt{J},\mathtt{j}-1}(\mathbb{R}^d)$ and $|\langle f, \bar{\mu}_t\rangle - \langle f, \bar{\mu}_s\rangle| \leq C|t-s|\|f\|_{\mathcal{H}^{\mathtt{J},\mathtt{j}-1}}$. This proves that $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathtt{J},\mathtt{j}-1}(\mathbb{R}^{d+1}))$. The last claim is obtained by a density argument and the fact that $\mathcal{H}^{\mathtt{J},\mathtt{j}-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\gamma_0-1}(\mathbb{R}^{d+1})$. □

Lemma 73. *Assume **A**. For all $T > 0$, we have*

$$\sup_{N\geq 1} \sup_{t\in[0,T]} \mathbf{E}[\|\Theta^N_t\|^2_{\mathcal{H}^{-\mathtt{J}_1,\mathtt{j}_1}} + \|\Upsilon^N_t\|^2_{\mathcal{H}^{-\mathtt{J}_1,\mathtt{j}_1}}] < +\infty.$$

In particular, $\sup_{N\geq 1} \sup_{t\in[0,T]} \mathbf{E}[\|\eta^N_t\|^2_{\mathcal{H}^{-\mathtt{J}_1,\mathtt{j}_1}}] < +\infty$.

Proof.

Let $T > 0$. Pick $t \in [0,T]$, $N \geq 1$, and $f \in \mathcal{H}^{\mathtt{J}_1,\mathtt{j}_1}(\mathbb{R}^d)$. On the one hand, since $(f(\bar{X}^j_t) - \langle f, \bar{\mu}_t\rangle)_{j=1,\ldots,N}$ are independent centered random variables, one deduces that $\mathbf{E}[\langle f, \Theta^N_t\rangle^2] \leq \frac{2}{N}\sum_{i=1}^N (\mathbf{E}\big[|f(\bar{X}^i_t)|^2\big] + |\langle f, \bar{\mu}_t\rangle|^2) \leq C_T\|f\|^2_{\mathcal{H}^{\mathtt{J}_0,\mathtt{j}_0}}$, where the last inequality is a consequence of (6.19) together with $\mathcal{H}^{\mathtt{J}_0,\mathtt{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{0,\mathtt{j}_0}(\mathbb{R}^{d+1})$ and $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathtt{J}_0,\mathtt{j}_0}(\mathbb{R}^{d+1}))$ (see Lemma 72). Using also the embedding $\mathcal{H}^{\mathtt{J}_1,\mathtt{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{\mathtt{J}_0,\mathtt{j}_0}(\mathbb{R}^{d+1})$ and considering an orthonormal basis of $\mathcal{H}^{\mathtt{J}_1,\mathtt{j}_1}(\mathbb{R}^{d+1})$, one deduces the desired upper bound on Θ^N.

Let us now derive the bound on the second order moment of Υ^N. To this end, introduce an orthonormal basis $(f_a)_{-a\geq 1}$ of $\mathcal{H}^{\mathtt{J}_1,\mathtt{j}_1}(\mathbb{R}^{d+1})$. One then has:

$$\|\Upsilon^N_t\|^2_{\mathcal{H}^{-\mathtt{J}_1,\mathtt{j}_1}} = \sum_{a\geq 1} \langle f_a, \Upsilon^N_t\rangle^2. \tag{6.22}$$

Recall $\mathcal{H}^{\mathtt{J}_0,\mathtt{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{2,\mathtt{j}_0}(\mathbb{R}^{d+1})$. We have, by **(S)** and the fact that $f \in \mathcal{C}^{2,\mathtt{j}_0}(\mathbb{R}^{d+1})$,

$$\begin{aligned}
\langle f, \bar{\mu}^N_t\rangle = \langle f, \bar{\mu}^N_0\rangle &- \kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}^N_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \langle \nabla f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \bar{\mu}^N_s\rangle \mathrm{d}s.
\end{aligned} \tag{6.23}$$

We now set for $k \geq 0$ and $g \in \mathcal{C}^{2,\mathtt{j}}(\mathbb{R}^{d+1})$ ($\mathtt{j} \geq 0$):

1. $\mathbf{D}^N_k[g] := -\frac{\kappa}{N^3}\sum_{i=1}^N \sum_{j=1, j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\langle \phi(\theta^j_k, \cdot, x), \gamma\rangle - y\big)\big\langle \nabla_\theta g(\theta^i_k) \cdot \nabla_\theta \phi(\theta^i_k, \cdot, x), \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y) - \frac{\kappa}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot,\cdot,x) - y)\nabla_\theta g \cdot \nabla_\theta \phi(\cdot,\cdot,x), \nu^N_k \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)$.
2. $\mathbf{M}^N_k[g] = -\frac{\kappa}{N^3}\sum_{i,j=1}^N (\phi(\theta^j_k, \mathsf{Z}^j_k, x_k) - y_k)\nabla_\theta g(\theta^i_k) \cdot \nabla_\theta \phi(\theta^i_k, \mathsf{Z}^i_k, x_k) - \mathbf{D}^N_k[g]$.
3. $\mathbb{R}^N_k[g] := \frac{1}{2N}\sum_{i=1}^N (\theta^i_{k+1} - \theta^i_k)^T \nabla^2 g(\widehat{\theta^i_k})(\theta^i_{k+1} - \theta^i_k)$ is the rest of the second order Taylor expansion of $\frac{1}{N}\sum_{k=1}^N f(\theta^i_{k+1}) - f(\theta^i_k)$ (the point $\widehat{\theta^i_k}$ lies in $[\theta^i_{k+1}, \theta^i_k]$).

Note that $\mathbf{D}_k^N[g]$ and $\mathbf{M}_k^N[g]$ are well defined for $g \in \mathcal{C}^{1,\mathfrak{j}}(\mathbb{R}^{d+1})$ ($\mathfrak{j} \geq 0$). For $t \geq 0$, we also define:

$$\mathbb{R}_t^N[g] := \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbb{R}_k^N[g] \text{ and } \mathbf{M}_t^N[g] := \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbf{M}_k^N[g]. \tag{6.24}$$

Let $t \geq 0$. With these definitions, we recall that from Eq. (53) in [Descours et al., 2023b], there exist $\widehat{\theta_k^i}$ ($i = 1, \dots, N$ and $k = 0, \dots, \lfloor Nt \rfloor - 1$) such that for $g \in \mathcal{C}^{2,\mathfrak{j}_0}(\mathbb{R}^{d+1})$:

$$\begin{aligned}
\langle g, \mu_t^N \rangle - \langle g, \mu_0^N \rangle = &-\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta g \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \langle \nabla_\theta g \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s \\
&+ \frac{\kappa}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta g \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \frac{\kappa}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta g \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&+ \mathbf{M}_t^N[g] + \mathbf{W}_t^N[g] + \mathbb{R}_t^N[g],
\end{aligned} \tag{6.25}$$

where $\mathbf{W}_t^N[f] := -\mathbf{V}_t^N[f] + \kappa \int_{\frac{\lfloor Nt \rfloor}{N}}^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s$ and

$$\begin{aligned}
\mathbf{V}_t^N[f] := &-\kappa \int_{\frac{\lfloor Nt \rfloor}{N}}^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&+ \frac{\kappa}{N} \int_{\frac{\lfloor Nt \rfloor}{N}}^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \frac{\kappa}{N} \int_{\frac{\lfloor Nt \rfloor}{N}}^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

Hence, since by definition $\Upsilon^N = \sqrt{N}(\mu^N - \bar{\mu}^N)$, one has for all $t \in \mathbb{R}_+$, using (6.23) and (6.25) together with the fact that $\langle f, \mu_0^N \rangle = \langle f, \bar{\mu}_0^N \rangle$:

$$\begin{aligned}
\langle f, \Upsilon_t^N \rangle = &-\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s) \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \Upsilon_s^N \rangle \mathrm{d}s \\
&+ \frac{\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \frac{\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&+ \sqrt{N}\mathbf{M}_t^N[f] + \sqrt{N}\mathbf{W}_t^N[f] + \sqrt{N}\mathbb{R}_t^N[f].
\end{aligned} \tag{6.26}$$

Using **II**, when $\mathfrak{j} > \frac{d+1}{2}$, one has $\mathfrak{H}(\cdot, x) \in \mathcal{H}^{\mathfrak{J},\mathfrak{j}}(\mathbb{R}^{d+1})$ for all $\mathfrak{J} \geq 0$, and it holds:

$$\sup_{x \in \mathsf{X}} \left\| \mathfrak{H}(.,x) \right\|_{\mathcal{H}^{\mathfrak{J},\mathfrak{j}}} < +\infty. \tag{6.27}$$

By Lemma B.3 in [Descours et al., 2022b], one has, for all $t \in \mathbb{R}_+$,

$$\langle f, \Upsilon_t^N \rangle^2 \leq \mathbf{A}_t^N[f] + \mathbf{B}_t^N[f], \tag{6.28}$$

where

$$\begin{aligned}
\mathbf{A}_t^N[f] = &-2\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle f, \Upsilon_s^N \rangle \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- 2\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle f, \Upsilon_s^N \rangle \langle \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- 2\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle f, \Upsilon_s^N \rangle \langle \phi(\cdot,\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s) \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- 2\kappa \int_0^t \langle f, \Upsilon_s^N \rangle \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \Upsilon_s^N \rangle \mathrm{d}s \\
&+ \frac{2\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle f, \Upsilon_s^N \rangle \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&- \frac{2\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle f, \Upsilon_s^N \rangle \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s
\end{aligned} \tag{6.29}$$

and

$$\begin{aligned}
\mathbf{B}_t^N[f] = &\sum_{k=0}^{\lfloor Nt \rfloor - 1} \Big[2\langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle \sqrt{N} \mathbb{R}_k^N[f] + 3N \mathbb{R}_k^N[f]^2 \Big] + \sum_{k=0}^{\lfloor Nt \rfloor - 1} \Big[2\langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle \sqrt{N} \mathbf{M}_k^N[f] + 3N \mathbf{M}_k^N[f]^2 \Big] \\
&+ \sum_{k=0}^{\lfloor Nt \rfloor - 1} \Big[2\langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle \mathbf{a}_k^N[f] + 3\mathbf{a}_k^N[f]^2 \Big] - 2\sqrt{N} \int_0^t \langle f, \Upsilon_s^N \rangle \mathbf{L}_s^N[f] \mathrm{d}s,
\end{aligned}$$

with, for $s \in [0,t]$,

$$\begin{aligned}
\mathbf{L}_s^N[f] = &-\kappa \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&+ \frac{\kappa}{N} \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&- \frac{\kappa}{N} \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&- \kappa \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle
\end{aligned}$$

and, for $0 \leq k < \lfloor Nt \rfloor$, $\mathbf{a}_k^N[f] = \sqrt{N} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \mathbf{L}_s^N[f] \mathrm{d}s$. By (6.22) and (6.28),

$$\|\Upsilon_t^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}} \leq \sum_{a \geq 1} \mathbf{A}_t^N[f_a] + \mathbf{B}_t^N[f_a]. \tag{6.30}$$

Using Lemma 76, one deduces that:

$$\sum_{a \geq 1} \mathbf{E}[\mathbf{A}_t^N[f_a] + \mathbf{B}_t^N[f_a]] \leq C_T + C_T \int_0^t \mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \mathrm{d}s \tag{6.31}$$

Hence, by (6.30) and (6.31),

$$\mathbf{E}[\|\Upsilon_t^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \leq C_T + C_T \int_0^t \mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \mathrm{d}s. \tag{6.32}$$

Using Gronwall's lemma yields the desired moment estimate on Υ^N. □

The following lemma provides the compact containment condition we need to prove that $(\eta^N)_{N \geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$.

Lemma 74. *Assume **A**. Then, for all $T > 0$, $\sup_{N\geq 1} \mathbf{E}[\sup_{t\in[0,T]} \|\eta_t^N\|^2_{\mathcal{H}^{-\mathrm{J}_2,\mathrm{j}_2}}] < +\infty$.*

Proof. Let $T > 0$ and $N \geq 1$. Consider an orthonormal basis $(f_a)_{a\geq 1}$ of $\mathcal{H}^{\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1})$ and $f \in \mathcal{H}^{\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1})$. From (6.26) and using Jensen's inequality,

$$
\begin{aligned}
\sup_{t\in[0,T]} \langle f, \Upsilon_t^N\rangle^2 \leq\ & C\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} |\langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma\rangle|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
& + C\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} |\langle \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma\rangle|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
& + C\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} |\langle \phi(\cdot,\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s) \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma\rangle|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
& + C\int_0^T |\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \Upsilon_s^N\rangle|^2 \mathrm{d}s \\
& + \frac{C}{N}\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
& + \frac{C}{N}\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
& + N \sup_{t\in[0,T]} \mathbf{M}_t^N[f]^2 + N \sup_{t\in[0,T]} \mathbf{W}_t^N[f]^2 + N \sup_{t\in[0,T]} \mathbb{R}_t^N[f]^2. \qquad (6.33)
\end{aligned}
$$

Let us now provide upper bounds on each term appearing in the right-hand side of (6.33). Let us consider the first term in the right-hand side of (6.33). By **II**, for all $\mathrm{J} \geq 1$ and $\mathrm{j} \geq 0$,

$$
\sup_{g\in\mathcal{H}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1}),\|g\|_{\mathcal{H}^{\mathrm{J},\mathrm{j}}}=1} \sup_{x\in\mathsf{X}} \big\|\nabla_\theta g \cdot \mathfrak{h}(\cdot,x)\big\|_{\mathcal{H}^{\mathrm{J}-1,\mathrm{j}}} < +\infty. \qquad (6.34)
$$

By (6.48), (6.34), the embedding $\mathcal{H}^{\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}(\mathbb{R}^{d+1})$ together with Lemma 73, we have, for all $s \in [0,T]$,

$$
\begin{aligned}
&\mathbf{E}\Big[\int_{\mathsf{X}\times\mathsf{Y}} |\langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma\rangle|^2 \pi(\mathrm{d}x,\mathrm{d}y)\Big] \\
&\leq C\mathbf{E}\Big[\int_{\mathsf{X}\times\mathsf{Y}} \big\|\nabla_\theta f \cdot \mathfrak{h}(\cdot,x)\big\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}} \|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}} \pi(\mathrm{d}x,\mathrm{d}y)\Big] \\
&\leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}} \mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}}. \qquad (6.35)
\end{aligned}
$$

Let us now deal with the second term in the right hand side of (6.33). Using (6.27), Lemma 73 and (6.50), and Sobolev embeddings, we have, for all $s \in [0,T]$,

$$
\mathbf{E}[|\langle \phi(\cdot,\cdot,x), \Upsilon_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma\rangle|^2] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}} \mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}},
$$

which provides the required upper bound.

We now consider the t third term in the r.h.s of (6.33). We have, using (6.50) and (6.51), together with the embedding $\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$, for all $s \in [0,T]$,

$$
\mathbf{E}[|\langle \phi(\cdot,\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s) \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s^N \otimes \gamma\rangle|^2] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}.
$$

We now turn to the fourth term in (6.33). Note first that by (6.16), we have that $\nabla_\theta g \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1) \in \mathcal{H}^{\mathrm{J}-1,\mathrm{j}+1}(\mathbb{R}^{d+1})$ for all $g \in \mathcal{H}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1})$, $\mathrm{J} \geq 1$, $\mathrm{j} \geq 0$. Moreover, we have

$$
\sup_{g\in\mathcal{H}^{\mathrm{J},\mathrm{j}}(\mathbb{R}^{d+1}),\|g\|_{\mathcal{H}^{\mathrm{J},\mathrm{j}}}=1} \|\nabla_\theta g \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1)\|_{\mathcal{H}^{\mathrm{J}-1,\mathrm{j}+1}} < +\infty. \qquad (6.36)
$$

Hence, using the embedding $\mathcal{H}^{\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}(\mathbb{R}^{d+1})$ (see the beginning of Section 6) and Lemma 73, we obtain, for all $s \in [0,T]$,

$$\mathbf{E}[|\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \Upsilon_s^N\rangle|^2] \leq \mathbf{E}[\|\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1)\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}} \|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}}.$$

By **A** and Lemma 69, the fifth and sixth terms in the r.h.s of (6.33) are bounded by $C\|f\|^2_{\mathcal{C}^{1,\mathrm{j}_0}}$ and thus by $C\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}$.

We now turn to the three last terms of (6.33). Note first that $t \in \mathbb{R}_+ \mapsto \mathbf{M}_t^N[f]$ is a $\mathcal{F}_t^N$-martingale, where $\mathcal{F}_t^N = \mathcal{F}^N_{\lfloor Nt \rfloor}$ (to see this, use the same computations as those used in the proof of Lemma 3.2 in [Descours et al., 2022b]). Now, using Equations (65), (61) and (62) in [Descours et al., 2023b], we obtain, using Doob's inequality and Sobolev embeddings,

$$\mathbf{E}[\sup_{t\in[0,T]} \mathbf{M}_t^N[f]^2] = \mathbf{E}[\mathbf{M}_T^N[f]^2] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}/N, \tag{6.37}$$

$$\mathbf{E}[\sup_{t\in[0,T]} \mathbf{W}_t^N[f]^2] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}/N,$$

$$\mathbf{E}[\sup_{t\in[0,T]} \mathbb{R}_t^N[f]^2] \leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}/N^2. \tag{6.38}$$

Collecting these bounds, we obtain

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \Upsilon_t^N\rangle^2\Big] \leq C(\|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}} + \|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}} + \|f\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}} + \|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}). \tag{6.39}$$

Hence, by Sobolev embeddings together with the embedding $\mathcal{H}^{\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}(\mathbb{R}^{d+1})$, one deduces that:

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \|\Upsilon_t^N\|^2_{\mathcal{H}^{-\mathrm{J}_2,\mathrm{j}_2}}\Big] \leq C. \tag{6.40}$$

We now turn to the study of $\mathbf{E}[\sup_{t\in[0,T]}\langle f, \Theta_t^N\rangle^2]$. Recall that $\Theta_t^N = \sqrt{N}(\bar{\mu}_t^N - \bar{\mu}_t)$. Using (6.23) and (6.13) (recall that by Lemma 72, one can use test functions $f \in \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$ in (6.13)), one has:

$$\begin{aligned}\langle f, \Theta_t^N\rangle = \langle f, \Theta_0^N\rangle &- \kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Theta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \kappa \int_0^t \langle \nabla f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \Theta_s^N\rangle \mathrm{d}s.\end{aligned} \tag{6.41}$$

By Jensen's inequality, together with (6.48) and Lemma 73, we obtain

$$\begin{aligned}&\mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \Theta_t^N\rangle^2\Big] \\ &\leq C\mathbf{E}[\langle f, \Theta_0^N\rangle^2] + C\mathbf{E}\Big[\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \langle \nabla f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Theta_s^N \otimes \gamma\rangle^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \\ &\quad + C\mathbf{E}\Big[\int_0^T \langle \nabla f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \Theta_s^N\rangle^2 \mathrm{d}s\Big] \\ &\leq C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}} + C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}} \int_0^T \mathbf{E}[\|\Theta_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}]\mathrm{d}s + C\|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}} \int_0^T \mathbf{E}[\|\Theta_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}]\mathrm{d}s \\ &\leq C(\|f\|^2_{\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}} + \|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}} + \|f\|^2_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}}).\end{aligned}$$

Hence, by Sobolev embeddings (see the very beginning of Section 6), we deduce that:

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \|\Theta_t^N\|^2_{\mathcal{H}^{-\mathrm{J}_2,\mathrm{j}_2}}\Big] \leq C. \tag{6.42}$$

Together with (6.40), this completes the proof of the lemma. □

The following lemma provides the regularity condition needed to prove that the sequence of fluctuation processes $(\eta^N)_{N\geq 1}$ is relatively compact in the space $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$.

Lemma 75. *Assume **A**. For all $T>0$, there exist $C>0$ such that for all $N\geq 1$, $\delta>0$, $0\leq r<t\leq T$ with $t-r\leq\delta$ and $f\in\mathcal{C}_c^\infty(\mathbb{R}^{d+1})$, $\mathbf{E}[|\langle f,\eta_t^N\rangle-\langle f,\eta_r^N\rangle|]\leq C(\sqrt{\delta}+\delta+(1+\delta)/\sqrt{N})\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}}$.*

Proof. From (6.26), $\langle f,\Upsilon_t^N\rangle-\langle f,\Upsilon_r^N\rangle$ is equal to:

$$
\begin{aligned}
&-\kappa\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\Upsilon_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&-\kappa\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),\Upsilon_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&-\kappa\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),\sqrt{N}(\bar{\mu}_s^N-\bar{\mu}_s)\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&-\kappa\int_r^t\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\Upsilon_s^N\rangle\mathrm{d}s\\
&+\frac{\kappa}{\sqrt{N}}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&-\frac{\kappa}{\sqrt{N}}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&+\sqrt{N}(\mathbf{M}_t^N[f]-\mathbf{M}_r^N[f])+\sqrt{N}(\mathbf{W}_t^N[f]-\mathbf{W}_r^N[f])+\sqrt{N}(\mathbb{R}_t^N[f]-\mathbb{R}_r^N[f]).
\end{aligned}
\tag{6.43}
$$

Using similar techniques as those used in the proof of Lemma 74, we obtain the following bounds:

$$
\begin{aligned}
&\mathbf{E}\Big[\Big|\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\Upsilon_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|\Big]\leq C\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{i}_1}}(t-r),\\
&\mathbf{E}\Big[\Big|\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),\Upsilon_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|\Big]\leq C\|f\|_{\mathcal{H}^{\mathrm{J}_1,\mathrm{i}_1}}(t-r),\\
&\mathbf{E}\Big[\Big|\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),\sqrt{N}(\bar{\mu}_s^N-\bar{\mu}_s)\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|\Big]\leq C\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{i}_0}}(t-r),\\
&\mathbf{E}\Big[\Big|\int_r^t\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\Upsilon_s^N\rangle\mathrm{d}s\Big|\Big]\leq C\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{i}_1-1}}(t-r),\\
&\mathbf{E}\Big[\Big|\frac{1}{\sqrt{N}}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|\Big]\leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{i}_0}}}{\sqrt{N}}(t-r),\\
&\mathbf{E}\Big[\Big|\frac{1}{\sqrt{N}}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|\Big]\leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{i}_0}}}{\sqrt{N}}(t-r).
\end{aligned}
$$

Let us now treat the three last terms appearing at the last line of Equation (6.43). From the proof of Lemma 21 in [Descours et al., 2023b], we have:

$$
\mathbf{E}[|\mathbf{M}_t^N[f]-\mathbf{M}_r^N[f]|]\leq C\frac{\sqrt{N\delta+1}}{N}\|f\|_{\mathcal{C}^{1,\mathrm{i}_0}}\text{ and }\mathbf{E}[|\mathbb{R}_t^N[f]-\mathbb{R}_r^N[f]|]\leq C\frac{\|f\|_{\mathcal{C}^{2,\mathrm{i}_0}}}{N}.
$$

Let us mention that the upper bound on $\mathbf{E}[|\mathbf{W}_t^N[f]-\mathbf{W}_r^N[f]|]$ provided in the proof of Lemma 21 in[Descours et al., 2023b] (which we recall implies that this term is control by $1/\sqrt{N}$) is not sharp enough. With straightforward computations, from the definition of $\mathbf{W}_t^N[f]$, we actually have:

$$
\mathbf{E}[|\mathbf{W}_t^N[f]-\mathbf{W}_r^N[f]|]\leq\mathbf{E}[|\mathbf{W}_t^N[f]|]+\mathbf{E}[|\mathbf{W}_t^N[f]|]\leq C\frac{\|f\|_{\mathcal{C}^{1,j_0}}}{N}.
$$

In conclusion, using Sobolev embeddings (see the very beginning of Section 6), we obtain

$$
\mathbf{E}[|\langle f,\Upsilon_t^N\rangle-\langle f,\Upsilon_r^N\rangle|]\leq C(\sqrt{\delta}+\delta+(1+\delta)/\sqrt{N})\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{i}_1-1}}.
\tag{6.44}
$$

Let us now consider $\Theta_t^N - \Theta_r^N$. By (6.41), one has:

$$\begin{aligned}\langle f, \Theta_t^N\rangle - \langle f, \Theta_r^N\rangle =& -\kappa \int_r^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Theta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ & - \kappa \int_r^t \langle \nabla f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \Theta_s^N\rangle \mathrm{d}s. \end{aligned} \tag{6.45}$$

By (6.48), (6.34) and (6.36), together with Lemma 73, it then holds:

$$\begin{aligned}\mathbf{E}\big[|\langle f, \Theta_t^N\rangle - \langle f, \Theta_r^N\rangle|\big] &\le C\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}} \int_r^t \mathbf{E}[\|\Theta_s^N\|_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}]\mathrm{d}s + C\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}} \int_r^t \mathbf{E}[\|\Theta_s^N\|_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}]\mathrm{d}s \\ &\le C\delta(\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1}} + \|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}}). \end{aligned} \tag{6.46}$$

Hence, by (6.44) and (6.46), and recalling that $\eta^N = \Upsilon^N + \Theta^N$, we get that $\mathbf{E}[|\langle f, \eta_t^N\rangle - \langle f, \eta_r^N\rangle|] \le C(\sqrt{\delta} + \delta + (1+\delta)/\sqrt{N})\|f\|_{\mathcal{H}^{\mathrm{J}_1+1,\mathrm{j}_1-1}}$. □

Lemma 76. *Assume **A**. Let $(f_a)_{a\ge1}$ be an orthonormal basis of $\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})$. Then, for all $T>0$, there exists $C>0$ such that for all $0\le t\le T$,*

(i)

$$\begin{aligned}&\sum_{a\ge1}\mathbf{E}\Big[-2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon_s^N\rangle\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\Upsilon_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \\ &\le C\int_0^t\mathbf{E}\Big[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big]\mathrm{d}s.\end{aligned}$$

(ii)

$$\sum_{a\ge1}\mathbf{E}\Big[-2\kappa\int_0^t\langle f_a,\Upsilon_s^N\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\Upsilon_s^N\rangle\mathrm{d}s\Big]\le C\int_0^t\mathbf{E}\Big[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big]\mathrm{d}s.$$

(iii)

$$\begin{aligned}&\sum_{a\ge1}\mathbf{E}\Big[-2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon_s^N\rangle\langle\phi(\cdot,\cdot,x),\Upsilon_s^N\otimes\gamma\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &\qquad -2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon_s^N\rangle\langle\phi(\cdot,\cdot,x),\sqrt{N}(\bar{\mu}_s^N-\bar{\mu}_s)\otimes\gamma\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \\ &\le C+C\int_0^t\mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}]\mathrm{d}s.\end{aligned}$$

(iv)

$$\begin{aligned}&\sum_{a\ge1}\mathbf{E}\Big[\frac{2\kappa}{\sqrt{N}}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon_s^N\rangle\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &\qquad -\frac{2\kappa}{\sqrt{N}}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon_s^N\rangle\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \\ &\le C+C\int_0^t\mathbf{E}\Big[|\|\Upsilon_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big]\mathrm{d}s.\end{aligned}$$

(v)

$$\sum_{a\ge1}\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[2\langle f_a,\Upsilon^N_{\frac{k+1}{N}-}\rangle\sqrt{N}\mathbf{M}_k^N[f_a]+3N\mathbf{M}_k^N[f_a]^2\Big]\Big]\le C.$$

(vi)

$$\sum_{a\geq 1}\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[2\langle f_a,\Upsilon^N_{\frac{k+1}{N}-}\rangle\sqrt{N}\mathbb{R}^N_k[f_a]+3N\mathbb{R}^N_k[f_a]^2\Big]\Big]\leq C+\int_0^t\mathbf{E}\Big[\|\Upsilon^N_s\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big]\mathrm{d}s.$$

(vii)

$$\sum_{a\geq 1}\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[2\langle f_a,\Upsilon^N_{\frac{k+1}{N}-}\rangle\mathbf{a}^N_k[f_a]+3\mathbf{a}^N_k[f_a]^2\Big]-2\sqrt{N}\int_0^t\langle f_a,\Upsilon^N_s\rangle\mathbf{L}^N_s[f]\mathrm{d}s\Big]\leq C.$$

Proof. Let $0\leq t\leq T$ and $N\geq 1$. Consider an orthonormal basis $(f_a)_{a\geq 1}$ of $\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})$ and a function $f\in\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})$. In what follows, $C>0$ will denote a constant independent of t, N, $s\in[0,t]$, f and $(f_a)_{a\geq 1}$, which can change from one occurrence to another. Let us prove item *(i)*. Introduce for $x\in\mathsf{X}$, the operator $\mathbf{T}_x:\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})\to\mathcal{H}^{\mathrm{J}_1-1,\mathrm{j}_1}(\mathbb{R}^{d+1})$ defined by

$$\theta\in\mathbb{R}^{d+1}\mapsto\mathbf{T}_x(f)(\theta)=\nabla_\theta f(\theta)\cdot\nabla_\theta\int_{\mathbb{R}^d}\phi(\theta,z,x)\gamma(z)\mathrm{d}z=\nabla_\theta f\cdot\mathfrak{H}(\cdot,x),\tag{6.47}$$

where we recall that $\phi(\theta,z,x)=s(\Psi_\theta(z),x)$. Note that $\mathbf{T}_x$ is well defined since the function $\mathfrak{H}(\cdot,x):\theta\mapsto\int_{\mathbb{R}^d}\phi(\theta,z,x)\gamma(z)\mathrm{d}z=\langle\phi(\theta,\cdot,x),\gamma\rangle$ is smooth and all its derivatives of non negative order are uniformly bounded w.r.t $x\in\mathsf{X}$ over $\mathbb{R}^{d+1}$ (this follows from **A1** and **A3**). Then, one has

$$\begin{aligned}
&\sum_{a\geq 1}-2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon^N_s\rangle\langle\phi(\cdot,\cdot,x)-y,\mu^N_s\otimes\gamma\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\Upsilon^N_s\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&=-2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu^N_s\otimes\gamma\rangle\sum_{a\geq 1}\langle f_a,\Upsilon^N_s\rangle\langle\mathbf{T}_x f_a,\Upsilon^N_s\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&=-2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu^N_s\otimes\gamma\rangle\langle\Upsilon^N_s,\mathbf{T}^*_x\Upsilon^N_s\rangle_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

Since the function ϕ is bounded and Y is compact, one has:

$$\exists C>0,\forall\nu\in\mathcal{P}(\mathbb{R}^{d+1}),\forall(x,y)\in\mathsf{X}\times\mathsf{Y},\ |\langle\phi(\cdot,\cdot,x)-y,\nu\otimes\gamma\rangle|\leq C.\tag{6.48}$$

By (6.48) and using Lemma B.2 in [Descours et al., 2022b] (note that $\Upsilon^N\in\mathcal{D}(\mathbb{R}_+,\mathcal{H}^{-\mathrm{J}_1+1,\mathrm{j}}(\mathbb{R}^{d+1}))$ by (6.21) together with the Sobolev embedding $\mathcal{H}^{\mathrm{J}_1-1,\mathrm{j}}(\mathbb{R}^{d+1})\hookrightarrow\mathcal{C}^{1,\mathrm{j}_1}(\mathbb{R}^{d+1})$, $\mathrm{j}\geq 0$), we have

$$\begin{aligned}
&\mathbf{E}\Big[\sum_{a\geq 1}-2\kappa\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle f_a,\Upsilon^N_s\rangle\langle\phi(\cdot,\cdot,x)-y,\mu^N_s\otimes\gamma\rangle\langle\nabla_\theta f_a\cdot\nabla_\theta\phi(\cdot,\cdot,x),\Upsilon^N_s\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big]\\
&\leq C\int_0^t\mathbf{E}\Big[\|\Upsilon^N_s\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big]\mathrm{d}s,
\end{aligned}$$

which is the desired estimate.

Introduce the operator $\mathbf{T}:\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})\to\mathcal{H}^{\mathrm{J}_1-1,\mathrm{j}_1+1}(\mathbb{R}^{d+1})$ defined by (see also (6.16))

$$\mathbf{T}(f):\theta\mapsto\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P^1_0),\tag{6.49}$$

Item *(ii)* is proved as the previous item, using now Lemma 77 below.

Item *(iii)* is obtained with exactly the same arguments as those used to derive the upper bounds on $\sum_{a\geq 1}\mathbf{J}^N_t[f_a]$ and $\sum_{a\geq 1}\mathbf{K}^N_t[f_a]$ in the proof of Lemma 3.1 in [Descours et al., 2022b] (it suffices indeed to change $\sigma(\cdot,x)$ there into $\mathfrak{H}(\cdot,x)$). In particular, by **II** and (6.19), it holds:

$$\sup_{x\in\mathsf{X}}|\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar{\mu}^N_s\otimes\gamma\rangle|\leq C\|f\|_{\mathcal{C}^{1,\mathrm{j}_0}},\tag{6.50}$$

and (see Equation (3.20) in [Descours et al., 2022b]),

$$\mathbf{E}[\langle \phi(\cdot,\cdot,x), (\bar{\mu}_s^N - \bar{\mu}_s) \otimes \gamma \rangle^2] \leq C/N. \tag{6.51}$$

Note also that by Lemma 69 and **I**, it holds:

$$\mathbf{E}\big[\langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \mu_s^N \otimes \gamma \rangle^2\big] \leq C \|f\|^2_{\mathcal{C}^{1,\mathrm{j}_0}} \tag{6.52}$$

Item *(iv)* follows from $\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$ and $\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$.

Let us prove item *(v)*. Since $\mathbf{E}[\mathbf{M}_k^N[f]|\mathcal{F}_k^N] = 0$, we have with the same arguments as those used to derive Equation (B.1) in [Descours et al., 2022b],

$$\sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbf{E}[\langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle \sqrt{N} \mathbf{M}_k^N[f]] = 0.$$

Moreover, we recall that by Lemma 69 (see Eqaution (60) in [Descours et al., 2023b]), one has $\mathbf{E}[\mathbf{M}_k^N[f]^2] \leq C\|f\|^2_{\mathcal{C}^{1,j_0}}/N^2$. Hence, we conclude, using again $\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$ and $\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$, that

$$\sum_{a\geq 1} \mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt \rfloor - 1} \Big[2\langle f_a, \Upsilon^N_{\frac{k+1}{N}-} \rangle \sqrt{N} \mathbf{M}_k^N[f_a] + 3N \mathbf{M}_k^N[f_a]^2 \Big] \Big] \leq C \sum_{a\geq 1} \|f_a\|^2_{\mathcal{C}^{1,j_0}} \leq C. \tag{6.53}$$

Let us prove item *(vi)*. We have

$$\sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle \sqrt{N} \mathbb{R}_k^N[f] \leq \sum_{k=0}^{\lfloor Nt \rfloor - 1} \frac{1}{N} \langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle^2 + \sum_{k=0}^{\lfloor Nt \rfloor - 1} N^2 \mathbb{R}_k^N[f]^2.$$

Recall that from the analysis performed at the end of the proof of Lemma B.1 in [Descours et al., 2023b], $\mathbf{E}[\mathbb{R}_k^N[f]^2] \leq C/N^4$ so that

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt \rfloor - 1} N^2 \mathbb{R}_k^N[f]^2 \Big] \leq C\|f\|^2_{\mathcal{C}^{2,\mathrm{j}_0}}/N.$$

Using (6.19) and Lemma 69, the same computations as those of the proof of item *(iv)* in Lemma B.1 in [Descours et al., 2022b] yield:

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt \rfloor - 1} \frac{1}{N} \langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle^2 \Big] \leq C\|f\|^2_{\mathcal{C}^{2,\mathrm{j}_0}} + \mathbf{E}\Big[\int_0^t \langle f, \Upsilon_s^N \rangle^2 \mathrm{d}s \Big].$$

Hence,

$$\sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle \sqrt{N} \mathbb{R}_k^N[f] \leq C\|f\|^2_{\mathcal{C}^{2,\mathrm{j}_0}} + \mathbf{E}\Big[\int_0^t \langle f, \Upsilon_s^N \rangle^2 \mathrm{d}s \Big] + C\|f\|^2_{\mathcal{C}^{2,\mathrm{j}_0}}/N. \tag{6.54}$$

Item *(vi)* then follows from $\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{2,\mathrm{j}_0}(\mathbb{R}^{d+1})$ and $\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$.

Let us prove item *(vii)*. Using Jensen's inequality together with Lemma 69 and (6.16), we have, for all $0 \leq s \leq t$,

$$\mathbf{E}[|\mathbf{L}_s^N[f]|^2] \leq C\|f\|^2_{\mathcal{C}^{1,j_0}}(1 + 1/N). \tag{6.55}$$

On the other hand, for all $s \in (\frac{k}{N}, \frac{k+1}{N})$, by (6.19) and the same computations as those used to derive Equation (B.5) in [Descours et al., 2022b], we have:

$$|\langle f, \Upsilon^N_{\frac{k+1}{N}-} \rangle - \langle f, \Upsilon_s^N \rangle| = \sqrt{N} \big| \langle f, \bar{\mu}_s^N \rangle - \langle f, \bar{\mu}^N_{\frac{k+1}{N}} \rangle \big| \leq C\|f\|_{\mathcal{C}^{2,\mathrm{j}_0}}. \tag{6.56}$$

Hence,

$$\begin{aligned}
&\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle \mathbf{a}^N_k[f]-\sqrt{N}\int_0^{\frac{\lfloor Nt\rfloor}{N}}\langle f,\Upsilon^N_s\rangle \mathbf{L}^N_s[f]\mathrm{d}s\Big]\\
&=\sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{E}\Big[\Big(\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle-\langle f,\Upsilon^N_s\rangle\Big)\mathbf{L}^N_s[f]\Big]\mathrm{d}s\\
&\leq C\|f\|_{\mathcal{C}^{2,\mathfrak{j}_0}}\int_0^{\frac{\lfloor Nt\rfloor}{N}}\mathbf{E}[|\mathbf{L}^N_s[f]|]\mathrm{d}s\leq C\|f\|^2_{\mathcal{C}^{2,\mathfrak{j}_0}}. \qquad (6.57)
\end{aligned}$$

We also have, using Lemma 69 and (6.19), it is straightforward to deduce that $\mathbf{E}[\langle f,\Upsilon^N_s\rangle^2]\leq CN\|f\|^2_{\mathcal{C}^{1,j_0}}$. Consequently, one has:

$$\mathbf{E}\Big[\sqrt{N}\Big|\int_{\frac{\lfloor Nt\rfloor}{N}}^t\langle f,\Upsilon^N_s\rangle\mathbf{L}^N_s[f]\mathrm{d}s\Big|\Big]\leq\sqrt{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^t\sqrt{\mathbf{E}[\langle f,\Upsilon^N_s\rangle^2]}\sqrt{\mathbf{E}[\mathbf{L}^N_s[f]^2]}\leq C\|f\|^2_{\mathcal{C}^{1,j_0}}. \qquad (6.58)$$

Finally,

$$\begin{aligned}
\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{a}^N_k[f]^2\Big]=N\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{L}^N_s[f]\mathrm{d}s\Big|^2\Big]&\leq\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{E}[\mathbf{L}^N_s[f]^2]\mathrm{d}s\\
&\leq C\|f\|^2_{\mathcal{C}^{1,j_0}}(1+1/N). \qquad (6.59)
\end{aligned}$$

Item *(vii)* follows from (6.57), (6.58) and (6.59). The proof of the lemma is complete. □

Lemma 77. *Let $\mathrm{J}\geq 1$ and $\mathfrak{j}\geq 0$. Recall the definition of $\mathbf{T}\in\mathcal{L}(\mathcal{H}^{\mathrm{J},\mathfrak{j}}(\mathbb{R}^{d+1}),\mathcal{H}^{\mathrm{J}-1,\mathfrak{j}+1}(\mathbb{R}^{d+1}))$ in* (6.49). *Then, there exists $C>0$ such that for any $\Upsilon\in\mathcal{H}^{-\mathrm{J}+1,\mathfrak{j}+1}(\mathbb{R}^{d+1})$,*

$$|\langle\Upsilon,\mathbf{T}^*\Upsilon\rangle_{\mathcal{H}^{-\mathrm{J},\mathfrak{j}}}|\leq C\|\Upsilon\|^2_{\mathcal{H}^{-\mathrm{J},\mathfrak{j}}}. \qquad (6.60)$$

Note that $\mathbf{T}^*\in\mathcal{L}(\mathcal{H}^{-\mathrm{J}+1,\mathfrak{j}+1}(\mathbb{R}^{d+1}),\mathcal{H}^{-\mathrm{J},\mathfrak{j}}(\mathbb{R}^{d+1}))$. Let us mention that the upper bound (6.60) is much better than the one which would be obtained applying the Cauchy-Schwarz inequality. **Proof.** The proof is inspired from the one of Lemma B.2 in [Descours et al., 2022b] (see also Lemma B1 in [Sirignano and Spiliopoulos, 2020c]). We will give the proof in dimension 1, i.e when $d=0$, the other cases are treated the same way. Let $\Upsilon\in\mathcal{H}^{-\mathrm{J}+1,\mathfrak{j}+1}(\mathbb{R})\hookrightarrow\mathcal{H}^{-\mathrm{J},\mathfrak{j}}(\mathbb{R})$. By the Riesz representation theorem, there exists a unique $\Psi\in\mathcal{H}^{\mathrm{J},\mathfrak{j}}(\mathbb{R})$ such that

$$\langle f,\Upsilon\rangle=\langle f,\Psi\rangle_{\mathcal{H}^{\mathrm{J},\mathfrak{j}}},\ \forall f\in\mathcal{H}^{\mathrm{J},\mathfrak{j}}(\mathbb{R}).$$

Define F by $F(\Upsilon)=\Psi$. The density of $\mathcal{C}^\infty_c(\mathbb{R})$ in $\mathcal{H}^{\mathrm{J},\mathfrak{j}}(\mathbb{R})$ implies that $\{\Upsilon\in\mathcal{H}^{-\mathrm{J},\mathfrak{j}}(\mathbb{R}):F(\Upsilon)\in\mathcal{C}^\infty_c(\mathbb{R})\}$ is dense in $\mathcal{H}^{-\mathrm{J},\mathfrak{j}}(\mathbb{R})$. It is thus sufficient to show (6.60) when $\Psi=F(\Upsilon)\in\mathcal{C}^\infty_c(\mathbb{R})$. We have

$$\langle\Upsilon,\mathbf{T}\Upsilon\rangle_{\mathcal{H}^{-\mathrm{J},\mathfrak{j}}}=\langle\Psi,\mathbf{T}^*\Upsilon\rangle=\langle\mathbf{T}\Psi,\Upsilon\rangle=\langle\mathbf{T}\Psi,\Psi\rangle_{\mathcal{H}^{\mathrm{J},\mathfrak{j}}}. \qquad (6.61)$$

Hence, to prove (6.60), it is enough to show $|\langle\mathbf{T}\Psi,\Psi\rangle_{\mathcal{H}^{\mathrm{J},\mathfrak{j}}}|\leq C\|\Psi\|^2_{\mathcal{H}^{\mathrm{J},\mathfrak{j}}}$ for $\Psi\in\mathcal{C}^\infty_c(\mathbb{R})$. We will only consider the case when $\mathrm{J}=\mathfrak{j}=1$, the other cases being treated very similarly. Recall the upper bounds (6.16). Let $\Psi\in\mathcal{C}^\infty_c(\mathbb{R})$.

We have, by integration by parts and using the fact that Ψ is compactly supported,

$$\begin{aligned}\langle \mathbf{T}\Psi, \Psi\rangle_{\mathcal{H}^{1,1}} &= \int_{\mathbb{R}} \Psi'(\theta)\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)\frac{\Psi(\theta)}{1+\theta^2}\mathrm{d}\theta + \int_{\mathbb{R}} (\Psi'(\theta)\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0))'\frac{\Psi'(\theta)}{1+\theta^2}\mathrm{d}\theta \\ &= \int_{\mathbb{R}} \Psi'(\theta)\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)\frac{\Psi(\theta)}{1+\theta^2}\mathrm{d}\theta + \int_{\mathbb{R}} \Psi''(\theta)\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)\frac{\Psi'(\theta)}{1+\theta^2}\mathrm{d}\theta \\ &\quad + \int_{\mathbb{R}} \mathscr{D}''_{\mathrm{KL}}(q^1_\theta|P^1_0)\frac{\Psi'(\theta)^2}{1+\theta^2}\mathrm{d}\theta \\ &= -\frac{1}{2}\int_{\mathbb{R}} \Psi(\theta)^2 \frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)}{1+\theta^2}\Big)\mathrm{d}\theta - \frac{1}{2}\int_{\mathbb{R}} \Psi'(\theta)^2 \frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)}{1+\theta^2}\Big)\mathrm{d}\theta \\ &\quad + \int_{\mathbb{R}} \mathscr{D}''_{\mathrm{KL}}(q^1_\theta|P^1_0)\frac{\Psi'(\theta)^2}{1+\theta^2}\mathrm{d}\theta. \end{aligned} \tag{6.62}$$

To bound the first two terms of (6.62), we use the bounds (6.16). More precisely, for all $\theta \in \mathbb{R}$,

$$\Big|\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)}{1+\theta^2}\Big)\Big| \le \frac{|\mathscr{D}''_{\mathrm{KL}}(q^1_\theta|P^1_0)(1+\theta^2)| + 2|\theta\mathscr{D}'_{\mathrm{KL}}(q^1_\theta|P^1_0)|}{(1+\theta^2)^2} \le \frac{C}{1+\theta^2} + \frac{C|\theta|(1+|\theta|)}{(1+\theta^2)^2} \le \frac{C}{1+\theta^2}.$$

Hence, we obtain, plugging this bound in (6.62),

$$|\langle \mathbf{T}\Psi, \Psi\rangle_{\mathcal{H}^{1,1}}| \le C\Big(\int_{\mathbb{R}} \frac{\Psi(\theta)^2}{1+\theta^2}\mathrm{d}\theta + \int_{\mathbb{R}} \frac{\Psi'(\theta)^2}{1+\theta^2}\mathrm{d}\theta\Big) \le C\|\Psi\|^2_{\mathcal{H}^{1,1}}.$$

This completes the proof of the lemma. □

We now collect the previous results to prove Proposition 70. **Proof.** [Proof of Proposition 70] The proof consists in applying Th. 4.6 in [Jakubowski, 1986] with $E = \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$ and $\mathbb{F} = \{\mathsf{H}_f,\ f \in \mathcal{C}^\infty_c(\mathbb{R}^{d+1})\}$ where

$$\mathsf{H}_f : \nu \in \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}) \mapsto \langle f, \nu\rangle.$$

Note that $\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}(\mathbb{R}^{d+1})$ is compactly embedded in $\mathcal{H}^{\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1})$. Hence, by Schauder's theorem, $\mathcal{H}^{-\mathrm{J}_2,\mathrm{j}_2}(\mathbb{R}^{d+1})$ is compactly embedded in $\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$. Thus, for all $C > 0$, the set $\{h \in \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}),\ \|h\|_{\mathcal{H}^{-\mathrm{J}_2,\mathrm{j}_2}} \le C\}$ is compact. Hence, Condition (4.8) in Th. 4.6 in [Jakubowski, 1986] follows from Lemma 74 and Markov's inequality. Let us now show that Condition (4.9) in [Jakubowski, 1986] is verified, i.e., that for all $f \in \mathcal{C}^\infty_c(\mathbb{R}^{d+1})$, the sequence $(\langle f, \eta^N\rangle)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathbb{R})$. To do this, it suffices to use Lemma 75 and Prop. A.1 in [Descours et al., 2022b] (with $\mathcal{H}_1 = \mathcal{H}_2 = \mathbb{R}$ there). In conclusion, according to Th. 4.6 in [Jakubowski, 1986], the sequence $(\eta^N)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$. □

6.2 Relative compactness of $(\sqrt{N}\mathbf{M}^N)_{N\ge1}$ and regularity of the limit points

Throughout this section, we that the $\{\theta^i_k, i \in \{1,\dots,N\}\}$'s are generated by the algorithm (6.7) (with straightforward modifications, one can check that all the results of this section are valid when the $\{\theta^i_k, i \in \{1,\dots,N\}\}$'s are generated by the algorithms (6.5) and (6.9)).

Lemma 78. *Assume **A**. Then, for all $T > 0$, $\sup_{N\ge1} \mathbf{E}\Big[\sup_{t\in[0,T]} \|\sqrt{N}\mathbf{M}^N_t\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big] < +\infty$.*

Proof.

Recall that by (6.37), there exists $C > 0$ such that for all $f \in \mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})$ and $N \ge 1$,

$$\mathbf{E}[\sup_{t\in[0,T]} |\sqrt{N}\mathbf{M}^N_t[f]|^2] \le C\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}.$$

Considering an orthonormal basis of $\mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$, one gets that $\mathbf{E}[\sup_{t\in[0,T]} \|\sqrt{N}\mathbf{M}^N_t\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}] \le C$ uniformly in $N \ge 1$.

□

We now turn to the regularity condition on the sequence $\{t \in \mathbb{R}_+ \mapsto \sqrt{N}\mathbf{M}^N_t[f]\}_{N\ge1}$, for $f \in \mathcal{C}^\infty_c(\mathbb{R}^{d+1})$.

Lemma 79. *Assume **A**. Then, for all $T > 0$, there exists $C > 0$ such that for all $N \geq 1$, $\delta > 0$, $0 \leq r < t \leq T$ such that $t - r \leq \delta$ and $f \in \mathcal{C}_c^\infty(\mathbb{R}^{d+1})$, it holds*

$$\mathbf{E}\big[|\sqrt{N}\mathbf{M}_t^N[f] - \sqrt{N}\mathbf{M}_r^N[f]|\big] \leq C\sqrt{N\delta + 1}\frac{\|f\|_{\mathcal{C}^{1,\mathrm{j}_0}}}{\sqrt{N}}.$$

Proof. From the proof of Lemma 21 in [Descours et al., 2023b], it holds

$$\mathbf{E}\big[|\mathbf{M}_t^N[f] - \mathbf{M}_r^N[f]|^2\big] \leq C(N\delta + 1)\frac{\|f\|_{\mathcal{C}^{1,\mathrm{j}_0}}^2}{N^2}.$$

This leads the desired result.

□

Proposition 80. *Assume **A**. Then, the sequence $\{t \in \mathbb{R}_+ \mapsto \sqrt{N}\mathbf{M}_t^N\}_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$.*

Proof. Recall that $\mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1})$. The same arguments as those used to prove Proposition 70 together with Lemmata 78 and 79 imply the result. □

We now turn to the regularity of the limit points of the sequence $(\eta^N)_{N\geq 1}$.

Lemma 81. *Assume **A**. Then, for all $T > 0$,*

$$\lim_{N\to\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \|\eta_t^N - \eta_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}^2\Big] + \mathbf{E}\Big[\sup_{t\in[0,T]} \|\sqrt{N}\mathbf{M}_t^N - \sqrt{N}\mathbf{M}_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}}^2\Big] = 0. \tag{6.63}$$

Any limit point of $(\eta^N)_{N\geq 1}$ (resp. of $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$) in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$ (resp. in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$) belongs a.s. to $\mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$ (resp. to $\mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$).

Proof. Let $T > 0$. Let us first consider the sequence $(\eta^N)_{N\geq 1}$. In what follows, $C > 0$ is a constant independent of $N \geq 1$, $k \in \{1, \ldots, \lfloor NT \rfloor\}$, and $f \in \mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}(\mathbb{R}^{d+1})$. We have

$$\sup_{t\in[0,T]} \|\eta_t^N - \eta_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}^2 \leq 2 \sup_{t\in[0,T]} \|\Upsilon_t^N - \Upsilon_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}^2 + 2 \sup_{t\in[0,T]} \|\Theta_t^N - \Theta_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}^2. \tag{6.64}$$

According to Lemma 72, one has, for all $t \in \mathbb{R}_+$ and $N \geq 1$, $\|\Theta_t^N - \Theta_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}} = 0$. In addition, since a.s. $\bar{\mu}^N \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1}))$, it follows, by definition of Υ^N, that a.s. for all $N \geq 1$,

$$\sup_{t\in[0,T]} \langle f, \Upsilon_t^N - \Upsilon_{t^-}^N\rangle^2 = N \sup_{t\in[0,T]} \langle f, \mu_t^N - \mu_{t^-}^N\rangle^2. \tag{6.65}$$

The function $t \in [0,T] \mapsto \langle f, \mu_t^N\rangle$ has exactly $\lfloor NT \rfloor$ discontinuities located at times $t_k = k/N$ ($k \in \{1, \ldots, \lfloor NT \rfloor\}$). In addition, from (6.26), for $k \in \{1, \ldots, \lfloor NT \rfloor\}$, its k-th discontinuity is bounded by

$$\begin{aligned}
\boldsymbol{\delta}_k^N[f] := {} & |\mathbf{M}_{k-1}^N[f]| + |\mathbb{R}_{k-1}^N[f]| \\
& + \kappa\Big|\int_{\frac{k-1}{N}}^{\frac{k}{N}} \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big| \\
& + \frac{\kappa}{N}\Big|\int_{\frac{k-1}{N}}^{\frac{k}{N}} \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big| \\
& + \frac{\kappa}{N}\Big|\int_{\frac{k-1}{N}}^{\frac{k}{N}} \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big| \\
& + \kappa\Big|\int_{\frac{k-1}{N}}^{\frac{k}{N}} \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N\rangle \mathrm{d}s\Big|.
\end{aligned}$$

Thus,

$$\sup_{t\in[0,T]} \langle f, \mu_t^N - \mu_{t^-}^N\rangle^2 \leq \max\{|\boldsymbol{\delta}_{k+1}^N[f]|^2,\, 0 \leq k < \lfloor NT\rfloor\}. \tag{6.66}$$

Using the bounds provided by the proof of Lemma 19 in [Descours et al., 2023b], we obtain, for $0 \leq k < \lfloor NT\rfloor$,

$$\mathbf{E}[|\mathbf{M}_k^N[f]|^4] \leq C\frac{\|f\|_{\mathcal{C}^{1,\mathrm{j}_0}}^4}{N^4} \leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}^4}{N^4}, \quad \mathbf{E}[|\mathbb{R}_k^N[f]|^4] \leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}^4}{N^8}, \tag{6.67}$$

and

$$\begin{aligned}
&\mathbf{E}\Big[\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|^4 \\
&+ \frac{1}{N}\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|^4 \\
&+ \frac{1}{N}\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|^4\Big] \leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}^4}{N^4}.
\end{aligned}$$

In addition, one also has (see Equation (57) in [Descours et al., 2023b]):

$$\mathbf{E}\Big[\Big|\int_{\frac{k-1}{N}}^{\frac{k}{N}} \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N\rangle \mathrm{d}s\Big|^4\Big] \leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}^4}{N^4}.$$

Consequently, it holds:

$$\mathbf{E}[\max\{|\boldsymbol{\delta}_{k+1}^N[f]^2,\, 0 \leq k < \lfloor NT\rfloor\}] \leq \Big|\sum_{k=0}^{\lfloor NT\rfloor - 1} \mathbf{E}[\boldsymbol{\delta}_k^N[f]^4]\Big|^{1/2} \leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}^2}{N^{3/2}}.$$

Hence

$$\mathbf{E}\Big[N \sup_{t\in[0,T]} \langle f, \mu_t^N - \mu_{t^-}^N\rangle^2\Big] \leq C\frac{\|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}^2}{\sqrt{N}}. \tag{6.68}$$

Since $\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}(\mathbb{R}^{d+1}) \hookrightarrow_{\mathrm{H.S.}} \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$, one deduces that $\mathbf{E}[\sup_{t\in[0,T]} \|\eta_t^N - \eta_{t^-}^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}^2] \to 0$ as $N \to +\infty$. The fact that any limit points of $(\eta^N)_{N\geq 1}$ is a.s. continuous follows from Condition 3.28 in Proposition 3.26 of [Jacod and Shiryaev, 1987].

The case of the sequence $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$ is treated very similarly. The proof of the lemma is complete. □

6.3 Convergence of $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$ to a $\mathfrak{G}$-process

In this section, we prove that the sequence $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$ converges towards a $\mathfrak{G}$-process (see Definition 66), see Proposition 85. The case when the $\{\theta_k^i, i \in \{1,\ldots,N\}\}$'s are generated by the algorithm (6.7) requires extra analysis compared to the cases when the $\{\theta_k^i, i \in \{1,\ldots,N\}\}$'s are generated by the algorithms (6.5) or (6.9) (see indeed the second part of the proof of Proposition 85 and Lemma 83 below).

Proposition 82. *Assume that the $\{\theta_k^i, i \in \{1,\ldots,N\}\}$'s are generated either by the algorithm* (6.5) *or by the algorithm* (6.7)*. Then, for every $f \in \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$, the sequence $\{t \in \mathbb{R}_+ \mapsto \sqrt{N}\mathbf{M}_t^N[f]\}_{N\geq 1}$ converges in distribution in $\mathcal{D}(\mathbb{R}_+,\mathbb{R})$ towards a process $\mathcal{X}^f \in \mathcal{C}(\mathbb{R}_+,\mathbb{R})$ that has independent Gaussian increments. Moreover, for all $t \in \mathbb{R}_+$,*

$$\mathbf{E}[\mathcal{X}_t^f] = 0 \text{ and } Var(\mathcal{X}_t^f) = \kappa^2 \int_0^t Var_\pi(\mathscr{Q}[f](x,y,\bar{\mu}_s))\mathrm{d}s,$$

where we recall $\mathscr{Q}[f](x,y,\bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$ (see Theorem 68).

Proof. We treat separately the two cases when the $\{\theta_k^i, i \in \{1,\ldots,N\}\}$'s are generated by the algorithm (6.5) or by the algorithm (6.7). Let $f \in \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$.

The case of the Idealized algorithm (6.5)**.** Let us assume that the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the algorithm (6.5). To prove the desired result, we apply the martingale central limit theorem 5.1.4 in [Ethier and Kurtz, 2009] to the sequence $\{t \in \mathbb{R}_+ \mapsto \sqrt{N}\mathbf{M}_t^N[f]\}_{N\geq 1}$. Let us first show that Condition (a) in Th. 7.1.4 in [Ethier and Kurtz, 2009] holds. First of all, by Remark 7.1.5 in [Ethier and Kurtz, 2009], the covariation matrix of $\sqrt{N}\mathbf{M}_t^N[f]$ is

$$\mathfrak{a}_t^N[f] = N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{M}_k^N[f]^2 \tag{6.69}$$

In particular, $\mathfrak{a}_t^N[f] - \mathfrak{a}_s[f] \geq 0$ when $t \geq s$. On the other hand, by (6.67) (which, we recall, also holds when the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the algorithm (6.5)), we have for all $T \geq 0$:

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} |\sqrt{N}\mathbf{M}_t^N[f] - \sqrt{N}\mathbf{M}_{t^-}^N[f]|\Big] = 0. \tag{6.70}$$

Thus Condition (a) in Th. 7.1.4 in [Ethier and Kurtz, 2009] is satisfied. Let us prove the last required condition in Theorem 7.1.4 of [Ethier and Kurtz, 2009], namely that for all $t \in \mathbb{R}_+$, $\lim_N \mathfrak{a}_t^N[f] = \mathfrak{c}_t[f]$ in $\mathbf{P}$-probability, where $\mathfrak{c}$ satisfies the assumptions of Th. 7.1.1 in [Ethier and Kurtz, 2009] (i.e., $t \in \mathbb{R}_+ \mapsto \mathfrak{c}_t[f]$ is continuous, $\mathfrak{c}_0[f] = 0$, and $\mathfrak{c}_t[f] - \mathfrak{c}_s[f] \geq 0$ if $t \geq s$). Let us consider and fix $t \geq 0$. We recall that when the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the algorithm (6.5), one has that for $k \geq 0$ (see Equation (21) in [Descours et al., 2023b]),

$$\begin{aligned}
\mathbf{D}_k^N[f] &= -\frac{\kappa}{N^3}\sum_{i=1}^N\sum_{j=1,j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\theta_k^j,\cdot,x) - y, \gamma\big\rangle \big\langle \nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\
&\quad - \frac{\kappa}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \nu_k^N\otimes\gamma\big\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\
&= -\frac{\kappa}{N^3}\sum_{i=1}^N\sum_{j=1}^N \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\theta_k^j,\cdot,x) - y, \gamma\big\rangle \big\langle \nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\
&\quad + \frac{\kappa}{N^3}\sum_{i=1}^N \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\theta_k^i,\cdot,x) - y, \gamma\big\rangle \big\langle \nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\
&\quad - \frac{\kappa}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \nu_k^N\otimes\gamma\big\rangle \pi(\mathrm{d}x,\mathrm{d}y),
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{M}_k^N[f] &:= -\frac{\kappa}{N^3}\sum_{i=1}^N\sum_{j=1,j\neq i}^N (\langle\phi(\theta_k^j,\cdot,x_k),\gamma\rangle - y_k)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x_k),\gamma\rangle \\
&\quad - \frac{\kappa}{N^2}\langle(\phi(\cdot,\cdot,x_k)-y_k)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x_k),\nu_k^N\otimes\gamma\rangle - \mathbf{D}_k^N[f].
\end{aligned}$$

Let us introduce, for any $\nu \in \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$,

$$\mathfrak{Q}[f](\nu) = \int_{\mathsf{X}\times\mathsf{Y}} \big(\big\langle\phi(\cdot,\cdot,x),\nu\otimes\gamma\big\rangle - y\big)\big\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y).$$

Let us also define for $k \geq 0$ and $N \geq 1$,

$$\begin{aligned}
\mathfrak{K}_k^N[f] &:= \frac{\kappa}{N^3}\sum_{i=1}^N (\langle\phi(\theta_k^i,\cdot,x_k),\gamma\rangle - y_k)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x_k),\gamma\rangle \\
&\quad - \frac{\kappa}{N^2}\langle(\phi(\cdot,\cdot,x_k)-y_k)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x_k),\nu_k^N\otimes\gamma\rangle \\
&\quad - \frac{\kappa}{N^3}\sum_{i=1}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\big\langle\phi(\theta_k^i,\cdot,x),\gamma\big\rangle - y\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y) \\
&\quad + \frac{\kappa}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y).
\end{aligned}$$

It then holds for all $k \geq 0$ and $N \geq 1$:

$$\mathbf{M}_k^N[f] = -\frac{\kappa}{N}\mathscr{Q}[f](x_k, y_k, \nu_k^N) + \frac{\kappa}{N}\mathfrak{Q}[f](\nu_k^N) + \mathfrak{K}_k^N[f]. \tag{6.71}$$

Hence, by (6.69) and (6.71), for all $t \in \mathbb{R}_+$,

$$\begin{aligned}\mathfrak{a}_t^N[f] = \frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\left[\mathscr{Q}[f](x_k, y_k, \nu_k^N) - \mathfrak{Q}[f](\nu_k^N)\right]^2 + 2\kappa\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathfrak{K}_k^N[f]\left[\mathfrak{Q}[f](\nu_k^N) - \mathscr{Q}[f](x_k, y_k, \nu_k^N)\right] \\ + N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathfrak{K}_k^N[f]^2.\end{aligned} \tag{6.72}$$

Fix $t \geq 0$. Recall that we want to identify the limit of $(\mathfrak{a}_t^N[f])_{N\geq 1} \in \mathbb{R}^{\mathbf{N}^*}$ in $\mathbf{P}$-probability. Using the following two upper bounds (which can be easily derived using **A** and Lemma 69)

$$\mathbf{E}\left[|\mathfrak{K}_k^N[f]|^2\right] \leq C\|f\|^2_{\mathcal{C}^{1,\mathfrak{i}_0}}/N^4 \text{ and } \mathbf{E}\left[|\mathfrak{Q}[f](\nu_k^N)|^2\right] + \mathbf{E}\left[|\mathscr{Q}[f](x_k, y_k, \nu_k^N)|^2\right] \leq C\|f\|^2_{\mathcal{C}^{1,\mathfrak{i}_0}},$$

one deduces that the two last terms of (6.72) converge to zero in L^1. Therefore, one just needs to determine the limit in $\mathbf{P}$-probability of

$$\begin{aligned}\frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\left[\mathscr{Q}[f](x_k, y_k, \nu_k^N) - \mathfrak{Q}[f](\nu_k^N)\right]^2 = \frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} Var_\pi(\mathscr{Q}[f](x, y, \nu_k^N)) \\ + \frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\left(\mathscr{Q}[f](x_k, y_k, \nu_k^N) - \mathfrak{Q}[f](\nu_k^N)\right)^2 \\ - \frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} Var_\pi(\mathscr{Q}[f](x, y, \nu_k^N)).\end{aligned} \tag{6.73}$$

On the one hand, using Theorem 65 together with the continuous mapping theorem and the dominated convergence theorem, one deduces that for all $t \geq 0$[2]:

$$\begin{aligned}\frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} Var_\pi(\mathscr{Q}[f](x, y, \nu_k^N)) &= \kappa^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}} Var_\pi(\mathscr{Q}[f](x, y, \mu_s^N))\mathrm{d}s \\ &= \kappa^2\int_0^t Var_\pi(\mathscr{Q}[f](x, y, \mu_s^N))\mathrm{d}s - \kappa^2\int_{\frac{\lfloor Nt\rfloor}{N}}^t Var_\pi(\mathscr{Q}[f](x, y, \mu_s^N))\mathrm{d}s \\ &\xrightarrow[N\to+\infty]{\mathbf{P}} \kappa^2\int_0^t Var_\pi(\mathscr{Q}[f](x, y, \bar{\mu}_s))\mathrm{d}s.\end{aligned}$$

Let us now deal with the two remainders terms in (6.73). Denoting by $\mathscr{L}_k^N = \left[\mathscr{Q}[f](x_k, y_k, \nu_k^N) - \mathfrak{Q}[f](\nu_k^N)\right]^2$, we notice that $Var_\pi(\mathscr{Q}[f](x, y, \nu_k^N)) = \mathbf{E}_{(x,y)\sim\pi}[\mathscr{L}_k^N]$. Moreover if $j < k$, since $\mathscr{L}_j^N$ is $\mathcal{F}_k^N$-measurable (see (6.6)) as well as ν_k^N, and $(x_k, y_k) \perp\!\!\!\perp \mathcal{F}_k^N$, one has:

$$\begin{aligned}\mathbf{E}\left[\left(\mathscr{L}_k^N - \mathbf{E}_\pi[\mathscr{L}_k^N]\right)\left(\mathscr{L}_j^N - \mathbf{E}_\pi[\mathscr{L}_j^N]\right)\right] &= \mathbf{E}\left[\left(\mathscr{L}_j^N - \mathbf{E}_\pi[\mathscr{L}_j^N]\right)\mathbf{E}\left[\left(\mathscr{L}_k^N - \mathbf{E}_\pi[\mathscr{L}_k^N]\right)|\mathcal{F}_k^N\right]\right] \\ &= \mathbf{E}\left[\left(\mathscr{L}_j^N - \mathbf{E}_\pi[\mathscr{L}_j^N]\right)\mathbf{E}_\pi\left[\left(\mathscr{L}_k^N - \mathbf{E}_\pi[\mathscr{L}_k^N]\right)\right]\right] \\ &= \mathbf{E}\left[\left(\mathscr{L}_j^N - \mathbf{E}_\pi[\mathscr{L}_j^N]\right)\times 0\right] = 0.\end{aligned}$$

[2]This is indeed the same proof as the one made just after Eq. (3.63) in [Descours et al., 2022b], changing σ there by $\mathfrak{H} = \langle\nabla_\theta\phi, \gamma\rangle$.

Thus, it holds:

$$\begin{aligned}
&\mathbf{E}\Big[\Big|\frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}[\mathscr{Q}[f](x_k,y_k,\nu_k^N)-\mathfrak{Q}[f](\nu_k^N)]^2-\frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}Var_\pi(\mathscr{Q}[f](x,y,\nu_k^N))\Big|^2\Big]\\
&=\frac{\kappa^4}{N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\Big[\Big|[\mathscr{Q}[f](x_k,y_k,\nu_k^N)-\mathfrak{Q}[f](\nu_k^N)]^2-Var_\pi(\mathscr{Q}[f](x,y,\nu_k^N))\Big|^2\Big]\\
&\leq\frac{C}{N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[|\mathscr{Q}[f](x_k,y_k,\nu_k^N)|^4]\leq\frac{C}{N}\|f\|^4_{\mathcal{C}^{1,\mathrm{i}_0}}\to 0.
\end{aligned}$$

We have thus shown that for all $t\geq 0$, $\mathfrak{a}_t^N[f]\to\kappa^2\int_0^t Var_\pi(\mathscr{Q}[f](x,y,\bar\mu_s))\mathrm{d}s$ in $\mathbf{P}$-probability and as $N\to+\infty$. Therefore, for $t\geq 0$, $\mathfrak{c}_t[f]=\kappa^2\int_0^t Var_\pi(\mathscr{Q}[f](x,y,\bar\mu_s))\mathrm{d}s$. This ends the proof of the proposition when the the $\{\theta_k^i, i\in\{1,\dots,N\}\}$'s are generated by the algorithm (6.5).

The case of the BbB algorithm (6.7)**.** Let us assume that the $\{\theta_k^i, i\in\{1,\dots,N\}\}$'s are generated by the algorithm (6.7). We will also apply the central limit theorem 7.1.4 in [Ethier and Kurtz, 2009] to the sequence $\{t\in\mathbb{R}_+\mapsto\sqrt{N}\mathbf{M}_t^N[f]\}_{N\geq 1}$. Again, we define, as in (6.69),

$$\mathfrak{a}_t^N[f]=N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]^2.$$

Condition (a) in Th. 7.1.4 in [Ethier and Kurtz, 2009] is satisfied and we will now prove the last required condition in Th. 7.1.4 in [Ethier and Kurtz, 2009]. Let us introduce the following random probability measures over $\mathbb{R}^{d+1}\times\mathbb{R}^d$:

$$\mathbf{r}_k^N=\frac{1}{N}\sum_{i=1}^N\delta_{(\theta_k^i,\mathsf{Z}_k^i)}\text{ and }\rho_t^N=\mathbf{r}_{\lfloor Nt\rfloor}^N,\ k\geq 0,\ t\geq 0. \tag{6.74}$$

We also set, for $(x,y)\in\mathsf{X}\times\mathsf{Y}$ and $\rho\in\mathcal{P}(\mathbb{R}^{d+1}\times\mathbb{R}^d)$,

$$\mathfrak{Q}[f](x,y,\rho)=\langle\phi(\cdot,\cdot,x)-y,\rho\rangle\langle\nabla_\theta f(\pi_{\mathbb{R}^{d+1}}(\cdot))\cdot\nabla_\theta\phi(\cdot,\cdot,x),\rho\rangle,$$

where, for $(\theta,\mathsf{Z})\in\mathbb{R}^{d+1}\times\mathbb{R}^d$, $\pi_{\mathbb{R}^{d+1}}$ is the projection onto $\mathbb{R}^{d+1}$: $\pi_{\mathbb{R}^{d+1}}(\theta,\mathsf{Z})=\theta\in\mathbb{R}^{d+1}$. By Item **2** in the proof of Lemma 73, one has for $k\geq 0$,

$$\begin{aligned}
\mathbf{M}_k^N[f]&=-\frac{\kappa}{N}\langle\phi(\cdot,\cdot,x_k)-y_k,\mathbf{r}_k^N\rangle\langle\nabla_\theta f(\pi_{\mathbb{R}^{d+1}}(\cdot))\cdot\nabla_\theta\phi(\cdot,\cdot,x_k),\mathbf{r}_k^N\rangle-\mathbf{D}_k^N[f]\\
&=-\frac{\kappa}{N}\mathfrak{Q}[f](x_k,y_k,\mathbf{r}_k^N)-\mathbf{D}_k^N[f]=\mathbf{F}^N(x_k,y_k,\mathbf{r}_k^N)-\mathbf{D}_k^N[f]
\end{aligned}$$

where

$$\mathbf{F}^N(x_k,y_k,\mathbf{r}_k^N)=-\frac{\kappa}{N}\mathfrak{Q}[f](x_k,y_k,\mathbf{r}_k^N).$$

Fix $t\geq 0$. Let us identify the limit in probability as $N\to+\infty$ of the sequence $(\mathfrak{a}_t^N[f])_{N\geq 1}\subset\mathbb{R}$. We define at iteration $k\geq 1$ a larger σ-algebra than $\mathcal{F}_k^N$ (see (6.8)), in which, contrary to $\mathcal{F}_k^N$, the sequence $\{\mathsf{Z}_k^j, j=1,\dots,N\}$ is considered:

$$\Sigma_k^N=\boldsymbol{\sigma}\Big(\theta_0^i,\mathsf{Z}_{q'}^j,(x_q,y_q),1\leq i,j\leq N,0\leq q\leq k-1,0\leq q'\leq k\}\Big).$$

We rewrite $\mathfrak{a}_t^N[f]$ as follows:

$$\mathfrak{a}_t^N[f]=N\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big(\mathbf{E}\big[\mathbf{M}_k^N[f]^2\big|\Sigma_k^N\big]+\mathbf{M}_k^N[f]^2-\mathbf{E}\big[\mathbf{M}_k^N[f]^2\big|\Sigma_k^N\big]\Big). \tag{6.75}$$

By (6.67), it holds:

$$\mathbf{E}\Big[\Big(N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]^2-\mathbf{E}[\mathbf{M}_k^N[f]^2|\Sigma_k^N]\Big)^2\Big]=N^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\Big[\Big(\mathbf{M}_k^N[f]^2-\mathbf{E}[\mathbf{M}_k^N[f]^2|\Sigma_k^N]\Big)^2\Big]$$
$$\leq CN^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[\mathbf{M}_k^N[f]^4]\leq CN^2\|f\|^4_{\mathcal{C}^{1,\mathfrak{i}_0}}/N^3\to 0.$$

Hence, the two last terms of (6.75) converge to zero in L^2, i.e.:

$$N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]^2-\mathbf{E}\big[\mathbf{M}_k^N[f]^2\big|\Sigma_k^N\big]\xrightarrow[N\to\infty]{L^2}0. \tag{6.76}$$

Therefore, the limit in $\mathbf{P}$-probability $\mathfrak{c}_t[f]$ of $\mathfrak{a}_t^N[f]$ is given by the limit in $\mathbf{P}$-probability of

$$N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\big[\mathbf{M}_k^N[f]^2\big|\Sigma_k^N\big]=N\sum_{k=0}^{\lfloor Nt\rfloor-1}Var_\pi(\mathbf{F}^N(x,y,\mathbf{r}_k^N)),$$

where the equality holds since $(x_k,y_k)\perp\!\!\!\perp\Sigma_k^N$ and the $(\theta_k^j,\mathsf{Z}_j^k)$'s are Σ_k^N-measurable. We then write:

$$\begin{aligned}
&N\sum_{k=0}^{\lfloor Nt\rfloor-1}Var_\pi(\mathbf{F}^N(x,y,r_k^N))\\
&=\frac{\kappa^2}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}Var_\pi(\mathfrak{Q}[f](x,y,r_k^N))\\
&=\kappa^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}Var_\pi(\mathfrak{Q}[f](x,y,\rho_s^N))\mathrm{d}s\\
&=\kappa^2\int_0^t Var_\pi(\mathfrak{Q}[f](x,y,\rho_s^N))\mathrm{d}s-\kappa^2\int_{\frac{\lfloor Nt\rfloor}{N}}^t Var_\pi(\mathfrak{Q}[f](x,y,\rho_s^N))\mathrm{d}s.
\end{aligned}\tag{6.77}$$

For this fix time $t\geq 0$, we would like now to pass to the limit $N\to+\infty$ (in $\mathbf{P}$-probability) in (6.77). We recall the standard result: $(X^N)_{N\geq 1}$ converges to X in $\mathbf{P}$-probability if for any subsequence N' there exists a subsequence $N^\star$ of N' such that a.s. $X^{N^\star}\to X$. We will use such a result. Let us thus consider a subsequence N'. Let us show that there exists a subsequence $N^\star$ of N' such that a.s.

$$N^\star\sum_{k=0}^{\lfloor N^\star t\rfloor-1}Var_\pi(\mathrm{F}^{N^\star}(x,y,\mathbf{r}_k^{N^\star}))\to\kappa^2\int_0^t Var_\pi(\mathscr{Q}[f](x,y,\bar\mu_s))\mathrm{d}s.$$

Since $\mathfrak{q}_0:=2\max(\mathfrak{j}_0,\mathfrak{p}_0)>1+(d+1)/2$, by Theorem 65, in $\mathbf{P}$-probability, $\lim_{N'}\mu^{N'}=\bar\mu$ in the space $\mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$. Hence, there exists a subsequence N'' of N' such that $\mu^{N''}$ converges a.s. to $\bar\mu$ in $\mathcal{D}(\mathbb{R}_+,\mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$. By Lemma 83 below, it holds a.s. for all $s\geq 0$,

$$\rho_s^{N''}\to\bar\mu_s\otimes\gamma\text{ as }N''\to+\infty\text{ in }\mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}\times\mathbb{R}^d). \tag{6.78}$$

We now claim that a.s. for all $s\geq 0$

$$Var_\pi(\mathfrak{Q}[f](x,y,\rho_s^{N''}))\to Var_\pi(\mathfrak{Q}[f](x,y,\bar\mu_s\otimes\gamma))\text{ as }N''\to+\infty. \tag{6.79}$$

Let us prove this claim. We recall that by definition:

$$Var_\pi(\mathfrak{Q}[f](x,y,\rho_s^{N''}))=\mathbf{E}_{(x,y)\sim\pi}[|\mathfrak{Q}[f](x,y,\rho_s^{N''})|^2]-\mathbf{E}_{(x,y)\sim\pi}[\mathfrak{Q}[f](x,y,\rho_s^{N''})]^2, \tag{6.80}$$

where

$$\mathcal{Q}[f](x,y,\rho_s^{N''}) = \langle \phi(\cdot,\cdot,x) - y, \rho_s^{N''}\rangle \langle \nabla_\theta f(\pi_{\mathbb{R}^{d+1}}(\cdot)) \cdot \nabla_\theta \phi(\cdot,\cdot,x), \rho_s^{N''}\rangle. \tag{6.81}$$

Since $(\theta,z) \mapsto \phi(\theta,z,x) - y$ is continuous and bounded (uniformly over θ,z,x,y), it holds a.s. for all $s \geq 0$, $x,y \in \mathsf{X}\times\mathsf{Y}$, $\langle \phi(\cdot,\cdot,x) - y, \rho_s^{N''}\rangle \to \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle$ as $N'' \to +\infty$. On the other hand, since the function $(\theta,z) \mapsto \langle \nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta,z,x)$ is continuous and bounded by $C\|f\|_{\mathcal{C}^{1,\mathfrak{i}_0}}(1+|\theta|^{\mathfrak{j}_0})\mathfrak{b}(z)$. Since $(1+|\theta|^{\mathfrak{j}_0})\mathfrak{b}(z)$ is bounded by the function $\mathfrak{D}_{\mathfrak{q}_0}(\theta,z) = 1 + |\theta|^{\mathfrak{q}_0} + |z|^{\mathfrak{q}_0}$ (recall that by **A1**, $\mathfrak{b}(z) = 1 + |z|^{\mathfrak{p}_0}$), one has from (6.78), as $N'' \to +\infty$, a.s. for all $s \geq 0$, $x \in \mathsf{X}$,

$$\langle \nabla_\theta f(\pi_{\mathbb{R}^{d+1}}(\cdot)) \cdot \nabla_\theta \phi(\cdot,\cdot,x), \rho_s^{N''}\rangle \to \langle \nabla_\theta f(\pi_{\mathbb{R}^{d+1}}(\cdot)) \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle.$$

Note also that by the previous analysis, we have a.s. for all $s \geq 0$, $x,y \in \mathsf{X}\times\mathsf{Y}$,

$$\begin{aligned} |\mathcal{Q}[f](x,y,\rho_s^{N''})| \leq \sup_{x,y} |\mathcal{Q}[f](x,y,\rho_s^{N''})| &\leq C\|f\|_{\mathcal{C}^{1,\mathfrak{i}_0}} \langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^{N''}\rangle \\ &\leq C\|f\|_{\mathcal{C}^{1,\mathfrak{i}_0}} \sup_{N''\geq 1} \langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^{N''}\rangle < +\infty \end{aligned} \tag{6.82}$$

where the last inequality follows e.g. from the fact that $(\langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^{N''}\rangle)_{N''}$ is a converging sequence. Together with the dominated convergence theorem, one deduces (6.79).

Let us now consider the random variable $\int_0^t Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N''}))\mathrm{d}s$ appearing in the r.h.s of (6.77). By (6.80), (6.81), (6.82), and (6.89), it holds a.s. for all $s \geq 0$ and $x,y \in \mathsf{X}\times\mathsf{Y}$,

$$\begin{aligned} Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N''})) &\leq C\|f\|^2_{\mathcal{C}^{1,\mathfrak{i}_0}} |\langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^{N''}\rangle|^2 \\ &\leq C\|f\|^2_{\mathcal{C}^{1,\mathfrak{i}_0}} \sup_{N''\geq 1} \sup_{s\in[0,t]} |\langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^{N''}\rangle|^2 < +\infty. \end{aligned}$$

Therefore, using also (6.79) and the dominated convergence theorem, for this fix $t \geq 0$, one has:

$$\kappa^2 \int_0^t Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N''}))\mathrm{d}s \xrightarrow[N''\to\infty]{a.s.} \kappa^2 \int_0^t Var_\pi(\mathcal{Q}[f](x,y,\bar{\mu}_s\otimes\gamma))\mathrm{d}s.$$

Let us now consider the last term in (6.77). We have using (6.90),

$$\begin{aligned} \mathbf{E}\Big[\Big|\int_{\frac{\lfloor N''t\rfloor}{N''}}^t Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N''}))\mathrm{d}s\Big|\Big] &= \mathbf{E}\Big[\Big|\int_0^t Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N''}))\mathbf{1}_{s\in\left[\frac{\lfloor N''t\rfloor}{N''},t\right]}\mathrm{d}s\Big|\Big] \\ &\leq \frac{1}{N''}\mathbf{E}\Big[\sup_{s\in\left[\frac{\lfloor N''t\rfloor}{N''},t\right]} Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N''}))\Big] \\ &\leq \frac{C\|f\|^2_{\mathcal{C}^{1,\mathfrak{i}_0}}}{N''}\mathbf{E}\Big[\sup_{s\in\left[\frac{\lfloor N''t\rfloor}{N''},t\right]} |\langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^{N''}\rangle|^2\Big] \\ &\leq \frac{C\|f\|^2_{\mathcal{C}^{1,\mathfrak{i}_0}}}{N''} \xrightarrow[N''\to\infty]{} 0. \end{aligned}$$

Therefore, there exists $N^\star \subset N''$ such that

$$\int_{\frac{\lfloor N^\star t\rfloor}{N^\star}}^t Var_\pi(\mathcal{Q}[f](x,y,\rho_s^{N^\star}))\mathrm{d}s \xrightarrow[N^\star\to\infty]{a.s.} 0.$$

Thus, we have found a subsequence $N^\star \subset N'$ such that a.s.

$$N^\star \sum_{k=0}^{\lfloor N^\star t\rfloor - 1} Var_\pi(\mathbf{F}^{N^\star}(x,y,\mathbf{r}_k^{N^\star})) \xrightarrow[N^\star\to\infty]{a.s.} \mathfrak{c}_t[f] := \kappa^2 \int_0^t Var_\pi(\mathcal{Q}[f](x,y,\bar{\mu}_s\otimes\gamma))\mathrm{d}s.$$

Consequently

$$N \sum_{k=0}^{\lfloor Nt \rfloor - 1} Var_\pi(\mathbf{F}^N(x,y,\mathbf{r}_k^N)) \xrightarrow[N\to\infty]{\mathbf{P}} \mathfrak{c}_t[f].$$

This is the desired result since $\mathfrak{Q}[f](x,y,\bar{\mu}_s \otimes \gamma) = \mathscr{Q}[f](x,y,\bar{\mu}_s)$. The proof of the proposition is complete. □

Lemma 83. *Assume that the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the algorithm* (6.7)*. Assume also* ***A*** *and let $\mathfrak{q}_0 \in 2\mathbf{N}$ such that $\mathfrak{q}_0 > 1 + (d+1)/2$. Assume that along some subsequence $\mathcal{N}$, $(\mu^{\mathcal{N}})_{\mathcal{N}}$ converges a.s. to $\bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$. Then, it holds a.s. for all $s \geq 0$:*

$$\lim_{\mathcal{N}\to+\infty} \rho_s^{\mathcal{N}} = \bar{\mu}_s \otimes \gamma \ \text{ in } \mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1} \times \mathbb{R}^d).$$

Proof. In the following, we simply denote $\mathcal{N}$ by N. Assume that $\mu^N \xrightarrow{a.s.} \bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$. Recall that $\mathfrak{D}_{\mathfrak{q}_0}(\theta,z) = 1 + |\theta|^{\mathfrak{q}_0} + |z|^{\mathfrak{q}_0}$. According to Th. 6.0 in [Villani, 2009], to prove the lemma it is enough to show that a.s. for all $s \geq 0$,

$$\lim_{N\to+\infty} \rho_s^N = \bar{\mu}_s \otimes \gamma \ \text{ in } \mathcal{P}(\mathbb{R}^{d+1} \times \mathbb{R}^d) \text{ and } \lim_{N\to+\infty} \langle \mathfrak{D}_{\mathfrak{q}_0}, \rho_s^N \rangle = \langle \mathfrak{D}_{\mathfrak{q}_0}, \bar{\mu}_s \otimes \gamma \rangle. \tag{6.83}$$

We have for any continuous fonction $h : \mathbb{R}^{d+1} \times \mathbb{R}^d \to \mathbb{R}$ and $s \geq 0$,

$$\begin{aligned} \langle h, \rho_s^N \rangle - \langle h, \bar{\mu}_s \otimes \gamma \rangle &= \frac{1}{N} \sum_{i=1}^N \Big(h(\theta^i_{\lfloor Ns \rfloor}, \mathsf{Z}^i_{\lfloor Ns \rfloor}) - \int_{\mathbb{R}^d} h(\theta^i_{\lfloor Ns \rfloor}, z) \gamma(z) \mathrm{d}z \Big) \\ &\quad + \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}^d} h(\theta^i_{\lfloor Ns \rfloor}, z) \gamma(z) \mathrm{d}z - \langle h, \bar{\mu}_s \otimes \gamma \rangle, \end{aligned} \tag{6.84}$$

as soon as the $\int_{\mathbb{R}^d} h(\theta,z)\gamma(z)\mathrm{d}z$'s ($\theta \in \mathbb{R}^{d+1}$) and $\langle h, \bar{\mu}_s \otimes \gamma \rangle$ are well defined.

Step 1. We start by proving the first statement in (6.83). Let $t \geq 0$. We pick $g \in \mathcal{C}_b(\mathbb{R}^{d+1} \times \mathbb{R}^d)$. Note that in this case (6.84) holds with $h = g$. For ease of notation, we set $\mathscr{S}_k^i(g) = g(\theta_k^i, \mathsf{Z}_k^i) - \int_{\mathbb{R}^d} g(\theta_k^i, z)\gamma(z)\mathrm{d}z$, and we will also simply denote $\mathscr{S}_k^i(g)$ by $\mathscr{S}_k^i$. Note that since g is bounded, for all $m \in \mathbf{N}^*$, $\mathbf{E}[|\mathscr{S}_k^i|^m] \leq C$ for some $C > 0$ independent of $i \in \{1,\dots,N\}$, $N \geq 1$, and $k \geq 0$. Let us consider $i_j \in \{0,\dots,6\}$, $j = 1,\dots,6$ such that $\sum_{j=1}^6 i_j = 6$. Assume that there exists $j_0 \in \{1,\dots,6\}$ such that $i_{j_0} = 1$ and $i_{j_0} \neq i_l$ for all $l \neq j_0$. Then, it holds:

$$\mathbf{E}\Big[\prod_{j=1}^6 \mathscr{S}_k^{i_j} \Big] = 0.$$

Therefore, it holds:

$$\begin{aligned} &\mathbf{E}\Big[\sup_{s\in[0,t]} \Big| \frac{1}{N} \sum_{i=1}^N g(\theta^i_{\lfloor Ns \rfloor}, \mathsf{Z}^i_{\lfloor Ns \rfloor}) - \int_{\mathbb{R}^d} g(\theta^i_{\lfloor Ns \rfloor}, z)\gamma(z)\mathrm{d}z \Big|^6 \Big] \\ &\leq \sum_{k=0}^{\lfloor Nt \rfloor} \mathbf{E}\Big[\Big| \frac{1}{N} \sum_{i=1}^N \mathscr{S}_k^i \Big|^6 \Big] \\ &= \frac{1}{N^6} \sum_{k=0}^{\lfloor Nt \rfloor} \sum_{i=1}^N \mathbf{E}[|\mathscr{S}_k^i|^6] + \frac{1}{N^6} \sum_{k=0}^{\lfloor Nt \rfloor} \sum_{i\neq j} \mathbf{E}[(\mathscr{S}_k^i)^3 (\mathscr{S}_k^j)^3] + \frac{1}{N^6} \sum_{k=0}^{\lfloor Nt \rfloor} \sum_{i\neq j} \mathbf{E}[|\mathscr{S}_k^i|^4 |\mathscr{S}_k^j|^2] \\ &\quad + \frac{1}{N^6} \sum_{k=0}^{\lfloor Nt \rfloor} \sum_{i\neq j\neq \ell} \mathbf{E}[|\mathscr{S}_k^i|^2 |\mathscr{S}_k^j|^2 |\mathscr{S}_k^\ell|^2] \leq \frac{C}{N^2}, \end{aligned}$$

where $\sum_{i\neq j\neq \ell}$ is a short notation for the sum over the triples (i,j,ℓ) such that $i \neq j$, $j \neq \ell$, and $\ell \neq i$.

By Borel-Cantelli lemma, one deduces that, for all $t \geq 0$ it holds a.s.

$$\sup_{s\in[0,t]} \Big| \frac{1}{N} \sum_{i=1}^{N} g(\theta^i_{\lfloor Ns \rfloor}, \mathsf{Z}^i_{\lfloor Ns \rfloor}) - \int_{\mathbb{R}^d} g(\theta^i_{\lfloor Ns \rfloor}, z)\gamma(z)\mathrm{d}z \Big| \to 0 \text{ as } N \to +\infty. \tag{6.85}$$

Considering $t \in \mathbf{N}$, on deduces that a.s. for all $t \geq 0$, (6.85) holds. Let us now show that a.s. for all $s \in \mathbb{R}_+$,

$$\frac{1}{N} \sum_{i=1}^{N} \int_{\mathbb{R}^d} g(\theta^i_{\lfloor Ns \rfloor}, z)\gamma(z)\mathrm{d}z - \langle g, \bar{\mu}_s \otimes \gamma \rangle \to 0 \text{ as } N \to +\infty. \tag{6.86}$$

Since $\mathsf{W}_1 \leq \mathsf{W}_{\mathfrak{q}_0}$, we have that $\mu^N \xrightarrow{a.s.} \bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$. As $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^{d+1}))$, it holds a.s. for all $t \in \mathbb{R}_+$, $\mu^N_t \to \bar{\mu}_t$ in $\mathcal{P}_1(\mathbb{R}^{d+1})$. Let us define the function $G : \theta \in \mathbb{R}^{d+1} \mapsto \int_{\mathbb{R}^d} g(\theta, z)\gamma(z)\mathrm{d}z$, which is bounded continuous. We have a.s. for all $s \in \mathbb{R}_+$, $\langle G, \mu^N_s \rangle \to \langle G, \bar{\mu}_s \rangle$. This is exactly (6.86).

Considering (6.84) together with (6.85) and (6.86), we have shown that for all $g \in \mathcal{C}_b(\mathbb{R}^{d+1} \times \mathbb{R}^d)$, it holds a.s. for all $s \geq 0$:

$$\langle g, \rho^N_s \rangle \to \langle g, \bar{\mu}_s \otimes \gamma \rangle. \tag{6.87}$$

We now would like to prove that it holds a.s. for all $g \in \mathcal{C}_b(\mathbb{R}^{d+1} \times \mathbb{R}^d)$ and all $s \geq 0$: $\langle g, \rho^N_s \rangle \to \langle g, \bar{\mu}_s \otimes \gamma \rangle$ (which would exactly implies the first statement in (6.83)). To this end, by Remark 5.1.6 in [Ambrosio et al., 2008], it is sufficient to show that a.s. for all $s \geq 0$ and $g \in \mathcal{C}_c(\mathbb{R}^{d+1} \times \mathbb{R}^d)$ (the space of continuous functions with compact support), $\langle g, \rho^N_s \rangle \to_{N\to\infty} \langle g, \bar{\mu}_s \otimes \gamma \rangle$. Since the space $\mathcal{C}_c(\mathbb{R}^{d+1} \times \mathbb{R}^d)$ is separable, this last statement follows from (6.87) and a standard continuity argument. Hence, we have proved that a.s. for all $s \geq 0$, $\rho^N_s \to \bar{\mu}_s \otimes \gamma$. The proof of the first statement in (6.83) is complete.

Step 2. Let us now prove the second statement in (6.83). Fix $t \geq 0$. Note first that by **A1**, γ has moments of every order. Thus, $\int_{\mathbb{R}^d} \mathfrak{D}_{\mathfrak{q}_0}(\theta, z)\gamma(z)\mathrm{d}z = 1 + |\theta|^{\mathfrak{q}_0} + \langle |\cdot|^{\mathfrak{q}_0}, \gamma \rangle$ and $\langle \mathfrak{D}_{\mathfrak{q}_0}, \bar{\mu}_t \otimes \gamma \rangle = 1 + \langle |\cdot|^{\mathfrak{q}_0}, \bar{\mu}_t \rangle + \langle |\cdot|^{\mathfrak{q}_0}, \gamma \rangle$ are well defined. Thus (6.84) holds with $h = \mathfrak{D}_{\mathfrak{q}_0}$. From the analysis carried out in the first step, (6.85) holds with g is replaced by $\mathfrak{D}_{\mathfrak{q}_0}$ if for all $m \geq 1$, $i \in \{1, \ldots, N\}$ and $k \in \{1, \ldots, \lfloor Nt \rfloor\}$, $\mathbf{E}[|\mathscr{S}^i_k(\mathfrak{D}_{\mathfrak{q}_0})|^m] \leq C$ ($C > 0$ independent of i, k, and N), which is the case if

$$\mathbf{E}[|\mathfrak{D}_{\mathfrak{q}_0}(\theta^i_k, \mathsf{Z}^i_k)|^m] + \mathbf{E}\Big[\Big| \int_{\mathbb{R}^d} \mathfrak{D}_{\mathfrak{q}_0}(\theta^i_k, z)\gamma(z)\mathrm{d}z \Big|^m\Big] \leq C.$$

On the one hand, we have $\mathbf{E}[|\mathfrak{D}_{\mathfrak{q}_0}(\theta^i_k, \mathsf{Z}^i_k)|^m] = \mathbf{E}[|1 + |\theta^i_k|^{\mathfrak{q}_0} + |\mathsf{Z}^i_k|^{\mathfrak{q}_0}|^m] \leq C_m(1 + \mathbf{E}[|\theta^i_k|^{\mathfrak{q}_0 m}] + \mathbf{E}[|\mathsf{Z}^i_k|^{\mathfrak{q}_0 m}]) \leq C$ (see Lemma 69). With similar computations, $\mathbf{E}[|\int_{\mathbb{R}^d} \mathfrak{D}_{\mathfrak{q}_0}(\theta^i_k, z)\gamma(z)\mathrm{d}z|^m] < +\infty$. Thus, (6.85) holds with g is replaced by $\mathfrak{D}_{\mathfrak{q}_0}$, i.e it holds a.s. for all $t \geq 0$:

$$\sup_{s\in[0,t]} \Big| \frac{1}{N} \sum_{i=1}^{N} \mathfrak{D}_{\mathfrak{q}_0}(\theta^i_{\lfloor Ns \rfloor}, \mathsf{Z}^i_{\lfloor Ns \rfloor}) - \int_{\mathbb{R}^d} \mathfrak{D}_{\mathfrak{q}_0}(\theta^i_{\lfloor Ns \rfloor}, z)\gamma(z)\mathrm{d}z \Big| \to 0 \text{ as } N \to +\infty. \tag{6.88}$$

Let us now prove that (6.86) holds with g replaced there by $\mathfrak{D}_{\mathfrak{q}_0}$. Consider the function $D_0 : \theta \in \mathbb{R}^{d+1} \mapsto \int_{\mathbb{R}^d} \mathfrak{D}_{\mathfrak{q}_0}(\theta, z)\gamma(z)\mathrm{d}z = 1 + |\theta|^{\mathfrak{q}_0} + \langle |\cdot|^{\mathfrak{q}_0}, \gamma \rangle$. The function D_0 is continuous over $\mathbb{R}^{d+1}$ and clearly $\theta \mapsto D_0(\theta)/(1 + |\theta|^{\mathfrak{q}_0})$ is bounded. Consequently, since $\mu^N \xrightarrow{a.s.} \bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$ and $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$, it holds a.s. for all $s \in \mathbb{R}_+$, $\langle D_0, \mu^N_s \rangle \to \langle D_0, \bar{\mu}_s \rangle$, which is exactly (6.86) when g is replaced by $\mathfrak{D}_{\mathfrak{q}_0}$. This achieves the proof of the second statement in (6.83). The proof of the lemma is therefore complete.

We end the proof of the lemma by deriving two extra estimates (namely (6.89) and (6.90) below) which will be useful in the proof of Proposition 82 when the algorithm (6.7) is considered. Since $\mu^N \xrightarrow{a.s.} \bar{\mu}$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{\mathfrak{q}_0}(\mathbb{R}^{d+1}))$, using e.g. Proposition 5.3 in Chapter 3 of [Ethier and Kurtz, 2009], one has a.s. for all $t \geq 0$,

$$\sup_{N\geq 1} \sup_{s\in[0,t]} |\langle D_0, \mu^N_s \rangle| < +\infty.$$

Say that the previous inequality holds for all $\omega \in \Omega^*$ where $\mathbf{P}(\Omega^*) = 1$. By (6.88), there exists Ω' with $\mathbf{P}(\Omega') = 1$ and such that for all $\omega \in \Omega'$ and $t \geq 0$, it holds as $N \to +\infty$

$$\sup_{s\in[0,t]}\left|\langle \mathfrak{D}_{\mathfrak{q}_0},\rho_s^N(\omega)\rangle - \langle D_0,\mu_s^N(\omega)\rangle\right| \to 0.$$

Therefore, for all $\omega\in\Omega'\cap\Omega^*$, there exists $N_1(\omega)\geq 1$ such that that for all $N\geq N_1(\omega)$ and $t\geq 0$,

$$\begin{aligned}\sup_{s\in[0,t]}\left|\langle \mathfrak{D}_{\mathfrak{q}_0},\rho_s^N(\omega)\rangle\right| &\leq 1+\sup_{s\in[0,t]}|\langle D_0,\mu_s^N(\omega)\rangle \\ &\leq 1+\sup_{N\geq 1}\sup_{s\in[0,t]}|\langle D_0,\mu_s^N(\omega)\rangle < +\infty.\end{aligned}$$

Therefore, for all $\omega\in\Omega'\cap\Omega^*$ and $t\geq 0$

$$\sup_{N\geq 1}\sup_{s\in[0,t]}\left|\langle \mathfrak{D}_{\mathfrak{q}_0},\rho_s^N(\omega)\rangle\right| < +\infty, \tag{6.89}$$

i.e. (6.89) holds a.s. for all $t\geq 0$ (since $\mathbf{P}(\Omega'\cap\Omega^*)=1$). Finally, it holds that for all $m\geq 1$ and $0\leq t_1\leq t_2$,

$$\sup_{s\in[t_1,t_2]}\frac{1}{N}\sum_{i=1}^N|\mathsf{Z}^i_{\lfloor Ns\rfloor}|^m \leq \sum_{k=\lfloor Nt_1\rfloor}^{\lfloor Nt_2\rfloor}\frac{1}{N}\sum_{i=1}^N|\mathsf{Z}^i_k|^m.$$

Since the Z^i_k's are i.i.d. with moments of all order (see **A1**), one deduces that:

$$\mathbf{E}\Big[\sup_{s\in[t_1,t_2]}\frac{1}{N}\sum_{i=1}^N|\mathsf{Z}^i_{\lfloor Ns\rfloor}|^m\Big] \leq (\lfloor Nt_2\rfloor-\lfloor Nt_1\rfloor+1)\mathbf{E}_\gamma[|\mathsf{Z}|^m].$$

Consequently, using also Lemma 19 in [Descours et al., 2023b], one has:

$$\begin{aligned}&\mathbf{E}\Big[\sup_{s\in[t_1,t_2]}\Big|\frac{1}{N}\sum_{i=1}^N\mathfrak{D}_{\mathfrak{q}_0}(\theta^i_{\lfloor Ns\rfloor},\mathsf{Z}^i_{\lfloor Ns\rfloor})\Big|^m\Big] \\ &\leq \mathbf{E}\Big[\sup_{s\in[t_1,t_2]}\frac{1}{N}\sum_{i=1}^N C_m[1+|\theta^i_{\lfloor Ns\rfloor}|^{\mathfrak{q}_0 m}+|\mathsf{Z}^i_{\lfloor Ns\rfloor}|^{\mathfrak{q}_0 m}]\Big] \\ &\leq C_m\mathbf{E}\Big[\sup_{s\in[0,t_2]}\langle 1+|.|^{\mathfrak{q}_0 m},\mu_t^N\rangle\Big]+C_m(\lfloor Nt_2\rfloor-\lfloor Nt_1\rfloor+1)\mathbf{E}_\gamma[|\mathsf{Z}|^{\mathfrak{q}_0 m}] \\ &\leq C+C(\lfloor Nt_2\rfloor-\lfloor Nt_1\rfloor+1)\mathbf{E}_\gamma[|\mathsf{Z}|^{\mathfrak{q}_0 m}],\end{aligned}$$

where $C>0$ is independent of $N\geq 1$. In particular, when $t_2-t_1\leq 1/N$, it holds

$$\mathbf{E}\Big[\sup_{s\in[t_1,t_2]}\left|\langle \mathfrak{D}_{\mathfrak{q}_0},\rho_s^N\rangle\right|^m\Big]\leq C, \tag{6.90}$$

where $C>0$ is independent of $N\geq 1$. □

With the same arguments as those used to prove Proposition 82 when the $\{\theta^i_k, i\in\{1,\dots,N\}\}$'s are generated by the algorithm (6.5), we obtain

Proposition 84. *Assume that the $\{\theta^i_k, i\in\{1,\dots,N\}\}$'s are generated by the algorithm* (6.9)*. Assume also **A**. Then, for every $f\in\mathcal{C}^{2,\mathfrak{j}_0}(\mathbb{R}^{d+1})$, the sequence $\{t\in\mathbb{R}_+\mapsto\sqrt{N}\mathbf{M}_t^N[f]\}_{N\geq 1}$ converges in distribution in $\mathcal{D}(\mathbb{R}_+,\mathbb{R})$ towards a process $\mathcal{X}^f\in\mathcal{C}(\mathbb{R}_+,\mathbb{R})$ that has independent Gaussian increments. Moreover, for all $t\in\mathbb{R}_+$,*

$$\mathbf{E}[\mathcal{X}_t^f]=0 \text{ and } Var(\mathcal{X}_t^f)=\kappa^2\int_0^t Var_{\pi\otimes\gamma^{\otimes 2}}(\mathscr{Q}[f](x,y,z^1,z^2,\bar{\mu}_s))\mathrm{d}s,$$

where we recall $\mathscr{Q}[f](x,y,z^1,z^2,\bar{\mu}_v)=\langle\phi(\cdot,z^1,x)-y,\bar{\mu}_v\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,z^2,x),\bar{\mu}_v\rangle$ (see Theorem 68).

Proposition 85. *Assume that the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated either by the algorithm* (6.5)*,* (6.7)*, or* (6.9)*. Assume also* ***A****. Then, $(\sqrt{N}\mathbf{M}^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$ to a $\mathfrak{G}$-process $\mathscr{G} \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$ (see Definition 66) with covariance structure given by: for all $1 \leq i,j \leq k$, $f_1,\dots,f_k \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1})$ and $0 \leq s \leq t$,*

- *When the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated either by the algorithm* (6.5) *and* (6.7)*,*

$$Cov(\mathscr{G}_t[f_i], \mathscr{G}_s[f_j]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f_i](x,y,\bar{\mu}_v), \mathscr{Q}[f_j](x,y,\bar{\mu}_v))\mathrm{d}v,$$

  where we recall $\mathscr{Q}[f](x,y,\bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$ (see Theorem 68).

- *When the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated either by the algorithm* (6.9)*,*

$$Cov(\mathscr{G}_t[f_i], \mathscr{G}_s[f_j]) = \eta^2 \int_0^s Cov(\mathscr{Q}[f_i](x,y,z^1,z^2,\bar{\mu}_v), \mathscr{Q}[f_j](x,y,z^1,z^2,\bar{\mu}_v))\mathrm{d}v,$$

  where we recall $\mathscr{Q}[f](x,y,z^1,z^2,\bar{\mu}_v) = \langle \phi(\cdot,z^1,x) - y, \bar{\mu}_v\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z^2,x), \bar{\mu}_v\rangle$ (see Theorem 68).

Proof. The proof of Proposition 85 relies on the same arguments as those used to prove Prop. 3.13 in [Descours et al., 2022b]. □

6.4 On the limit points of $(\eta^N, \sqrt{N}\mathbf{M}^N)_{N\geq 1}$

In this section, we come back to the case when the $\{\theta_k^i, i \in \{1,\dots,N\}\}$'s are generated by the algorithm (6.7). The other two cases (namely (6.5) and (6.9)) are treated similarly, and all the results of this section also holds for each of these other two algorithms.

Let us derive the pre-limit equation for the fluctuation process η^N, see (6.91) just below. On the one hand, one has for all $N \geq 1$, $t \geq 0$ and $f \in \mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}(\mathbb{R}^{d+1})$,

$$\begin{aligned}
&\sqrt{N}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \sqrt{N}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&= -\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \frac{1}{\sqrt{N}}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

Hence, using (6.25) and (6.13), we obtain the following pre-limit equation for η^N:

$$\begin{aligned}
\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle &= -\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \frac{\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \kappa \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \eta_s^N\rangle \mathrm{d}s \\
&+ \frac{\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \frac{\kappa}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&+ \sqrt{N}\mathbf{M}_t^N[f] + \sqrt{N}\mathbf{W}_t^N[f] + \sqrt{N}\mathbb{R}_t^N[f]. \qquad (6.91)
\end{aligned}$$

The aim of this section is to pass to the limit $N \to +\infty$ in (6.91). We start with the following lemma whose proof, identical to the one of Lemma 3.16 in [Descours et al., 2022b], is omitted.

Lemma 86. *Assume **A**. Then, the sequence $(\eta_0^N)_{N\geq 1}$ converges in distribution in $\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$ towards a variable ν_0 which is the unique (in distribution) $\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$-valued random variable such that for all $k \geq 1$ and $f_1,\dots,f_k \in \mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}(\mathbb{R}^{d+1})$, $(\langle f_1,\nu_0\rangle,\dots,\langle f_k,\nu_0\rangle)^T \sim \mathcal{N}(0,\mathfrak{C}(f_1,\dots,f_k))$, where $\mathfrak{C}(f_1,\dots,f_k)$ is the covariance matrix of the vector $(f_1(\theta_0^1),\dots,f_k(\theta_0^1))^T$.*

Let us now set

$$\mathscr{E} = \mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})) \times \mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1})). \qquad (6.92)$$

According to Propositions 70 and 80, $(\eta^N, \sqrt{\mathbf{M}}^N)$ is tight in $\mathscr{E}$. Let $(\eta^\star, \mathscr{G}^*)$ be one of its limit point in $\mathscr{E}$. Along some subsequence N', it holds:

$$(\eta^{N'}, \sqrt{N'}\mathbf{M}^{N'}) \to (\eta^\star, \mathscr{G}^\star), \text{ as } N' \to \infty.$$

Considering the marginal distributions, and according to Lemma 81, it holds a.s.

$$\eta^\star \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})) \text{ and } \mathscr{G}^\star \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1})). \qquad (6.93)$$

By uniqueness of the limit in distribution, using Lemma 86 (together with the fact that the function $m \in \mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})) \mapsto m_0 \in \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$ is continuous) and Proposition 85, it also holds:

$$\eta_0^\star \overset{\mathscr{L}}{=} \nu_0 \text{ and } \mathscr{G}^\star \overset{\mathscr{L}}{=} \mathscr{G}. \qquad (6.94)$$

Proposition 87. *Assume **A**. Then, $\eta^\star$ is a weak solution of **(EqL)** with initial distribution ν_0.*

Proof. Let us introduce, for $\Phi \in \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$, $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$, and $s \geq 0$:

$$\mathfrak{U}_s[f](\Phi) = \kappa \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \Phi \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y), \qquad (6.95)$$

$$\mathfrak{V}_s[f](\Phi) = \kappa \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \Phi \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y), \qquad (6.96)$$

and

$$\mathfrak{W}_s[f](\Phi) = \kappa \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \Phi\rangle \qquad (6.97)$$

The term $\mathfrak{U}_s[f](\Phi)$ is well defined because $f \in \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1})$. Since $\mathrm{j}_3 > (d+1)/2$, using (6.27) and because $\bar{\mu}_s \in \mathcal{P}_{\mathrm{j}_0}(\mathbb{R}^{d+1})$ ($f \in \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$), $\mathfrak{V}_s[f](\Phi)$ is well defined. The term $\mathfrak{W}_s[f](\Phi)$ is well defined because of (6.36). Equation (6.91) can be rewritten as follows:

$$\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle + \int_0^t (\mathfrak{U}_s[f](\eta_s^N) + \mathfrak{V}_s[f](\eta_s^N) + \mathfrak{W}_s[f](\eta_s^N))\mathrm{d}s - \sqrt{N}\mathbf{M}_t^N[f] = \mathbf{e}_t^N[f], \tag{6.98}$$

where

$$\begin{aligned}
\mathfrak{Re}_t^N[f] = &-\frac{\kappa}{\sqrt{N}}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N\otimes\gamma\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \eta_s^N\otimes\gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&+ \frac{\kappa}{\sqrt{N}}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x)-y,\gamma\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle, \mu_s^N\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \frac{\kappa}{\sqrt{N}}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \mu_s^N\otimes\gamma\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&+ \sqrt{N}\mathbf{W}_t^N[f] + \sqrt{N}\mathbb{R}_t^N[f].
\end{aligned}$$

Fix $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$ and $t \in \mathbb{R}_+$.

Step 1. In this step we study the continuity of the mapping

$$\mathfrak{B}_t[f] : m \in \mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})) \mapsto \langle f, m_t\rangle + \int_0^t (\mathfrak{U}_s[f](m_s) + \mathfrak{V}_s[f](m_s) + \mathfrak{W}_s[f](m_s))\mathrm{d}s \tag{6.99}$$

Let $(m^N)_{N\geq 1}$ such that $m^N \to m$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$. Using (6.34), it holds, for all $N \geq 1$, $s \in [0,t]$ and $x \in \mathsf{X}$,

$$\begin{aligned}
&|\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s\otimes\gamma\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), m_s^N\otimes\gamma\rangle| \\
&\leq C\big\|\nabla_\theta f\cdot\nabla_\theta \mathfrak{H}(\cdot,x)\big\|_{\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}} \sup_{N\geq 1}\sup_{s\in[0,t]} \|m_s^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}} \\
&\leq C\|f\|_{\mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3}} \sup_{N\geq 1}\sup_{s\in[0,t]} \|m_s^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}} < +\infty.
\end{aligned}$$

We also have, by (6.27) and the embedding $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{C}^{1,\mathrm{j}_0}(\mathbb{R}^{d+1})$ and the fact that $\bar{\mu} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_{j_0}(\mathbb{R}^{d+1}))$,

$$\begin{aligned}
|\langle \phi(\cdot,\cdot,x), m_s^N\otimes\gamma\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \bar{\mu}_s\otimes\gamma\rangle| \leq C &\sup_{N\geq 1}\sup_{s\in[0,t]} \|m_s^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}} \\
&\times \|f\|_{\mathcal{C}^{1,\mathrm{j}_0}} \sup_{s\in[0,t]} \langle 1+|\cdot|^{\mathrm{j}_0}, \bar{\mu}_s\rangle < +\infty.
\end{aligned}$$

Finally, using (6.36),

$$\begin{aligned}
|\langle \nabla_\theta f\cdot\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), m_s^N\rangle| &\leq \|\nabla_\theta f\cdot\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1)\|_{\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}} \sup_{N\geq 1}\sup_{s\in[0,t]} \|m_s^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}} \\
&\leq C\|f\|_{\mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}} \sup_{N\geq 1}\sup_{s\in[0,t]} \|m_s^N\|_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}} < +\infty.
\end{aligned}$$

These bounds allow to apply the dominated convergence theorem to obtain that $\mathfrak{B}_t[f](m^N) \to \mathfrak{B}_t[f](m)$, as soon as t is a continuity point of m. Consequently, using (6.93) and the continuous mapping theorem 2.7 in [Billingsley, 1999], it holds, for all $t \in \mathbb{R}_+$ and $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$,

$$\mathfrak{B}_t[f](\eta^{N'}) - \langle f, \eta_0^{N'}\rangle - \sqrt{N'}\mathbf{M}_t^{N'}[f] \xrightarrow[N'\to\infty]{\mathscr{L}} \mathfrak{B}_t[f](\eta^*) - \langle f, \eta_0^*\rangle - \mathscr{G}_t^*[f]. \tag{6.100}$$

Step 2. In this step, we prove that for any $t \in \mathbb{R}_+$ and $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$:

$$\mathbf{E}\big[|\Re\mathfrak{e}_t^N[f]|\big] \to_{N\to\infty} 0. \tag{6.101}$$

By (6.34)-(6.27), the embedding $\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}(\mathbb{R}^{d+1}) \hookrightarrow \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$ and Lemma 73, it holds

$$\begin{aligned}
&\mathbf{E}\Big[\Big|\frac{1}{\sqrt{N}}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x),\eta_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\eta_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|\Big]\\
&\le \frac{C\|f\|_{\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}}}{\sqrt{N}}\int_0^t\mathbf{E}\Big[\|\eta_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}\Big]\mathrm{d}s \le \frac{C\|f\|_{\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}}}{\sqrt{N}}\int_0^t\mathbf{E}\Big[\|\eta_s^N\|^2_{\mathcal{H}^{-\mathrm{J}_1,\mathrm{j}_1}}\Big]\mathrm{d}s \le \frac{C\|f\|_{\mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}}}{\sqrt{N}}.
\end{aligned}$$

By Lemma 69, we have

$$\begin{aligned}
&\mathbf{E}\Big[\frac{1}{\sqrt{N}}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big|\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\Big|\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&+\frac{\kappa}{\sqrt{N}}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big|\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\Big|\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \le \frac{C\|f\|_{\mathcal{C}^{1,\mathrm{j}_0}}}{\sqrt{N}}.
\end{aligned}$$

In addition, from (6.38), $\mathbf{E}[\sqrt{N}|\mathbb{R}_t^N[f]|] \le \|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}/\sqrt{N}$. Moreover, it is straightforward to prove that $\mathbf{E}[|\mathbf{W}_t^N[f]|] \le \|f\|_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}/N$. Hence, we have proved (6.101).

Step 3. End of the proof of Proposition 87. By (6.98), (6.100) and (6.101), we deduce that for all $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$, and $t \in \mathbb{R}_+$, it holds a.s. $\mathfrak{B}_t[f](\eta^\star) - \langle f,\eta_0^\star\rangle - \mathscr{G}_t^\star[f] = 0$. Since $\mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$ and $\mathbb{R}_+$ are separable, we conclude by a standard continuity argument (and using that every Hilbert-Schmidt embedding is continuous) that a.s. for all $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$ and $t \in \mathbb{R}_+$, $\mathfrak{B}_t[f](\eta^\star) - \langle f,\eta_0^\star\rangle - \mathscr{G}_t^\star[f] = 0$. Hence, $\eta^\star$ is a weak solution of **(EqL)** with initial distribution ν_0 (see (6.94)). This ends the proof of Proposition 87. □

6.5 Pathwise uniqueness and proof of Theorem 68

Throughout this section, we consider algorithm (6.7), but we recall that all our statements are valid for algorithms (6.5) and (6.9).

Proposition 88. *Assume **A**. Then strong (pathwise) uniqueness holds for **(EqL)**. Namely, on a fixed probability space, given a $\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1})$-valued random variable ν and a $\mathfrak{G}$-process $\mathscr{G} \in \mathcal{C}(\mathbb{R}_+,\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}(\mathbb{R}^{d+1}))$, there exists at most one $\mathcal{C}(\mathbb{R}_+,\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$-valued process η solution to **(EqL)** with $\eta_0 = \nu$ almost surely.*

Proof.

By linearity of the involved operators in **(EqL)**, it is enough to consider a $\mathcal{C}(\mathbb{R}_+,\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$-valued process η solution to **(EqL)** when a.s. $\nu = 0$ and $\mathscr{G} = 0$, i.e., for every $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$ and $t \in \mathbb{R}_+$,

$$\begin{cases}\langle f,\eta_t\rangle + \int_0^t(\mathfrak{U}_s[f](\eta_s) + \mathfrak{V}_s[f](\eta_s) + \mathfrak{W}_s[f](\eta_s))\mathrm{d}s = 0,\\ \langle f,\eta_0\rangle = 0,\end{cases} \tag{6.102}$$

where we recall that $\mathfrak{U}$, $\mathfrak{V}$ and $\mathfrak{W}$ are defined respectively in (6.95), (6.96) and (6.97). Pick $T > 0$. By (6.102), we have, a.s. for all $f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$ and $t \in [0,T]$,

$$\langle f,\eta_t\rangle^2 = -2\int_0^t(\mathfrak{U}_s[f](\eta_s) + \mathfrak{V}_s[f](\eta_s) + \mathfrak{W}_s[f](\eta_s))\langle f,\eta_s\rangle\mathrm{d}s. \tag{6.103}$$

Since $\sup_{s\in[0,T]}\langle 1+|\cdot|^{\mathrm{j}_0},\bar\mu_s\rangle < +\infty$, and using (6.27),

$$\begin{aligned}
&-2\int_0^t\mathfrak{V}_s[f](\eta_s)\langle f,\eta_s\rangle\mathrm{d}s\\
&\le 2\kappa\int_0^t\Big[\langle f,\eta_s\rangle^2 + \int_{\mathsf{X}\times\mathsf{Y}}|\langle\phi(\cdot,\cdot,x),\eta_s\otimes\gamma\rangle|^2|\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar\mu_s\otimes\gamma\rangle|^2\pi(\mathrm{d}x,\mathrm{d}y)\Big]\mathrm{d}s\\
&\le C\int_0^t\Big[\langle f,\eta_s\rangle^2 + \|\eta_s\|^2_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3}}\|f\|^2_{\mathcal{C}^{1,\mathrm{j}_0}}\Big]\mathrm{d}s \le C\int_0^t\Big[\langle f,\eta_s\rangle^2 + \|\eta_s\|^2_{\mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}}\|f\|^2_{\mathcal{H}^{\mathrm{J}_0,\mathrm{j}_0}}\Big]\mathrm{d}s.
\end{aligned}$$

Consider an orthonormal basis $\{f_a\}_{a\geq 1}$ of $\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$. Recall that $\mathbf{T}_x : f \in \mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1}) \mapsto \int_{\mathbb{R}^d} \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z,x)\gamma(z)\mathrm{d}z \in \mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3-1}(\mathbb{R}^{d+1})$ (see (6.47)). By Lemma B.2 in [Descours et al., 2022b], one deduces that:

$$\begin{aligned}
-2\sum_{a\geq 1}\int_0^t \mathfrak{U}_s[f_a](\eta_s)\langle f_a,\eta_s\rangle \mathrm{d}s &= -2\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \sum_{a\geq 1} \langle \mathbf{T}_x f_a, \eta_s\rangle \langle f_a,\eta_s\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&= -2\kappa \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \eta_s, \mathbf{T}_x^* \eta_s\rangle_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}} \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\leq C\int_0^t \|\eta_s\|^2_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}}\mathrm{d}s.
\end{aligned}$$

Using the operator $\mathbf{T} : f \in \mathcal{H}^{\mathrm{J}_3,\mathrm{j}_3-1}(\mathbb{R}^{d+1}) \mapsto \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1) \in \mathcal{H}^{\mathrm{J}_3-1,\mathrm{j}_3}(\mathbb{R}^{d+1})$ (see (6.49)) together with Lemma 77, we obtain

$$\begin{aligned}
\sum_{a\geq 1} -2\int_0^t \mathfrak{W}_s[f_a](\eta_s)\langle f_a,\eta_s\rangle \mathrm{d}s = -2\kappa \int_0^t \sum_{a\geq 1} \langle \mathbf{T} f_a, \eta_s\rangle \langle f_a,\eta_s\rangle \mathrm{d}s &= -2\kappa \int_0^t \langle \eta_s, \mathbf{T}^*\eta_s\rangle_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}}\mathrm{d}s \\
&\leq C\int_0^t \|\eta_s\|^2_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}}\mathrm{d}s
\end{aligned}$$

Hence, using (6.103), one deduces that a.s. for all $t \in [0,T]$,

$$\|\eta_t\|^2_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}} = \sum_{a\geq 1}\langle f_a,\eta_t\rangle^2 \leq C\int_0^t \|\eta_s\|^2_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}}\mathrm{d}s.$$

By Gronwall's lemma, a.s. for all $t \in [0,T]$, $\|\eta_t\|_{\mathcal{H}^{-\mathrm{J}_3,\mathrm{j}_3-1}} = 0$. This concludes the proof of Proposition 88. □

We are now in position to conclude the proof of Theorem 68.

Proof. [Proof of Theorem 68] Let us consider the case when the θ_k^i's are generated by the algorithm (6.7) (the proofs of Theorem 68 are exactly the same when they are generated by the algorithms (6.5) or the algorithm (6.9)). By Proposition 70, (η^N) admits a limit point. Assume that it admits two limit points. Let $\ell \in \{1,2\}$ and N_ℓ be such that in distribution $\eta^{N_\ell} \to \eta^\ell$ in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$. Recall that from Lemma 81, we have a.s. $\eta^\ell \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$. Let us now consider a limit point $(\eta^{\ell,\star}, \mathscr{G}^{\ell,\star})$ of $(\eta^{N_\ell}, \sqrt{N_\ell}\mathbf{M}^{N_\ell})$ in $\mathscr{E}$ (see (6.92)). Up to extracting a subsequence from N_ℓ, we assume

$$(\eta^{N_\ell}, \sqrt{N_\ell}\mathbf{M}^{N_\ell}) \xrightarrow[N_\ell\to\infty]{\mathscr{L}} (\eta^{\ell,\star}, \mathscr{G}^{\ell,\star}) \text{ in } \mathscr{E}.$$

Considering the marginal distributions, we then have by uniqueness of the limit in distribution, for $\ell = 1,2$,

$$\eta^{\ell,\star} \overset{\mathscr{L}}{=} \eta^\ell \text{ and } \mathscr{G}^{\ell,\star} \overset{\mathscr{L}}{=} \mathscr{G}. \tag{6.104}$$

where $\mathscr{G}$ is a G-process given by Proposition 85. Recall also that from Proposition 87, both $\eta^{1,\star}$ and $\eta^{2,\star}$ are two weak solutions of **(EqL)** with initial distribution ν_0 (see also Lemma 86). Since strong uniqueness for **(EqL)** (see Proposition 88) implies weak uniqueness for **(EqL)**, we deduce that $\eta^{1,\star} = \eta^{2,\star}$ in law. By (6.104), this implies $\eta^1 = \eta^2$ in law. Consequently, the whole sequence $(\eta^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}(\mathbb{R}_+, \mathcal{H}^{-\mathrm{J}_3+1,\mathrm{j}_3}(\mathbb{R}^{d+1}))$. Denoting by $\eta^\star$ its limit, we have proved that $\eta^\star$ has the same distribution as the unique weak solution of **(EqL)** with initial distribution ν_0. The proof Theorem 68 is complete. □

CHAPTER 7

Theoretical Guarantees for Variational Inference with Fixed-Variance Mixture of Gaussians

Chapter abstract: *Variational inference (VI) is a popular approach in Bayesian inference, that looks for the best approximation of the posterior distribution within a parametric family, minimizing a loss that is typically the (reverse) Kullback-Leibler (KL) divergence. Despite its empirical success, the theoretical properties of VI have only received attention recently, and mostly when the parametric family is the one of Gaussians. This work aims to contribute to the theoretical study of VI in the non-Gaussian case by investigating the setting of Mixture of Gaussians with fixed covariance and constant weights. In this view, VI over this specific family can be casted as the minimization of a Mollified relative entropy, i.e. the KL between the convolution (with respect to a Gaussian kernel) of an atomic measure supported on Diracs, and the target distribution. The support of the atomic measure corresponds to the localization of the Gaussian components. Hence, solving variational inference becomes equivalent to optimizing the positions of the Diracs (the particles), which can be done through gradient descent and takes the form of an interacting particle system. We study two sources of error of variational inference in this context when optimizing the mollified relative entropy. The first one is an optimization result, that is a descent lemma establishing that the algorithm decreases the objective at each iteration. The second one is an approximation error, that upper bounds the objective between an optimal finite mixture and the target distribution.*

1 Introduction

A fundamental problem in computational statistics and machine learning is to compute integrals with respect to some target probability distribution $\mu^\star$ on $\mathbb{R}^d$ whose density is known only up to a normalization constant. For instance in Bayesian inference, $\mu^\star$ is the posterior distribution over the parameters of complex models. The general goal of sampling methods is thus to provide an approximate distribution for which the integrals are easily computed. A large number of methods have been developed to tackle this problem. The classical approach is to sample the posterior using Markov Chain Monte Carlo (MCMC) algorithms, in which a Markov chain designed to converge to $\mu^\star$ is simulated for a sufficiently long time [Roberts and Rosenthal, 2004]. These methods use the discrete measure over past iterates of the algorithm as an approximation of the posterior to compute integrals of interest. However, MCMC algorithms are generally computationally expensive, and it is an open problem to diagnose their convergence in practice [Moins et al., 2023]. Variational inference (VI) [Blei et al., 2017] has emerged as a powerful and versatile alternative in Bayesian inference. By framing the problem as an optimization task, VI aims to find an approximate candidate distribution within a parametric family of distributions $\mathcal{G}$ that minimizes the (reverse) Kullback-Leibler

(KL) divergence to the target:

$$\hat{\nu} := \operatorname*{argmin}_{\mu \in \mathcal{G}} \mathrm{KL}(\mu|\mu^\star), \tag{7.1}$$

where $\mathrm{KL}(\mu|\mu^\star) = \int \log(\mathrm{d}\mu/\mathrm{d}\mu^\star)\mathrm{d}\mu$ if μ is absolutely continuous with respect to $\mu^\star$ denoting $\mathrm{d}\mu/\mathrm{d}\mu^\star$ its Radon-Nikodym density, and $+\infty$ else; and $\hat{\nu}$ is referred to as the optimal approximation within the variational family.

While VI methods can only return an approximation of the target, they are much more tractable in the large scale setting, since they benefit from efficient optimization methods, e.g. parallelization or stochastic optimization [Zhang et al., 2018b]. Hence, VI has proven effective in numerous applications and is a popular paradigm especially in high-dimensional scenarios. Still, the understanding of its theoretical properties remains a challenging and active area of research. Fundamentally, there are two sources of errors in VI: the *approximation* error that quantifies how far $\hat{\nu}$ is from $\mu^\star$, and the *optimization* error that comes from the optimization of the objective in (7.1) to approach $\hat{\nu}$.

Even among the recent literature on theoretical guarantees for VI, most efforts have been concentrated in the case where $\mathcal{G}$ is the set of non-degenerate Gaussian distributions. Recently, [Katsevich and Rigollet, 2023] studied the approximation quality (in total variation) of the approximate posterior $\hat{\nu}$, i.e., minimizers of the objective (7.1), and show that it better estimates the true mean and covariance of the posterior than the well-known Laplace approximation [Helin and Kretschmann, 2022]. Regarding the optimization of (7.1), still restricted to Gaussians, several recent works leverage the geometry of Wasserstein gradient flows, more precisely the equivalence between Bures-Wasserstein gradient flows on the space of probability distributions and Euclidean flows on the space of parameters of the variational approximation. They derive novel algorithms with convergence guarantees e.g. through gradient-descent [Lambert et al., 2022b] or forward-backward [Diao et al., 2023, Domke et al., 2023] time discretizations; and precise connections with Black-Box Variational Inference (BBVI) [Yi and Liu, 2023].

However, to the best of our knowledge, the study of approximation and computational guarantees when $\mathcal{G}$ is a set of mixture of Gaussians has not been tackled yet. Mixture models are a widely used class of probabilistic models that capture complex and multi-modal data distributions by combining simpler components. Moreover, they are dense in the space of probability distributions with p bounded moments in the Wasserstein-p metric [Delon and Desolneux, 2020, Lemma 3.1].

In this study, we propose to consider a simplified setting where the Gaussian components have equal weights and share the same diagonal covariance. This regime breaks down the complexity of the problem, and is still theoretically challenging, but remains a practically relevant scenario. In this setting, variational inference aims to optimize the locations of the means of the Gaussian mixture to approximate the target distribution.

In the following, we assume that $\mu^\star$ admits a density proportional to $\exp(-V)$ with respect to the Lebesgue measure over $\mathbb{R}^d$.

2 The mollified relative entropy

Writing $\mu^\star = e^{-V}/Z$ with Z the unknown normalization constant, the (reverse) Kullback-Leibler divergence (or relative entropy) can be written as

$$\begin{aligned}\mathrm{KL}(\mu|\mu^\star) &= \int V \mathrm{d}\mu + \int \log(\mu)\mathrm{d}\mu + \log(Z) \\ &:= \mathcal{E}_V(\mu) + \mathcal{U}(\mu) + \log(Z),\end{aligned}$$

for μ absolutely continuous with respect to $\mu^\star$, and $+\infty$ else. Hence, it decomposes as the sum of a potential energy $\mathcal{E}_V$, i.e. a linear functional, and the negative entropy $\mathcal{U}$, up to an additive constant that is fixed in the optimization problem.

We now consider the minimization problem of Variational Inference (7.1) for mixture of Gaussians. We will study a specific setting where the variational family is the set of mixture of n Gaussians with equally weighted components, and where these components have the same diagonal covariance $\epsilon^2 \mathrm{I}_d$, for some $n \in \mathbb{N}^*, \epsilon > 0$.

$$\mathcal{G}_n = \left\{ \frac{1}{n}\sum_{i=1}^n q_i,\ q_i = \mathcal{N}(x^i, \epsilon^2 \mathrm{I}_d),\ x^i \in \mathbb{R}^d \right\},$$

where I_d denotes the d-dimensional identity matrix. In our setting, only the positions (the means) of the mixture components will be optimized. Hence, searching for the optimal distribution in the variational family

approximating the target μ^* consists in finding the optimal locations of the Gaussian components in $\mathbb{R}^d$. We will denote k_ϵ the normalized Gaussian kernel, i.e. $k_\epsilon(x) = \exp\big(-\|x\|^2/(2\epsilon^2)\big) Z_\epsilon^{-1}$, where $\int k_\epsilon(x)dx = 1$ and $Z_\epsilon \propto (\epsilon^2)^{d/2}$. It is a specific example of mollifiers, i.e. smooth approximations of the Dirac delta at the origin, as introduced in [Friedrichs, 1944]. For μ a given probability distribution on $\mathbb{R}^d$, we denote by $k_\epsilon \star \mu$ its convolution with the Gaussian kernel that writes $k_\epsilon \star \mu = \int k_\epsilon(\cdot - x)\mathrm{d}\mu(x)$. Equipped with these notations, we can write $\mathcal{G}_n = \big\{k_\epsilon \star \mu_n,\ \mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i},\ x^1, \dots, x^n \in \mathbb{R}^d\big\}$.

Irrespective of the number of components n, VI with Gaussian mixtures whose components share the same variance can be written more generally as minimizing (7.1) restricted to the family $\mathcal{G} = \big\{k_\epsilon \star \mu,\ \mu \in \mathcal{P}(\mathbb{R}^d)\big\}$. The latter problem can be then reformulated as the optimization over $\mathcal{P}(\mathbb{R}^d)$ of the following objective functional, that we will refer to as the *mollified relative entropy* (or mollified KL):

$$\begin{aligned}\mathcal{F}_\epsilon(\mu) &= \int V d(k_\epsilon \star \mu) + \int \log(k_\epsilon \star \mu) d(k_\epsilon \star \mu) \\ &:= \mathcal{E}_{V_\epsilon}(\mu) + U_\epsilon(\mu),\end{aligned} \tag{7.2}$$

where $\mathcal{E}_{V_\epsilon}$ is a potential energy with respect to a convoluted potential $V_\epsilon = k_\epsilon \star V$ (using the associativity of the convolution operation), and $U_\epsilon(\mu) = \mathcal{U}(k_\epsilon \star \mu)$ is a functional that we will refer to as the mollified negative entropy. In contrast with the negative entropy defined above, the mollified one is well-defined for discrete measures.

2.1 Algorithm

We now discuss the optimization of the mollified relative entropy, starting from the continuous time dynamics to the practical discrete-time particle scheme.

A Wasserstein gradient flow of $\mathcal{F}_\epsilon$ [Ambrosio et al., 2008] can be described by the following continuity equation:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot (\mu_t \nabla_{W_2} \mathcal{F}_\epsilon(\mu_t)), \quad \nabla_{W_2} \mathcal{F}_\epsilon(\mu_t) := \nabla \mathcal{F}'_\epsilon(\mu_t), \tag{7.3}$$

where $\mathcal{F}'_\epsilon$ denotes the first variation of $\mathcal{F}_\epsilon$. Recall that if it exists, the first variation of a functional $\mathcal{F}$ at ν is the function $\mathcal{F}'(\nu) : \mathbb{R}^d \to \mathbb{R}$ s.t. for $\nu, \mu \in \mathcal{P}(\mathbb{R}^d)$: $\lim_{\epsilon \to 0} 1/\epsilon[\mathcal{F}(\nu + \epsilon(\mu - \nu)) - \mathcal{F}(\nu)] = \int \mathcal{F}'(\nu)(x)(\mathrm{d}\mu(x) - \mathrm{d}\nu(x))$. Wasserstein gradient flows are paths of steepest descent with respect to the W_2 metric, and can be seen as analog to Euclidean gradient flows on the space of probability distributions [Santambrogio, 2017].

Starting from some initial distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, and for some given step-size $\gamma > 0$, a forward (or explicit) time-discretization of (7.3) corresponds to the Wasserstein gradient descent algorithm, and can be written at each discrete time iteration $l \in \mathbb{N}$ as:

$$\mu_{l+1} = (\mathrm{I}_d - \gamma \nabla \mathcal{F}'_\epsilon(\mu_l))_{\#} \mu_l \tag{7.4}$$

where I_d is the identity map in $L^2(\mu_l)$.

For discrete measures $\mu_n = 1/n \sum_{i=1}^n \delta_{x^i}$, we can define the finite-dimensional objective $F(X^n) := \mathcal{F}_\epsilon(\mu_n)$ where $X^n = (x^1, \dots, x^n)$, since the functional $\mathcal{F}_\epsilon$ is well defined for discrete measures. The Wasserstein gradient descent dynamics of $\mathcal{F}_\epsilon$ (7.4) then correspond to standard gradient descent of the (finite-dimensional) function F, i.e., gradient descent on the position of the particles. In that setting, we recall that particles correspond to the means of the Gaussian components of the mixture. The gradient of F is readily obtained as

$$\nabla_{x^j} F(X^n) = \int_{\mathbb{R}^d} \nabla V(y) k_\epsilon(y - x^j) dy + \int_{\mathbb{R}^d} \frac{\sum_{i=1}^n \nabla k_\epsilon(y - x^i)}{\sum_{i=1}^n k_\epsilon(y - x^i)} k_\epsilon(y - x^j) dy. \tag{7.5}$$

Notice that the gradient above involves integrals over $\mathbb{R}^d$. However, using a Gaussian kernel k_ϵ, since $\nabla k_\epsilon(x) = -\frac{x}{\epsilon^2} k_\epsilon(x)$, these integrals can be easily approximated through Monte Carlo using Gaussian samples. A particle version of (7.4), e.g., starting with μ_0 discrete, can then be written as the following gradient descent iterates:

$$x_{l+1}^j = x_l^j - \gamma \nabla_{x_l^j} F(X_l^n) \tag{7.6}$$

for $j = 1, \dots, n$ and where $X_l^n = (x_l^1, \dots, x_l^n)$. Hence, minimizing $\mathcal{F}_\epsilon$ on discrete measures results in a particle system that interact through the gradient of the objective. The reader may refer to Subsection 7.1 for the detailed computations leading to the particle scheme. Notice that it recovers the scheme mentioned in [Lambert et al., 2022b, Section 5] where the covariance of the mixture components are fixed, see Subsection 7.2 for a detailed discussion.

Remark 89. *Notice that the Wasserstein gradient at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ of the mollified KL in 7.3, $\nabla \mathcal{F}'_\epsilon(\mu_t) : \mathbb{R}^d \to \mathbb{R}^d$ writes for any $w \in \mathbb{R}^d$:*

$$\nabla \mathcal{F}'_\epsilon(\mu_t)(w) = k_\epsilon \star \nabla V(w) + k_\epsilon \star \nabla \log(k_\epsilon \star \mu)(w), \tag{7.7}$$

see Subsection 7.1. Hence, it differs from the Wasserstein gradient of the (standard) KL w.r.t. $\mu^\star \propto e^{-V}$, i.e. $\mathrm{KL}(\cdot|\mu^\star)$ evaluated at the convoluted distribution that writes as $\nabla \log\left(k_\epsilon \star \mu / \mu^\star\right)$, see [Wibisono, 2018, Section 3.1.3].

2.2 Non-smoothness of the KL

In Euclidean optimization, it is standard that the convergence of gradient descent is guaranteed when the objective function is convex and smooth, which relates to a lower bound and upper bound on the Hessian of the objective when the latter is twice differentiable [Garrigos and Gower, 2023]. Analogously, when optimizing a functional on the Wasserstein space, lower and upper bounds on the Hessian characterize respectively convexity and smoothness on the functional $\mathcal{F}$ with respect to the Wasserstein-2 geometry (see [Villani, 2009, Proposition 16.2]). The Wasserstein space has a Riemannian geometry [Otto, 2001], where one can define for any μ the tangent space $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla \psi,\ \psi \in C_c^\infty(\mathsf{X})\}} \subset L^2(\mu)$ [Ambrosio et al., 2008, Definition 8.4.1]. The W_2 Hessian of a functional $\mathcal{F}$, denoted $H\mathcal{F}_{|\mu}$ is an operator over $\mathcal{T}_\mu \mathcal{P}_2(\mathsf{X})$ verifying $\left\langle H\mathcal{F}_{|\mu} v_t, v_t \right\rangle_{L^2(\mu)} = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathcal{F}(\rho_t)$ if $t \mapsto \rho_t$ is a geodesic starting at μ with vector field $t \mapsto v_t$. Considering $\psi \in C_c^\infty(\mathsf{X})$ and the path ρ_t from μ to $(I + \nabla\psi)_\# \mu$ given by: $\rho_t = (\mathrm{I}_d + t\nabla\psi)_\# \mu$, for all $t \in [0,1]$, the Hessian of $\mathcal{F}$ at μ, $H\mathcal{F}_{|\mu}$, is defined as a symmetric bilinear form on $C_c^\infty(\mathsf{X})$ associated with the quadratic form $Hess_\mu \mathcal{F}(\psi,\psi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathcal{F}(\rho_t)$.

We now recall the formula of the Wasserstein Hessian of the (standard) Kullback-Leibler divergence (or relative entropy).

Proposition 90. *[Villani, 2021, Section 9.1.2]. Assume that $\mu^\star$ has a density $\mu^\star \propto e^{-V}$ where the potential $V : X \to \mathbb{R}$ is $C^2(\mathsf{X})$. The Hessian of $\mathrm{KL}(\cdot|\mu^\star)$ at μ is given, for any $\psi \in C_c^\infty(\mathsf{X})$, by:*

$$Hess_\mu \mathrm{KL}(\psi,\psi) = \int \left[\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|_{HS}^2\right] \mathrm{d}\mu(x) \tag{7.8}$$

$$= Hess_\mu \mathcal{E}_V(\psi,\psi) + Hess_\mu \mathcal{U}(\psi,\psi), \tag{7.9}$$

where H_V is the Hessian of V.

The proof of Proposition 90 is provided in Subsection 7.5.1 for completeness. The reader may also refer to [Korba et al., 2021, Duncan et al., 2023] for similar computations on Wasserstein Hessians.

The KL divergence inherits the convexity of the target potential V in the Wasserstein geometry. Indeed, if $\mathrm{H}_V \succeq \lambda \mathrm{I}_d$, then $\mathrm{KL}(\cdot|\mu^\star)$ is λ-displacement convex, i.e. it is λ-convex along Wasserstein-2 geodesics, the underlying geometry for Wasserstein gradient flows. Yet, the Kullback-Leibler divergence is not a smooth objective in the Wasserstein sense, since its (Wasserstein) Hessian is not upper bounded even if the potential V is smooth. Indeed, assume $\mathrm{H}_V \preceq M\mathrm{I}_d$, i.e., the potential of the target distribution is M-smooth. This enables to control the first term in (7.8) by $M\|\nabla\psi\|^2_{L^2(\mu)}$, but the second term due to the negative entropy cannot be controlled similarly for any ψ [Wibisono, 2018, Korba et al., 2020].

Hence in this context, it is not possible to prove a descent lemma along (Wasserstein) gradient descent for the KL, unless restricting to smooth directions [Korba et al., 2020]. The non-smoothness of the KL is also the reason why many algorithms aiming to minimize the KL in the Wasserstein geometry rely on splitting-schemes such as the forward-backward algorithm, to perform a gradient descent (explicit) step on the potential energy part, and a JKO (implicit) step on the entropy part [Salim et al., 2020, Diao et al., 2023, Domke et al., 2023]. In contrast, we will leverage the fact that the *mollified* KL enjoys some smoothness properties that will allow us to derive a descent lemma in Section 3, at the price of loosing some convexity.

Still, we next show that $\mathcal{F}_\epsilon$ recovers displacement convexity (of the standard KL) as $\epsilon \to 0$, since its Hessian recovers the one of the KL.

Proposition 91. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. For any $\psi \in C_c^\infty(\mathbb{R}^d)$, the Wasserstein Hessian of $\mathcal{F}_\epsilon$ converges to the one of the regular KL, i.e:*

$$Hess_\mu \mathcal{F}_\epsilon(\psi,\psi) \xrightarrow[\varepsilon \to 0]{} Hess_\mu \mathrm{KL}(\psi,\psi). \tag{7.10}$$

The proof of Proposition 91 can be found in Subsection 7.5.2; the main technical difficulties arise when dealing with the negative entropy term. This result shows that as $\epsilon \to 0$, one can recover the geometric properties of the KL.

Proposition 91 serves as an auxiliary finding within our study, not directly influencing other results, yet it enables us to illustrate key conceptual distinctions. Specifically, it demonstrates that while the standard Kullback-Leibler (KL) divergence is convex in the Wasserstein geometry for log-concave targets—exhibiting even strong convexity for targets that are strongly log-concave—it loses this convexity when mollified, although it gains smoothness with a positive ϵ. This transition is typically delineated through lower and upper bounds on the Hessians within the Wasserstein framework. Getting a non-asymptotic, quantitative bounds on the Hessian of the mollified KL in terms of ϵ is the subject of future work. Such research could potentially offer insights into how small ϵ may be selected relative to the strong convexity constant of the target potential, ensuring the optimization objective maintains convexity.

3 Optimization Guarantees

We now turn to the analysis of the optimization error for VI in our setting, i.e. the optimization of $\mathcal{F}_\epsilon$. Under a smoothness assumption on the target potential, as well as moment conditions on the trajectory, one can obtain a descent lemma for the Wasserstein gradient descent iterates.

Assumption 3. *The potential V is L-smooth, i.e. for any $x, y \in \mathbb{R}^d$, $\|\nabla V(x) - \nabla V(y)\| \leq L\|x - y\|$.*

Assumption 4. *μ_0 is supported on n Diracs, and the second moments of $(\mu_l)_{l\geq 0}$ are bounded by $h > 0$ along gradient descent iterations, i.e. $\int \|x\|^2 \mathrm{d}\mu_l(x) < h$, $\forall l \geq 0$.*

Bounded moment assumptions such as these are commonly used in stochastic optimization, for instance in some analysis of the stochastic gradient descent [Moulines and Bach, 2011]. We also verified empirically this assumption in a specific setting outlined afterwards. The target $\mu^\star$ is a mixture of 100 Gaussians that we approximate with a mixture of 10 Gaussians. Then we run (7.6) (equivalently (7.4)) for 1000 iterations. The expectations in (7.5) with respect to the Gaussian kernel are estimated by Monte Carlo with 100 samples. Figure 7.1 displays the second moments of the particle distributions along iterations, for various dimensions. The 95% confidence interval displayed in Figure 7.1 is calculated based on 50 runs, and represents the randomness corresponding to Monte Carlo approximations, initialization of the target and initialization of our mixture. Our experiment shows that Assumption 4 holds for any dimension, i.e., the second moment of the particles distribution is bounded along the (discrete-time) flow. Further details on the setup are provided in Section 7.6. We now turn to one of our main results regarding the optimization of the mollified KL.

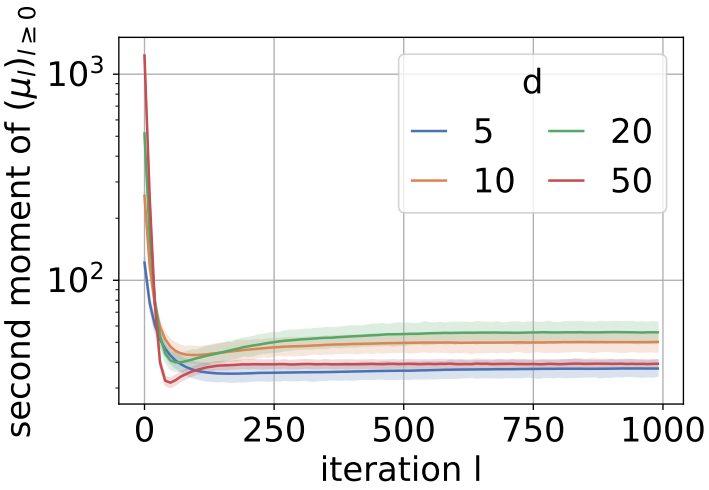


Fig. 7.1 Second moment along Wasserstein gradient descent iterations.

Proposition 92. *Suppose Assumptions 3 and 4 hold. Consider the sequence of iterates of Wasserstein gradient descent of $\mathcal{F}_\epsilon$ defined by* (7.4)*. Then, the following inequality holds:*

$$\mathcal{F}_\epsilon(\mu_{l+1}) - \mathcal{F}_\epsilon(\mu_l) \leq -\gamma\left(1 - \frac{\gamma}{2}M\right)\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}.$$

where $M = L + K_{\epsilon,n,h}$, and $K_{\epsilon,n,h}$ is a constant depending on ϵ, n, h.

Hence, for a small enough step-size γ, the latter proposition shows that the objective decreases at each iteration. We now provide a proof for this result, using similar techniques as [Arbel et al., 2019, Korba et al., 2020]. The main technical difficulties are left in the appendix and are related to showing the descent for the mollified entropy part, see 7.4 for details.

Proof. [Proof of Proposition 92] Consider a path between μ_l and μ_{l+1} of the form $\rho_t = (\psi_t)_{\#}\mu_l$ with $\psi_t = (\mathrm{I}_d + t\nabla\mathcal{F}'_\epsilon(\mu_l))$. We have $\frac{\partial\rho_t}{\partial t} = \boldsymbol{\nabla}\cdot(\rho_t v_t)$ with $v_t = -\nabla\mathcal{F}'_\epsilon(\mu_l)\circ\psi_t^{-1}$. The latter continuity equation holds in the sense of distributions [Ambrosio et al., 2008, Chapter 8] and holds for discrete measures. The function $t \mapsto \mathcal{F}_\epsilon(\rho_t)$ is differentiable and hence absolutely continuous. Therefore one can write:

$$\mathcal{F}_\epsilon(\rho_\gamma) = \mathcal{F}_\epsilon(\rho_0) + \gamma\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}_\epsilon(\rho_t) + \int_0^\gamma\left[\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\epsilon(\rho_t) - \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}_\epsilon(\rho_t)\right]\mathrm{d}t. \tag{7.11}$$

Moreover, using the chain rule in the Wasserstein space, we have successively:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\epsilon(\rho_t) = \left\langle\nabla\mathcal{F}'_\epsilon(\rho_t), v_t\right\rangle_{L^2(\rho_t)}, \text{ and } \quad \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}_\epsilon(\rho_t) = -\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}. \tag{7.12}$$

Then, since $\mathcal{F}_\epsilon = U_\epsilon + \mathcal{E}_{V_\epsilon}$, we have first under Assumption 3 that $k_\epsilon \star V$ is L-smooth and by Proposition 99 that:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_{V_\epsilon}(\rho_t) - \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_{V_\epsilon}(\rho_t)\Big|_{t=0} \leq L\,t\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}, \tag{7.13}$$

and by Proposition 100 and Assumption4:

$$\frac{d}{dt}\,U_\epsilon(\rho_t) - \frac{d}{dt}\,U_\epsilon(\rho_t)\Big|_{t=0} \leq K_{\epsilon,n,h}t\|\phi\|^2_{L^2(\mu_l)},$$

where $K_{\epsilon,n,h} = 1/\epsilon^2 + 2\sqrt{hn}/\epsilon^3 + \sqrt{n}/\epsilon^2 + n\sqrt{h}/2\epsilon^3$. Hence, the result follows directly by applying the above expressions to (7.11) where $M = L + K_{\epsilon,n,h}$. □

As a corollary, we obtain the convergence of the average of squared gradient norms along iterations.

Corollary 93. *Let $c_\gamma = \gamma(1 - \frac{\gamma M}{2})$. Under the assumptions of Proposition 92, one has*

$$\frac{1}{L}\sum_{l=1}^{L}\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)} \leq \frac{\mathcal{F}_\epsilon(\mu_0)}{2c_\gamma L}. \tag{7.14}$$

In contrast with the KL that is non-smooth as explained in 2.2, the mollified KL is smooth, which is why we can prove the descent lemma in Proposition 92 and the rate on average gradients. The descent lemma and its corollary imply that the sequence of squared gradient norms is summable and hence converges to zero.

We illustrate the validity of the rate derived in Corrolary 93 with simple experiments. The variational family there is a family of Gaussian mixtures with 10 components, while the target is a Gaussian mixture with 100 components. Figure 7.2 shows the convergence of the cumulative sum $\frac{1}{L}\sum_{l=1}^{L}\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}$ along iterations, for various dimensions and in log scale. Similarly to the previous experiment, the expectations involved in the gradient descent schemes are estimated using Monte Carlo with 100 samples. The 95% confidence interval displayed in Figure 7.2 is computed based on 50 runs, representing the randomness due to Monte Carlo approximations, randomization of the target and ininital distribution for the scheme. The term $\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}$ also involves expectations that are estimated by Monte Carlo with 1000 samples. Figure 7.2 illustrates that the cumulative sum is indeed of order $\frac{1}{L}$ as stated by Corollary 93. A detailed description of the experimental setup can be found in Section 7.6.

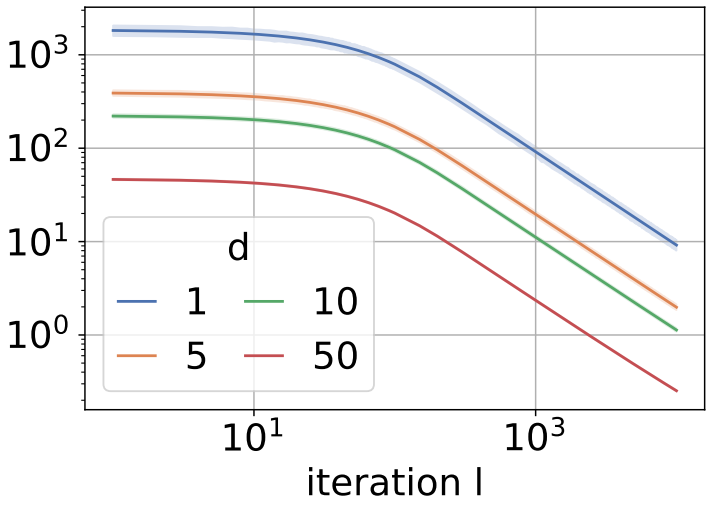


Fig. 7.2 Illustration of the rate of $\frac{1}{L}\sum_{l=1}^{L}\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}$ derived in Corollary 93

Remark 94. *Non-convex rates similar to our Corollary 93 have been obtained for Langevin Monte-Carlo [Balasubramanian et al., 2022] or Stein Variational Gradient Descent (SVGD) algorithm [Korba et al., 2020] leveraging similar techniques and smoothness of the potential. However, since Langevin Monte Carlo and SVGD optimizes the (standard) KL divergence, the squared gradient norm correspond to the Fisher Divergence and Kernel Stein Discrepancy respectively, that are valid probability divergences. In our setting, Corollary 93 implies the following. If μ_l converges weakly to some distribution μ_∞ (up to a subsequence) as $l \to \infty$, the Wasserstein gradient of $\mathcal{F}_\epsilon$ given in (7.7) is zero on the support of μ_∞, assuming $\mu \mapsto \|\nabla\mathcal{F}'_\epsilon(\mu)\|^2_{L^2(\mu)}$ is lower semi continuous with respect to the weak topology of measures. This can be rewritten $\nabla k_\epsilon \star (\log(k_\epsilon \star \mu_\infty) - \log(\mu^\star)) = 0$ μ_∞-a.e, i.e. $k_\epsilon \star (\log(k_\epsilon \star \mu_\infty) - \log(\mu^\star)) = c$ $\mu_\infty-$a.e. for some constant c.*

4 Approximation Guarantees

In this section, we investigate the approximation accuracy of a finite mixture of Gaussians to the posterior, i.e. minimizers of the objective functional $\mathcal{F}_\epsilon$ (assuming we are able to find these minimizers, e.g. after optimization). We obtain non-asymptotic rates with respect to the number of components in the mixture. For ease of notation, we will denote by $k^x_\epsilon := k_\epsilon(\cdot - x)$ for any $x \in \mathbb{R}^d$. We first consider the following assumption on the target distribution.

Assumption 5. *The target posterior distribution $\mu^\star$ has a mixture representation form, i.e. there exists P on $\mathbb{R}^d$ such that*

$$\mu^\star = \int_\Theta k^w_\epsilon \mathrm{d}P(w).$$

Notice that Assumption 5 is a relatively weak assumption, as mixture of Gaussians are dense in the space of probability distributions [Delon and Desolneux, 2020]. We now state our second main result.

Theorem 95. *Suppose Assumption 5 holds and define*

$$C^2_{\mu^\star} = \int \frac{\int (k^m_\epsilon(x))^2 \mathrm{d}P(m)}{\int k^w_\epsilon(x)\mathrm{d}P(w)} dx. \tag{7.15}$$

Define $\mathcal{G}_n = \left\{k_\epsilon \star \mu_n,\ \mu_n \in \mathcal{P}_n(\mathbb{R}^d)\right\}$, where $\mathcal{P}_n(\mathbb{R}^d)$ is the set of discrete probability distributions supported on n Dirac masses. Then,

$$\min_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(k_\epsilon \star \mu_n | \mu^\star) \leq C^2_{\mu^\star} \frac{\log(n)+1}{n}.$$

Our result is novel and quantifies the approximation quality of the family of mixtures of n Gaussian distributions (with equal weight and constant covariance) in the (reverse) Kullback-Leibler sense.

A major limitation in the use of Gaussian distributions in VI arises from the inherent simplicity of this family. In particular, the unimodality of the Gaussian distribution becomes a critical stumbling block when the target distribution is multimodal. A notable exception exists in the work of Katsevich & Rigollet (2023), which provides an error bound for cases where the target is a posterior distribution in the Bayesian inference context. As the sample size goes to infinity, the Bernstein Von-Mises theorem shows that the posterior distribution asymptotically converges to a Gaussian distribution, thereby lending some predictability to the approximation error in this specific scenario. In stark contrast, Theorem 95 offers a more versatile result, applicable to any target distribution, including those encountered in Bayesian inference with a fixed sample size. It shows that increasing the number of components in a Gaussian mixture can significantly mitigate the limitations of Gaussian VI. As we expand the mixture, the approximation error not only decreases, it converges to zero. This result highlights the potential of complexifying the variational family to achieve more accurate approximations of the target distribution.

The proof of Theorem 95 follows the steps of [Li and Barron, 1999], that proved similar guarantees for the forward KL (akin to likelihood maximization), while we focus on the reverse KL, i.e. the one considered in variational inference. Hence our proof requires non-trivial different inequalities and intermediate lemmas that are deferred to Subsection 7.3.

Proof. [Proof of Theorem 95] We will prove the previous result by induction. We denote by ν_n the minimizer of the Kullback-Leibler divergence to the target within this family, i.e.,

$$\nu_n := \operatorname*{argmin}_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(k_\epsilon \star \mu_n | \mu^\star),$$

and $D_n = \mathrm{KL}(k_\epsilon \star \mu_n | \mu^\star)$. For any $m \in \mathbb{R}^d$, we consider the distribution $\rho^m_{n+1} \in \mathcal{C}_{n+1}$ defined as

$$\rho^m_{n+1} = (1-\alpha)(k_\epsilon \star \mu_n) + \alpha k^m_\epsilon$$

where $\alpha = 1/n+1$. Therefore we have $D_{n+1} = \mathrm{KL}(k_\epsilon \star \mu_{n+1} | \mu^\star) \leq \mathrm{KL}(\rho^m_{n+1} | \mu^\star)$. By definition of the Kullback-Leibler divergence, denoting $f(x) = x \log x$, we have

$$\mathrm{KL}(\rho^m_{n+1} | \mu^\star) = \int f(r_{n+1}) d\mu^\star \,,$$

where we define r_{n+1} and r_0 as:

$$r_{n+1} := \frac{\rho^m_{n+1}}{\mu^\star} = (1-\alpha)\frac{(k_\epsilon \star \mu_n)}{\mu^\star} + \alpha \frac{k^m_\epsilon}{\mu^\star} := r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star}.$$

Define $B(x) = (x \log x - x + 1)/(x-1)^2$ for $x \in [0, +\infty[$. Note that $r_{n+1}(x) \geq r_0(x)$ for any x, then using that B is decreasing (see Lemma 96), we have $B(r_{n+1}(x)) \leq B(r_0(x))$. It follows that

$$\begin{aligned}
r_{n+1}\log(r_{n+1}) &\leq r_{n+1} - 1 + B(r_0)(r_{n+1}-1)^2 \\
&= r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star} - 1 + B(r_0)\left(r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star} - 1\right)^2 \\
&= r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star} - 1 + B(r_0)\big\{(r_0-1)^2 + \left(\alpha \frac{k^m_\epsilon}{\mu^\star}\right)^2 + 2\alpha(r_0-1)\frac{k^m_\epsilon}{\mu^\star}\big\} \\
&= \alpha \frac{k^m_\epsilon}{\mu^\star} + r_0 \log(r_0) + \left(\alpha \frac{k^m_\epsilon}{\mu^\star}\right)^2 B(r_0) + 2\alpha B(r_0)(r_0-1)\frac{k^m_\epsilon}{\mu^\star}. \qquad (7.16)
\end{aligned}$$

Moreover, we have the following inequality:

$$\begin{aligned}
D_{n+1} &= \int D_{n+1} dP(m) \\
&\leq \int \mathrm{KL}(\rho_{n+1}^m | \mu) dP(m) \\
&= \alpha + \int r_0(x) \log(r_0(x)) d\mu^\star(x) + \alpha^2 \iint \frac{k_\epsilon^m(x)^2}{\mu^\star(x)^2} B(r_0(x)) d\mu^\star(x) dP(m) \\
&+ 2\alpha \int B(r_0(x))(r_0(x) - 1) d\mu^\star(x),
\end{aligned}$$

where we used (7.16) in the last equality. We now focus on bounding each term on the r.h.s. of the previous inequality. By definition of r_0, the second term can be rewritten

$$\int r_0 \log(r_0) d\mu^\star = (1-\alpha)\log(1-\alpha) + (1-\alpha) D_n.$$

We now turn to the third term. For any $x \in \mathbb{R}^+$, since B is monotone decreasing, $B(r_0(x)) \leq B(0) = 1$. Under Assumption5 , it follows that

$$\int \int \frac{k_\epsilon^m(x)^2}{\mu^\star(x)^2} B(r_0(x)) d\mu^\star(x) dP(m) \leq \int \int \frac{k_\epsilon^m(x)^2}{\mu^\star(x)} dP(m) dx = C_{\mu^\star}^2.$$

Finally let's focus on the last term. We have $B(x)(x-1) \leq \sqrt{x} - 1$ using Lemma 97. Denoting $H^2(f,g) = 1 - \int \sqrt{f(x)g(x)} dx \in [0,1]$ the squared Hellinger distance between f and g, we have

$$\begin{aligned}
\int B(r_0)(r_0 - 1) \mathrm{d}\mu^\star &\leq \int (\sqrt{r_0} - 1) \mathrm{d}\mu^\star \\
&= \sqrt{1-\alpha}(1 - H^2(k_\epsilon \star \mu_n, \mu^\star)) - 1 \\
&\leq \sqrt{1-\alpha} - 1.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
D_{n+1} &\leq \alpha + (1-\alpha)\log(1-\alpha) + (1-\alpha) D_n + \alpha^2 C_{\mu^\star}^2 + 2\alpha(\sqrt{1-\alpha} - 1) \\
&\leq (1-\alpha) D_n + \alpha^2 C_{\mu^\star}^2,
\end{aligned}$$

where the last inequality uses that $-\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} \leq 0$ (see Lemma 98).

Now, recall that $\alpha = 1/(n+1)$. Denoting $U_n = nD_n$, our previous computations imply that $U_{n+1} \leq U_n + C_{\mu^\star}^2/n+1$, which by telescoping yields $U_n - U_0 \leq C_{\mu^\star}^2 H_n$, where H_n denotes the harmonic number and is upper bounded by $1 + \log(n)$. The rate on D_n follows. □

Our result is analog to the one of [Li and Barron, 1999] that bounds the forward Kullback-Leibler divergence to the target. Indeed under Assumption 5, their Theorem 1 states that

$$\operatorname*{argmin}_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(\mu^\star | k_\epsilon \star \mu_n) \leq \frac{C_{\mu^\star}^2 h}{n} \tag{7.17}$$

where $h = 4\log(3\sqrt{e} + a)$ is a constant depending on ϵ since $a = \sup_{m_1, m_2 \in \mathbb{R}^d} \log\left(k_\epsilon^{m_1}(x)/k_\epsilon^{m_2}(x)\right)$. In our case, the constant in the rate does not involve h as we do not rely on the same functions (our $B \leq 1$ instead of $B \leq h$ in their case).

When Assumption 5 does not hold, they also show in Theorem 2 that for every $g_P = \int k_\epsilon(\cdot - w) dP(w)$,

$$\operatorname*{argmin}_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(\mu^\star | k_\epsilon \star \mu_n) \leq \mathrm{KL}(\mu^\star | g_P) + \frac{C_{\mu^\star, P}^2 h}{n} \tag{7.18}$$

where $C^2_{\mu^\star,P} = \int \frac{\int k_\epsilon^m(x)^2 \mathrm{d}P(m)}{(\int k_\epsilon^w(x)\mathrm{d}P(w))^2}\mathrm{d}\mu^\star(x)$. However, they can easily obtain this result as a consequence of their first theorem along with the linearity of the forward KL. In constrast, the reverse KL does not verify linearity nor triangular inequality hence we cannot obtain readily such a generalization.

Notice that the forward KL rate obtained in [Li and Barron, 1999] is of order $1/n$, outpacing the one we attained. This is due to our chosen variational family, which is a Gaussian mixture with fixed weights. However, considering non-fixed weights (i.e. non-equally weighted mixtures) allows us to set $\alpha = 2/(n+1)$, thus achieving the exact same rate as [Li and Barron, 1999] for the reverse KL.

Since the Total Variation can be written as an Integral probability metric over measurable functions $f : \mathbb{R}^d \to [-1, 1]$, we deduce from Pinsker's inequality and Theorem 95 that a minimizer μ_n of $\mathrm{KL}(k_\epsilon \star \cdot|\mu^\star)$ achieves the following bound for the integral approximation error among this set of functions:

$$\left|\int f \,\mathrm{d}(k_\epsilon \star \mu_n) - \int f \mathrm{d}\mu^\star\right| \leq \sqrt{\frac{C^2_{\mu^\star}(\log(n)+1)}{2n}}.$$

The latter is then comparable to the integral approximation error of MCMC methods which is known to be of order $\mathcal{O}(n^{-\frac{1}{2}})$ when using n particles [Łatuszyński et al., 2013].

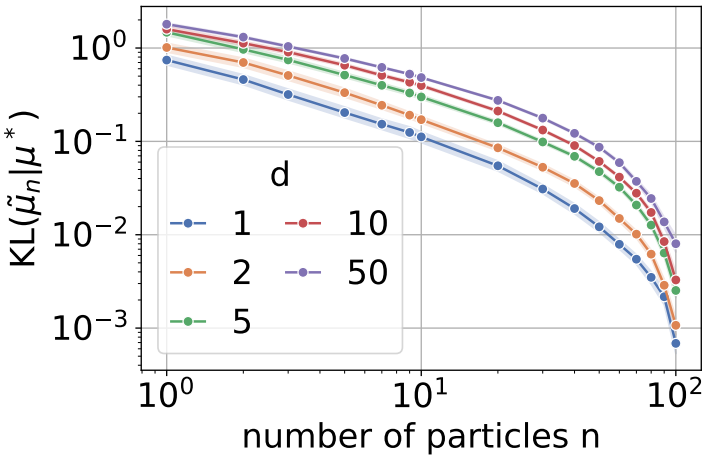


Fig. 7.3 Illustration of the rates of Theorem 95, where $\nu_n = \mathrm{argmin}_{\nu\in\mathcal{C}_n} \mathrm{KL}(\nu|\mu^\star)$ is approximated by $\tilde{\nu}_n$.

We finally test numerically the validity of Theorem 95 in a simple setting. The target distribution considered is a Gaussian mixture with 100 components. We denote by $(x_i^\star)_{i\leq 100}$ the mean of these components. For any $n \in [1, 100]$, the objective is to solve (7.1) and find $\nu_n := \mathrm{argmin}_{\nu\in\mathcal{C}_n} \mathrm{KL}(\nu|\mu^\star)$, where $\mathcal{Q}_n$ represents the family of Gaussian mixtures with n components. This minimizer is approximated by selecting only the first n components $(x_i^\star)_{i\leq n}$ of $\mu^\star$, and we denote $\tilde{\nu}_n$ the resulting approximate distribution. Note that in that specific setting, the variational family $\mathcal{C}_n$ and the target distribution $\mu^\star$ share the same standard deviation. Figure 7.3 shows the convergence rate of $\mathrm{KL}(\tilde{\nu}_n|\mu^\star)$ with respect to the number of components n, for various dimensions. The objective is estimated by Monte Carlo with 1000 samples. The 95% confidence interval displayed in Figure 7.3 is approximated based on 100 samples, representing the randomness corresponding to Monte Carlo approximation of the KL, and the initialization of the target. Figure 7.3 illustrates that the Kullback-Leibler divergence between $\tilde{\nu}_n$ and the $\mu^\star$ is indeed decreasing linearly with n. This result proves the validity of the rates derived in Theorem 95 for this specific setting. A full description of the experimental setup can be found in Subsection 7.6.

5 Related work

In this section we discuss relevant related work.

Theoretical guarantees for Variational Inference. For the variational inference optimization problem in (7.1), frequently employed constraint sets $\mathcal{G}$ in existing literature encompass the set of non-degenerate Gaussian distributions, location-scale families, mixtures of Gaussian components, and the set of product measures. In the Gaussian setting, [Lambert et al., 2022b, Diao et al., 2023] have been the first to leverage the geometry of Wasserstein gradient flows to study the convergence properties of variational inference, and provide convergence rates when the target $\mu^* \propto e^{-V}$ has a smooth and strongly convex potential V. In Mean-Field Variational inference (MFVI), the space $\mathcal{G}$ in (7.1) is taken to be the class of product measures over $\mathbb{R}^d$, written $\mathcal{P}(\mathbb{R})^{\otimes d}$. Several works have proposed algorithms in this context via wasserstein gradient flows [Yao and Yang, 2022, Lacker, 2023]. [Jiang et al., 2023] consider a smaller subset of $\mathcal{G}$, namely a polyhedral subset for which they can derive optimization and approximation guarantees. However the previous work do not tackle mixture of Gaussians for the variational family.

Mollified Relative entropies. A closely related line of work to this paper is the one of [Carrillo et al., 2019, Craig et al., 2023a,b, Carrillo et al., 2024] that study Wasserstein gradient flows of mollified relative entropies and the associated particle systems, that are of particular interest in the literature of partial differential equations and kinetic theory. In [Carrillo et al., 2019], the authors mention the mollified negative entropy U_ϵ that we define in (7.2) as a regularization of the negative entropy $\mathcal{U}(\mu) = \int \log(\mu) \mathrm{d}\mu$ (or entropy of order 1), but they do not study it. Instead, they focus on a closely related functional, defined as $\tilde{\mathcal{U}}_\epsilon(\mu) = \int \log(k_\epsilon \star \mu) \mathrm{d}\mu$ (i.e. with only one convolution inside the logarithm, while U_ϵ involves two convolutions). While they mention the possible choice of U_ϵ as a regularization of the entropy $\mathcal{U}$, they choose to study the alternative regularization $\tilde{\mathcal{U}}_\epsilon(\mu)$ for numerical reasons, as the Wasserstein gradient of the latter functional writes as an integral over the distribution of the particles, while the one of U_ϵ (hence $\mathcal{F}_\epsilon$) writes as an integral over the whole space w.r.t. Lebesgue measure, as explained in Section 2.1. Hence their results on λ-convexity[1] of the functional or the particle system differ from our setting. [Craig et al., 2023a] focus on a mollified chi-square divergence that corresponds to a weighted second order entropy; [Li et al., 2022] studies another mollified approximation of the chi-square divergence. [Carrillo et al., 2024] also study λ-convexity of entropies but only for entropies of order strictly greater than 1. Finally [Craig et al., 2023b] study functionals of the form $\int f_\epsilon(k_\epsilon \star \mu) \mathrm{d}\mathcal{L}$ as approximations of $\int f(\mu) \mathrm{d}\mathcal{L}$ where $\mathcal{L}$ denotes the Lebesgue measure. In their case f_ϵ is a specific function depending on f and ϵ, which excludes U_ϵ and thus also differs from our setting.

Variational inference for mixtures. Several works have tackled VI on mixtures on a computational aspect. [Gershman et al., 2012] optimize (with L-BFGS, that is a quasi Newton method) an approximate ELBO (recall that the ELBO is the reverse KL we consider up to an additive constant), using several consecutive approximations of ELBO terms for the case of mixture of Gaussians. In the end, their optimization objective differs a lot from the original KL objective from VI, that is a valid divergence between probability distributions - in contrast with their objective. [Arenz et al., 2018] adopt an Expectation-Maximization (EM) approach. As noted in [Aubin-Frankowski et al., 2022, Kunstner et al., 2021] EM can be seen as mirror descent scheme on the KL. Also, this algorithm can be seen as an Euler discretization of the gradient flow of the KL in the Fisher-Rao geometry [Domingo-Enrich and Pooladian, 2023, Chopin et al., 2024]. The parallel can be seen from eq (5) or (8) in [Arenz et al., 2018], that take a similar form as eq (6) in [Chopin et al., 2024], i.e. a geometric update on the distributions, i.e. that act directly on updating densities (in a "vertical" manner), equivalently weights. In contrast, we focus on gradient descent dynamics, that correspond to a time discretization of the KL gradient flow in the Wasserstein geometry. This correspond to "horizontal" updates, where particles are displaced at each iteration. [Lin et al., 2019] use natural gradient updates for VI in the natural parameter space (e.g. means for Gaussians). However, from [Raskutti and Mukherjee, 2015, Kunstner et al., 2021], it is known that this is equivalent to mirror descent on the exponential family parameters, which again is related to Fisher-Rao dynamics on the space of probability distributions (see eq (13) in [Chopin et al., 2024]).

6 Conclusion

The goal of this chapter is to improve our theoretical understanding of variational inference algorithms in the non-Gaussian case. We consider here a specific family of distributions, a mixture of Gaussians with constant covariance and equally weighted components, that enables us to derive novel results for the approximation and optimization error for Variational Inference. We derive theoretical guarantees regarding gradient descent of the objective (i.e. a descent lemma proving that the objective decreases at each iteration) leveraging smoothness of the objective and the Wasserstein geometry. We also derive novel approximation results for minimizers of the objective.

[1] λ-convexity for $\lambda \geq 0$ recovers displacement convexity, while $\lambda \leq 0$ recovers smoothness.

In our study, we chose to simplify our exploration of Variational Inference (VI) within the context of Gaussian Mixtures by assuming uniform weights for each Gaussian component and by fixing the covariances. Extending our findings to more complex scenarios, where the weights of each Gaussian are dynamically optimized and the covariances are variable, represents a significant challenge that goes beyond the scope of our current research. For instance, the task of optimizing the weights attached to each Gaussian component introduces a shift from the Wasserstein dynamics, which are central to our current discussion, to Fisher-Rao dynamics. Achieving a counterpart to our current optimization result Proposition 92 under these conditions would not only require the adoption of alternative proof techniques but also a deep dive into the intricacies of Fisher-Rao dynamics, which diverge significantly from those of Wasserstein. Furthermore, the optimization of covariance matrices introduces another level of complexity. Such an endeavor requires a unique analytical framework, primarily due to the constraints imposed by the requirement that these matrices be positive definite. While this aspect of the analysis is crucial for a comprehensive understanding of VI in Gaussian mixtures, it requires a specialized approach that our current methodology does not cover. The exploration of dynamic weight optimization and variable covariance matrices within the context of Gaussian Mixtures in VI presents a rich avenue for future work.

7 Appendix

The appendix is organized as follows. Subsection 7.1 details the computations for the Wasserstein gradients and the particle scheme corresponding to the optimization of the mollified relative entropy. Subsection7.2 discusses the connection with the algorithm and framework presented in [Lambert et al., 2022b]. Subsection 7.3 contains the intermediate lemmas needed for the proof of Theorem 95. Subsection 7.5 contains the proofs of the Propositions regarding Wasserstein Hessians. Subsection 7.6 outlines the setup used for the numerical experiments.

7.1 Particle implementation of the gradient flow

A solution of (7.3) is implemented by the Mac-Kean Vlasov process:

$$\dot{m}_t = -(\nabla\, U'_\epsilon(\mu_t)(m_t) + \nabla \mathcal{E}'_{V_\epsilon}(\mu_t)(m_t)). \tag{7.19}$$

Here we detail the computation of the vector field in (7.19) and its particle implementation.

For the negative entropy part, we can rewrite $U_\epsilon(\mu) = \int U(k_\epsilon \star \mu(\theta))\mathrm{d}\theta$ where $U : x \mapsto x\log(x)$. We have that

$$U'_\epsilon(\mu)(\cdot) = k_\epsilon \star (U' \circ (k_\epsilon \star \mu))(\cdot) = \int_{\mathbb{R}^d} k_\epsilon(\theta - \cdot)U'\left(\int k_\epsilon(\theta - y)\mathrm{d}\mu(y)\right)d\theta$$

where $U' : x \mapsto \log(x) + 1$. Hence, computing U'_ϵ requires an integration over $\mathbb{R}^d$. Then, we have, since k_ϵ is smooth, using an integration by parts with $\nabla U'(x) = \nabla \log(x)$ and symmetry of k_ϵ, for any $w \in \mathbb{R}^d$ we have:

$$\begin{aligned}
\nabla_w\, U'_\epsilon(\mu)(w) &= \nabla_w k_\epsilon \star (U' \circ (k_\epsilon \star \mu))(w) \\
&= \int_{\mathbb{R}^d} \nabla_w k_\epsilon(\theta - w)U'\left(\int k_\epsilon(\theta - y)\mathrm{d}\mu(y)\right)\mathrm{d}\theta \\
&= -\int_{\mathbb{R}^d} \nabla_\theta k_\epsilon(\theta - w)U'\left(\int k_\epsilon(\theta - y)\mathrm{d}\mu(y)\right)\mathrm{d}\theta \\
&= +\int_{\mathbb{R}^d} k_\epsilon(\theta - w)\nabla_\theta U'\left(\int k_\epsilon(\theta - y)\mathrm{d}\mu(y)\right)\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} k_\epsilon(\theta - w)\nabla_\theta \log\left(\int k_\epsilon(\theta - y)\mathrm{d}\mu(y)\right)\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} k_\epsilon(\theta - w)\frac{\int \nabla k_\epsilon(\theta - y)\mathrm{d}\mu(y)}{\int k_\epsilon(\theta - y)\mathrm{d}\mu(y)}\mathrm{d}\theta.
\end{aligned} \tag{7.20}$$

Finally, if μ_t is an atomic measure of the form $\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{m_t^{(i)}}$, then a particle implementation of (7.22) reduces to solve a system of ordinary differential equations for the locations of the Dirac masses:

$$\dot{m}_t^{(j)} = -\int_{\mathbb{R}^d} \nabla V(y) k_\epsilon(y - m_t^{(j)}) \mathrm{d}y - \int_{\mathbb{R}^d} \frac{\sum_{i=1}^N \nabla k_\epsilon(y - m_t^{(i)})}{\sum_{i=1}^N k_\epsilon(y - m_t^{(i)})} k_\epsilon(y - m_t^{(j)}) dy. \tag{7.21}$$

For the potential energy part, we can rewrite

$$\begin{aligned}
\mathcal{E}_{V_\epsilon}(\mu) &= \int_{\mathbb{R}^d} V(\theta) d(k_\epsilon \star \mu)(\theta) \\
&= \int_{\mathbb{R}^d} V(\theta) \int k_\epsilon(\theta - m) d\mu(m) d\theta \\
&= \iint_{\mathbb{R}^d} k_\epsilon(\theta - m) V(\theta) d\theta d\mu(m) \\
&:= \int_{\mathbb{R}^d} V_\epsilon(m) d\mu(m),
\end{aligned}$$

where $V_\epsilon(m) = \int_{\mathbb{R}^d} k_\epsilon(\theta - m) V(\theta) \mathrm{d}\theta = k_\epsilon \star V(m)$. Hence, we have for any $w \in \mathbb{R}^d$

$$\mathcal{E}'_{V_\epsilon}(\mu)(w) = V_\epsilon(w)$$

and successively

$$\nabla_w k_\epsilon(\theta - w) V(\theta) \mathrm{d}\theta = -\int_{\mathbb{R}^d} \nabla_\theta k_\epsilon(\theta - w) V(\theta) \mathrm{d}\theta = \int_{\mathbb{R}^d} k_\epsilon(\theta - \cdot) \nabla V(\theta) \mathrm{d}\theta$$

using again an integration by parts. Hence, (7.19) becomes:

$$\dot{m}_t = -\int_{\mathbb{R}^d} k_\epsilon(\theta - m_t) \nabla V(\theta) \mathrm{d}\theta - \int_{\mathbb{R}^d} k_\epsilon(\theta - m_t) \frac{\int \nabla k_\epsilon(\theta - y) \mathrm{d}\mu_t(y)}{\int k_\epsilon(\theta - y) \mathrm{d}\mu_t(y)} \mathrm{d}\theta. \tag{7.22}$$

7.2 Mixture of Gaussians optimization

[Lambert et al., 2022b] consider a Gaussian approximation of the Langevin diffusion given by Saarka's heuristic, i.e. $X_t \sim \mu_t$ where μ_t is the solution of Fokker-Planck equation is replaced by $Y_t \sim \mathcal{N}(m_t, \Sigma_t)$ where

$$\begin{aligned}
\dot{m}_t &= -\mathbb{E}[\nabla V(Y_t)] \\
\dot{\Sigma}_t &= 2\mathrm{I}_d - \mathbb{E}[\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes \nabla V(Y_t)]
\end{aligned}$$

They prove that the law of Y_t is the gradient flow of the KL on the Bures-Wasserstein manifold $\mathrm{BW}(\mathbb{R}^d) \cong \mathbb{R}^d \times S_d^{++}$ (the space of Gaussians equipped with the W_2 distance); which is a submanifold of $\mathcal{P}_2(\mathbb{R}^d)$. It can be seen as "Projected WG" where the Wasserstein gradient of the KL is projected onto the tangent space of the submanifold; another way to view it is to see that its the GF of the KL on the Bures-Wasserstein manifold.

Then, they propose to write a Gaussian mixture ρ on $\mathbb{R}^d$ as $\rho_\nu(\theta) = \int_{\mathrm{BW}(\mathbb{R}^d)} p(\theta) \mathrm{d}\nu(p)$[2] where ν is a measure over $\mathrm{BW}(\mathbb{R}^d)$; hence MOG is isomorphic to $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$. Then the WGF of $\nu \mapsto \mathrm{KL}(\rho_\nu | \mu^\star)$, ie the GF of this functional over $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$ is implemented through a particle system $\nu_t = \frac{1}{N}\sum_{i=1}^N \delta_{(m_t^{(i)}, \Sigma_t^{(i)})}$:

$$\dot{m}_t^{(i)} = -\mathbb{E}\left[\nabla \log\left(\frac{\rho_{\nu_t}}{\mu^\star}\right)\left(Y_t^{(i)}\right)\right] \tag{7.23}$$

$$\dot{\Sigma}_t^{(i)} = -\mathbb{E}\left[\nabla^2 \log\left(\frac{\rho_{\nu_t}}{\mu^\star}\right)\left(Y_t^{(i)}\right)\right]\Sigma_t^{(i)} - \Sigma_t^{(i)}\mathbb{E}\left[\nabla^2 \log\left(\frac{\rho_{\nu_t}}{\mu^\star}\right)\left(Y_t^{(i)}\right)\right] \tag{7.24}$$

[2] We can rewrite it as $\int_{\mathbb{R}^d \times S_d^{++}} p_{y,\Sigma}(\theta) d\nu(y, \Sigma)$

where $Y_t^{(i)} \sim \mathcal{N}(m_t^{(i)}, \Sigma_t^{(i)})$.

In contrast, in this work we restrict ourselves to Gaussian mixtures ρ that write $\rho_\mu = \int_{\mathbb{R}^d} k_\epsilon(\theta - y)\mathrm{d}\mu(y)$ where μ is a measure over $\mathbb{R}^d$. Then the WGF of $\mu \mapsto \mathrm{KL}(\rho_\mu|\pi)$, i.e. the GF of this functional over $\mathcal{P}_2(\mathbb{R}^d)$ is equivalent to the update above from [Lambert et al., 2022b]. Indeed if we fix $\nu = \mu \otimes \delta_{\epsilon \mathrm{I}_d}$, a Gaussian mixture writes $\rho_\nu(\theta) = \rho_\mu(\theta) = \int_{\mathbb{R}^d} k_\epsilon(\theta - y)\mathrm{d}\mu(y)$. In this case we consider the particle system $\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{m_t^{(i)}}$, we have $\rho_{\mu_t}(\theta) = \frac{1}{N}\sum_{i=1}^N k_\epsilon(\theta - m_t^{(i)})$. The update (7.23) becomes:

$$\begin{aligned}
\dot{m}_t^{(j)} &= -\mathbb{E}\left[\nabla \log\left(\frac{\rho_{\nu_t}}{\pi}\right)\left(Y_t^{(j)}\right)\right] \\
&= -\mathbb{E}[\nabla V(Y_t^{(j)})] - \mathbb{E}\left[\nabla \log\left(\rho_{\mu_t}\right)\left(Y_t^{(j)}\right)\right] \\
&= -\mathbb{E}[\nabla V(Y_t^{(j)})] - \mathbb{E}\left[\frac{\sum_{i=1}^N \nabla k_\epsilon(Y_t^{(j)} - m_t^{(i)})}{\sum_{i=1}^N k_\epsilon(Y_t^{(j)} - m_t^{(i)})}\right] \\
&= -\int \nabla V(y) k_\epsilon(y - m_t^{(j)})\mathrm{d}y - \int \frac{\sum_{i=1}^N \nabla k_\epsilon(y - m_t^{(i)})}{\sum_{i=1}^N k_\epsilon(y - m_t^{(i)})} k_\epsilon(y - m_t^{(j)})\mathrm{d}y
\end{aligned}$$

since $Y_t^{(j)} \sim \mathcal{N}(m_t^{(j)}, \epsilon \mathrm{I}_d)$ has density $k_\epsilon(\cdot - m_t^{(j)})$ hence we obtain the same update as (7.21).

7.3 Lemmas for the proof of Theorem 95

Lemma 96. *For any $x \in \mathbb{R}^+$, the function defined on $[0, +\infty[$ by*

$$B(x) = \frac{x\log x - x + 1}{(x-1)^2} \text{ if } x > 0,$$

and $B(0) = 1$ is monotone decreasing in r.

Proof. Firstly, the derivative of f is given by

$$B'(x) = \frac{2 - \frac{x+1}{x-1}\log(x)}{(x-1)^2}.$$

Recall some inequalities of the log function derived in [Topsøe, 2007]:

$$\begin{aligned}
&\forall x \in [1, +\infty[\,, \quad \frac{2(x-1)}{x+1} \leq \log(x), \\
&\forall x \in [0, 1]\,, \quad \log(x) \leq \frac{2(x-1)}{x+1}.
\end{aligned}$$

Consequently, combining those two inequalities and multiplying by $1/(x-1)$ which is positive on $[1, +\infty[$ and negative on $[0, 1[$ we obtain for any $x \in [0, +\infty[$

$$\frac{\log(x)}{x-1} \geq 2(x-1)$$

It implies that the derivative $f'(x)$ is always negative and f is monotone decreasing. □

Lemma 97. *For any $x \in \mathbb{R}^+$ we have*

$$C(x) = B(x)(x-1) = \frac{x\log(x) - x + 1}{x-1} \leq \sqrt{x} - 1$$

Proof. Recall the inequalities derived in [Topsøe, 2007]

1. for any $x \in [1, +\infty[$, $\log(x) \leq \frac{x-1}{\sqrt{x}}$
2. for any $x \in [0, 1]$, $\log(x) \geq \frac{x-1}{\sqrt{x}}$.

Combining those inequalities and multiplying by $1/(x-1)$ which is positive on $[1, +\infty[$ and negative on $[0, 1[$, we obtain for any $x \in [0, \infty[$,

$$\frac{\log(x)}{x-1} \leq \frac{1}{\sqrt{x}}.$$

Moreover,

$$C(x) - \sqrt{x} - 1 = \frac{x \log(x)}{x-1} - \sqrt{x}.$$

Consequently, by multiplying the previous inequality by x, we obtain that $C(x) - (\sqrt{x}+1) \leq 0$ □

Lemma 98. *For any $\alpha \in [0, 1]$, we have*

$$-\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} \leq 0.$$

Proof. Let's start by applying the classical inequality $\forall x > -1$, $\log(1+x) \leq x$ at $x = -\alpha$, we obtain $\log(1-\alpha) \leq -\alpha$. Hence,

$$\begin{aligned} -\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} &\leq -\alpha - \alpha(1-\alpha) + 2\alpha\sqrt{1-\alpha} \\ &= \alpha(2\sqrt{1-\alpha} - 2 + \alpha) \\ &:= \alpha g(\alpha) \end{aligned}$$

Moreover,

$$g(\alpha) = 2\sqrt{1-\alpha} - 2 + \alpha \text{ and } g'(\alpha) = \frac{-1}{\sqrt{1-\alpha}} - 1 \leq 0,$$

hence g is decreasing. Consequently,

$$-\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} \leq \alpha g(0) \leq 0.$$

□

7.4 Proof of Proposition 92

We first deal with the potential energy term. Notice that under Assumption 3, V_ϵ is also L-smooth, since for any $x, y \in \mathbb{R}^d$

$$\|\nabla V_\epsilon(x) - \nabla V_\epsilon(y)\| \leq \int k_\epsilon(\theta)\|\nabla V(x-\theta) - \nabla V(y-\theta)\| \mathrm{d}\theta \leq L\|x-y\| \tag{7.25}$$

since $\int k_\epsilon(\theta)\mathrm{d}\theta = 1$. Hence we have the following.

Proposition 99. *Let ρ in $\mathcal{P}_2(\mathbb{R}^d)$ and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{I}_d + t\phi$.*

$$\frac{d}{dt}\mathcal{E}_{V_\epsilon}(\rho_t) - \frac{d}{dt}\mathcal{E}_{V_\epsilon}(\rho_t)\Big|_{t=0} \leq Lt\|\phi\|^2_{L^2(\rho)}.$$

Proof. By the chain rule in Wasserstein space we have

$$\frac{d}{dt}\mathcal{E}_{V_\epsilon}(\rho_t) = \langle \nabla \mathcal{E}'_{V_\epsilon}(\rho_t), v_t \rangle_{L^2(\rho_t)} = \langle \nabla V_\epsilon, v_t \rangle_{L^2(\rho_t)}.$$

Hence, using the transfer Lemma and Cauchy-Schwarz successively,

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_{V_\epsilon}(\rho_t) - \frac{d}{dt}\mathcal{E}_{V_\epsilon}(\rho_t)\Big|_{t=0} &= \langle \nabla V_\epsilon, v_t\rangle_{L^2(\rho_t)} - \langle \nabla V_\epsilon, \phi\rangle_{L^2(\rho)} \\ &= \langle \nabla V_\epsilon \circ T_t - \nabla V_\epsilon, \phi\rangle_{L^2(\rho)} \\ &\leq \mathbb{E}_{w\sim\rho}[L\|T_t(x) - x\|\|\phi(w)\|] \leq Lt\|\phi\|^2_{L^2(\rho)}\,. \end{aligned} \tag{7.26}$$

□

We now turn to the mollified entropy term that is the most challenging.

Proposition 100. *Let ρ denote a mixture of n Diracs and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{I}_d + t\phi$. We have:*

$$\frac{d}{dt}\,U_\epsilon(\rho_t) - \frac{d}{dt}\,U_\epsilon(\rho_t)\Big|_{t=0} \leq \left(\frac{1}{\epsilon^2} + \frac{\sqrt{m_2(\rho)n}}{\epsilon^3} + \frac{\sqrt{n}}{\epsilon^2} + \frac{n\sqrt{m_2(\rho)}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)}$$

where $m_2(\rho)$ denotes the second moment of ρ.

Proof. By the chain rule in Wasserstein space, we have

$$\frac{d}{dt}\,U_\epsilon(\rho_t) = \langle \nabla\, U'_\epsilon(\rho_t), v_t\rangle_{L^2(\rho_t)}.$$

Consequently,

$$\begin{aligned}\frac{d}{dt}\,U_\epsilon(\rho_t) - \frac{d}{dt}\,U_\epsilon(\rho_t)\Big|_{t=0} &= \langle \nabla\, U'_\epsilon(\rho_t), v_t\rangle_{L^2(\rho_t)} - \langle \nabla\, U'_\epsilon(\rho), \phi\rangle_{L^2(\rho)} \\ &= \langle \nabla\, U'_\epsilon(T_t\#\rho)\circ T_t - \nabla\, U'_\epsilon(\rho), \phi\rangle_{L^2(\rho)} \\ &\leq \mathbb{E}_{w\sim\rho}[\|\nabla\, U'_\epsilon(T_t\#\rho)(T_t(w)) - \nabla\, U'_\epsilon(\rho)(w)\|\|\phi(w)\|] \end{aligned} \tag{7.27}$$

where in the second line we have used the transfer Lemma and in the last inequality Cauchy-Schwarz. Now, let's focus on the term $\|\nabla\, U'_\epsilon(T_t\#\rho)(T_t(w)) - \nabla\, U'_\epsilon(\rho)(w)\|$, that we will decompose as

$$\begin{aligned}\nabla\, U'_\epsilon(T_t\#\rho)\circ T_t - \nabla\, U'_\epsilon(\rho) &= \nabla\, U'_\epsilon(T_t\#\rho)(T_t(w)) - \nabla\, U'_\epsilon(\rho)(T_t(w)) + \nabla\, U'_\epsilon(\rho)(T_t(w)) - \nabla\, U'_\epsilon(\rho)(w) \\ &:= \mathcal{B}_{T_t(w)}(\rho_t, \rho) + \mathcal{A}_\rho(T_t(w), w). \end{aligned} \tag{7.28}$$

In the rest of the proof, we will show the Lipschitzness on w for $\mathcal{A}$ and on ρ for $\mathcal{B}$.

Using Proposition 101 and Proposition 102, we have

$$\begin{aligned}\frac{d}{dt}\,U_\epsilon(\rho_t) - \frac{d}{dt}\,U_\epsilon(\rho_t)\Big|_{t=0} &\leq \mathbb{E}_{w\sim\rho}\left[(\|\mathcal{A}_\rho(T_t(w), w)\| + \|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\|)\|\phi(w)\|\right] \\ &\leq \left(\frac{1}{\epsilon^2} + \frac{\sqrt{m_2(\rho)n}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)} + \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{nm_2(\rho)}}{2\epsilon^3} + \frac{n\sqrt{m_2(\rho)}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)} \\ &\leq \left(\frac{1}{\epsilon^2} + \frac{2\sqrt{m_2(\rho)n}}{\epsilon^3} + \frac{\sqrt{n}}{\epsilon^2} + \frac{n\sqrt{m_2(\rho)}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)}. \end{aligned}$$

□

Proposition 101. *Let ρ denote a mixture of n Diracs and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{I}_d + t\phi$. It holds that*

$$\|\mathcal{A}_\rho(T_t(w), w)\| \leq \left(\frac{1}{\epsilon^2} + \frac{\sqrt{2m_2(\rho)n}}{\epsilon^3}\right) t\|\phi(w)\| \tag{7.29}$$

Proof. Recalling the definition of $\nabla U'_\epsilon$ in (7.20), we obtain

$$\nabla U'_\epsilon(\rho)(w) = \int k_\epsilon(\theta - w) \frac{\int \nabla k_\epsilon(\theta - y) \mathrm{d}\rho(y)}{\int k_\epsilon(\theta - y) \mathrm{d}\rho(y)} \mathrm{d}\theta = \frac{1}{\epsilon^2} \int k_\epsilon(\theta - w) \frac{\int y k_\epsilon(\theta - y) \mathrm{d}\rho(y)}{\int k_\epsilon(\theta - y) \mathrm{d}\rho(y)} \mathrm{d}\theta - \frac{w}{\epsilon^2}. \tag{7.30}$$

Then from the definition of $\mathcal{A}$ in (7.28) we have

$$\begin{aligned}
\|\mathcal{A}_\rho(T_t(w), w)\| &= \frac{1}{\epsilon^2} \left\| \int (k_\epsilon(\theta - T_t(w)) + k_\epsilon(\theta - w)) \frac{\int y k_\epsilon(\theta - y) \mathrm{d}\rho(y)}{k_\epsilon \star \rho(\theta)} \mathrm{d}\theta - T_t(w) + w \right\| \\
&\leq \frac{1}{\epsilon^2} \int |k_\epsilon(\theta - T_t(w)) - k_\epsilon(\theta - w)| \frac{\int \|y\| k_\epsilon(\theta - y) \mathrm{d}\rho(y)}{k_\epsilon \star \rho(\theta)} \mathrm{d}\theta + \frac{t\|\phi(w)\|}{\epsilon^2}
\end{aligned}$$

Moreover, recall that ρ is a mixture of n Diracs. Therefore, we have

$$\begin{aligned}
\int \|y\| k_\epsilon(\theta - y) \mathrm{d}\rho(y) &\leq \sqrt{\int \|y\|^2 \mathrm{d}\rho(y)} \sqrt{\int k_\epsilon(\theta - y)^2 \mathrm{d}\rho(y)} \\
&= \sqrt{m_2(\rho)} \sqrt{\frac{1}{n} \sum_{i=1}^n k_\epsilon(\theta - y_i)^2} \\
&\leq \sqrt{m_2(\rho)} \frac{1}{\sqrt{n}} \sum_{i=1}^n k_\epsilon(\theta - y_i) \\
&\leq \sqrt{m_2(\rho) n}\, k_\epsilon \star \rho(\theta).
\end{aligned} \tag{7.31}$$

Consequently,

$$\begin{aligned}
\|\mathcal{A}_\rho(T_t(w), w)\| &\leq \frac{\sqrt{m_2(\rho) n}}{\epsilon^2} \, 2\,\mathrm{TV}(\mathcal{N}(T_t(w), \epsilon^2 \mathrm{I}_d), \mathcal{N}(w, \epsilon^2 \mathrm{I}_d)) + \frac{t\|\phi(w)\|}{\epsilon^2} \\
&\leq \frac{\sqrt{m_2(\rho) n}}{\epsilon^2} \sqrt{2\mathrm{KL}(\mathcal{N}(T_t(w), \epsilon^2 \mathrm{I}_d), \mathcal{N}(w, \epsilon^2 \mathrm{I}_d))} + \frac{t\|\phi(w)\|}{\epsilon^2} \\
&= \left(\frac{1}{\epsilon^2} + \frac{\sqrt{2 m_2(\rho) n}}{\epsilon^3} \right) t\|\phi(w)\|.
\end{aligned}$$

□

Proposition 102. *Let ρ denote a mixture of n Diracs and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{I}_d + t\phi$. We have:*

$$\mathbb{E}_{w \sim \rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\| \|\phi(w)\|] \leq \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{n m_2(\rho)}}{\epsilon^3} + \frac{n\sqrt{m_2(\rho)}}{\epsilon^3} \right) t\|\phi\|^2_{L^2(\rho)}.$$

Proof. Recalling the definition of (7.20), we have

$$\begin{aligned}
&\mathbb{E}_{w \sim \rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\| \|\phi(w)\|] \\
&\quad = \int \left\| \frac{1}{\epsilon^2} \int k_\epsilon(\theta - T_t(w)) \frac{\int y k_\epsilon(\theta - y) \mathrm{d}\rho_t(y)}{k_\epsilon \star \rho_t(\theta)} - \frac{\int y k_\epsilon(\theta - y) \mathrm{d}\rho(y)}{k_\epsilon \star \rho(\theta)} \right\| . \|\phi(w)\| \mathrm{d}\theta \mathrm{d}\rho(w) \\
&\quad \leq \frac{1}{\epsilon^2} \int \Big(\int k_\epsilon(\theta - T_t(w)) \|\phi(w)\| \mathrm{d}\rho(w) \Big) \left\| \frac{\int y k_\epsilon(\theta - y) \mathrm{d}\rho_t(y)}{k_\epsilon \star \rho_t(\theta)} - \frac{\int y k_\epsilon(\theta - y) \mathrm{d}\rho(y)}{k_\epsilon \star \rho(\theta)} \right\| \mathrm{d}\theta.
\end{aligned} \tag{7.32}$$

Then, we can use Cauchy Schwarz inequality and that ρ_t is supported on n Diracs to obtain

$$\int \|\phi(w)\| k_\epsilon(\theta - T_t(w)) \mathrm{d}\rho(w) \leq \|\phi\|_{L^2(\rho)} \sqrt{\int k_\epsilon(\theta - w)^2 \mathrm{d}\rho_t(w)} = \sqrt{n} \|\phi\|_{L^2(\rho)} \, k_\epsilon \star \rho_t(\theta). \tag{7.33}$$

Moreover, recall that

$$\begin{aligned}
&\Big\| \frac{\int y k_\epsilon(\theta - y)\mathrm{d}\rho_t(y)}{k_\epsilon \star \rho_t(\theta)} - \frac{\int y k_\epsilon(\theta - y)\mathrm{d}\rho(y)}{k_\epsilon \star \rho(\theta)} \Big\| \\
&\quad \leq \frac{\Big\| \int y k_\epsilon(\theta - y)\mathrm{d}\rho_t(y) - \int y k_\epsilon(\theta - y)\mathrm{d}\rho(y) \Big\|}{k_\epsilon \star \rho_t(\theta)} + \Big\| \int y k_\epsilon(\theta - y)\mathrm{d}\rho(y) \Big\| \Big| \frac{1}{k_\epsilon \star \rho_t(\theta)} - \frac{1}{k_\epsilon \star \rho(\theta)} \Big| \\
&\quad \leq \frac{\int \|T_t(y) k_\epsilon(\theta - T_t(y)) - y k_\epsilon(\theta - y)\| \mathrm{d}\rho(y)}{k_\epsilon \star \rho_t(\theta)} + \int \|y\| k_\epsilon(\theta - y)\mathrm{d}\rho(y) \Big| \frac{k_\epsilon \star \rho(\theta) - k_\epsilon \star \rho_t(\theta)}{k_\epsilon \star \rho(\theta)\, k_\epsilon \star \rho_t(\theta)} \Big| \\
&\quad := \mathcal{C}_1(\theta) + \mathcal{C}_2(\theta).
\end{aligned} \tag{7.34}$$

We can now combine inequalities 7.32, 7.33 and 7.34 to obtain

$$\mathbb{E}_{w\sim\rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\| \|\phi(w)\|] \leq \frac{\sqrt{n}\|\phi\|_{L^2(\rho)}}{\epsilon^2} \int k_\epsilon \star \rho_t(\theta)(\mathcal{C}_1(\theta) + \mathcal{C}_2(\theta))\mathrm{d}\theta.$$

We first focus on the $\mathcal{C}_1$ term:

$$\begin{aligned}
\int k_\epsilon \star \rho_t(\theta)\mathcal{C}_1(\theta)\mathrm{d}\theta &= \int\int \|T_t(y) k_\epsilon(\theta - T_t(y)) - y k_\epsilon(\theta - y)\| \mathrm{d}\rho(y)\mathrm{d}\theta \\
&\leq \int\int \|T_t(y) - y\| k_\epsilon(\theta - T_t(y)) + \|y\|\, |k_\epsilon(\theta - T_t(y)) - k_\epsilon(\theta - y)| \mathrm{d}\rho(y)\mathrm{d}\theta \\
&\leq t\mathbb{E}_{y\sim\rho}[\|\phi(y)\|] + \int \|y\| 2\,\mathrm{TV}(\mathcal{N}(T_t(y), \epsilon^2 \mathrm{I}_d), \mathcal{N}(y, \epsilon^2 \mathrm{I}_d))\mathrm{d}\rho(y) \\
&\leq t\|\phi\|_{L^2(\rho)} + \int \frac{t\|y\|\, \|\phi(y)\|}{\epsilon} \mathrm{d}\rho(y) \\
&\leq t\|\phi\|_{L^2(\rho)} + \frac{t\|\phi\|_{L^2(\rho)}}{2\epsilon} \sqrt{\int \|y\|^2 \mathrm{d}\rho(y)} \\
&= \left(1 + \frac{\sqrt{m_2(\rho)}}{\epsilon}\right) t\|\phi\|_{L^2(\rho)}.
\end{aligned}$$

Finally, we focus on the $\mathcal{C}_2$ term. We obtain using the same computations as in (7.31):

$$\begin{aligned}
\int k_\epsilon \star \rho_t(\theta)\mathcal{C}_2(\theta)\mathrm{d}\theta &= \int \frac{\int \|y\| k_\epsilon(\theta - y)\mathrm{d}\rho(y)}{k_\epsilon \star \rho(\theta)} |k_\epsilon \star \rho(\theta) - k_\epsilon \star \rho_t(\theta)| \mathrm{d}\theta \\
&\leq \sqrt{m_2(\rho)n} \int |k_\epsilon \star \rho(\theta) - k_\epsilon \star \rho_t(\theta)| \mathrm{d}\theta \\
&\leq \sqrt{m_2(\rho)n} \int\int |k_\epsilon(\theta - y) - k_\epsilon(\theta - T_t(y))| \mathrm{d}\rho(y)\mathrm{d}\theta \\
&\leq \sqrt{m_2(\rho)n} \int\int \frac{t\|\phi\|_{L^2(\rho)}}{\epsilon} \mathrm{d}\rho(y)\mathrm{d}\theta \\
&\leq \frac{t\sqrt{m_2(\rho)n}}{\epsilon} \|\phi\|_{L^2(\rho)}.
\end{aligned}$$

Combining the previous inequalities, we obtain

$$\mathbb{E}_{w\sim\rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\| \|\phi(w)\|] \leq \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{nm_2(\rho)}}{\epsilon^3} + \frac{n\sqrt{m_2(\rho)}}{\epsilon^3} \right) t\|\phi\|^2_{L^2(\rho)}.$$

□

7.5 Wasserstein Hessians of relative entropies

7.5.1 Proof of Proposition 90

Proof. Let $\mu_t = (\mathrm{I}_d + t\nabla\psi)_{\#}\mu$ where $\psi \in C_c^\infty(\mathbb{R}^d)$. Let μ_t, μ^* be the densities of μ_t and μ^* respectively. We denote by $\phi_t = \mathrm{I}_d + tg$ where $g = \nabla\psi$. Hence we have $\mathrm{J}\phi_t = \mathrm{I}_d + t\mathrm{J}g$. Time derivatives are denoted as $\phi_t' = \frac{d\phi_t}{dt}$. Notice that $(\mathrm{J}\phi_t)' = \mathrm{J}\phi_t' = \mathrm{J}g = Hess\psi$.

For any f-divergence,

$$h_\mu(t) = \int f\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\mu^*(x)\mathrm{d}x = \int \tilde{f}\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\mu_t(x)\mathrm{d}x.$$

where $\tilde{f}(t) = f(t)/t$. By the transfer lemma and change of variables formula, we have

$$h_\mu(t) = \int \tilde{f}\left(\frac{\mu(x)}{\mu^*(\phi_t(x))|\mathrm{J}\phi_t(x)|}\right)\mathrm{d}\mu(x).$$

Let us rewrite

$$h_\mu(t) = \int \tilde{f}\left(\mu(x)e^{n_t(x)}\right)\mathrm{d}\mu(x), \quad \text{where} \quad n_t(x) = V(\phi_t(x)) - \log|\mathrm{J}\phi_t(x)|.$$

We have consecutively:

$$h_\mu'(t) = \int \tilde{f}'\left(\mu(x)e^{n_t(x)}\right)\mu(x)n_t'(x)e^{n_t(x)}\mathrm{d}\mu(x)$$
$$h_\mu''(t) = \int \left[\tilde{f}''(\mu(x)e^{n_t(x)})\left(\mu(x)n_t'(x)e^{n_t(x)}\right)^2\tilde{f}'(\mu(x)e^{n_t(x)})\left(n_t''(x) + n_t'(x)^2\right)\mu(x)e^{n_t(x)}\right]\mathrm{d}x$$

where

$$n_t'(x) = \left\langle \nabla V(\phi_t(x)), \phi_t'(x)\right\rangle - \mathrm{Tr}\left((\mathrm{J}\phi_t(x))^{-1}\mathrm{J}\phi_t'(x)\right),$$
$$n_t''(x) = \left\langle \mathrm{H}_V(\phi_t(x))\phi_t'(x), \phi_t'(x)\right\rangle + \mathrm{Tr}\left((\mathrm{J}\phi_t(x))^{-1}\mathrm{J}\phi_t'(x)^2\right),$$

since $\phi_t'' = 0$. At time $t = 0$, we have

$$n_0(x) = V(x) = -\log(\mu^*(x))$$
$$n_0'(x) = \langle \nabla V(x), \nabla\psi(x)\rangle - \Delta\psi(x),$$
$$n_0''(x) = \langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|_{HS}^2$$

since $\mathrm{Tr}(\mathrm{H}\psi) = \Delta\Psi$ and $\mathrm{Tr}\left((\mathrm{H}\psi)^2\right) = \|\mathrm{H}\psi\|_{HS}^2$. Notice that $n_0'(x) = \mathcal{L}_{\mu^*}\psi(x)$ where $\mathcal{L}_{\mu^*} : \psi \mapsto \langle \nabla V, \nabla\psi\rangle - \Delta\psi$ denotes the (negative) generator of the standard Langevin diffusion with stationary distribution μ^* with density $\mu^* \propto e^{-V}$, see [Pavliotis, 2014, Section 4.5].

Now we get at time $t = 0$:

$$h_\mu''(0) = \int \left[\left(\tilde{f}''\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\frac{\mu(x)}{\mu^*(x)}\right)^2 + \tilde{f}'\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\frac{\mu(x)}{\mu^*(x)}\right)\right)(\mathcal{L}_{\mu^*}\psi(x))^2 \right.$$
$$\left. + \tilde{f}'\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|_{HS}^2\right)\right]\mu(x)\mathrm{d}x.$$

Hence if V is convex, and that $\min(\tilde{f}'(t), t\tilde{f}'(t) + t^2\tilde{f}''(t)) \geq 0$, then $h_\mu''(0) \geq 0$. Now let $f(t) = t\log t - t$, then $h_\mu(t) = \mathrm{KL}(\mu_t|\mu^*) - 1$. Then, $\tilde{f}(t) = \log(t) - 1$; $\tilde{f}'(t) = 1/t$, $\tilde{f}''(t) = -1/t^2$, hence $t\tilde{f}'(t) + t^2\tilde{f}''(t) = 0$ and we obtain more precisely:

$$Hess_\mu \mathrm{KL}(\psi, \psi) = \int \left[\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|_{HS}^2\right]\mu(x)\mathrm{d}x.$$

□

7.5.2 Hessian of the mollified relative entropy

Recall that $\mathcal{F}_\epsilon(\mu) = \mathcal{E}_{V_\epsilon}(\mu) + U_\epsilon(\mu)$. Hence, for any $\psi \in C_c^\infty(\mathbb{R}^d)$, $Hess_\mu \mathcal{F}_\epsilon(\psi,\psi) = Hess_\mu \mathcal{E}_{V_\epsilon}(\psi,\psi) + Hess_\mu\, U_\epsilon(\psi,\psi)$. We directly have for the potential energy part that

$$\frac{d^2\mathcal{E}_{V_\epsilon}(\rho_t)}{\mathrm{d}t^2}\Big|_{t=0} = \int \langle \mathrm{H}_{V_\epsilon}(x)\nabla\psi(x), \nabla\psi(x)\rangle \,\mathrm{d}\mu(x). \tag{7.35}$$

using again our computation from Subsection 7.5.1. Since $\mathrm{H}_{V_\epsilon} = k_\epsilon \star \mathrm{H}_V$ and k_ϵ converges to a Dirac at origin as ϵ goes to zero, we get $Hess_\mu \mathcal{E}_{V_\epsilon}(\psi,\psi) \xrightarrow[\varepsilon\to 0]{} \int \langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle \,\mathrm{d}\mu(x)$.

We now turn to the mollified entropy part. We rewrite it along a geodesic $(\rho_t, v_t)_{t\in[0,1]}$ as

$$U_\epsilon(\rho_t) = \int \log(k_\epsilon \star \rho_t)\,\mathrm{d}(k_\epsilon \star \rho_t) = \int_\theta U(k_\epsilon \star \rho_t(\theta))\mathrm{d}\mathcal{L}_d(\theta),$$

denoting $U : x \mapsto x\log(x)$. The first time derivative of $t \mapsto U_\epsilon(\rho_t)$ is:

$$\frac{d\, U_\epsilon(\rho_t)}{dt} = \int U'(k_\epsilon \star \rho_t(\theta)) \frac{\mathrm{d}}{\mathrm{d}t} k_\epsilon \star \rho_t(\theta)\mathrm{d}\theta \tag{7.36}$$

$$= \int (1 + \log(k_\epsilon \star \rho_t(\theta))) \int \langle \nabla k_\epsilon(\theta - x), v_t(x)\rangle \,\mathrm{d}\rho_t(x)\mathrm{d}\theta \tag{7.37}$$

Since by an integration by parts,

$$\frac{\mathrm{d}}{\mathrm{d}t} k_\epsilon \star \rho_t(\theta)\mathrm{d}\theta = \int k_\epsilon(\theta - x)\frac{\partial \rho_t(x)}{\partial t}\mathrm{d}x = \int \nabla k_\epsilon(\theta - x)\rho_t(x)v_t(x)\mathrm{d}x.$$

From 7.36 we obtain

$$\begin{aligned}\frac{d^2\, U_\epsilon(\mu_t)}{dt^2} &= \int \left[U''(k_\epsilon \star \rho_t(\theta)) \left(\frac{dk_\epsilon \star \rho_t(\theta)}{dt} \right)^2 + U'(k_\epsilon \star \rho_t(\theta)) \frac{d^2 k_\epsilon \star \rho_t(\theta)}{dt^2} \right] \mathrm{d}\theta \\ &= \int_\theta \left[(k_\epsilon \star \rho_t(\theta))^{-1} \left(\frac{dk_\epsilon \star \rho_t(\theta)}{dt} \right)^2 + (1 + \log(k_\epsilon \star \rho_t(\theta))) \frac{d^2 k_\epsilon \star \rho_t(\theta)}{dt^2} \right] \mathrm{d}\theta. \end{aligned} \tag{7.38}$$

The first term in (7.38) is always positive but the second may not because of the logarithmic term. However, as $\epsilon \to 0$, we recover the geodesic convexity of the negative entropy, as stated in the following proposition.

Proposition 103. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Let $\psi \in C_c^\infty(\mathbb{R}^d)$. As $\epsilon \to 0$, the Wasserstein Hessian of the regularized entropy U_ϵ converges to the one of the regular negative entropy $\mathcal{U}(\mu) = \int \log(\mu)\mathrm{d}\mu$, i.e:*

$$Hess_\mu\, U_\epsilon(\psi,\psi) \xrightarrow[\varepsilon\to 0]{} Hess_\mu \mathcal{U}(\psi,\psi) = \int \|\mathrm{H}\psi(x)\|_{HS}^2 \mathrm{d}\mu(x). \tag{7.39}$$

Proof.

For each term, we will first take the limit as $t \to 0$ to recover the definition of the Hessian at μ (limiting distribution of ρ_t as t goes to 0), then $\epsilon \to 0$ to recover the case of the standard (non-regularized) relative entropy. Denote $h_\mu^\epsilon(t,\theta) = k_\epsilon \star \rho_t(\theta) = \int k_\epsilon(\theta - x)d\rho_t(x) = \int k_\epsilon(\theta - \phi_t(x))d\mu(x)$ by the transfer lemma. We have

$$\begin{aligned}\frac{d^2\, U_\epsilon(\mu_t)}{dt^2} &= \int \left[U''(h_\mu^\epsilon(t,\theta)) \left(\frac{dh_\mu^\epsilon(t,\theta)}{dt} \right)^2 + U'(h_\mu^\epsilon(t,\theta)) \frac{d^2 h_\mu^\epsilon(t,\theta)}{dt^2} \right] \mathrm{d}\theta \\ &= \int_\theta \left[(h_\mu^\epsilon(t,\theta))^{-1} \left(\frac{dh_\mu^\epsilon(t,\theta)}{dt} \right)^2 + \big(1 + \log\big(h_\mu^\epsilon(t,\theta)\big)\big) \frac{d^2 h_\mu^\epsilon(t,\theta)}{dt^2} \right] \mathrm{d}\theta. \end{aligned} \tag{7.40}$$

We firstly have

$$h_\mu^\epsilon(t,\theta) = k_\epsilon \star \rho_t(\theta) \xrightarrow[t\to 0]{} k_\epsilon \star \mu(\theta) \xrightarrow[\varepsilon\to 0]{} \mu(\theta).$$

Recall that the continuity equation along Wasserstein geodesics write:

$$\frac{\partial \rho_t(x)}{\partial t} + \boldsymbol{\nabla} \cdot \big(\rho_t(x) \nabla \psi \circ \phi_t^{-1}(x)\big)\big) = 0. \tag{7.41}$$

Then, using $\boldsymbol{\nabla} \cdot (aB) = \langle \nabla a, B \rangle + a \boldsymbol{\nabla} \cdot (B)$, the first time derivative of $t \mapsto h_\mu^\epsilon(t, \theta)$ writes

$$\begin{aligned}
\frac{dh_\mu^\varepsilon(t,\theta)}{dt} &= \int k_\epsilon(\theta - x) \frac{\partial \rho_t(x)}{\partial t} \mathrm{d}x && (7.42)\\
&= -\int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot \big(\rho_t(x) \nabla \psi(\phi_t^{-1}(x))\big) \mathrm{d}x && (7.43)\\
&\xrightarrow[t \to 0]{} -\int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot (\mu(x) \nabla \psi(x)) \mathrm{d}x, && (7.44)
\end{aligned}$$

Hence for the first term in (7.40) we have:

$$\begin{aligned}
&\int (h_\mu^\epsilon(t,\theta))^{-1} \left(\frac{dh_\mu^\varepsilon(t,\theta)}{dt} \right)^2 \mathrm{d}\theta && (7.45)\\
&\qquad \xrightarrow[t \to 0]{} \int (k_\epsilon \star \mu(\theta))^{-1} \left(\int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot (\mu(x) \nabla \psi(x)) \mathrm{d}x \right)^2 \mathrm{d}\theta && (7.46)\\
&\qquad \xrightarrow[\varepsilon \to 0]{} \int \mu(\theta)^{-1} \boldsymbol{\nabla} \cdot (\mu(\theta) \nabla \psi(\theta))^2 \mathrm{d}\theta && (7.47)\\
&\qquad = \int \mu(\theta)^{-1} \langle \nabla \mu(\theta), \nabla \psi(\theta) \rangle^2 d\theta + \int \Delta \psi(\theta)^2 d\mu(\theta) + 2 \int \Delta \psi(\theta) \left\langle \nabla \mu(\theta), \nabla \psi(\theta), \mathrm{d} \right\rangle \theta && (7.48)\\
&\qquad = (a) + (b) + (c). && (7.49)
\end{aligned}$$

We now turn to second term in (7.40). Using (7.42), the second time derivative of h_μ^ϵ writes:

$$\begin{aligned}
\frac{d^2 h_\mu^\varepsilon(t,\theta)}{dt^2} &= -\int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot \left(\frac{d}{dt} (\rho_t(x) \nabla \psi(\phi_t^{-1}(x))) \right) \mathrm{d}x \\
&= -\int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot \left(\frac{\partial \rho_t(x)}{\partial t} \nabla \psi(\phi_t^{-1}(x)) \right) \mathrm{d}x - \int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot \left(\rho_t(x) \frac{d \nabla \psi(\phi_t^{-1}(x))}{dt} \right) \mathrm{d}x \\
&= (d) + (e).
\end{aligned}$$

Then using $\frac{\partial \rho_t(x)}{\partial t} = -\boldsymbol{\nabla} \cdot \big(\rho_t(x) \nabla \psi(\phi_t^{-1}(x))\big) = -\langle \nabla \rho_t(x), \nabla \psi(\phi_t^{-1}(x)) \rangle - \rho_t(x) \Delta \psi(\phi_t^{-1}(x))$, we have

$$\begin{aligned}
(d) &= \int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot \big(\langle \nabla \rho_t(x), \nabla \psi(\phi_t^{-1}(x)) \rangle + \rho_t(x) \Delta \psi(\phi_t^{-1}(x)) \big) \nabla (\psi(\phi_t^{-1}(x))) \mathrm{d}x \\
&\xrightarrow[t \to 0]{} \int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot (\langle \nabla \mu(x), \nabla \psi(x) \rangle + \mu(x) \Delta \psi(x)) \nabla (\psi(x)) \mathrm{d}x \\
&\xrightarrow[\epsilon \to 0]{} \boldsymbol{\nabla} \cdot \Big(\langle \nabla \mu(\theta), \nabla \psi(\theta) \rangle \nabla \psi(\theta) \Big) + \boldsymbol{\nabla} \cdot \Big(\mu(\theta) \Delta \psi(\theta) \nabla (\psi(\theta)) \Big)
\end{aligned}$$

Now, using $\phi_t^{-1} \approx \mathrm{I}_d - t \nabla \psi$ for $t \approx 0$:

$$\begin{aligned}
(e) &= -\int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot \left(\rho_t(x) \frac{d}{dt} (\nabla \psi(\phi_t^{-1}(x))) \right) \mathrm{d}x \\
&\xrightarrow[t \to 0]{} \int k_\epsilon(\theta - x) \boldsymbol{\nabla} \cdot (\mu(x) \mathrm{H}\psi(x) \nabla \psi(x)) \mathrm{d}x \\
&\xrightarrow[\epsilon \to 0]{} \boldsymbol{\nabla} \cdot (\mu(\theta) \mathrm{H}\psi(\theta) \nabla \psi(\theta)) \mathrm{d}x.
\end{aligned}$$

Finally, for the second term in (7.40) we have

$$\begin{aligned}
&\int (1+\log\big(h_\mu^\epsilon(t,\theta)\big))\frac{d^2h_\mu^\epsilon(t,\theta)}{dt^2}\mathrm{d}\theta \\
&\quad\xrightarrow[t,\epsilon\to 0]{} \int (1+\log(\mu(\theta)))\Big\{\boldsymbol{\nabla}\cdot(\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\nabla\psi(\theta)+\mu(\theta)\Delta\psi(\theta)\nabla\psi(\theta)+\mu(\theta)\mathrm{H}\psi(\theta)\nabla\psi(\theta))\Big\}\mathrm{d}\theta \\
&\quad= -\int\langle\nabla\log(\mu(\theta)),\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\nabla\psi(\theta)+\mu(\theta)\Delta\psi(\theta)\nabla\psi(\theta)+\mu(\theta)\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta \\
&\quad= -\int\mu(\theta)^{-1}\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle^2\mathrm{d}\theta-\int\Delta\psi(\theta)\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\mathrm{d}\theta-\int\langle\nabla\mu(\theta),\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta \\
&\quad= -(a)-\frac{1}{2}(c)-\int\langle\nabla\mu(\theta),\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta.
\end{aligned}$$

Moreover, by an integration by parts, using the divergence of matrix vector product $\boldsymbol{\nabla}\cdot(Ab)=\boldsymbol{\nabla}\cdot(A)b+\mathrm{Tr}(A\nabla b)$:

$$\begin{aligned}
-\int\langle\nabla\mu(\theta),\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta &= \int\boldsymbol{\nabla}\cdot(\mathrm{H}\psi(\theta)\nabla\psi(\theta))\mathrm{d}\mu(\theta) \\
&= \int\langle\boldsymbol{\nabla}\cdot\mathrm{H}\psi(\theta),\nabla\psi(\theta)\rangle+\mathrm{Tr}(\mathrm{H}\psi(\theta)\mathrm{H}\psi(\theta))^\top\mathrm{d}\mu(\theta) \\
&= \int\langle\nabla(\Delta\psi(\theta)),\nabla\psi(\theta)\rangle+\|\mathrm{H}\psi(\theta)\|_F^2\mathrm{d}\mu(\theta),
\end{aligned}$$

where

$$\begin{aligned}
\int\langle\nabla(\Delta\psi(\theta)),\nabla\psi(\theta)\rangle d\mu(\theta) &= -\int\Delta\psi(\theta)\boldsymbol{\nabla}\cdot(\mu(\theta)\nabla\psi(\theta))\mathrm{d}\theta \\
&= -\int\Delta\psi(\theta)(\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle-\mu(\theta)\Delta\psi(\theta))\mathrm{d}\theta \\
&= -\frac{1}{2}(c)-(b).
\end{aligned}$$

Consequently,

$$\int(1+\log\big(h_\mu^\epsilon(t,\theta)\big))\frac{d^2h_\mu^\epsilon(t,\theta)}{dt^2}\mathrm{d}\theta\xrightarrow[t,\epsilon\to 0]{}-(a)-\frac{1}{2}(c)-\frac{1}{2}(c)-(b)+\int\|\mathrm{H}\psi(\theta)\|_F^2\mathrm{d}\mu(\theta). \tag{7.50}$$

Finally combining (7.49) and (7.50) we get the result. □

7.6 Experimental setting

The target distribution The target distribution $\mu^\star$ is chosen to be a Gaussian mixture with 100 components:

$$\mu^\star=\frac{1}{100}\sum_{i=1}^{100}\mathcal{N}(x_i^\star,\epsilon^2\mathrm{I}_d)$$

The components $(x_i^\star)_{i\leq 100}$ are randomly sampled from a normal distribution $\mathcal{N}(0,\sigma^2\mathrm{I}_d)$, where $\sigma=5$ in all experiments. The standard deviation of the target is set to $\epsilon=\epsilon_0\sqrt{d}$, where $\epsilon_0=1$ in our setting. This standard deviation scales with $\sqrt{d}$ because the term $\|x_i^\star\|_2$ also scales with $\sqrt{d}$. Without this scaling, the term $\mathcal{N}(x_i^\star,\epsilon^2\mathrm{I}_d)$ would be very close to a Dirac mass in high dimensions.

Variational family The variational family used for the experiments is the family of Gaussian mixtures with 10 components:

$$\mathcal{G}=\left\{\frac{1}{10}\sum_{i=1}^{1}0\mathcal{N}(x_i,\epsilon^2\mathrm{I}_d),\ x_i\in\mathbb{R}^d\right\}$$

At the beginning of the training, the mean of each component $(x_i)_{i\leq 10}$ is randomly initialized, sampled from a normal distribution $\mathcal{N}(0, \zeta^2 \mathrm{I}_d)$, where $\zeta = 15$ in all experiments. Note that these components are initialized further than the parameters $(x_i^\star)_{i\leq 100}$ of the target $\mu^\star$, ie, $\zeta \geq \sigma$. It seems that this setting allows to slightly improve the performances and the mode coverage of the algorithm. For simplicity, the variational family shares the same standard deviation ϵ than the target.

Training parameters The step-size is set as $\gamma = \gamma_0 \cdot d$, where $\gamma_0 = 0.01$. According to Proposition 92, the step-size should satisfy $\gamma \leq 2/M$ to ensure a decrease in the objective of each iteration, where the constant M scales inversely with d. Therefore, we opted for γ to scale with d accordingly.

Monte Carlo approximation of the cumulative mean Let μ a Gaussian mixture with n components. We denote by $(x_i)_{i\leq n}$ the mean of those components. Therefore, the term $\|\nabla \mathcal{F}_\epsilon'(\mu)\|^2_{L^2(\mu)}$ can be approximated by Monte Carlo with B samples using

$$\begin{aligned}
\|\nabla \mathcal{F}_\epsilon'(\mu)\|^2_{L^2(\mu)} &= \int \|\nabla \mathcal{F}_\epsilon'(\mu)(w)\|_2^2 d\mu(w) \\
&= \frac{1}{n}\sum_{i=1}^n \|\nabla \mathcal{F}_\epsilon'(\mu)(x_i)\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^n \| \int \nabla \log\left(\frac{\mu(y)}{\mu^\star(y)}\right) dk_\epsilon^{x_i}(y)\|_2^2 \\
&\approx \frac{1}{n}\sum_{i=1}^n \| \frac{1}{B}\sum_{j=1}^B \nabla \log\left(\frac{\mu(y_j^i)}{\mu^\star(y_j^i)}\right) \|_2^2 \,,
\end{aligned}$$

where $y_j^i \sim \mathcal{N}(x_i, \epsilon^2 \mathrm{I}_d)$.

Monte Carlo approximation of the KL Let ν_n a Gaussian mixture with n components. We denote by $(x_i)_{i\leq n}$ the mean of those components. Therefore, the Kullback-Leibler divergence between ν_n and the target $\mu^\star$ can be approximated by Monte Carlo with B samples using

$$\begin{aligned}
\mathrm{KL}(\nu_n, \mu^\star) &= \int \log\left(\frac{\nu_n(y)}{\mu^\star(y)}\right) d\mu(y) \\
&= \frac{1}{n}\sum_{i=1}^n \int \log\left(\frac{\nu_n(y)}{\mu^\star(y)}\right) dk_\epsilon^{x_i}(y) \\
&\approx \frac{1}{B\cdot n}\sum_{i=1}^n\sum_{j=1}^B \log\left(\frac{\nu_n(y_j^i)}{\mu^\star(y_j^i)}\right) \,,
\end{aligned}$$

where $y_j^i \sim \mathcal{N}(x_i, \epsilon^2 \mathrm{I}_d)$.

Part II

Thompson Sampling for Multi-Armed Bandit problems

CHAPTER 8

VARIATIONAL INFERENCE THOMPSON SAMPLING FOR CONTEXTUAL BANDITS

Chapter abstract: *In this chapter, we introduce and analyze a variant of the Thompson sampling (TS) algorithm for contextual bandits. At each round, traditional TS requires samples from the current posterior distribution, which is usually intractable. To circumvent this issue, approximate inference techniques can be used and provide samples with distribution close to the posteriors. However, current approximate techniques yield to either poor estimation (Laplace approximation) or can be computationally expensive (MCMC methods, Ensemble sampling...). In this paper, we propose a new algorithm, Varational Inference TS (*`VITS`*), based on Gaussian Variational Inference. This scheme provides powerful posterior approximations which are easy to sample from, and is computationally efficient, making it an ideal choice for TS. In addition, we show that* `VITS` *achieves a sub-linear regret bound of the same order in the dimension and number of round as traditional TS for linear contextual bandit. Finally, we demonstrate experimentally the effectiveness of* `VITS` *on both synthetic and real world datasets.*

1 Introduction

In traditional Multi-Armed Bandit (MAB) problems, an agent, has to sequentially choose between several actions (referred to as "arms"), from which he receives a reward from the environment. The arm selection process is induced by a sequence of policies, which is inferred and refined at each round from past observations. These policies are designed to optimize the cumulative rewards over the entire process. The main challenge in this task is to effectively manage a suitable exploitation and exploration trade-off [Robbins, 1952, Katehakis and Veinott, 1987, Berry and Fristedt, 1985, Auer et al., 2002, Lattimore and Szepesvári, 2020, Kveton et al., 2020a]. Here, exploitation refers to selecting an arm that is currently believed to be the best based on past observations, while exploration refers to selecting arms that have not been selected frequently in the past in order to gather more information.

Contextual bandit problems is a particular instance of MAB problem, which supposes, at each round, that the set of arms and the corresponding reward depend on a d-dimensional feature vector called a contextual vector or context. This scenario has been extensively studied over the past decades and learning algorithms have been developed to address this problem [Langford and Zhang, 2007a, Abbasi-Yadkori et al., 2011a, Agrawal and Goyal, 2013a, Kveton et al., 2020a], and they have been successfully applied in several real-world problem such as recommender systems, mobile health and finance [Li et al., 2010, Agarwal et al., 2016a, Tewari and Murphy, 2017, Bouneffouf et al., 2020]. The existing algorithms for addressing contextual bandit problems can be broadly categorized into two groups. The first category is based on maximum likelihood and the principle of optimism in the face of uncertainty (OFU) and has been studied in [Auer et al., 2002, Chu et al., 2011, Abbasi-Yadkori et al., 2011b, Li

et al., 2017a, Ménard and Garivier, 2017, Zhou et al., 2020, Foster and Rakhlin, 2020, Zenati et al., 2022]. The second category consists in randomized probability matching algorithms, which is based on Bayesian belief and posterior sampling. Thompson Sampling (TS) is one of the most famous algorithms that fall into this latter category. Since its introduction by [Thompson, 1933], it has been widely studied, both theoretically and empirically [Agrawal and Goyal, 2012, Kaufmann et al., 2012a, Agrawal and Goyal, 2013a, Russo and Van Roy, 2014, 2016, Lu and Van Roy, 2017, Riquelme et al., 2018, Jin et al., 2021a]. Despite the fact that OFU algorithms offer better theoretical guarantees compared to classic TS-based algorithms, traditional TS methodologies still appeal to us due to their straightforward implementation and empirical advantages. In [Agrawal and Goyal, 2012], the authors claimed that: "In applications like display advertising and news article recommendation, TS is competitive with or better than popular methods such as UCB". Similarly, [Chapelle and Li, 2011] has examined the empirical performances of TS on both simulated and real data. Their experiments demonstrate that TS outperforms OFU methods, leading them to conclude: "In any case, TS is very easy to implement and should thus be considered as a standard baseline". Taking all these factors into account, we have decided to focus on TS-based algorithms for addressing contextual bandit problems.

Despite its relative simplicity, effectiveness and convergence guarantees, TS comes with a computational burden which is to sample, at each iteration $t \in \mathbb{N}$, from an appropriate Bayesian posterior distribution $\hat{\mathrm{p}}_t$ defined from the previous observations. Indeed, these posteriors are usually intractable and approximate inference methods have to be used to obtain samples with distributions "close" to the posterior. The family of TS methods using approximate inference methods will be referred to as approximate inference TS in the sequel. Among the simplest approximate inference methods, Laplace approximation has been proposed for TS in [Chapelle and Li, 2011]. This method consists in approximating the posterior distribution $\hat{\mathrm{p}}_t$ by a Gaussian distribution with a carefully chosen mean and covariance matrix. More precisely, the mean is a mode of the target distribution which is typically found using an optimization algorithm, while the covariance matrix is taken to be the negative Hessian matrix of the log posterior at the considered mode. Despite this method is easy to implement, it may lead to poor posterior representations. Indeed, while Laplace method achieves minimal optimality in terms of regret [Faury et al., 2022], it doesn't dictate the posterior convergence rate. More precisely, in [Katsevich and Rigollet, 2023] it has been demonstrated that VI outperforms Laplace in terms of mean convergence by a factor of $1/n$. It is worth noting that the covariance rates remain the same for both methods. This discrepancy can lead to inadequate approximations, especially in high-dimensional settings, as highlighted in section I.4 of [Katsevich and Rigollet, 2023]. Another class of popular approximate inference methods are Markov Chain Monte Carlo (MCMC) methods, such as Metropolis or Langevin Monte Carlo (LMC) algorithms. In the bandit literature, LMC has been proposed to get approximate samples from TS posteriors for solving traditional bandit problem in [Mazumdar et al., 2020a] and for contextual bandit problems in [Xu et al., 2022, Huix et al., 2023]. Also, [Lu and Van Roy, 2017] have proposed to adapt Ensemble Methods to the bandit setting. Roughly, the idea here is to maintain and incrementally update an ensemble of statistically plausible models and to draw a uniform sample from this family at each iteration. Finally, [Zhang et al., 2020] suggests a TS method based on Neural Tangent Kernel. While this performs well on relative datasets, their method is much more expensive than previously mentioned approaches, as it requires training a neural network.

Finally, Variational Inference (VI) [Blei et al., 2017] is another class of approximate method that could be used to get samples from the posterior distribution. The core concept behind VI is to find a distribution $\tilde{\mathrm{q}}$, referred to as the variational posterior, to closely match the true posterior $\hat{\mathrm{p}}$ in terms of Kullback-Leibler divergence (KL) within a predefined family of distributions known as the variational family $\mathcal{G}$. In general, the variational family is chosen to make the optimization of the KL tractable and to be easy to sample from. In their work [Urteaga and Wiggins, 2018] propose the mean-field mixture of Gaussian variational family for TS. This family of distributions is quite extensive and provides an accurate approximation for a wide range of posterior distributions. However, in our perspective, it might not be the most suitable choice for TS. Firstly, the optimization algorithm at each time step can be computationally expensive. Secondly, the mean-field assumption assumes that the parameters are independent, a premise that holds true in the regime of large, overparameterized models. In our perspective, this regime may not align with the Bandit problem, which often operates in a setting where the number of data points tends towards infinity in comparison to the model size.

In this chapter, we develop an efficient VI method that makes use of the whole family of non-degenerate Gaussian distributions. This choice of VI family is supported by the Bernstein-Von Mises theorem [Van der Vaart, 2000]. This theorem, subject to specific regularity conditions, asserts that a properly scaled version of the posterior converges to a Gaussian as the sample size grows. When applied to contextual bandits, the data points progressively accumulate

over time, leading to the gradual concentration of the posterior around a dominant mode. As a consequence, the Gaussian approximation becomes increasingly suitable for representing the posterior in this particular setting. Furthermore, the covariance of the rescaled posterior distribution tends to converge towards the inverse Fisher information matrix, which may not necessarily be diagonal, thus justifying the need for a non-mean-field hypothesis.

The theoretical foundations of TS for linear contextual bandits were initially explored by [Agrawal and Goyal, 2013a]. In this paper, the authors establish a sub-linear cumulative regret bound $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ for Linear TS (Lin-TS). Compared to this study, our method achieves a similar regret bound in the linear framework. However, it should be noted that Lin-TS is a specialized algorithm that can be only used when the posterior is known and can be efficiently sampled from.

As mentioned previously, VI has been suggested for TS in [Urteaga and Wiggins, 2018]. This paper introduces a TS algorithm called VTS that utilizes a mixture of mean-field Gaussian distributions to approximate the sequence of posteriors. In comparison to this work, the setting and the variational family we consider are richer than [Urteaga and Wiggins, 2018]. A more detailed comparison is postponed in Section 8. Moreover, the methodology developed in [Urteaga and Wiggins, 2018] does not come with any convergence guarantees. An empirical and theoretical study of using LMC as approximate inference method for TS for contextual bandit problems was carried out in [Xu et al., 2022]. This paper establishes that the resulting algorithm, called LMC-TS, achieves a state-of-the-art sub-linear cumulative regret for linear contextual bandits. Compared to this method, our approach yields a similar sub-linear regret in the same setting.

2 Thompson sampling for contextual bandits

Contextual bandit: We now present in more details the contextual bandit framework. Let X be a contextual space and consider $\mathcal{A} : \mathsf{X} \to 2^{\mathsf{A}}$ a set-valued action map, where 2^{A} stands for the power set of the action space A. For simplicity, we assume here that $\sup_{x\in\mathsf{X}} \mathrm{Card}(\mathcal{A}(x)) < +\infty$. A (deterministic or random) function $\pi : \mathsf{X} \to \mathsf{A}$ is said to be a policy if for any $x \in \mathsf{X}$, $\pi(x) \in \mathcal{A}(x)$. Then, for a fixed horizon $T \in \mathbb{N}$, a contextual bandit process can be defined as follows: at each iteration $t \in [T]$ and given the past observations $\mathrm{D}_{t-1} = \{(x_s, a_s, r_s)\}_{s<t}$:

- The agent receives a contextual feature $x_t \in \mathsf{X}$;
- The agent chooses an action $a_t = \pi_t(x_t)$ where π_t is a policy sampled from $\mathbb{Q}_t(\cdot|\mathrm{D}_{t-1})$;
- Finally, the agent receives a reward r_t sampled from $\mathrm{R}(\cdot|x_t, a_t)$ given D_{t-1}. Here, R is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$, where $\mathsf{R} \subset \mathbb{R}$

For a fixed family of conditional distributions $\mathbb{Q}_{1:T} = \{\mathbb{Q}_t\}_{t\leq T}$, this process defines a random sequence of policies, $\pi_{1:T} = \{\pi_t\}_{t\leq T}$ with distribution still denoted by $\mathbb{Q}_{1:T}$ by abuse of notation. Let's defined the optimal expected reward for a contextual vector $x \in \mathsf{X}$ as

$$f_\star(x) = \max_{a\in\mathcal{A}(x)} f(x, a)\,, \tag{8.1}$$

and the expected reward given x and any action $a \in \mathcal{A}(x)$ as

$$f(x, a) = \int r\mathrm{R}(\mathrm{d}r|x, a)\,. \tag{8.2}$$

The main challenge of a contextual bandit problem is to find the distribution $\mathbb{Q}_{1:T}$ that minimizes the cumulative regret defined as

$$\mathrm{CREG}(\mathbb{Q}_{1:T}) = \sum_{s\leq T} \mathrm{Regret}_s^{\pi_s} \tag{8.3}$$

where the regret is defined as

$$\mathrm{Regret}_s^{\pi_s} = f_\star(x_s) - f(x_s, \pi_s(x_s))\,. \tag{8.4}$$

The main difficulty in the contextual bandit problem, comes from the fact that the reward distribution R is intractable and must be inferred to find the best policy to minimize the instantaneous regret $\pi \mapsto f_\star(x) - f(x, \pi(x))$

for a context $x \in \mathsf{X}$. However, the estimation of R may be in contradiction with the primary objective to minimize the cumulative regret (8.3), since potential non-effective arms has to be chosen to obtain a complete description of R. Therefore, bandit learning algorithms have to achieve an appropriate trade-off between exploitation of arms which have been confidently learned and exploration of misestimated arms.

Thompson sampling: To achieve such a trade-off, we consider the popular Thompson Sampling (TS) algorithm. Consider a parametric model $\{\mathrm{R}_\theta : \theta \in \mathbb{R}^d\}$ for the reward distribution, where for any θ, R_θ is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$ parameterized by $\theta \in \mathbb{R}^d$. We assume in this paper that R_θ admits a density with respect to some dominating measure λ_{ref}. An important example are generalized linear bandits [Filippi et al., 2010, Kveton et al., 2020a]. In particular, it assumes that $\{\mathrm{R}_\theta(\cdot|x,a) : \theta \in \Theta\}$ is an exponential family with respect to λ_{ref}, i.e., for $x \in \mathsf{X}$ and $a \in \mathsf{A}$,

$$\frac{\mathrm{dR}_\theta}{\mathrm{d}\lambda_{\mathrm{ref}}}(r|x,a) = \mathrm{h}(r)\exp(g(\theta,x,a)\mathrm{T}(r) - \mathrm{C}(\theta,x,a)), \tag{8.5}$$

for $h : \mathsf{R} \to \mathbb{R}_+$, natural parameter and log-partition function $g, \mathrm{C} : \mathbb{R}^d \times \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ and sufficient statistics $\mathrm{T} : \mathsf{R} \to \mathbb{R}$. The family is said to be in canonical form if $g(\theta,x,a) = \langle \phi(x,a), \theta \rangle$ for some feature map $\phi : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ and $\mathrm{C}(\theta,x,a) = \sigma(\langle \phi(x,a), \theta \rangle)$ for some link function σ. Linear contextual bandits [Chu et al., 2011, Abbasi-Yadkori et al., 2011b] fall into this model taking $\lambda_{\mathrm{ref}} = \mathrm{Leb}$, T equals to the identity function,

$$\mathrm{h}(r) = \exp\bigl(-\eta r^2/2\bigr) \quad \text{and} \quad g(\theta,x,a) = \eta \langle \phi(x,a), \theta \rangle, \tag{8.6}$$

for some $\eta > 0$.

As a result, $\mathrm{R}_\theta(\cdot|x,a)$ is simply the Gaussian distribution with mean $\langle \phi(x,a), \theta \rangle$ and variance $1/\eta$. Finally [Riquelme et al., 2018, Zhou et al., 2020, Xu et al., 2020] introduced an extension of linear contextual bandits, referred to as linear neural contextual bandits where g is a neural network with weights θ and taking as input a pair (x,a). With the introduced notations, the likelihood function associated to the observations D_t at step $t > 1$ is given by

$$\mathrm{L}_t(\theta) \propto \exp\left\{ \sum_{s=1}^{t-1} \ell(\theta|x_s,a_s,r_s) \right\}, \tag{8.7}$$

where the log-likelihood is given by $\ell(\theta|x_s,a_s,r_s) = \log(\mathrm{dR}_\theta/\mathrm{d}\lambda_{\mathrm{ref}})(r_s|x_s,a_s)$. Choosing a prior on θ with density p_0 with respect to Leb, and applying Bayes formula, the posterior distribution at round $t \in [T]$ is given by

$$\hat{\mathrm{p}}_t = \mathrm{L}_t(\theta)\mathrm{p}_0(\theta)/\mathfrak{Z}_t \tag{8.8}$$

where $\mathfrak{Z}_t = \int \mathrm{L}_t(\theta)\mathrm{p}_0(\theta)\mathrm{d}\theta$ denotes the normalizing constant and we used the convention that $\hat{\mathrm{p}}_1 = \mathrm{p}_0$. Moreover we define the potential function $U(\theta) \propto -\log \hat{\mathrm{p}}_t(\theta)$. Then, at each iteration $t \in [T]$, TS consists in sampling a sample θ_t from the posterior $\hat{\mathrm{p}}_t$ and from it, use as a policy, $\pi_t^{(\mathrm{TS})}(x)$ defined for any x by

$$\pi_t^{(\mathrm{TS})}(x) = a^{\theta_t}(x)\,, \quad a^\theta(x) = \operatorname*{argmax}_a \int r \mathrm{R}_\theta(\mathrm{d}r|x,a) \tag{8.9}$$

Since $\mathfrak{Z}_t$ is generally intractable, sampling from the posterior distribution is not in general an option.

Variational inference TS: To address this challenge, practitioners often employ approximate inference methods to generate samples from a distribution that is expected to be 'close' to the actual posterior distribution. In this context, we specifically concentrate on the application of VI. In this scenario, we consider a variational family $\mathcal{G}$ which is a set of probability densities with respect to the Lebesgue measure, from which it is typically easy to sample from. Then ideally, at each round $t \in [T]$, the posterior distribution $\hat{\mathrm{p}}_t$ is approximated by the variational posterior distribution $\tilde{\mathrm{q}}_t$ which is defined as:

$$\tilde{\mathrm{q}}_t = \operatorname*{argmin}_{\mathrm{p} \in \mathcal{G}} \mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}}_t)\,, \tag{8.10}$$

where KL is the Kullback-Leibler divergence.

However, we have to determine at each round a solution to the problem specified in (8.10). In this paper, we consider as variational family the set of non-degenerate Gaussian distribution $\mathcal{G} = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_+^*\}$ where $\mathcal{N}(\mu, \Sigma)$ is the Gaussian distribution with mean μ and covariance matrix Σ and $\mathcal{S}_+^*$ is the set of symmetric positive definite matrices. As explained in the introduction, this Gaussian variational family is particularly relevant in bandit framework according to Bernstein-Von Mises theorem.

Presentation of `VITS-I`: As we will see, this choice of variational family will allow to derive an efficient method for solving (8.10) using the Riemannian structure of $\mathcal{G}$. As noted in [Lambert et al., 2022b], $\mathcal{G}$ equipped with the Wasserstein distance of order 2 is a complete metric space as a closed subset of $\mathcal{P}_2(\mathbb{R}^d)$, the set of probability distributions with finite second moment. Recall that for two Gaussian distributions $p_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $p_1 = \mathcal{N}(\mu_1, \Sigma_1)$, their Wasserstein distance has a closed form:

$$W_2^2(p_0, p_1) = \|\mu_0 - \mu_1\|^2 + \mathrm{tr}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}) .$$

This Wasserstein distance on $\mathcal{G}$ allows to derive a Riemannian metric denoted $\mathfrak{g}$. The corresponding geodesic is given through the exponential map. More precisely, for a Gaussian distribution $p = \mathcal{N}(\mu_p, \Sigma_p)$, this map is defined as

$$\begin{aligned}\exp_p(\mu_v, \Sigma_v) =& (\mu_p + \mu_v + (\Sigma_v + \mathrm{I}_d)(\cdot - \mu_p))_{\#}p \\ =& \mathcal{N}(\mu_p + \mu_v,\ (\Sigma_v + \mathrm{I}_d)\ \Sigma_p\ (\Sigma_v + \mathrm{I}_d)) .\end{aligned} \tag{8.11}$$

With all these preliminaries, we can now present and motivate the algorithm developed in [Lambert et al., 2022b] to efficiently solve (8.10). This method can be formalized as a Riemannian gradient descent scheme on $\mathcal{G}$. Firstly, we define the loss function $\mathcal{F}_t : p \to \mathrm{KL}(p|\hat{\mathrm{p}}_t)$. Then, following [Lambert et al., 2022b], we derive the gradient operator of $\mathcal{F}_t$ on $\mathcal{G}$ equipped with $\mathfrak{g}$ as

$$\nabla_{\mathfrak{g}}\mathcal{F}_t(p) = (\int \nabla U_t(\theta)\mathrm{d}p(\theta),\ \int \nabla^2 U_t(\theta)\mathrm{d}p(\theta) - \Sigma_p^{-1}) \tag{8.12}$$

where Σ_p is the covariance matrix of p. From this expression, the corresponding Riemannian gradient descent [Bonnabel, 2013] using a step size $h_t > 0$ defines the sequence of iterates $\{\mathrm{q}_{t,k}\}_{k=1}^{K_t}$ recursively as:

$$\mathrm{q}_{t,k+1} = \exp_{\mathrm{q}_{t,k}}(-h_t \nabla_{\mathfrak{g}}\mathcal{F}_t(\mathrm{q}_{t,k})) .$$

At each time step t, this sequence is initialized with variational posterior at the previous step, ie, $\mathrm{q}_{t,0} = \mathrm{q}_{t-1,K_{t-1}}$. Please note that this warm initialization of the posterior results in an efficient algorithm and has been directly used in our main theoretical result (see (8.32)). Combining (8.11) and (8.12), this recursion amounts defining a sequence of means $\{\mu_{t,k}\}_{k=1}^{K_t}$ and covariance matrices $\{\Sigma_{t,k}\}_{k=1}^{K_t}$ by the recursions

$$\mu_{t,k+1} = \mu_{t,k} - h_t \int \nabla U_t(\theta)\mathrm{d}\mathrm{q}_{t,k}(\theta) , \quad \Sigma_{t,k+1} = A_{t,k}\Sigma_{t,k}A_{t,k} , \quad \mathrm{q}_{t,k+1} = \mathcal{N}(\mu_{t,k+1}, \Sigma_{t,k+1}) ,$$

where $A_{t,k} = \mathrm{I}_d - h_t(\int \nabla^2 U_t(\theta)\mathrm{d}\mathrm{q}_{t,k}(\theta) - \Sigma_{t,k}^{-1})$.

The main computational challenge in this recursion stems is that the integrals involved are typically intractable. To overcome this issue, we employ a Monte Carlo procedure to approximate these integrals. Subsequently, we consider a sequence of mean values denoted as $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ and covariance matrices $\{\tilde{\Sigma}_{t,k}\}_{k=1}^{K_1}$ such that:

$$\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k}) , \quad \tilde{\Sigma}_{t,k+1} = \tilde{A}_{t,k}\tilde{\Sigma}_{t,k}\ \tilde{A}_{t,k} ,$$

where $\tilde{A}_{t,k} = \mathrm{I}_d - h_t(\nabla^2 U_t(\tilde{\theta}_{t,k}) - \tilde{\Sigma}_{t,k}^{-1})$ and $\tilde{\theta}_{t,k} \sim \mathcal{N}(\tilde{\mu}_{t,k}, \tilde{\Sigma}_{t,k})$. Consequently, following [Lambert et al., 2022b] we obtain an algorithm capable of addressing the problem defined in (8.10). However, this algorithm exhibits computational inefficiency, particularly in high-dimensional scenarios. This inefficiency arises from the necessity to sample from a Gaussian distribution with a non-diagonal covariance matrix during each updating step $k \in [K_t]$. As a result, it becomes impractical for use in a contextual bandit problem, where, at each time step t, we must solve

the problem described in (8.10). This chapter introduces an improved version of the earlier algorithm, designed to efficiently address the problem presented in (8.10). To achieve this, we begin by examining a sequence of matrices denoted as $B_{t,k}$, defined by the following

$$B_{t,k+1} = \left\{\mathrm{I}_d - h_t \nabla^2 U_t(\tilde{\theta}_{t,k})\right\} B_{t,k} + h_t (B_{t,k}^{-1})^\top . \tag{8.13}$$

It is important to note that $B_{t,k}$ is a square-root matrix of the covariance of the variational distribution $\tilde{\Sigma}_{t,k}$, ie, $B_{t,k}B_{t,k}^\top = \tilde{\Sigma}_{t,k}$. Then we can sample efficiently from the variational distribution using $B_{t,k}$ with $\tilde{\theta}_{t,k} = \tilde{\mu}_{t,k} + B_{t,k}\epsilon_{t,k}$, $\epsilon \sim \mathcal{N}(0, \mathrm{I}_d)$. As a result, note that our method does not require any Cholesky decomposition, which has a complexity of $\mathcal{O}(d^3)$, contrary to the algorithm derived in [Lambert et al., 2022a] and also in LinTS. The updating strategies for the new sequence of $\tilde{\mu}_{t,k}$ and $B_{t,k}$ are given by

$$\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k}) , \quad B_{t,k+1} = \left\{\mathrm{I}_d - h_t \nabla^2 U_t(\tilde{\theta}_{t,k})\right\} B_{t,k} + h_t (B_{t,k}^{-1})^\top , \quad \tilde{\theta}_{t,k} \sim \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k}) .$$

From this methodology, we can now complete the description of our first algorithm, referred to as `VITS-I`. At each step t, we consider the variational distribution $\tilde{\mathrm{q}}_t = \tilde{\mathrm{q}}_{t,K_t} = \mathcal{N}(\tilde{\mu}_{t,K_t}, B_{t,k}^\top B_{t,k})$ which approximates the solution of (8.10). Then, at round $t+1$, `VITS-I` consists in sampling $\tilde{\theta}_{t+1}$ according to $\tilde{\mathrm{q}}_t$ and choosing

$$\pi_{t+1}^{\texttt{VITS-I}}(x) = \underset{a \in \mathcal{A}(x)}{\operatorname{argmax}}\, a^{\tilde{\theta}_{t+1}}(x) . \tag{8.14}$$

As in TS, the likelihood function and the posterior distribution $\hat{\mathrm{p}}_{t+1}$ are updated following equations (8.7) and (8.8) using the new observed reward r_{t+1} distributed according to $\mathrm{R}(\cdot|x_{t+1}, a_{t+1})$ with $a_{t+1} = \pi_{t+1}^{\texttt{VITS-I}}(x)$. The round $t+1$ is then concluded by solving $\tilde{\mathrm{q}}_{t+1} = \tilde{\mathrm{q}}_{t+1,K_{t+1}}$. The pseudo-code associated with this algorithm is given in Algorithm 3 and Algorithm 4.

Algorithm 3 `VITS` algorithm

Parameter:
variance parameters λ and η, time horizon T
Initialize:
$B_{1,1} = \mathrm{I}_d/\sqrt{\lambda\eta}$, $\tilde{W}_{1,1} = \mathrm{I}_d/(\eta\lambda)$, $\tilde{\mu}_{1,1} \sim \mathcal{N}(0, \tilde{W}_{1,1})$
for $t = 1, \ldots, T$ **do**
receive $x_t \in \mathsf{X}$
sample $\tilde{\theta}_t$ from $\tilde{\mathrm{q}}_{t,K_t} = \mathcal{N}(\tilde{\mu}_{t,K_t}, B_{t,K_t}^\top B_{t,k})$
choose $a_t = \pi^{(\mathrm{VITS})}(x_t)$ presented in (8.14)
receive $r_t \sim \mathrm{R}(\cdot|x_t, a_t)$
update $\tilde{\mathrm{q}}_{t+1,K_{t+1}}$ using Alg. 4 or 5.
end for

Algorithm 4 `VITS-I`

Parameter:
step-size h_t, number of iterations K_t
Initialize:
$\tilde{\mu}_{t,1} \leftarrow \tilde{\mu}_{t-1,K_{t-1}}$, $B_{t,1} \leftarrow B_{t-1,K_{t-1}}$
for $k = 1, \ldots, K_t$ **do**
draw $\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k} = \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$
$\tilde{\mu}_{t,k+1} \leftarrow \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k})$
$B_{t,k+1} \leftarrow \left\{\mathrm{I}_d - h_t \nabla^2 (U_t(\tilde{\theta}_{t,k}))\right\} B_{t,k} + h_t (B_{t,k}^{-1})^\top$
end for

Presentation of VITS-II: In high dimension, the computational cost of the recursion of mean values and covariance matrices may be prohibitive since at each iteration $k \in [K_t]$, it requires inverting the matrix $B_{t,k}$. To tackle this computational issue, we propose a new version of VITS. More precisely, the inverse of the square root covariance matrix $B_{t,k}^{-1}$ can be approximated using a first order Taylor expansion in h_t; see Section 7 for more details. We denote by $C_{t,k}$ the approximation of $B_{t,k}^{-1}$, and we obtain the following recursions for the sequence of $\{C_{t,k}\}_{k\leq K_t}$ and $\{B_{t,k}\}_{k\leq K_t}$:

$$C_{t,k+1} = C_{t,k}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - \nabla^2 U_t(\tilde{\theta}_{t,k}))\}\,, \quad B_{t,k+1} = (\mathrm{I}_d - h_t\nabla^2 U_t(\tilde{\theta}_{t,k}))B_{t,k} + h_t C_{t,k}^\top\,.$$

This trick reduces the complexity from $\mathcal{O}(d^3)$ to $\mathcal{O}(d^2)$ for the computation of the inverse. This version of VITS is referred to as VITS − II and is given in Algorithm 3 and 5.

Presentation of VITS − II Hessian-free: The most computationally intensive step in VITS − II remains the computation of the Hessian of U_t. In scenarios with a large number of data points and high dimensions, this step can become highly demanding. To avoid computing the Hessian of U_t, we suggest to use the following property of Gaussian distribution which is the result of a simple integration by part:

$$\int \nabla^2 U_t \mathrm{d}\mathcal{N}(\mu,\Sigma) = \int \Sigma^{-1}(\mathrm{I}_d - \mu)\nabla U_t^\top \,\mathrm{d}\mathcal{N}(\mu,\Sigma)\,. \tag{8.15}$$

After approximating this right side integral using Monte Carlo, we derive a new sequence of square-root covariance matrix $\{B_{t,k}\}_{k\leq K_t}$ and inverse square-root covariance matrix $\{C_{t,k}\}_{k\leq K_t}$, defined recursively by:

$$C_{t,k+1} = C_{t,k}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - A_{t,k})\}\,, \quad B_{t,k+1} = (\mathrm{I}_d - h_t A_{t,k})B_{t,k} + h_t C_{t,k}^\top\,,$$

where $A_{t,k} = C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t^\top(\tilde{\theta}_{t,k})$ and $\tilde{\theta}_{t,k} \sim \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$. This last version of VITS is referred to as VITS − II Hessian-free and its pseudo-code is given in Algorithm 3 and Algorithm 5.

Algorithm 5 VITS − II / VITS − II Hessian-free

Parameter:
step-size h_t, number of iterations K_t
Initialize:
$\tilde{\mu}_{t,1} \leftarrow \tilde{\mu}_{t-1,K_{t-1}}$, $B_{t,1} \leftarrow B_{t-1,K_{t-1}}$
for $k = 1,\dots,K_t$ **do**
draw $\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k} = \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$
$\tilde{\mu}_{t,k+1} \leftarrow \tilde{\mu}_{t,k} - h_t\nabla U_t(\tilde{\theta}_{t,k})$
$A_{t,k} = \begin{cases} \nabla^2(U_t(\tilde{\theta}_{t,k})) & \text{(Hessian)} \\ C_{t,k}^2(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})(\nabla U_t(\tilde{\theta}_{t,k}))^\top & \text{(Hessian free)} \end{cases}$.
$B_{t,k+1} \leftarrow \{\mathrm{I}_d - h_t A_{t,k}\}B_{t,k} + h_t C_{t,k}^\top$
$C_{t,k+1} \leftarrow C_{t,k}(\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - A_{t,k}))$
end for

The computational complexity of all methods has been experimentally studied in a simple case, as discussed in Section 10.3.

3 Main results

3.1 Linear Bandit

In this section, we are interested in convergence guarantees for VITS − I applied to the linear contextual bandit framework. This framework consists in assuming that R_θ has form (8.5) with $\lambda_{\mathrm{ref}} = \mathrm{Leb}$, T is the identity function and h and g are specified by (8.6):

$$\frac{\mathrm{dR}_\theta}{\mathrm{dLeb}}(r|x,a) \propto \exp\Big[\eta(r - \langle\phi(x,a),\theta\rangle)^2/2\Big] \,. \tag{8.16}$$

Assumption on the reward kernel R is the following:

Assumption 6. *(Sub-Gaussian Reward Distribution) There exists $R > 1$ such that for any $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$, $\rho > 0$,*

$$\log \int \exp\{\rho(r - f(x,a))\}\mathrm{R}(\mathrm{d}r|x,a) \leq R\rho^2 \,,$$

where f is defined in 8.1

We could only assume that $R > 0$ in Assumption 6 since if a distribution is R-sub-Gaussian, it is also R'-sub-Gaussian for any $R' \geq R$, however, we choose to set $R \geq 1$ to ease the presentation of our main results. We also assume that the model is well-specified.

Assumption 7. *There exists $\theta^\star$ such that $\mathrm{R} = \mathrm{R}_{\theta^\star}$ and satisfying $\|\theta^\star\|_2 \leq 1$. Feature map ϕ satisfies the boundedness condition.*

Assumption 8. *For any contextual vector $x \in \mathbb{R}^d$ and action $a \in \mathcal{A}(x)$, it holds that $\|\phi(x,a)\|_2 \leq 1$.*

Uniform boundedness condition on the feature map is relatively common for obtaining regret bounds for linear bandit problems [Agrawal and Goyal, 2013a, Xu et al., 2022, Kveton et al., 2020a, Abbasi-Yadkori et al., 2011b]. Note that Assumption 8 is equivalent to $\sup_{x\in\mathsf{X}, a\in\mathcal{A}(a)}\|\phi(x,a)\|_2 \leq M_\phi$ for some arbitrary but fixed constant $M_\phi > 0$, changing the feature map ϕ by ϕ/M_ϕ. Finally, we specify the prior distribution.

Assumption 9. *The prior distribution is assumed to be zero-mean Gaussian distribution with variance $1/(\lambda\eta)$, where η also appears in the definition R_θ in* (8.16)*,*

While our theoretical results can readily be extended to accommodate a non-zero mean Gaussian prior, for the sake of simplicity, we have chosen to center the prior. Under Assumption 9, combining (8.8) and (8.16), the negative log posterior $-\log \hat{\mathrm{p}}_t$ denoted by U_t is given by

$$\begin{aligned} U_t(\theta) &= \frac{\eta}{2}\left(\sum_{s=1}^{t-1}(\phi(a_s,x_s)^\top\theta - r_s)^2 + \lambda\|\theta\|_2^2\right) \\ &= \frac{\eta}{2}(\theta^\top V_t\theta - 2\theta^\top b_t + \sum_{s=1}^{t-1} r_s^2) \,, \end{aligned} \tag{8.17}$$

where $V_t = \lambda \mathrm{I}_d + \sum_{s=1}^{t-1}\phi_s\phi_s^\top \in \mathbb{R}^{d\times d}$ and $b_t = \sum_{s=1}^{t-1} r_s\phi_s \in \mathbb{R}^{d\times 1}$. Therefore, it follows that the gradient of U_t is given by $\nabla U_t(\theta) = \eta(V_t\theta - b_t)$ and its hessian matrix is equal to $\nabla^2 U_t(\theta) = \eta V_t$. Consequently, we recover the well-known fact that the posterior is a Gaussian distribution with mean $\hat{\mu}_t = V_t^{-1}b_t$ and covariance matrix $\hat{\Sigma}_t = (\eta V_t)^{-1}$. Denote by $\tilde{\mathbb{Q}}_{1:T}$ the distribution on the sequence of policies induced by the sequence of variational posterior $\{\tilde{\mathrm{q}}_t = \mathbb{N}(\tilde{\mu}_{t,K_t}, B_{t,K_t}^\top B_{t,K_t})\}_{t\in[T]}$ obtained with `VITS − I`

We now state our main result on the cumulative regret associated to `VITS − I` for linear contextual bandit, where a the proof is provided in Section 6.

Theorem 104. *Assume Assumptions 6 to 9 hold. For the choice of hyperparameters $\{K_t, h_t\}_{t\in[T]}$ and η specified in Section 6.2, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the cumulative regret is bounded by*

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) \leq \frac{CR^2 d\sqrt{dT}\log(3T^3)}{\lambda^2}\log\Big(\frac{(1+T/\lambda d)}{\delta}\Big)$$

where $C \geq 0$ is a constant independent of the problem. Our main result shows that the distribution of the sequence of policies generated by VITS − I or results in a cumulative regret of order $\tilde{\mathcal{O}}(d\sqrt{dT})$. It is in the same order as the state-of-the-art cumulative regret obtained in [Agrawal and Goyal, 2013a] for LinTS. The number of optimization steps K_t we found are of order $\kappa_t^2 \log(dT\log(T))$ where $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t)$. Following [Hamidi and Bayati, 2020, Wu et al., 2020], if the diverse context assumption holds, the condition number is $\kappa_t = \mathcal{O}(1)$. Therefore, under this previous assumption, VITS − I and require a number of optimization steps that scale as $\log(dT\log(T))$. Finally, [Xu et al., 2022] derived similar bounds for TS using LMC for linear contextual bandit problems. Although our proof is based on the linear case, it could be extended to more general cases insofar as our updates remain Gaussian by definition of the variational family. This allows the use of Gaussian (anti) concentration bound in the theoretical analysis. This is in contrast to other approximation methods, which do not possess this advantage.

Comparison table. In this paragraph we have added a comparison table between Linear TS (LinTS), Linear UCB (LinUCB), Feel-Good TS [Huix et al., 2023, Zhang, 2022a], VITS − I, VITS − II (VITS-I/II), VITS − II Hessian-free (VITS-II HF), Langevin Monte Carlo TS (LMCTS) and Variational TS (VTS). The column "Regret" corresponds to the theoretical regret bound obtained by the algorithm, more precisely the symbol $(++)$ corresponds to a regret $O(\sqrt{dT})$, $(+)$ to $O(d^{3/2}\sqrt{T})$ and $(-)$ to no existing regret bound. "Complexity" is the computational complexity. "Linear" is set to Yes when the algorithm is designed only for the Linear Bandit setting and No for general setting including Linear. The "Conditioning" column describes the algorithm's robustness against the conditioning of the problem.

	Regret	Complexity	Linear	Conditioning
LinTS	+	++	Yes	++
LinUCB	++	++	Yes	++
FG-TS	++		No	
VITS-I/II	+	+	No	+
VITS-II HF	—	+ +	No	+
LMC-TS	+	++	No	—
VTS	—	—	No	

4 Numerical experiments

4.1 Linear and quadratic bandit

Our initial investigation focused on a toy setting where contextual vectors are sampled from a Gaussian distribution. However, in this specific setting, the contextual vectors exhibit high diversity, resulting in a posterior covariance matrix with a condition number of $O(1)$. This condition makes the optimization problem overly simplistic, as a result, all approximation methods seem to perform identically in this simple well-conditioned problem. So we introduce a novel setting in which the diversity of arms is controlled by a parameter, denoted as ζ. Firstly, we consider a fixed pool of arms denoted as $P = [\tilde{x}_1, \ldots, \tilde{x}_n]$ with $n = 50$, where each arm $\tilde{x}_i$ follows a normal distribution $\mathcal{N}(0_d, \mathrm{I}_d)$. This fixed pool is relevant in real-world scenarios, such as in a Recommender system, where this pool corresponds to the concept of a meta-user. Then, at each step $t \in [T]$, for every arm, we randomly sample a vector $\tilde{x}_i$ from the pool P, and the contextual vector associated with this arm is defined as $x = \tilde{x}_i + \zeta\epsilon$, where $\epsilon \sim \mathcal{N}(0_d, \mathrm{I}_d)$. When ζ has a high value, the corresponding user is far from the meta-user. Consequently, the diversity among arms is high, resulting in a well-conditioned problem. However, in cases where ζ is low, the problem is ill-conditioned and the optimization becomes challenging.

We consider the linear bandit and the quadratic bandit problems. In both settings, the bandit environment is simulated using a random vector $\theta^\star$ sampled from a normal distribution $\mathcal{N}(0_d, \sigma^\star \mathrm{I}_d)$. We opted for $\sigma^\star = 1/d$ to ensure that the variance of the scalar product $x^\top \theta^\star$ remains independent of the dimension d. The parameter dimension d is set to 20 and we consider a number of arms $K = 50$. In the linear bandit setting, the reward associated with the contextual vector x, is $r = x^\top \theta^\star + \alpha\epsilon$ where $\epsilon \sim \mathcal{N}(0_d, \mathrm{I}_d)$. However, to maintain problem complexity

independent of ζ, we have set the signal-to-noise ratio to a fixed value of 1, meaning $\mathbb{E}[(x^\top\theta^\star)^2]/\mathbb{E}[(\alpha\epsilon)^2] = 1$. This implies that $\sqrt{1+\zeta^2} = \alpha$. See Subsection 9.2 for more details about the setting. In these experiments, we have chosen to compare VITS − II, VITS − II Hessian-free, Linear TS (LinTS), and LMC-TS, with 10 and 50 iterations of Langevin diffusion at each step. For VITS based algorithm, we have only used 10 updating steps. We have omitted the performance of VITS − I since it experimentally performs identically to VITS − II. For the algorithm VITS − II Hessian-free, we approximate the integral presented in (8.15) using 20 Monte Carlo samples. This choice is made due to the observed instability caused by the Monte Carlo error when considering high values of η. However, in our setting, even with 20 Monte Carlo samples, VITS − II Hessian-free remains a faster method compared to VITS − II. We also attempted to assess the performance of VTS, but, in the ill-conditioned setting, it exhibited a linear and notably high cumulative regret. Consequently, we have opted to exclude it from the figure for the sake of clarity and visibility. The mean and standard error are reported for all experiments over 50 runs. The hyperparameter is provided in Subsection 9.1.

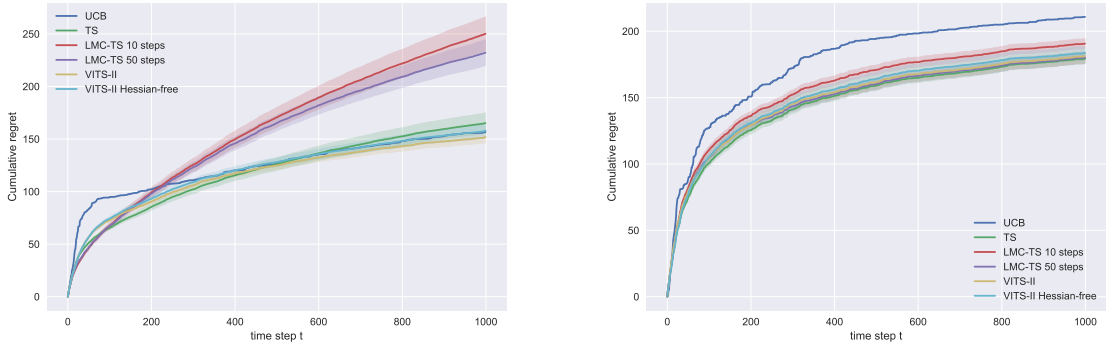


Fig. 8.1 Linear bandits, $\zeta = 0.1$ (left), $\zeta = 1$ (right).

Figure 8.1 illustrates the cumulative regret with respect to the time step t for a well-conditioned problem ($\zeta = 1$) and a ill-conditioned problem ($\zeta = 0.1$). Firstly, for $\zeta = 1$, it appears that all methods exhibit similar performance, with the exception of LMC-TS with 10 steps, which slightly underperforms. However, for $\zeta = 0.1$, the optimization problem becomes harder and LMC-TS underperforms even with 50 Langevin steps. This behaviour was expected in our setting, because LMC requires a lot of iterations to converge to the posterior compared to VI. A more complete explanation of this phenomenon can be found in Subsection 10.1. Finally, we can conclude that VITS − II performs similarly to LinTS and that its Hessian-free version slightly underperforms but is computationally more efficient.

For Quadratic bandit in Fig 8.2, the reward is $r = (x^\top\theta^\star)^2 + \alpha\epsilon$. This setting is similar to the Linear setting, but we ensure the condition $\mathbb{E}[(x^\top\theta^\star)^4]/\mathbb{E}[(\alpha\epsilon)^2] = 1$ to still get the signal-to-noise ratio equals to 1. This implies a slight different condition $\alpha = (\zeta^2+1)\sqrt{3+6/d}$, see Subsection 9.2. Moreover, a simple MLP with two hidden layers of 20 neurons is used for LMC, VITS − II, and its Hessian-free version as neural network architecture. Performance in Fig 8.2 are similar to linear bandits where VITS − II slightly performs better than its Hessian-free version but outperforms both LMC and LinTS algorithms as LinTS is not adapted for this setting. The gap between LMC and our algorithm is smaller in the well-conditioned setting than in the ill-conditioned, which was also expected. Finally, additional experience on non-contextual bandits can also be found in Appendix 10.2.

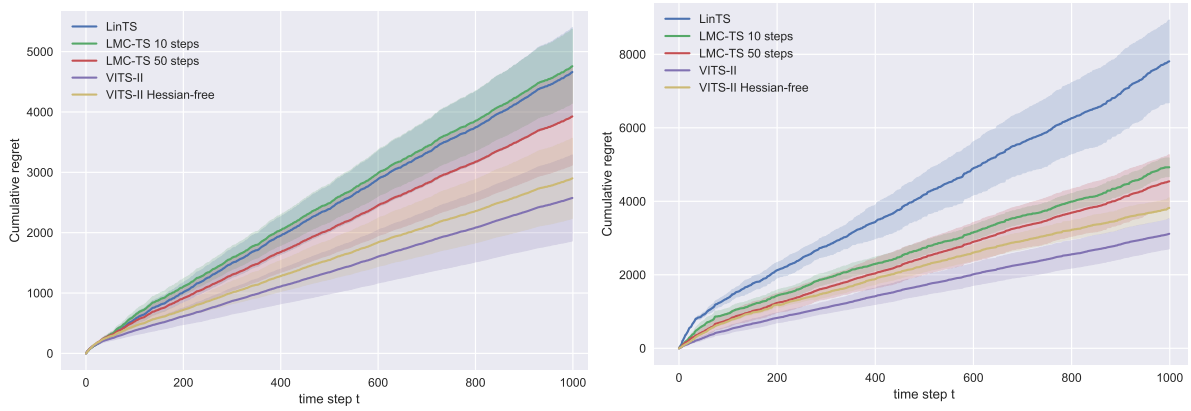


Fig. 8.2 Quadratic bandit, $\zeta = 0.1$ (left), $\zeta = 1$ (right).

4.2 MovieLens Dataset

In this section, we evaluate `VITS` on the MovieLens dataset, consisting of one million ratings by 6040 users for 3952 movies. We adopt the setup proposed in [Aouali et al., 2022], involving a low-rank factorization of the rating matrix to yield 5-dimensional representations for users ($x_j \in \mathbb{R}^5$) and movies ($\theta_i \in \mathbb{R}^5$). Movies are treated as potential actions, and context x_t is uniformly sampled from the pool of user vectors. We consider logistic rewards, sampled from $\text{Ber}(\mu(x_j^\top \theta_i))$, where μ is the sigmoid function. We conduct 50 simulations, each involving 100 randomly selected movies.Our prior distribution employs a Gaussian distribution with mean μ_0 and covariance $\Sigma_0 = \text{diag}(\sigma_0)$. Here, μ_0 and σ_0 represent the mean and variance of movie vectors across all dimensions. This setting deviates somewhat from our theoretical framework, where we consider a unified posterior distribution for all arms using a feature map function ϕ representing context-action pairs. In the MovieLens context, each arm possesses an individual posterior distribution. These two settings closely align when the feature map is the vector concatenation function. In practice, we can apply `VITS` or LMC at each arm to obtain posterior samples. In this experiment, we compare LinTS against LMC-TS, `VITS - II`, and the `VITS - II Hessian-free` variant. LMC-TS uses 10 Langevin updating steps. It's crucial to note that for each time step t and each arm a, LMC-TS requires running Langevin diffusion to obtain a new parameter with low correlation to the previous one. This leads to a high computational complexity for LMC-TS. In contrast, `VITS` for each arm only involves sampling from a low-dimensional Gaussian distribution and updating the variational posterior corresponding to the chosen arm. This approach offers significant computational efficiency.

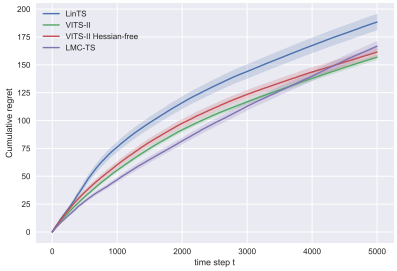


Fig. 8.3 Cumulative regret for MovieLens dataset.

Figure 8.3 reveals that LinTS is ill-suited for this particular task, as it assumes rewards to be linear while the approximated algorithms outperform LinTS, as they specifically target the logistic posterior. Remarkably, `VITS` appears to slightly outperform LMC-TS, despite its computational efficiency advantages.

5 Conclusion and perspectives

This paper presents a novel TS algorithms called VITS − I, that use VI as an approximation method. This algorithms provide robust theoretical guarantees, in particular a cumulative regret bound of $\tilde{\mathcal{O}}(d\sqrt{dT})$ in the linear setting. One limitation of our theoretical analysis is that the regret bound derived is limited to the linear setting while the interest of our algorithm relies on nonlinear tasks. Additionally, we introduce two other algorithms, named VITS − II and VITS − II Hessian-free, which offer enhanced computational efficiency. The VITS − II algorithm uses a Taylor expansion to remove the computation of inverse matrix and the VITS − II Hessian-free algorithm removes the computations of Hessian, resulting in faster execution. Finally, all algorithms have been extensively evaluated in both simulated and real problems.

6 Proof of the regret bound

6.1 Proof of Theorem 104

While [Lambert et al., 2022b] establishes quantitative bounds on the bias introduced by Algorithm 4 for the VI of the posterior. Combining this result with the one derived in [Agrawal and Goyal, 2013a] for TS leads to sub-optimal regret bounds. It is similar to LMC-TS [Xu et al., 2022] which had to make a clever adaptation of [Agrawal and Goyal, 2013a]. Similar to this work, we need here to revise the proof of [Agrawal and Goyal, 2013a] to VITS. We give in this section the main steps of our proofs. Each step is based on Lemmas which are stated and proved in the next Subsections. First, we define the filtration $(\mathcal{F}_t)_{t\leq T-1}$ such that for any $t\in[T]$, $\mathcal{F}_{t-1}$ is the σ-field generated by $\mathcal{H}_{t-1}$ and x_t where $\mathcal{H}_{t-1}=\{(x_s,a_s,r_s)\}_{s\leq t-1}$ is the observations up to $t-1$ and x_t is the contextual vector at step t. For some feature map $\phi:\mathsf{X}\times\mathsf{A}\to\mathbb{R}$ and for any $t\in[T]$, we denote by

$$\overset{*}{\phi}_t=\phi(x_t,a_t^\star)\ \text{ ,and }\quad \phi_t=\phi(x_t,a_t)\,,$$

the features vector of the best arm $a_t^\star$ and the features vector of the arm a_t chosen by VITS at time t respectively. the difference between the best expected reward and the expected reward obtained by VITS is denoted by

$$\Delta_t=\overset{*}{\phi}_t^\top\theta^\star-\phi_t^\top\theta^\star\,.$$

At each round $t\in[T]$, we consider the set of saturated arms $\mathcal{S}_t$ and unsaturated arms $\mathcal{U}_t$ defined by

$$\mathcal{S}_t=\bigcap_{a\in\mathcal{A}(x_t)}\{\Delta_t(a)>g(t)\,\|\phi(x_t,a)\|_{V_t^{-1}}\}\,, \tag{8.18}$$

and $\mathcal{U}_t=\mathcal{A}(x_t)\setminus\mathcal{S}_t$ where V_t^{-1} is defined in (8.17) and

$$g(t)=CR^2d\sqrt{\log(t)\log(T)}/\lambda^{3/2}\,,$$

for some constant $C\geq 0$ independent of d, t and T. In addition, consider the events $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ such that

$$\bigcap_{a\in\mathcal{A}(x_t)}\{|\phi(x_t,a)^\top\hat{\mu}_t-\phi(x_t,a)^\top\theta^\star|\leq\ g_1(t)\ \ \|\phi(x_t,a)\|_{V_t^{-1}}\}\subset\mathrm{E}_t^{\text{true}}$$

$$\mathrm{E}_t^{\text{var}}=\bigcap_{a\subset\mathcal{A}(x_t)}\{|\phi(x_t,a)^\top\tilde{\theta}_t-\phi(x_t,a)^\top\hat{\mu}_t|\leq\ g_2(t)\ \ \|\phi(x_t,a)\|_{V_t^{-1}}\}\,,$$

where $\hat{\mu}_t$ is given by $\hat{\mu}_t=V^{-1}b_t$ is the mean of the posterior distribution at time t and b_t is given in (8.17). The specific definitions of $\mathrm{E}_t^{\text{true}}$, g, g_1 and g_2 are given in Subsection 6.3. Nevertheless, by definition, it holds that $g_1(t)+g_2(t)\leq g(t)$.

1. For ease of notation, the conditional expectation $\mathbb{E}_{\pi_{1:T}\sim\mathbb{Q}_{1:T}}[\cdot]$ and probabilities $\mathbb{P}_{\pi_{1:T}\sim\mathbb{Q}_{1:T}}(\cdot)$ with respect to the σ-field $\mathcal{F}_{t-1}$ are denoted by $\mathbb{E}_t[\cdot]$ and $\mathbb{P}_t(\cdot)$ respectively. Therefore, with these notations, we have by definition of the cumulative regret:

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T})=\sum_{t=1}^{T}\Delta_t\,.$$

We now bound for any $t \in [T]$, with high probability, $\Delta_t(a_t)$. To this end, in the next step of the proof, we show that the stochastic process $(X_t)_{t\in[T]}$ defined below is a $(\mathcal{F}_t)_{t\in[T]}$ super-martingale.

$$X_t = \sum_{s=1}^{t} Y_s$$
$$\text{with} \quad Y_s = \Delta_s - cg(s)\frac{\|\phi_s\|_{V_s^{-1}}}{p} - \frac{2}{s^2}\,,$$

where $p \in (0,1)$ and c is a sufficiently large real number, independent of d, T and s.

2. **Showing that $(X_t)_{t\in[T]}$ is a super-martingale.** We consider the following decomposition

$$\begin{aligned}\mathbb{E}_t[\Delta_t(a_t)] &= \mathbb{E}_t[\Delta_t(a_t)\mathbb{1}_{\mathrm{E}_t^{\text{true}}}] + \mathbb{E}_t[\Delta_t(a_t)|\bar{\mathrm{E}}_t^{\text{true}}]\mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{true}}) \\ &\leq \mathbb{E}_t[\Delta_t(a_t)\mathbb{1}_{\mathrm{E}_t^{\text{true}}}] + \mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{true}})\,, \end{aligned} \tag{8.19}$$

where we used for the last inequality that $\|\theta^\star\|_2 \leq 1$ and Assumption 8. Then, since $\mathrm{E}_t^{\text{true}} \in \mathcal{F}_{t-1}$, we have,

$$\begin{aligned}\mathbb{E}_t[\Delta_t(a_t)\mathbb{1}_{\mathrm{E}_t^{\text{true}}}] &= \mathbb{1}_{\mathrm{E}_t^{\text{true}}}\mathbb{E}_t[\Delta_t(a_t)|\mathrm{E}_t^{\text{var}}]\mathbb{P}_t(\mathrm{E}_t^{\text{var}}) + \mathbb{1}_{\mathrm{E}_t^{\text{true}}}\mathbb{E}_t[\Delta_t(a_t)|\bar{\mathrm{E}}_t^{\text{var}}]\mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{var}}) \\ &\leq \mathbb{1}_{\mathrm{E}_t^{\text{true}}}[\mathbb{E}_t[\Delta_t(a_t)|\mathrm{E}_t^{\text{var}}] + \mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{var}})] \end{aligned} \tag{8.20}$$

where in the last line we have used that $\Delta_t(a_t) \leq 1$ again. Denote by $\bar{a}_t = \mathrm{argmin}_{a\in\mathcal{U}_t} \|\phi(x_t,a)\|_{V_t^{-1}}$ and $\overline{\phi}_t = \phi(x_t,\bar{a}_t)$. Then, given $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ we have

$$\begin{aligned}\Delta_t(a_t) &= \overset{*}{\phi}_t^\top\theta^\star - \phi_t^\top\theta^\star \\ &= \overset{*}{\phi}_t^\top\theta^\star - \overline{\phi}_t^\top\theta^\star + \overline{\phi}_t^\top\theta^\star - \phi_t^\top\theta^\star \\ &\overset{(a)}{\leq} g(t)\|\overline{\phi}_t\|_{V_t^{-1}} + \overline{\phi}_t^\top\theta^\star - \phi_t^\top\theta^\star \\ &\overset{(b)}{\leq} g(t)\|\overline{\phi}_t\|_{V_t^{-1}} + (\overline{\phi}_t^\top\tilde{\theta}_t + g(t)\|\overline{\phi}_t\|_{V_t^{-1}}) - (\phi_t^\top\tilde{\theta}_t - g(t)\|\phi_t\|_{V_t^{-1}}) \\ &\overset{(c)}{\leq} (2\|\overline{\phi}_t\|_{V_t^{-1}} + \|\phi_t\|_{V_t^{-1}})g(t) \end{aligned} \tag{8.21}$$

where inequality (a) is due to $\bar{a}_t \in \mathcal{U}_t$, and therefore $\Delta_t(\bar{a}_t) \leq g(t)\|\overline{\phi}_t\|_{V_t^{-1}}$, inequality (b) uses that given $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$, for any $\phi \in \mathbb{R}^d$, $|\phi^\top\tilde{\theta}_t - \phi^\top\theta^\star| \leq g(t)\|\phi\|_{V_t^{-1}}$ since by definition $g_1(t) + g_2(t) \leq g(t)$; finally, the arm a_t maximizes the quantity $\phi(x_t,a_t)^\top\tilde{\theta}_t$, $\overline{\phi}_t^\top\tilde{\theta}_t - \phi_t^\top\tilde{\theta}_t$ is obviously negative, which implies inequality (c).

Moreover, given $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$,

$$\begin{aligned}\mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}] &= \mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}|a_t \in \mathcal{U}_t]\mathbb{P}_t(a_t \in \mathcal{U}_t) + \mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}|a_t \in \mathcal{S}_t]\mathbb{P}_t(a_t \in \mathcal{S}_t) \\ &\overset{(a)}{\geq} \|\overline{\phi}_t\|_{V_t^{-1}}\mathbb{P}_t(a_t \in \mathcal{U}_t) \\ &\overset{(b)}{\geq} (p - 1/t^2)\|\overline{\phi}_t\|_{V_t^{-1}} \end{aligned}$$

where (a) is due to the definition of $\overline{\phi}_t$, i.e. for any $a \in \mathcal{U}_t$, $\|\overline{\phi}_t\|_{V_t^{-1}} \leq \|\phi(x_t,a)\|_{V_t^{-1}}$, and (b) uses Lemma 108 with $p \in (0,1)$. Here is one of the main differences with the proof conducted by [Agrawal and Goyal, 2013a]. Indeed, to obtain such a bound, we need to carefully dig into the convergence of the the sequence of means

$\{\tilde{\mu}_{t,K_t}\}_{k\in[1,K_t]}$ and covariance matrices $\{\tilde{\Sigma}_{t,K_t}\}_{k\in[1,K_t]}$ to obtain a fine-grained analysis of the distribution of $\tilde{\mathrm{q}}_t$. Therefore, using equations (8.20) and (8.21)

$$\begin{aligned}\mathbb{1}_{\bar{\mathrm{E}}_t^{\text{true}}}\mathbb{E}_t[\Delta_t(a_t)] &\leq \Big(\frac{2}{p-1/t^2}+1\Big)g(t)\mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}] + \frac{1}{t^2} \\ &\leq \frac{cg(t)}{p}\mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}] + \frac{1}{t^2}\,,\end{aligned}$$

where c is a sufficiently large real number independent of the problem. Plugging this bounds in 8.19, we obtain

$$\mathbb{E}_t[\Delta_t(a_t)] \leq \frac{cg(t)}{p}\mathbb{E}_t[\|\phi\|_{V_t^{-1}}] + \frac{1}{t^2} + \mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{true}})$$

Applying Lemma 107 yields

$$\mathbb{E}_t[\Delta_t(a_t)] \leq \frac{cg(t)}{p}\mathbb{E}_t[\|\phi\|_{V_t^{-1}}] + \frac{2}{t^2}\,.$$

This is another important difference with the original proof of [Agrawal and Goyal, 2012] which uses our precise convergence study for $\{\tilde{\mu}_{t,K_t}\}_{k\in[1,K_t]}$. Then, it follows that $(X_t)_{t\in[T]}$ is a $(\mathcal{F}_t)_{t\in[T]}$-super martingale.

3. **Concentration for $(X_t)_{t\in[T]}$.** Note that $(X_t)_{t\in[T]}$ is a super-martingale with bounded increments: for any $t\in[T]$

$$\begin{aligned}|X_{t+1}-X_t| &= |Y_{t+1}| \\ &= |\Delta_t(a_t) - \frac{cg(t)}{p}\|\phi_t\|_{V_t^{-1}} - \frac{2}{t^2}| \\ &\overset{(a)}{\leq} |\Delta_t(a_t)(a_t) - \frac{cg(t)}{\sqrt{\lambda}p} - \frac{2}{t^2}| \\ &\leq \frac{3cg(t)}{\sqrt{\lambda}p}\,,\end{aligned}$$

where in (a) we have used that

$$\|\phi_t\|_{V_t^{-1}} \leq \|\phi_t\|_{V_1^{-1}} \leq 1/\sqrt{\lambda}\,,$$

and inequality (b) is due to $\Delta_t(a_t)\leq 1$, $2/t^2\leq 2$ and $3cg(t)/(p\sqrt{\lambda}) > 2$ for an appropriate choice of the numerical constant c. Therefore, applying Azuma-Hoeffding inequality (Lemma (122)), with probability $1-\delta$ it holds that

$$X_T \leq \sqrt{2\log(1/\delta)\sum_{s=1}^{T}\frac{9c^2g(s)^2}{p^2\lambda}} \leq \sqrt{18\log(1/\delta)\frac{c^2}{p^2\lambda}g(T)^2T}\,,$$

using that $g(T)\geq g(t)$.

4. **Conclusion.** The super-martingale $(X_t)_{t=1}^T$ is directly linked to the cumulative regret by

$$\begin{aligned}X_T &= \sum_{t=1}^{T} Y_t \\ &= \sum_{s=1}^{T}\Delta_t - cg(t)\frac{\|\phi_t\|_{V_t^{-1}}}{p} - \frac{2}{t^2} \\ &= \mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) - \sum_{t=1}^{T} cg(t)\frac{\|\phi_t\|_{V_t^{-1}}}{p} + \frac{2}{t^2}\end{aligned}$$

then taking the expectation and using the super-martingale previous argument of the proof, we obtain the following upper bound for the cumulative regret:

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) \leq \sum_{t=1}^{T} \frac{cg(t)}{p} \|\phi_t\|_{V_t^{-1}} + \sqrt{18\log(1/\delta)\frac{c^2}{p^2\lambda} g(T)^2 T} + \frac{\pi^2}{3} \,.$$

using that $\sum_{t=1}^{+\infty} 1/t^2 \leq \pi^2/6$. As a result, applying Lemma 109 yields

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) \leq \frac{cg(T)}{p}\sqrt{2dT\log\Big(1+\frac{T}{\lambda d}\Big)} + \frac{cg(T)}{p\sqrt{\lambda}}\sqrt{18\log(1/\delta)T} + \frac{\pi^2}{3} \,.$$

Using the definition of $g(T)$ in (8.19), we get

$$\mathrm{CREG}(\tilde{\mathbb{Q}}_{1:T}) \leq \frac{CR^2 d}{\lambda^2}\log\big(3T^3\big)\sqrt{dT\log\big(1+\frac{T}{\lambda d}\big)\log\big(1/\delta\big)} \,,$$

where $C \geq 0$ is a constant, independent of the problem, which completes the proof.

6.2 Hyperparameters choice and values

In this section, we define and discuss the values of the main hyperparameters.

Parameter η : is the inverse of the temperature. The lower is η, the better is the exploration. For theoretical reasons, it is fixed to

$$\eta = 4\lambda^2/(81R^2 d\log\big(3T^3\big)) \leq 1 \tag{8.22}$$

Parameter λ : it is used in the standard deviation of the prior distribution. It controls the regularization. The lower is λ, the better is the exploitation. This parameter is fixed but lower than 1.

Parameter h_t : is the step size used in all Algorithms. It is fixed to

$$h_t = \frac{\lambda_{\min}(V_t)}{2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)} \tag{8.23}$$

Parameter K_t : is the number of gradient descent steps performed at each steps. It is fixed to

$$K_t = 1 + 2(1+2\kappa_t^2)\log\Big(2R\kappa_t d^2 T^2 \log^2(3T^3)\Big) \,. \tag{8.24}$$

Therefore the number of gradient descent steps is $K_t \leq \mathcal{O}(\kappa_t^2 \log(dT\log(T)))$.

6.3 Useful definitions

Definition 105. *(Variational approximation)*

Recall that $\hat{\mathrm{p}}_t(\theta) \propto \exp(-U_t(\theta))$ is the posterior distribution. And $\tilde{\mathrm{q}}_t$ is the variational posterior distribution in the sense that

$$\tilde{\mathrm{q}}_t = \operatorname*{argmin}_{\mathrm{p}\in\mathcal{G}} \mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}}_t) \,,$$

where $\mathcal{G}$ is a variational family. In this paper we focus on the Gaussian variational family and we denote by $\tilde{\mu}_t$ and B_t respectively the mean and the square root covariance matrix of the variational distribution, ie,

$$\tilde{\mathrm{q}}_t = \mathcal{N}(\tilde{\mu}_t, B_t B_t^\top) \,.$$

The values of $\tilde{\mu}_t$ and B_t are obtained after running K_t steps of algorithm 4 or 5. Note that the sequence of means $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ is defined recursively by

$$\begin{aligned}\tilde{\mu}_{t,k+1} &= \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k}) \\ &= \tilde{\mu}_{t,k} - h_t \eta V_t(\tilde{\theta}_{t,k} - \hat{\mu}_t)\end{aligned}$$

where $\tilde{\theta}_{t,k} \sim \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$ and we have used that $\nabla U_t(\theta) = \eta(V_t\theta - b_t)$ (see equation (8.17)). Consequently, $\tilde{\mu}_{t,k}$ is also Gaussian and we denote by $\tilde{m}_{t,k}$ and $\tilde{W}_{t,k}$ its mean and covariance matrix, ie, $\tilde{\mu}_{t,k} \sim \mathcal{N}(\tilde{m}_{t,k}, \tilde{W}_{t,k})$. Furthermore, the sequence of square root covariance matrix $\{B_{t,k}\}_{k=1}^{K_t}$ is defined recursively in Algorithm 4 by

$$\begin{aligned}B_{t,k+1} &= \left\{\mathrm{I}_d - h_t \nabla^2(U_t(\tilde{\theta}_{t,k}))\right\} B_{t,k} + (B_{t,k}^\top)^{-1} \\ &= \left\{\mathrm{I}_d - \eta h_t V_t\right\} B_{t,k} + h_t (B_{t,k}^\top)^{-1}\end{aligned}$$

where we have used that $\nabla^2(U_t(\theta)) = \eta V_t$ for the linear bandit case (see (8.17)). Let denote by $\tilde{\Sigma}_{t,k} = B_{t,k} B_{t,k}^\top$ the covariance of the variational posterior $\tilde{\mathrm{q}}_{t,k}$. For ease of notation we denote by $A_t = \mathrm{I}_d - \eta h_t V_t$, it follows that

$$\tilde{\Sigma}_{t,k+1} = A_t \tilde{\Sigma}_{t,k} A_t + 2h_t A_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1}$$

If $\Lambda_{t,k} = \tilde{\Sigma}_{t,k} - 1/\eta V_t^{-1}$ denotes the difference between the covariance matrix of the varational posterior and the true posterior, therefore it holds that

$$\begin{aligned}\Lambda_{t,k+1} &= A_t \tilde{\Sigma}_{t,k} A_t + 2h_t A_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} - 1/\eta V_t^{-1} \\ &= A_t \Lambda_{t,k} A_t + 2h_t A_t - 2h_t \mathrm{I}_d + \eta h_t^2 V_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} \\ &= A_t \Lambda_{t,k} A_t - \eta h_t^2 V_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} \\ &= A_t \Lambda_{t,k} A_t - h_t^2 \eta V_t \Lambda_{t,k} \tilde{\Sigma}_{t,k}^{-1}\end{aligned}$$

Definition 106. *(Concentration events)*

The main challenge for the proof of Theorem 104, is to control the probability of the following events: for any $t \in [T]$ we define

- $\widehat{\mathrm{E}_t^{true}} = \left\{ \text{for any } a \in \mathcal{A}(x_t) : \ |\phi(x_t,a)^\top \hat{\mu}_t - \phi(x_t,a)^\top \theta^\star| \leq \ g_1(t) \ \|\phi(x_t,a)\|_{V_t^{-1}} \right\}$
- $\mathrm{E}_t^{true} = \widehat{\mathrm{E}_t^{true}} \bigcap \left\{ |\xi_t| < R\sqrt{1 + \log 3t^2} \right\} \bigcap \left\{ \|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\| \leq \sqrt{4d \log 3t^3} \right\}$
- $\mathrm{E}_t^{var} = \left\{ \text{for any } a \in \mathcal{A}(x_t) : \ |\phi(x_t,a)^\top \tilde{\theta}_t - \phi(x_t,a)^\top \hat{\mu}_t| \leq \ g_2(t) \ \|\phi(x_t,a)\|_{V_t^{-1}} \right\}$,

where $g_1(t) = R\sqrt{d \log(3t^3)} + \sqrt{\lambda}$ and $g_2(t) = 10\sqrt{d \log(3t^3)/(\eta\lambda)}$ and ξ_t is the R-sub Gaussian noise of the reward definition defined by the relation

$$r_t = \phi_t^\top \theta^\star + \xi_t \ . \tag{8.25}$$

The first event $\widehat{\mathrm{E}_t^{true}}$ controls the concentration of $\phi(x_t,a)^\top \hat{\mu}_t$ around its mean. Similarly, event E_t^{var} controls the concentration of $\phi(x_t,a)^\top \tilde{\theta}_t$ around its mean. Note that compared to [Agrawal and Goyal, 2013a], in our case, it is important to include within E_t^{true}, the concentration of the distributions ξ_t and $\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})$. Consequently, conditionally on E_t^{true} and E_t^{var} it holds that: for any $a \in \mathcal{A}(x_t)$

$$\begin{aligned}|\phi(x_t,a)^\top \tilde{\theta}_t - \phi(x_t,a)^\top \theta^\star| &\leq \left(R\sqrt{d \log(3t^3)} + \sqrt{\lambda} + 10\sqrt{d \log(3t^3)/(\eta\lambda)} \right) \|\phi(x_t,a)\|_{V_t^{-1}} \\ &\leq 12R\sqrt{d \log(3t^3)/(\eta\lambda)} \ \|\phi(x_t,a)\|_{V_t^{-1}} \\ &\overset{(a)}{=} \frac{108 d R^2}{\lambda^{3/2}} \sqrt{\log(3t^3)\log(3T^3)} \|\phi(x_t,a)\|_{V_t^{-1}} \\ &:= g(t) \|\phi(x_t,a)\|_{V_t^{-1}} \ ,\end{aligned} \tag{8.26}$$

where in (a), we have used that $\eta = 4\lambda/\Big(81R^2d\log\big(3T^3\big)\Big)$ and in the last inequality we have used that $g(t) = CR^2d\sqrt{\log(t)\log(T)}/\lambda^{3/2}$.

6.4 Main lemmas

Lemma 107. *(Concentration lemma for $\hat{\mu}_t$)*
Recall the definition of the event E_t^{true} in (106)*. Therefore, for any $t \in [T]$, it holds that*

$$\mathbb{P}(\mathrm{E}_t^{true}) \geq 1 - \frac{1}{t^2} \tag{8.27}$$

This lemma shows that the mean of the posterior distribution $\hat{\mu}_t$ is concentrated around the true parameter $\theta^\star$ with high probability.

Proof. Firstly, we apply Lemma 123, with $m_t = \phi_t/\sqrt{\lambda} = \phi(x_t, a_t)/\sqrt{\lambda}$ and $\epsilon_t = \big(r_{a_t}(t) - \phi_t^\top\theta^\star\big)/\sqrt{\lambda}$, where $r_{a_t}(t)$ is sampled from the R-sub-Gaussian reward distribution of mean $\phi_t^\top\theta^\star$. Let's define the filtration $\mathcal{F}'_t = \{a_{\tau+1}, m_{\tau+1}, \epsilon_\tau\}_{\tau\leq t}$. By the definition of $\mathcal{F}'_t$, m_t is $\mathcal{F}'_{t-1}$-measurable. Moreover, ϵ_t is conditionally $R/\sqrt{\lambda}$-sub-Gaussian due to Assumption 6 and is a martingale difference process because $\mathbb{E}[\epsilon_t|\mathcal{F}'_{t-1}] = 0$. If we denote by

$$M_t = \mathrm{I}_d + \frac{1}{\lambda}\sum_{\tau=1}^{t} m_\tau m_\tau^\tau = \frac{1}{\lambda}V_{t+1}\ ,$$

and

$$\zeta_t = \sum_{\tau=1}^{t} m_\tau\epsilon_\tau\ ,$$

Then, Lemma 123 shows that $\|\zeta_t\|_{M_t^{-1}} \leq R/\sqrt{\lambda}\sqrt{d\log\big(\frac{t+1}{\delta'}\big)}$ with probability at least $1-\delta'$. Moreover, note that

$$\begin{aligned}
M_{t-1}^{-1}(\zeta_{t-1} - \theta^\star) &= M_t^{-1}(\frac{1}{\lambda}b_t - \frac{1}{\lambda}\sum_{\tau=1}^{t-1}\phi_\tau\phi_\tau^\top\theta^\star - \theta^\star) \\
&= M_{t-1}^{-1}(\frac{1}{\lambda}b_t - M_{t-1}\theta^\star) \\
&= \hat{\mu}_t - \theta^\star\ .
\end{aligned}$$

Note that $||\theta^\star||_{M_{t-1}^{-1}} = \|\theta^\star M_{t-1}^{-1/2}\|_2 \leq \|\theta^\star\|_2\|M_{t-1}^{-1/2}\|_2 \leq \|\theta^\star\|_2$, where the last inequality is due to Assumption 7. Then, for any arm $a \in \mathcal{A}(x_t)$ we have

$$\begin{aligned}
|\phi(x_t,a)^\top\hat{\mu}_t - \phi(x_t,a)^\top\theta^\star| &= |\phi(x_t,a)M_{t-1}^{-1}(\xi_{t-1} - \theta^\star)| \\
&\leq \|\phi(x_t,a)\|_{M_{t-1}^{-1}}\|\xi_{t-1} - \theta^\star\|_{M_{t-1}^{-1}} \\
&\leq \|\phi(x_t,a)\|_{M_{t-1}^{-1}}(\|\xi_{t-1}\|_{M_{t-1}^{-1}} + \|\theta^\star\|_{M_{t-1}^{-1}}) \\
&\leq \sqrt{\lambda}\left(\frac{R}{\sqrt{\lambda}}\sqrt{d\log\left(\frac{t}{\delta'}\right)} + 1\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\
&= \sqrt{\lambda}\left(\frac{R}{\sqrt{\lambda}}\sqrt{d\log(3t^3)} + 1\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\
&= \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda}\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\
&:= g_1(t)\|\phi(x_t,a)\|_{V_t^{-1}}\ .
\end{aligned}$$

This inequality holds with probability at least $\delta' = 1/(3t^2)$.

Moreover, recall the definition of the R-subGaussian noise of the reward definition in Definition 106

$$r_t = \phi_t^\top \theta^\star + \xi_t$$

Then it holds that $\mathbb{P}(|\xi_t| > x) \leq \exp\big(1 - x^2/R^2\big)$. It follows that $\mathbb{P}(|\xi_t| \leq R\sqrt{1+\log 3t^2}) \geq 1 - 1/(3t^2)$, for any $t \leq 1$. Finally, recall the definition of $\tilde{W}_{t,k}$, $\tilde{\mu}_{t,k}$ and $\tilde{m}_{t,k}$ in Subsection 6.3. Consequently, the term $\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})$ is gaussian with mean 0 and an identity covariance matrix. Therefore, it holds that

$$\mathbb{P}\Big(\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\| \leq \sqrt{4d\log 3t^3}\Big) \geq 1 - 1/(3t^2) \tag{8.28}$$

Consequently, we have

$$\mathbb{P}\Big(\widehat{\mathrm{E}_t^{\text{true}}} \bigcap \Big\{|\xi_i| < R\sqrt{1+\log 3t^2}\Big\} \bigcap \Big\{\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\| \leq \sqrt{4d\log 3t^3}\Big\}\Big) \geq 1 - \frac{1}{t^2}\,,$$

where $\widehat{\mathrm{E}_t^{\text{true}}}$ is defined is Definition 106

□

Lemma 108. *(Probability of playing an unsaturated arm)*
Given E_t^{true} defined in Definition (106)*, the conditional probability of playing an unsaturated arm is strictly positive and is lower bounded as*

$$\mathbb{1}_{\mathrm{E}_t^{true}}\mathbb{P}_t(a_t \in \mathcal{U}_t) := \mathbb{P}(a_t \in \mathcal{U}_t|\mathcal{F}_{t-1}) \geq \mathbb{1}_{\mathrm{E}_t^{true}}(p - 1/t^2)\,, \tag{8.29}$$

where $p = 1/\sqrt{2\pi \mathrm{e}}$ and $\mathcal{U}_t$ is defined in (8.18)*.*

Proof. If we suppose that $\forall a \in \mathcal{S}_t,\ \ \phi(x_t, a_t^\star)^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t$, then $a_t \in \mathcal{U}_t$. Indeed, The optimal arm $a_t^\star$ is obviously in the unsaturated arm set ($\mathcal{U}_t$) and $\phi(x_t, a_t)^\top \tilde{\theta}_t \geq \phi(x_t, a_t^\star)^\top \tilde{\theta}_t$ by construction of the algorithm. Hence we have

$$\mathbb{P}(a_t \in \mathcal{U}_t|\mathcal{F}_{t-1}) \geq \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t, \forall a \in \mathcal{S}_t|\mathcal{F}_{t-1})$$

Subsequently, given events $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ we have

$$\Big\{\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t, \forall a \in \mathcal{S}_t\Big\} \supset \Big\{\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star\Big\}\,.$$

Indeed, for any $a \in \mathcal{S}_t$,

$$\begin{aligned}\phi(x_t, a)^\top \tilde{\theta}_t &\overset{(a)}{\leq} \phi(x_t, a)^\top \theta^\star + g(t)\|\phi(x_t, a)\|_{\hat{\Sigma}_t} \\ &\overset{(b)}{\leq} \overset{*}{\phi}{}^\top \theta^\star\,,\end{aligned}$$

where (a) uses that $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ hold. And in inequality (b) we have used that $a \in \mathcal{S}_t$, ie, $\overset{*}{\phi}{}_t^\top \theta^\star - \phi(x_t, a)^\top \theta^\star := \Delta_t(a) > g(t)\|\phi(x_t, a)\|_{\hat{\Sigma}_t}$.

Consequently,

$$\begin{aligned}\mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star|\mathcal{F}_{t-1}) &= \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star|\mathcal{F}_{t-1}, \mathrm{E}_t^{\text{var}})\mathbb{P}(\mathrm{E}_t^{\text{var}}) + \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star|\mathcal{F}_{t-1}, \mathrm{E}_{\mathrm{t}}^{\bar{\text{var}}})\mathbb{P}(\mathrm{E}_{\mathrm{t}}^{\bar{\text{var}}}) \\ &\leq \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t, \forall a \in \mathcal{S}_t|\mathcal{F}_{t-1}) + \mathbb{P}(\mathrm{E}_{\mathrm{t}}^{\bar{\text{var}}})\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}(a_t \in \mathcal{U}_t|\mathcal{F}_{t-1}) &\geq \mathbb{P}(\overset{*}{\phi}{}_t^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}_t \theta^\star|\mathcal{F}_{t-1}) - \mathbb{P}(\mathrm{E}_{\mathrm{t}}^{\bar{\text{var}}}) \\ &\geq p - \frac{1}{t^2}\,,\end{aligned}$$

where the last inequality is due to Lemma 121 and Lemma 120 with $p = 1/(2\sqrt{2\pi e})$. □

Lemma 109. *(Upper bound of $\sum_{t=1}^{T}\|\phi_t\|_{\hat{\Sigma}_t}$) The following lemma will be useful in the derivation of the regret bound later in the proof.*

$$\sum_{t=1}^{T}\|\phi_t\|_{V_t^{-1}} \leq \sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)}$$

Proof. Recall the relation between the 1-norm and 2-norm for a d-dimensional vector, ie, $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$. Hence, it follows that

$$\sum_{t=1}^{T}\|\phi_t\|_{V_t^{-1}} \leq \sqrt{T\sum_{t=1}^{T}\|\phi_t\|^2_{V_t^{-1}}}$$

Firs, recall the definition of $V_t = \lambda \mathrm{I}_d \sum_{s=1}^{t-1}\phi_s\phi_s^\top$ in (8.17). Therefore, we apply Lemma 11 and Lemma 10 of [Abbasi-Yadkori et al., 2011b], then we have

$$\begin{aligned}\sum_{t=1}^{T}\|\phi_t\|^2_{V_t^{-1}} &\leq 2\log\frac{\det V_t}{\det\lambda\mathrm{I}_d}\\ &\leq 2\log\frac{(\lambda+T/d)^d}{\lambda^d}\\ &= 2d\log\left(1+\frac{T}{\lambda d}\right).\end{aligned}$$

Consequently,

$$\sum_{t=1}^{T}\|\phi_t\|_{V_t^{-1}} \leq \sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)}$$

□

6.5 Technical Lemmas

6.5.1 Upper bound of variational mean concentration term

In this section the objective is to bound the mean variational concentration term, ie, $|\phi^\top(\tilde{m}_{t,k}-\hat{\mu}_t)|$.

Lemma 110. *Given E_t^{true} defined in Definition* (106)*, the expected mean of the variational posterior at time step t after K_t steps of gradient descent $\tilde{m}_{t,K_t}$, defined in Subsection 6.3, is equal to:*

$$\tilde{m}_{t,K_t} = \sum_{j=1}^{t-1}\prod_{i=j}^{t-1}A_i^{K_i-1}(\hat{\mu}_j-\hat{\mu}_{j+1})+\hat{\mu}_t \tag{8.30}$$

where $A_i = \mathrm{I}_d - \eta h_i V_i$.

Proof. Recall the definitions of $\tilde{\mu}_{t,k}$ and $\tilde{m}_{t,k}$ in Subsection 6.3. Moreover, this Subsection also presents the sequence $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ defined recursively in Algorithm 3 by:

$$\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - h_t\eta V_t(\tilde{\theta}_{t,k}-\hat{\mu}_t)\,.$$

Note that $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ is a sequence of Gaussian samples with mean and covariance matrix $\tilde{m}_{t,k}$ and $\tilde{W}_{t,k+1}$ respectively (see 6.3). Then, we have,

$$\begin{aligned}\tilde{m}_{t,k+1} &= \mathbb{E}[\tilde{\mu}_{t,k+1}]\\ &= \tilde{m}_{t,k} - \eta h_t V_t(\tilde{m}_{t,k}-\hat{\mu}_t)\\ &= (\mathrm{I}_d - h_t\eta V_t)\tilde{m}_{t,k} + \eta h_t V_t\hat{\mu}_t\end{aligned}$$

Now, we recognise an arithmetico-geometric sequence, therefore the solution is:

$$\tilde{m}_{t,k} = (\mathrm{I}_d - h_t \eta V_t)^{k-1}(\tilde{m}_{t,1} - \hat{\mu}_t) + \hat{\mu}_t$$

Moreover, in the algorithm we use that $\tilde{\mu}_{t,1} = \tilde{\mu}_{t-1,K_{t-1}}$, which implies that $\tilde{m}_{t,1} = \tilde{m}_{t-1,k_{t-1}}$ and $W_{t,1} = W_{t-1,K_{t-1}}$. Hence, we have

$$\tilde{m}_{t,K_t} = \prod_{i=1}^{t} (\mathrm{I}_d - \eta h_i V_i)^{K_i - 1}(\tilde{m}_{1,1} - \hat{\mu}_1) + \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} (\mathrm{I}_d - \eta h_i V_i)^{K_i - 1}(\hat{\mu}_j - \hat{\mu}_{j+1}) + \hat{\mu}_t \tag{8.31}$$

Fianlly, the mean of the variational posterior is initialized at $\tilde{\mu}_{1,1} = 0_d$, then the expected mean of the variational posterior $\tilde{m}_1 = \hat{\mu}_1 = 0_d$. Therefore the first term of (8.31) is null. □

Lemma 111. *Given E_t^{true}, for any $\phi \in \mathbb{R}^d$, it holds that*

$$|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} (1 - \eta h \lambda_{min}(V_i))^{K_i - 1} \|\phi\|_{V_j^{-1}} \|\phi_j\|_{V_j^{-1}} \left(\frac{g_1(t)}{\sqrt{\lambda}} + R\sqrt{1 + \log(3t^2)} \right)$$

where $\tilde{m}_{t,K_t}$ is the expected mean of the variational posterior at time step t after K_t steps of gradient descent, ie, $\tilde{m}_{t,K_t} = \mathbb{E}[\tilde{\mu}_{t,K_t}]$, (see Subsection 6.3). Recall that $g_1(t) = R\sqrt{d \log(3t^3)} + \sqrt{\lambda}$ (see Definition: 106).

Proof. Lemma 110 gives us that $\tilde{m}_{t,K_t} = \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} A_i^{K_i - 1}(\hat{\mu}_j - \hat{\mu}_{j+1}) + \hat{\mu}_t$ where $A_i = \mathrm{I}_d - \eta h_i V_i$. Then, for any $\phi \in \mathbb{R}^d$, the term we want to upper bound is:

$$|\phi^\top(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \sum_{j=1}^{t-1} |\phi^\top \prod_{i=j}^{t-1} A_i^{K_i - 1}(\hat{\mu}_j - \hat{\mu}_{j+1})| \,, \tag{8.32}$$

We can notice that the previous term only depends on the difference between the mean posterior at time j and the one at time j+1, which can be upper bounded. Recall the different relations between V_j, b_j, r_j, ϕ_j and $\hat{\Sigma}_j$ in the linear bandit setting (see equation (8.17)): $V_{j+1} = V_j + \phi_j \phi_j^\top$, $b_{j+1} = b_j + r_j \phi_j$ and $\hat{\mu}_j = V_j^{-1} b_j$, then by Sherman-Morrison formula we have:

$$V_{j+1}^{-1} = (V_j + \phi_j \phi_j^\top)^{-1} = V_j^{-1} - \frac{V_j^{-1} \phi_j \phi_j^\top V_j^{-1}}{1 + \phi_j^\top V_j^{-1} \phi_j} \tag{8.33}$$

The difference between the mean posterior at time $j+1$ and the one at time j becomes:

$$\begin{aligned}
\hat{\mu}_{j+1} - \hat{\mu}_j &= V_{j+1}^{-1} b_{j+1} - V_j^{-1} b_j \\
&= (V_j^{-1} - \frac{V_j^{-1} \phi_j \phi_j^\top V_j^{-1}}{1 + \phi_j^\top V_j^{-1} \phi_j})(b_j + r_j \phi_j) - V_j^{-1} b_j \\
&= r_j V_j^{-1} \phi_j - \frac{V_j^{-1} \phi_j \phi_j^\top V_j^{-1}}{1 + \phi_j^\top V_j^{-1} \phi_j}(b_j + r_j \phi_j) \\
&= \frac{V_j^{-1} \phi_j}{1 + \phi_j^\top V_j^{-1} \phi_j} \left\{ -\phi_j^\top \hat{\mu}_j - r_j \phi_j^\top V_j^{-1} \phi_j + r_j (1 + \phi_j^\top V_j^{-1} \phi_j^\top) \right\} \\
&= \frac{V_j^{-1} \phi_j (r_j - \phi_j^\top \hat{\mu}_j)}{1 + \phi_j^\top V_j^{-1} \phi_j} \\
&\overset{(a)}{=} \frac{V_j^{-1} \phi_j (\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)}{1 + \phi_j^\top V_j^{-1} \phi_j} \\
&\overset{(b)}{\leq} V_j^{-1} \phi_j (\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)
\end{aligned} \tag{8.34}$$

where in (a) we have used that $r_j = \phi_j^\top \theta^\star + \xi_j$ with ξ_j is sampled from a R-Subgaussian distribution. Inequality (b) is due to $\phi_j^\top V_j^{-1} \phi_i = \|\phi_j\|^2_{V_j^{-1}} > 0$.

Subsequently, combining equations (8.32) and (8.34), we obtain the following upper bound

$$
\begin{aligned}
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| &\leq \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} |\phi^\top A_i^{K_i-1}(\hat{\mu}_j - \hat{\mu}_{j+1})| \\
&\overset{(a)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} \left|\phi^\top A_i^{K_i-1} V_j^{-1}\phi_j(\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)\right| \\
&\overset{(b)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h \lambda_{\min}(V_i))^{K_i-1} \phi^\top V_j^{-1/2} V_j^{-1/2}\phi_j |(\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)| \\
&\overset{(c)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h \lambda_{\min}(V_i))^{K_i-1} \|\phi\|_{V_j^{-1}} \|\phi_j\|_{V_j^{-1}} |(\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)| \\
&\overset{(d)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h \lambda_{\min}(V_i))^{K_i-1} \|\phi\|_{V_j^{-1}} \|\phi_j\|_{V_j^{-1}} \left(\frac{g_1(t)}{\sqrt{\lambda}} + R\sqrt{1+\log(3t^2)}\right)
\end{aligned}
$$

In the inequality (a) we have used equation (8.34), in (b) the relation $A_i^{K_i-1} = (\mathrm{I}_d - \eta h_i V_i)^{K_i-1} \preceq (1-\eta h_i \lambda_{\min}(V_i))^{K_i-1}\, \mathrm{I}_d$, in (c) the definition of $\|\phi\|_{V_t^{-1}} = \sqrt{\phi^\top V_t^{-1}\phi} = \sqrt{\phi^\top V_t^{-1/2}}\sqrt{V_t^{-1/2}\phi} = \phi^\top V_t^{-1/2}$, and finally (d) is due to $|\xi_i| < R\sqrt{1+\log 3t^2}$ as $\mathrm{E}_t^{\text{true}}$ holds and $|\phi_i^\top(\theta^\star - \hat{\mu}_t)| \leq g_1(t)\left\|\phi_i^\top\right\|_{V_t^{-1}} \leq g_1(t)/\sqrt{\lambda}$

□

Lemma 112. *Given E_t^{true}, for any $\phi \in \mathbb{R}^d$, for $t \geq 2$, if the number of gradient descent of Algorithm 4 is such that*

$$
K_t \geq 1 + 2(1+2\kappa_t^2)\log\left(4R\sqrt{dT\log(3T^3)}\right),
$$

then it holds that

$$
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \frac{2\|\phi\|_{V_t^{-1}}}{\lambda}
$$

This lemma provides the upper bound for variational mean concentration term.

Proof.

Firstly, we can apply Lemma 111, it gives us

$$
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h_t \lambda_{\min}(V_t))^{K_i-1} \|\phi\|_{V_j^{-1}} \|\phi_j\|_{V_j^{-1}} \left(g_1(t)/\sqrt{\lambda} + R\sqrt{1+\log(3t^2)}\right)
$$

where $g_1(t) = R\sqrt{d\log(3t^3)} + \sqrt{\lambda}$. Moreover, for $t \geq 2$,

$$
\begin{aligned}
R\sqrt{1+\log 3t^2} + g_1(t)/\sqrt{\lambda} &\leq R\sqrt{\log 3t^2} + R\sqrt{d\log\left(3t^3\right)/\lambda} + R + 1 \\
&\leq 4R\sqrt{d\log 3t^2/\lambda}
\end{aligned} \tag{8.35}
$$

where we have used that $R \geq 1$ and $\lambda \leq 1$. Moreover, for any $j \in [1,t]$ we have

$$
\begin{aligned}
\|\phi\|_{V_j^{-1}} &\leq \|\phi\|_2/\sqrt{\lambda} \\
&\leq \lambda_{\max}(V_t)^{1/2}\|\phi\|_{V_t^{-1}}/\sqrt{\lambda} \\
&= \lambda_{\max}(V_t)^{1/2}\|\phi\|_{V_t^{-1}}/\lambda^{1/2}
\end{aligned} \tag{8.36}
$$

Let's define $\epsilon = \left(4R\sqrt{dt\log(3t^2)}\right)^{-1} \leq 1/2$ and let's take K_i such that $(1 - h_t\lambda_{\min}(V_t))^{K_i - 1} \leq \epsilon$, this condition will be explained later in the proof. It follows that

$$
\begin{aligned}
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| &\leq \sum_{j=1}^{t-1}\prod_{i=j}^{t-1}(1 - \eta h_t\lambda_{\min}(V_t))^{K_i-1}\|\phi\|_{V_j^{-1}}\|\phi_j\|_{V_j^{-1}}\epsilon^{-1} \\
&\overset{(a)}{\leq} \frac{\|\phi\|_{V_t^{-1}}\lambda_{\max}(V_t)^{1/2}}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j}^{t-1}(1 - h_t\lambda_{\min}(V_T))^{K_i-1}\epsilon^{-1} \\
&\overset{(b)}{\leq} \frac{\|\phi\|_{V_t^{-1}}}{\lambda}\sum_{j=1}^{t-1}\epsilon^{t-j-1} \\
&\overset{(c)}{\leq} \frac{\|\phi\|_{V_t^{-1}}}{\lambda} \times \frac{1}{1-\epsilon} \\
&\overset{(d)}{\leq} \frac{2\|\phi\|_{V_t^{-1}}}{\lambda}\,,
\end{aligned}
$$

where (a) comes from equations (8.36) and (8.35). The point (b) comes that $\lambda_{\max}(V_t) \leq \sqrt{t}$ because $\lambda \leq 1$ and definition of ϵ, then (c) from the geometric series formula. Finally, in (d), we have used $\epsilon \leq 1/2$.

Now, let's focus on that condition on K_i presented previously. For any $i \in [t]$, recall the definition of the step size h_t in 6.3.

$$
h_i = \frac{\lambda_{\min}(V_i)}{2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_i)^2)}\,,
$$

and define $\kappa_i = \lambda_{\max}(V_i)/\lambda_{\min}(V_i)$. Therefore, it holds that

$$
(1 - \eta h_i\lambda_{\min}(V_i))^{K_i-1} = \left(1 - \frac{1}{2(1+2\kappa_i^2)}\right)^{K_i-1}
$$

For any $\epsilon > 0$, we want that $(1 - h_i\lambda_{\min}(V_i))^{K_i-1} \leq \epsilon$. Hence we deduce that

$$
K_i \geq 1 + \frac{\log(1/\epsilon)}{\log\big(1 - 1/(2(1+2\kappa_i^2))\big)}\,.
$$

Moreover, if $0 < x < 1$ then we have $-x > \log(1-x)$, if follows that

$$
K_i \geq 1 + 2(1+2\kappa_i^2)\log(1/\epsilon)\,.
$$

We note that,

$$
\begin{aligned}
\log(1/\epsilon) &= \log\left(4R\sqrt{dt\log 3t^3}\right) \\
&\leq \log\left(4R\sqrt{dT\log 3T^3}\right)\,.
\end{aligned}
$$

Finally, taking $K_i \geq 1 + 2(1+2\kappa_i^2)\log\left(4R\sqrt{dT\log(3T^3)}\right)$, we obtain the condition

$$
(1 - \eta h_i\lambda_{\min}(V_t))^{K_i-1} \leq \epsilon\,,
$$

which concludes the proof. □

6.5.2 Control the variational covariance matrix

The objective of this Subsection is to control the following term: $|\phi^\top \left(\tilde{\theta}_{\mathbf{t},\mathbf{k}} - \tilde{\mu}_{\mathbf{t},\mathbf{k}}\right)|$. As $\tilde{\theta}_{t,k}$ is a sample from a Gaussian distribution with mean $\tilde{\mu}_{t,k}$, the previous term will be controlled using Gaussian concentration and an upper bound of the norm of the variational covariance matrix $\tilde{\Sigma}_{t,k}$. Recall the definitions of parameters $\hat{\Sigma}_t$, $B_{t,k}$, $\tilde{\theta}_{t,k}$ and $\tilde{\mu}_{t,k}$ in Definition 6.3.

Lemma 113. *For any $t \in [T]$ and $k \in [K_t]$, the following relation holds:*

$$(\mathrm{H}):\ \tilde{\Sigma}_{t,k} \succeq \frac{1}{2\eta} V_t^{-1}\,. \tag{8.37}$$

Proof. The sequence $\{\tilde{\Sigma}_{t,n}\}_{t\in[T],\, n\in[k_t]}$ is initialized by $\tilde{\Sigma}_{1,1} = \mathrm{I}_d/(\lambda\eta) = V_1^{-1}/\eta \succeq V_t^{-1}/(2\eta)$. Hence, (H) holds for the pair $t = 1$ and $k = 1$. Therefore, to conclude the proof, we have to show that the following transitions are true:

- for any $t \in [T]$, if (H) holds at step (t, K_t) then it stays true at step $(t+1, 1)$ (`recursion in t`),
- for any $k \in [K_t]$, if (H) holds at step (t, k) then it stays true at step $(t, k+1)$ (`recursion in k`).

Firstly, let's focus on the first implication and suppose that (H) holds at step (t, K_t). Therefore we have

$$\tilde{\Sigma}_{t+1,1} \overset{(a)}{=} \tilde{\Sigma}_{t,K_t} \overset{(b)}{\succeq} \frac{1}{2\eta} V_t^{-1} \overset{(c)}{\succeq} \frac{1}{2\eta} V_{t+1}^{-1}$$

where (a) comes from the initialization of the sequence $\{\tilde{\Sigma}_{t,k}\}_{k\in[k_t]}$, (b) from the hypothesis (H) at step (t, K_t). And finally, (c) is due to $V_{t+1} = V_t + \phi_t \phi_t^\top \succeq V_t$. Then we can conclude that (H) holds at step $(t+1, 1)$.

Now we focus on the second implication and we suppose that (H) holds at step (t, k). For ease of notation we denote by $Z_{t,k} := \tilde{\Sigma}_{t,k} - V_t^{-1}/(2\eta)$. Therefore using the recursive definition of $\tilde{\Sigma}_{t,k}$ given in Subsection (6.3), we have

$$\begin{aligned}
Z_{t,k+1} &= A_t \tilde{\Sigma}_{t,k} A_t + 2h_t A_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} - V_t^{-1}/(2\eta) \\
&= A_t Z_{t,k} A_t + 2h_t A_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} - h_t \mathrm{I}_d + \eta h_t^2 V_t/2 \\
&= A_t Z_{t,k} A_t + h_t \mathrm{I}_d - 3h_t^2 \eta V_t/4 + h_t^2 \tilde{\Sigma}_{t,k}^{-1}
\end{aligned}$$

where in the last inequalities we have used that $A_t = (\mathrm{I}_d - \eta h_t^2 V_t)$. Moreover, all terms in the previous inequality are positive semi-definite. Indeed, as (H) holds at step (t, k), we know that $Z_{t,k} \succeq 0$ and then that $A_t Z_{t,k} A_t \succeq 0$. Moreover, $\tilde{\Sigma}_{t,k} \succeq V_t^{-1}/(2\eta) \succeq 0$, so $\tilde{\Sigma}_{t,k}^{-1} \succeq 0$. Finally, recall the definition of h_t in Subsection (6.3)

$$\begin{aligned}
h_t &\leq \frac{\lambda_{\min}(V_t)}{2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)} \\
&= \frac{1/\kappa_t}{2\eta\lambda_{\max}(V_t)((1/\kappa_t)^2 + 1)} \\
&= \frac{4}{3\eta\lambda_{\max}(V_t)} \times \frac{3/\kappa_t}{8(1 + (1/\kappa_t^2))} \\
&\leq \frac{4}{3\eta\lambda_{\max}(V_t)}\,,
\end{aligned}$$

where $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t) \geq 1$. Consequently, the matrix $\mathrm{I}_d - 3\eta h_t V_t/4$ is also positive semi-definite. Subsequently, we have

$$Z_{t,k+1} \succeq 0\,.$$

□

Lemma 114. *For any $\phi \in \mathbb{R}^d$, let $B_{t,k}$ the square root of the covariance matrix defined in Algorithm* (4)*. It holds that*

$$\|B_{t,K_t}\phi\|_2 \leq 1/\sqrt{\eta}\Big(1+\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$
$$\textit{and} \quad \|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1-\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$

where $C_t = \frac{1}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\Big(1-\frac{3h_t\eta}{2}\lambda_{min}(V_i)\Big)^{K_i-1}$.

Proof. Recall the recursive relation of $\Lambda_{t,k}$ defined in Section (6.3).

$$\Lambda_{t,k+1} = A_t\Lambda_{t,k}A_t - \eta h_t^2 V_t\Lambda_{t,k}\tilde{\Sigma}_{t,k}^{-1}\,,$$

Hence, we have the following relation on the norm of $\Lambda_{t,k+1}$:

$$\begin{aligned}
\|\Lambda_{t,k+1}\|_2 &\leq \|A_t\|_2\|\Lambda_{t,k}\|_2\|A_t\|_2 + \eta h_t^2\|V_t\|_2\|\Lambda_{t,k}\|_2\|\tilde{\Sigma}_{t,k}^{-1}\|_2 \\
&= \Big(\lambda_{\max}(A_t)^2 + \eta h_t^2\lambda_{\max}(V_t)\lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})\Big)\|\Lambda_{t,k}\|_2 \\
&\overset{(a)}{=} \Big(1-2\eta h_t\lambda_{\min}(V_t) + \eta^2 h_t^2\lambda_{\min}(V_t)^2 + \eta h_t^2\lambda_{\max}(V_t)\lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})\Big)\|\Lambda_{t,k}\|_2 \\
&= \left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t) + \eta h_t\{h_t(\eta\lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)\lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})) - \lambda_{\min}(V_t)/2\}\right)\|\Lambda_{t,k}\|_2 \\
&\overset{(b)}{\leq} \left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)\|\Lambda_{t,k}\|_2\,,
\end{aligned}$$

where (a) uses that $\lambda_{\max}(A_t) = 1-\eta h\lambda_{\min}(V_t)$. Finally, inequality (b) is due to: $\tilde{\Sigma}_{t,k} \succeq V_t^{-1}/(2\eta)$ (Lemma 113). Indeed it implies that

$$\begin{aligned}
h_t(\eta\lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)\lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})) &\leq h_t(\eta\lambda_{\min}(V_t)^2 + 2\eta\lambda_{\max}(V_t)^2) \\
&\leq \frac{\lambda_{\min}(V_t)}{2}\,,
\end{aligned}$$

where the inequality comes from the definition of the step size: $h_t \leq \lambda_{\min}(V_t)/\Big(2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)\Big)$.

Subsequently,

$$
\begin{aligned}
\|\Lambda_{t,K_t}\|_2 &\leq \left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)\|\Lambda_{t,K_t-1}\|_2 \\
&\leq \left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)^{K_t-1}\|\tilde{\Sigma}_{t,1}-1/\eta V_t^{-1}\|_2 \\
&= \left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)^{K_t-1}\|\tilde{\Sigma}_{t-1,k_{t-1}}-1/\eta V_t^{-1}\|_2 \\
&\leq \prod_{i=1}^{t-1}\left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}\|\tilde{\Sigma}_{1,1}-1/\eta V_1^{-1}\|_2 \qquad (8.38)\\
&+\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}\|1/\eta V_j^{-1}-1/\eta V_{j+1}^{-1}\|_2 \\
&\overset{(a)}{\leq} \frac{1}{\eta}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}\|V_j^{-1}-V_{j+1}^{-1}\|_2 \\
&\overset{(b)}{\leq} \frac{1}{\lambda\eta}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}, \\
&:= C_t/\eta
\end{aligned}
$$

where in (a) we have used that $\tilde{\Sigma}_{1,1}=\frac{1}{\lambda\eta}\mathrm{I}_d=1/\eta V_1^{-1}$. Moreover $\|V_j^{-1}-V_{j+1}^{-1}\|_2=\|\big(V_j^{-1}\phi_j\phi_j^\top V_j^{-1}\big)/\big(1-\phi_j^\top V_j^{-1}\phi_j\big)\|_2$ see result (8.33). It implies that $\|V_j^{-1}-V_{j+1}^{-1}\|_2\leq\|V_j^{-1}\|_2^2\leq\|V_1^{-1}\|_2^2=1/\lambda$.

Finally, for any $\phi\in\mathbb{R}^d$,

$$
\begin{aligned}
\|B_{t,K_t}\phi\|_2 &= \sqrt{\phi^\top B_{t,K_t}^\top B_{t,K_t}\phi} \\
&= \sqrt{\phi^\top\tilde{\Sigma}_{t,K_t}\phi} \\
&= \sqrt{\phi^\top(\tilde{\Sigma}_{t,K_t}-1/\eta V_t^{-1})\phi+1/\eta\phi^\top V_t^{-1}\phi} \\
&\leq \|\phi\|_2\sqrt{\|\tilde{\Sigma}_{t,K_t}-1/\eta V_t^{-1}\|_2}+1/\sqrt{\eta}\|\phi\|_{V_t^{-1}}
\end{aligned}
$$

where the last inequality comes from the fact that for $a,b>0,\sqrt{a+b}<\sqrt{a}+\sqrt{b}$, Moreover,

$$
\begin{aligned}
\|\phi\|_2 &= \|\phi V_t^{-1/2}V_t^{1/2}\|_2 \\
&\leq \|\phi\|_{V_t^{-1}}\|V_t^{1/2}\|_2
\end{aligned}
$$

Consequently, we have

$$
\|B_{t,K_t}\phi\|_2\leq 1/\sqrt{\eta}\Big(1+\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}
$$

The lower bound of this lemma

$$
\|B_{t,K_t}\phi\|_2\geq 1/\sqrt{\eta}\Big(1-\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}
$$

is obtained because

$$\begin{aligned}\|B_{t,K_t}\phi\|_2 &= \sqrt{\phi^\top B_{t,K_t}^\top B_{t,K_t}\phi} \\ &= \sqrt{\phi^\top \tilde{\Sigma}_{t,K_t}\phi} \\ &= \sqrt{\phi^\top(\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1})\phi + 1/\eta\phi^\top V_t^{-1}\phi} \\ &\geq -\|\phi\|_2\sqrt{\|\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1}\|_2} + 1/\sqrt{\eta}\|\phi\|_{V_t^{-1}} \quad \leq \|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}},\end{aligned}$$

where the first inequality comes from remarkable identity $\sqrt{a} - \sqrt{b} < \sqrt{a+b}$ for $a, b > 0$.

□

Lemma 115. *For any $t \in [T]$ and $a \in \mathcal{A}(x_t)$, if the number of gradient descent steps of Algorithm 4 is $K_t \geq 1 + 4(1+2\kappa_t^2)\log(2T)/(3\eta)$, therefore it holds that*

$$\begin{aligned}&\|\phi(x_t,a)^\top B_{t,K_t}\|_2 \leq 1/\sqrt{\eta}\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} \\ \textit{and} \quad &\|\phi(x_t,a)^\top B_{t,K_t}\|_2 \geq 1/\sqrt{\eta}\Big(1 - 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}}\,.\end{aligned}$$

Proof. Firstly, Lemma 114, gives us

$$\begin{aligned}\|B_{t,K_t}\phi\|_2 &\leq 1/\sqrt{\eta}\Big(1 + \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}} \\ \|B_{t,K_t}\phi\|_2 &\geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}\end{aligned}$$

with $C_t = 1/\lambda\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t))^{K_i-1}$. Furthermore, for any $t \in [T]$, recall that

$$h_t = \frac{\lambda_{\min}(V_t)}{2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)}\,,$$

and define $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t)$. Therefore, it holds that

$$\left(1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t)\right)^{K_t-1} = \left(1 - \frac{3}{4(1+2\kappa_t^2)}\right)^{K_t-1}$$

For any $\epsilon > 0$, we want that $\Big(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t)\Big)^{K_t-1} \leq \epsilon$. Hence we deduce the following relation for K_t:

$$K_t \geq 1 + \frac{\log(\epsilon)}{\log\big(1 - 3\eta/(4(1+2\kappa_t^2))\big)}\,.$$

Moreover, if $0 < x < 1$ then we have $-x > \log(1-x)$, then we have

$$K_t \geq 1 + 4(1+2\kappa_t^2)\log(1/\epsilon)/3\,.$$

Subsequently, let's apply the last result to $\epsilon = 1/(2t)$. Then for $K_t \geq 1 + 4(1+2\kappa_t^2)\log(2T)/3$, we have

$$\begin{aligned}\frac{\|V_t\|_2}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t)\right)^{K_i-1} &\leq \frac{\|V_t\|_2\epsilon}{\lambda}\sum_{j=1}^{t-1}\epsilon^{t-j-1} \\ &\leq \frac{\|V_t\|_2\epsilon}{\lambda}\sum_{j=0}^{+\infty}\epsilon^j \\ &\overset{(a)}{\leq} \frac{\|V_t\|_2}{2\lambda t(1-\epsilon)} \\ &\overset{(b)}{\leq} \frac{1}{\lambda}\,,\end{aligned}$$

where (a) and (b) come from the geometric serie because $\epsilon \leq 1/2$ and we have used that $\|V_t\|_2 = \|\lambda \mathrm{I}_d + \sum_{s=1}^{t-1} \phi\phi^\top\|_2 \leq \lambda + t - 1 \leq t$, as $\lambda \leq 1$. Consequently, we have

$$\begin{aligned} &\|B_{t,K_t}\phi\|_2 \leq 1/\sqrt{\eta}\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} \\ \text{and} \quad &\|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} \end{aligned}$$

□

Lemma 116. *For any $t \in [T]$ and $a \in \mathcal{A}(x_t)$, if the number of gradient descent steps of Algorithm 4 is $K_t \geq 1 + 4(1 + 2\kappa_t^2)\log(2T)/3$, then with probability at least $1 - 1/t^2$, we have*

$$|\phi(x_t, a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| \leq \sqrt{4d\log(t^3)/\eta}\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi(x_t, a)\|_{V_t^{-1}}$$

Proof. For any $a \in \mathcal{A}(x_t)$, if $K_t \geq 1 + 4(1 + 2\kappa_t^2)\log(2T)/(3\eta)$, Lemma 115 gives us that

$$\begin{aligned} |\phi(x_t, a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| &\leq \|B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t})\|_2 \|\phi(x_t, a)^\top B_{t,K_t}\|_2 \\ &\leq \|B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t})\|_2 (1/\sqrt{\eta})\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} . \end{aligned}$$

where first inequality comes from classical matrix norm inequality and the second one is previous Lemma 115, recall that $\tilde{\theta}_{t,K_t} \sim \mathcal{N}(\tilde{\mu}_{t,K_t}, B_{t,K_t}B_{t,K_t}^\top)$, hence $B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}) \sim \mathcal{N}(0, \mathrm{I}_d)$. Therefore, with probability $1 - 1/t^2$ we have

$$B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}) \leq \sqrt{4d\log(t^3)} .$$

Finally, we conclude that with probability $1 - 1/t^2$, it holds that

$$|\phi(x_t, a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| \leq \sqrt{4d\log(t^3)/\eta}\left(1 + 1/\sqrt{\lambda}\right)\|\phi(x_t, a)\|_{V_t^{-1}} .$$

□

6.5.3 Concentration of the mean of the Variational posterior around its mean

In this section, the objective is the show to concentration of $\tilde{\mu}_{t,k}$ around its mean $\tilde{m}_{t,k}$. More precisely, we want an upper bound of $|\phi^\top \left(\tilde{\mu}_{\mathbf{t},\mathbf{k}} - \tilde{\mathbf{m}}_{\mathbf{t},\mathbf{k}}\right)|$.

Lemma 117. *For any $t \in [T]$ and $k \in [K_t]$, we have the following relation*

$$\tilde{W}_{t,k+1} = (\mathrm{I}_d - \eta h_t V_t)\tilde{W}_{t,k}(\mathrm{I}_d - \eta h_t V_t)^T + \eta^2 h_t^2 V_t \mathbb{E}[\tilde{\Sigma}_{t,k}]V_t$$

where the sequence $\{\tilde{W}_{t,k}\}_{k=1}^{K_t}$ is introduced in Section 6.3. (Recall: $\tilde{\mu}_{t,k} \sim \mathcal{N}(\tilde{m}_{t,k}, \tilde{W}_{t,k})$)

Proof. We focus on the covariance matrix $\tilde{W}_{t,k}$ (see definition 6.3), by definition we have

$$\begin{aligned} \tilde{W}_{t,k+1} &= \mathbb{E}[(\tilde{\mu}_{t,k+1} - \tilde{m}_{t,k+1})(\tilde{\mu}_{t,k+1} - \tilde{m}_{t,k+1})^\top] \\ &= \mathbb{E}[a_{t,k+1}a_{t,k+1}^\top] , \end{aligned}$$

where $a_{t,k}$ is the difference between $\tilde{\mu}_{t,k}$ and its mean. For ease of notation, let's define $\Omega_{t,k} := \tilde{\theta}_{t,k} - \tilde{m}_{t,k}$, then we have

$$a_{t,k+1} = \tilde{\mu}_{t,k} - \tilde{m}_{t,k} - \eta h_t V_t(\tilde{\theta}_{t,k} - \tilde{m}_{t,k})// \qquad = \tilde{\mu}_k - \tilde{m}_{t,k} - \eta h_t V_t \Omega_{t,k}.$$

Consequently,

$$a_{t,k+1}a_{t,k+1}^\top = (\tilde{\mu}_{t,k} - \tilde{m}_{t,k})(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})^\top - \eta h_t V_t \Omega_{t,k}(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})^\top \\ - \eta h_t(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})\Omega_{t,k}^\top V_t + \eta^2 h_t^2 V_t \Omega_{t,k}\Omega_{t,k}^\top V_t$$

Moreover, note that $\tilde{\theta}_{t,k} = \tilde{\mu}_{t,k} + B_{t,k}\epsilon_{t,k}$ where $\epsilon_{t,k} \sim \mathcal{N}(0, \mathrm{I}_d)$. Subsequently we have $\Omega_{t,k} = \tilde{\mu}_{t,k} - \tilde{m}_{t,k} + \tilde{\Sigma}_{t,k}^{1/2}\epsilon_{t,k}$. Then we have $\mathbb{E}[\Omega_{t,k}\Omega_{t,k}^\top] = \tilde{W}_{t,k} + \mathbb{E}[B_{t,k}B_{t,k}^\top]$, $\mathbb{E}[\Omega_{t,k}(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})^\top] = W_{t,k}$, and $\mathbb{E}[(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})\Lambda_{t,k}^\top] = \tilde{W}_{t,k}$. Finally we obtain that

$$\begin{aligned}\tilde{W}_{t,k+1} &= \mathbb{E}[a_{t,k+1}a_{t,k+1}^\top] \\ &= \tilde{W}_{t,k} - \eta h_t V_t \tilde{W}_{t,k} - \eta h_t \tilde{W}_{t,k} V_t + \eta^2 h_t^2 V_t \tilde{W}_{t,k} V_t + \eta^2 h_t^2 V_t \mathbb{E}[B_{t,k}B_{t,k}^\top] V_t \\ &= (\mathrm{I}_d - \eta h_t V_t)\tilde{W}_{t,k}(\mathrm{I}_d - \eta h_t V_t)^\top + \eta^2 h_t^2 V_t \mathbb{E}[B_{t,k}B_{t,k}^\top] V_t \,.\end{aligned}$$

□

Lemma 118. *Recall that $\tilde{\mu}_{t,K_t}$, the mean of the variational posterior after K_t steps of gradient descent, is a sample from the Gaussian with mean $\tilde{m}_{t,K_t}$ and covariance matrix $\tilde{W}_{t,K_t}$, ie, $\tilde{\mu}_{t,K_t} \sim \mathcal{N}(\tilde{m}_{t,K_t}, \tilde{W}_{t,K_t})$. Recall the definition of $\Lambda_{t,k} = \tilde{\Sigma}_{t,k} - 1/\eta V_t^{-1}$, and let denote by $\Gamma_{t,k} = \tilde{W}_{t,k} - J_t V_t^{-1}$, where $J_t = h_t(2\mathrm{I}_d - \eta h_t V_t)^{-1}V_t$.*

This Lemma shows that the 2-norm of Γ_{t,K_t} is controlled by

$$\|\Gamma_{t,K_t}\|_2 \leq \sum_{j=1}^{t} \frac{6\kappa_j h_j \|V_j\|_2}{\lambda} \sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i - 1} \,,$$

where $\kappa_j = \lambda_{max}(V_j)/\lambda_{min}(V_j)$ and $D_i = \mathrm{I}_d - 3\eta h_i \lambda_{min}(V_i)/2$.

Proof. Lemma (117) gives us that

$$\tilde{W}_{t,k+1} = A_t \tilde{W}_{t,k} A_t + \eta^2 h_t^2 V_t \tilde{\Sigma}_{t,k} V_t \,,$$

where $A_t = \mathrm{I}_d - \eta h_t V_t$.

Note that J_t and V_t commute, therefore we have

$$A_t J_t V_t^{-1} A_t = J_t V_t^{-1} - 2h_t \eta J_t + \eta^2 h_t^2 J_t V_t \,.$$

Consequently, by combining the two previous equations we obtain

$$\begin{aligned}\Gamma_{t,k+1} &= A_t \Gamma_{t,k} A_t - 2h_t \eta J_t + \eta^2 h_t^2 J_t V_t + \eta h_t^2 V_t + \eta^2 h_t^2 V_t \Lambda_{t,k} V_t \\ &= A_t \Gamma_{t,k} A_t + \eta^2 h_t^2 V_t \Lambda_{t,k} V_t - h_t \eta J_t(2\mathrm{I}_d - \eta h_t V_t) + \eta h_t^2 V_t \\ &= A_t \Gamma_{t,k} A_t + \eta^2 h_t^2 V_t \Lambda_{t,k} V_t \,.\end{aligned}$$

It follows that

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^2 \|\Gamma_{t,k}\|_2 + \eta^2 h_t^2 \|V_t\|_2^2 \|\Lambda_{t,k}\|_2$$

Therefore, iterating over k gives us

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^{2k} \|\Gamma_{t,k}\|_2 + \eta^2 h_t^2 \sum_{j=0}^{k-1} \|A_t\|_2^{2j} \|V_t\|_2^2 \|\Lambda_{t,k-j}\|_2 \,.$$

Moreover, Equation (8.38) is used to controls the following quantity

$$\|\Lambda_{t,k}\|_2 \leq \left(1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t)\right)^{k-1} \|\Lambda_{t,1}\|_2 \,.$$

Let's denote by $D_t = 1 - 3\eta h_t \lambda_{\min}(V_t)/2$, Subsequently

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^{2k}\|\Gamma_{t,k}\|_2 + \eta^2 h_t^2 \sum_{j=0}^{k-1} \|A_t\|_2^{2j}\|V_t\|_2^2 D_t^{k-j-1}\|\Lambda_{t,1}\|_2 \, .$$

However, $\|A_t\|_2^2 = (1-\eta h_t \lambda_{\min}(V_t))^2 < (1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t))$, because $\eta h_t \leq 1/(4\lambda_{\min}(V_t))$. Consequently, the geometric sum has a common ratio strictly lower than 1, then it is upper bounded by:

$$\begin{aligned}
\sum_{j=0}^{k-1}\left(\frac{\|A_t\|_2^2}{(1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t))}\right)^j &\leq \sum_{j=0}^{+\infty}\left(\frac{\|A_t\|_2^2}{(1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t))}\right)^j \\
&= \frac{1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t)}{1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t) - \|A_t\|_2^2} \\
&\leq \frac{1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t)}{1/2\,\eta h_t \lambda_{\min}(V_t) - \eta^2 h_t^2 \lambda_{\min}(V_t)^2} \\
&\leq \frac{6}{\eta h_t \lambda_{\min}(V_t)} \, , \qquad (8.39)
\end{aligned}$$

where in the first inequality we have used that the ratio of the previous sum is positive. In the last inequality we have used that $\eta h_t \leq 1/(6\lambda_{\min}(V_t))$ in the denominator and we can remove the negative part of the numerator. Therefore, it holds that

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^{2k}\|\Gamma_{t,k}\|_2 + 6\eta\kappa_t h_t D_t^{k-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \, ,$$

where the last inequality comes from (8.39) and the definition of $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t)$. Finally, iterating over t yields to:

$$\begin{aligned}
\|\Gamma_{t,k+1}\|_2 &\leq \|A_t\|_2^{2k}\prod_{j=1}^{t-1}\|A_j\|_2^{2(K_j-1)}\|\Gamma_{1,1}\|_2 + \sum_{j=1}^{t-1}\|A_t\|_2^{2k}\prod_{i=j+1}^{t-1}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) \\
&\quad + 6\eta\kappa_t h_t D_t^{k-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \\
&\leq \sum_{j=1}^{t-1}\|A_t\|_2^{2k}\prod_{i=j+1}^{t-1}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) + 6\eta\kappa_t h_t D_t^{k-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \, ,
\end{aligned}$$

where in the last inequalities we have used that $W_{1,1}$ is initialized such that $W_{1,1} = 1/(11\eta\lambda)\mathrm{I}_d$ and that $J_1 V_1 = h_1(2\mathrm{I}_d - \eta h_1 \lambda \mathrm{I}_d)^{-1} = 1/(11\eta\lambda)\mathrm{I}_d$ because $h_1 = 1/(6\eta\lambda)$. Finally, we can conclude

$$\begin{aligned}
\|\Gamma_{t,K_t}\|_2 &\leq \sum_{j=1}^{t-1}\|A_t\|_2^{2(K_t-1)}\prod_{i=j+1}^{t-1}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) + 6\eta\kappa_t h_t D_t^{K_t-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \\
&= \sum_{j=1}^{t}\prod_{i=j+1}^{t}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) \\
&\leq \sum_{j=1}^{t}\prod_{i=j}^{t} D_i^{K_i-1}\Big(6\eta\kappa_j h_j \|V_j\|_2\|\Lambda_{t,1}\|_2\Big) \, ,
\end{aligned}$$

where in the last inequality we have used that $\|A_t\|_2^2 \leq D_t$. Moreover, equation (8.38) gives us that $\|\Lambda_{j,1}\|_2 \leq 1/(\eta\lambda)\sum_{r=1}^{j}\prod_{l=r}^{j-1} D_l^{K_l-1}$. Consequently, it holds that

$$\begin{aligned}
\|\Gamma_{t,K_t}\|_2 &\leq \sum_{j=1}^{t}\frac{6\kappa_j h_j\|V_j\|_2}{\lambda}\sum_{r=1}^{j}\prod_{l=r}^{j-1} D_l^{K_l-1}\prod_{i=j}^{t} D_i^{K_i-1} \\
&= \sum_{j=1}^{t}\frac{6\kappa_j h_j\|V_j\|_2}{\lambda}\sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1} \, .
\end{aligned}$$

□

Lemma 119. *For any $t \geq 2$, given E_t^{true}, if the number of gradient descent steps is $K_t \geq 1+4(1+2\kappa_t^2)\log\left(2\kappa_t d^2 T \log^2(3T^3)\right)/3$, therefore it holds that*

$$|\phi^\top \left(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t}\right)| \leq \left(\sqrt{\frac{3}{\eta\lambda d \log(3t^3)}} + \sqrt{4d\log(3t^3)/(11\eta)}\right)\|\phi\|_{V_t^{-1}} .$$

Proof.

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

$$|\phi^\top \left(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t}\right)| \leq \|\phi^\top \tilde{W}_{t,K_t}^{1/2}\|_2 \|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\|_2 , \tag{8.40}$$

where $\tilde{W}_{t,K_t}^{1/2}$ is the unique symmetric square root of $\tilde{W}_{t,K_t}$. Firstly, given $\mathrm{E}_t^{\mathrm{true}}$, the term $\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\|_2 < \sqrt{4d\log(3t^3)}$.

Then, we observe that

$$\begin{aligned}\sqrt{4d\log(3t^3)}\|\tilde{W}_{t,K_t}^{1/2}\phi\|_2 &= \sqrt{4d\log(3t^3)\phi\tilde{W}_{t,K_t}\phi^\top} \\ &\leq \sqrt{4d\log(3t^3)\phi\Gamma_{t,K_t}\phi^\top} + \sqrt{4d\log(3t^3)\phi J_t V_t^{-1}\phi^\top} ,\end{aligned} \tag{8.41}$$

where $J_t = h_t(2\mathrm{I}_d - \eta h_t V_t)^{-1}V_t = (2V_t^{-1}/h_t - \eta \mathrm{I}_d)^{-1}$ and $\Gamma_{t,k} = \tilde{W}_{t,k} - J_t V_t^{-1}$.

Moreover,

$$\begin{aligned}\sqrt{\phi J_t V_t^{-1}\phi^\top} &= \|(J_t V_t^{-1})\|_2^{1/2}\phi \\ &\overset{(a)}{=} \|J_t^{1/2}V_t^{-1/2}\phi\|_2 \\ &\leq \|J_t^{1/2}\|_2\|\phi\|_{V_t^{-1}} ,\end{aligned}$$

where in inequality (a) we have used that J_t and V_t^{-1} commute.

Recall that V_t is a symmetric matrix, therefore we have $\lambda_{\min}(V_t)\mathrm{I}_d \preceq V_t \preceq \lambda_{\max}(V_t)\mathrm{I}_d$. It follows that

$$\frac{2}{h_t\lambda_{\max}(V_t)}\mathrm{I}_d \preceq \frac{2}{h_t}V_t^{-1} \preceq \frac{2}{h_t\lambda_{\min}(V_t)}\mathrm{I}_d.$$

Recall the definition of $h_t = \lambda_{\min}(V_t)/\left(2\eta(\lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)^2)\right)$. Consequently, the previous relation becomes

$$\left(\frac{4\eta(1+2\kappa_t^2)}{\kappa_t} - \eta\right)\mathrm{I}_d \preceq \frac{2}{h_t}V_t^{-1} - \eta\mathrm{I}_d \preceq \left(3\eta + 8\eta\kappa_t^2\right)\mathrm{I}_d . \tag{8.42}$$

The left hand term is obviously positive, therefore it holds that

$$\begin{aligned}\|J_t^{1/2}\|_2 &= \|(\frac{2}{h_t}V_t^{-1} - \eta\mathrm{I}_d)\|_2^{-1/2} \\ &\leq \sqrt{\frac{\kappa_t}{\eta(4+8\kappa_t^2-\kappa_t)}} \\ &\leq \frac{1}{\sqrt{\eta(4+7\kappa_t^2)}} \\ &\leq \frac{1}{\sqrt{11\eta}} .\end{aligned}$$

Finally,

$$\sqrt{4d\log(3t^3)\phi J_t V_t^{-1}\phi^\top} \leq \sqrt{4d\log(3t^3)/(11\eta)}\|\phi\|_{V_t^{-1}} .$$

Now, we focus on the first term of equation 8.41. Lemma 118 gives us that

$$\begin{aligned}\|\Gamma_{t,K_t}\|_2 &\leq \sum_{j=1}^{t}\frac{6\kappa_j h_j\|V_j\|_2}{\lambda}\sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1}\\ &\leq \sum_{j=1}^{t}\frac{\kappa_j}{\eta\lambda}\sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1} ,\end{aligned}$$

where in the last inequality we have used that $h_t\|V_t\|_2 = \kappa_t/\big(2\eta(1+2\kappa_t)\big) \leq 1/(6\eta)$. For any $j \in [2,t]$, let's define $\epsilon_j = 1/(2(\kappa_j d^2 t^2\log^2(3t^3)))$. Additionally, let's fix K_i such that $D_i^{K_i-1} \leq \epsilon_j$ (this condition will be explained later in the Lemma). Subsequently, we have

$$\begin{aligned}4d\log\big(3t^3\big)\|V_t\|_2/(\eta\lambda)\sum_{j=1}^{t}\kappa_j\sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1} &\leq 4d\log\big(3t^3\big)\|V_t\|_2/(\eta\lambda)\sum_{j=1}^{t}\kappa_j\sum_{r=1}^{j}\epsilon_j^{t-r+1}\\ &\leq \frac{2\|V_t\|_2}{t^2\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\sum_{j=r}^{t}\epsilon_j^{t-r}\\ &\overset{(a)}{\leq} \frac{2\|V_t\|_2}{t^2\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\sum_{j=r}^{t}\Big(\frac{1}{2d^2t^2\log^2(3t^3)}\Big)^{t-r}\\ &\leq \frac{2\|V_t\|_2}{t\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\frac{t-r+1}{t}\Big(\frac{1}{2d^2t^2\log^2(3t^3)}\Big)^{t-r}\\ &\overset{(b)}{\leq} \frac{2\|V_t\|_2}{t\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\Big(\frac{1}{2d^2t^2\log^2(3t^3)}\Big)^{t-r}\\ &\overset{(c)}{\leq} \frac{2\|V_t\|_2}{\eta t\lambda d\log(3t^3)}\sum_{u=0}^{t-1}\Big(\frac{1}{25}\Big)^{u}\\ &\leq \frac{3}{\eta\lambda d\log(3t^3)} ,\end{aligned}$$

where in (a) we have used that $\epsilon_j \leq 1/2(d^2t^2\log^2(3t^3))$. Inequality (b) is due to $t-r+1 \leq t$. The inequality (c) is obtained because $1/(2d^2t^2\log^2 3t^3) \leq 1/(4\times\log^2(8)) \leq 1/25$ and $u = t-r$. For the last inequality we have used the geometric series formula and $\|V_t\|_2 = \|\lambda \mathrm{I}_d + \sum_{s=1}^{t-1}\phi\phi^\top\|_2 \leq \lambda + t - 1 \leq t$, because $\lambda \leq 1$.

Consequently, as $\|\phi\|_2 \leq \|V_t^{1/2}\|_2\|\phi\|_{V_t^{-1}}$, we obtain

$$\sqrt{4d\log(3t^3)\phi\Gamma_{t,K_t}\phi^\top} \leq \sqrt{\frac{3}{\eta\lambda d\log(3t^3)}}\|\phi\|_{V_t^{-1}} . \tag{8.43}$$

Moreover, the previous inequalities hold if $(1-(3/2)\eta h_i\lambda_{\min}(V_i))^{K_i-1} \leq \epsilon$, following a similar reasoning than in Section 6.5.2, it follows that we need

$$K_t \geq 1 + 4(1+2\kappa_t^2)\log\Big(2\kappa_t d^2T^2\log^2(3T^3)\Big)/3$$

□

6.6 Concentration and anti-concentration

Lemma 120. *(Concentration lemma for $\tilde{\theta}_t$)*
For any $t \in [T]$, given E_t^{true}, the following event is controlled

$$\mathbb{P}(\mathrm{E}_t^{var} | \mathcal{F}_{t-1}) \geq 1 - \frac{1}{t^2}$$

Proof. Firstly, if $t = 1$, the condition is obvious because $\mathbb{P}(\mathrm{E}_t^{\text{var}} | \mathcal{F}_{t-1}) \geq 0$. For the rest of the proof, we assume that $t \geq 2$. Recall the definition of the event $\mathrm{E}_t^{\text{var}}$:

$$\mathrm{E}_t^{\text{var}} = \left\{ \text{for any } a \in \mathcal{A}(x_t), |\phi(x_t,a)^\top \tilde{\theta}_t - \phi(x_t,a)^\top \hat{\mu}_t| \leq g_2(t) \|\phi(x_t,a)\|_{V_t^{-1}} \right\} .$$

with $g_2(t) = 10\sqrt{d \log(3t^3)/(\eta\lambda)}$.
Let $a \in \mathcal{A}(x_t)$, it holds that

$$|\phi(x_t,a)^\top (\tilde{\theta}_t - \hat{\mu}_t)| \leq |\phi(x_t,a)^\top (\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t})| + |\phi(x_t,a)^\top (\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| + |\phi(x_t,a)^\top (\tilde{m}_{t,K_t} - \hat{\mu})| ,$$

where $\tilde{\theta}_t = \tilde{\theta}_{t,K_t}$ is a sample from the variational posterior distribution trained after K_t steps of Algorithm 4. $\tilde{\mu}_{t,K_t}$ and $\tilde{\Sigma}_{t,K_t}$ are, respectively, the mean and covariance matrix of the variational posterior. Moreover, $\tilde{\mu}_{t,K_t}$ is gaussian with mean $\tilde{m}_{t,K_t}$ and covariance matrix $\tilde{W}_{t,K_t}$ (see Section 6.3). If the number of gradient descent steps is $K_t^{(1)} \geq 1 + 4(1+2\kappa_t^2)\log(2T)/3$, then Lemma 116 shows that with probability at least $1 - 1/t^2$, we have

$$\begin{aligned} |\phi(x_t,a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| &\leq \sqrt{4d\log(t^3)/\eta}\left(1 + 1/\sqrt{\lambda}\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\ &\leq 4\sqrt{d\log(t^3)/(\eta\lambda)}\|\phi(x_t,a)\|_{V_t^{-1}}, \end{aligned}$$

where the last inequality is due to $\lambda \leq 1$.

Similarly, Lemma 119 shows that for any $t \geq 2$, given $\mathrm{E}_t^{\text{true}}$, if $K_t^2 \geq 1 + 4(1+2\kappa_t^2)\log\left(2\kappa_t d^2 T^2 \log^2(3T^3)\right)/3$, therefore we have

$$|\phi(x_t,a)^\top \left(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t}\right)| \leq \left(\sqrt{\frac{3}{\eta\lambda d\log(3t^3)}} + \sqrt{4d\log(3t^3)/(11\eta)}\right)\|\phi(x_t,a)\|_{V_t^{-1}}$$

where in the last simplification we have used $\lambda \leq 1$.

Finally, Given $\mathrm{E}_t^{\text{true}}$, let's apply Lemma 112 with a number of gradient descent steps such $K_t^{(3)} \geq 1 + 2(1 + 2\kappa_t^2)\log\left(4R\sqrt{dT\log(3T^3)}\right)$, we obtain that

$$|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq 2/\lambda \|\phi(x_t,a)\|_{V_t^{-1}}.$$

Note that $K_t = 1 + 2(1+2\kappa_t^2)\log\left(2R\kappa_t d^2 T^2 \log^2(3T^3)\right) \geq \max\{K_t^{(1)}, K_t^{(2)}, K_t^{(3)}\}$ (see Equation (8.24)), then with probability at least $1 - 1/t^2$ we have

$$\begin{aligned} |\phi(x_t,a)^\top \left(\tilde{\theta}_{t,k} - \hat{\mu}_t\right)| &\leq |\phi(x_t,a)^\top \left(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k}\right)| + |\phi(x_t,a)^\top \left(\tilde{\mu}_{t,k} - \tilde{m}_{t,k}\right)| + |\phi(x_t,a)^\top \left(\tilde{m}_{t,k} - \hat{\mu}\right)| \\ &\leq \left(4\sqrt{d\log(t^3)/(\eta\lambda)} + \sqrt{\frac{3}{\eta\lambda d\log(3t^3)}} + \sqrt{4d\log(3t^3)/(11\eta)} + 2/\lambda\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\ &\leq 10\sqrt{d\log(3t^3)/(\eta\lambda)}\|\phi(x_t,a)\|_{V_t^{-1}} \\ &\leq g_2(t)\|\phi(x_t,a)\|_{V_t^{-1}} . \end{aligned}$$

where the last inequality holds because $t \geq 2$, $\lambda \leq 1$ and $\eta \leq 1$. □

Lemma 121. *(Anti-concentration lemma)*

Given E_t^{true}*, if the number of gradient steps is* $K_t = 1 + 2(1+\kappa_t^2)\log\Big(2R\kappa_t d^2T^2\log^2(3T^3)\Big)$ *Therefore, it holds that*

$$\mathbb{P}\Big(\overset{*}{\phi}_t^\top\tilde{\theta}_{t,k} > \overset{*}{\phi}_t^\top\theta^\star\Big) \leq p\,,$$

where $p = 1/(2\sqrt{2\pi e})$

Proof. Firstly, note that

$$\mathbb{P}\Big(\overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} > \overset{*}{\phi}_t^\top\theta^\star\Big) = \mathbb{P}\left(\frac{\overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}}{\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_{t,K_t}\overset{*}{\phi}_t + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_t}} > \frac{\overset{*}{\phi}_t^\top\theta^\star - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}}{\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_t\overset{*}{\phi}_t + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_t}}\right)\,.$$

Recall that

$$\overset{*}{\phi}_t^\top\tilde{\mu}_t \sim \mathbb{N}\big(\overset{*}{\phi}_t^\top\tilde{m}_t,\ \overset{*}{\phi}_t\tilde{W}_{t,k}\overset{*}{\phi}_t^\top\big) \text{ and } \quad \overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} \sim \mathbb{N}\big(\overset{*}{\phi}_t^\top\tilde{\mu}_{t,k},\ \overset{*}{\phi}_t^\top\tilde{\Sigma}_t\overset{*}{\phi}_t\big)\,.$$

Therefore, using the conditional property of Gaussian vectors, we have

$$\overset{*}{\phi}_t^\top\tilde{\theta}_t \sim \mathbb{N}\big(\overset{*}{\phi}_t^\top\tilde{m}_t,\ \overset{*}{\phi}_t\tilde{\Sigma}_t\overset{*}{\phi}_t^\top + \overset{*}{\phi}_t\tilde{W}_t\overset{*}{\phi}_t^\top\big)\,.$$

Consequently, we have to control the term

$$Y_t := \big(\overset{*}{\phi}_t^\top\theta^\star - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}\big)/\big(\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_{t,K_t}\overset{*}{\phi}_{t,K_t} + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_{t,K_t}}\big)$$

and use the Gaussian anti-concentration lemma (Lemma 124). First, in this lemma, we suppose that $\mathrm{E}_t^{\text{true}}$ holds, therefore we have

$$\begin{aligned}|\overset{*}{\phi}_t^\top(\hat{\mu}_t - \theta^\star)| &\leq g_1(t)\|\overset{*}{\phi}_t\|_{V_t^{-1}} \\ &= \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda}\right)\|\overset{*}{\phi}_t\|_{V_t^{-1}}\,.\end{aligned}$$

Moreover, as the number of gradient descent, defined in Section 6.2 is upper than $K_t^{(1)} = 1 + 2(1 + 2\kappa_i^2)\log\Big(4R\sqrt{dT\log(3T^3)}\Big)$, then Lemma 112 gives us that

$$|\overset{*}{\phi}_t^\top(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \frac{2\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\lambda}.$$

Consequently, the numerator of Y_t is upper bounded by

$$\begin{aligned}|\phi_t^{*\top}(\theta^\star - \tilde{m}_{t,K_t})| &\overset{a}{\leq} |\overset{*}{\phi}_t^\top(\theta^\star - \hat{\mu}_{t,K_t})| + |\overset{*}{\phi}_t^\top(\hat{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| \\ &\overset{(b)}{\leq} \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda} + \frac{2}{\lambda}\right)\|\overset{*}{\phi}_t^\top\|_{V_t^{-1}}\end{aligned}$$

Regarding the denominator of Y_t, we need a lower bound for $\|B_{t,k}\overset{*}{\phi}_t\|_2$. Lemma 114 for gives us that

$$\|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\overset{*}{\phi}_t\|_{V_t^{-1}}$$

with

$$-C_t^{1/2} = -\Big(\frac{1}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t))^{K_i - 1}\Big)^{1/2}\,. \tag{8.44}$$

Finally, we find a lower bound to this quantity.

$$
\begin{aligned}
\|B_{t,K_t}\phi\|_2 &\geq 1/\sqrt{\eta}\Big(1-\sqrt{\|V_t\|_2 C_t}\Big)\|\overset{*}{\phi}_t\|_{V_t^{-1}} \\
&\overset{(a)}{=} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\Big(1-\sqrt{\|V_t\|_2}\Big(\frac{1}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\big(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t)\big)^{K_i-1}\Big)^{1/2}\Big) \\
&\overset{(b)}{=} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\left(1-\Big(\sum_{j=1}^{t-1}\epsilon^{t-j-1}\Big)^{1/2}\right) \\
&\overset{(c)}{=} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\left(1-\Big(\sum_{j=0}^{t-2}\epsilon^{j}\Big)^{1/2}\right) \\
&\overset{(d)}{\geq} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\Big(1-\frac{1}{9t^{1/4}}\Big) \\
&\overset{(e)}{\geq} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\Big(1-\frac{1}{9}\Big) \\
&= \frac{8\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{9\sqrt{\eta}}
\end{aligned}
$$

with (a) is 8.44. Where (b) we use $\|V_t\|_2 \leq t$ and setting $\epsilon = (4t)^{-1}$, point (c) comes from a change of variable, (d) comes from the fact that for any $t \geq 1$, $\sum_{j=0}^{t-2}\epsilon^j < 1/(81\sqrt{t})$. Finally, (e) comes from that $1/t$ by can be upper bounded by 1 for any t.

Finally, regrouping the nominator and the denominator, we have the following expression for Y_t:

$$
\begin{aligned}
Y_t &\leq \frac{\overset{*}{\phi}_t^\top\theta^\star - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}}{\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_{t,K_t}\overset{*}{\phi}_{t,K_t} + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_{t,K_t}}} \\
&\leq \frac{\overset{*}{\phi}_t^\top\theta^\star - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}}{\Big\|\overset{*}{\phi}_t\Big\|_{\tilde{\Sigma}_{t,K_t}}} \\
&\leq \frac{R\sqrt{d\log(3t^3)} + \sqrt{\lambda} + \frac{2}{\lambda}}{8/(9\sqrt{\eta})} \\
&\leq \frac{9R\sqrt{d\log(3t^3)}\sqrt{\eta}}{2\lambda}
\end{aligned}
$$

Recall the definition of η in Section 6.2

$$
\eta = \frac{4\lambda^2}{81R^2 d\log(3T^3)}
$$

Consequently, it yields that $|Y_t| \leq 1$.

Finally, Lemma 124 gives us that

$$
\mathbb{P}\Big(\overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} > \overset{*}{\phi}_t^\top\theta^\star\Big) \geq \frac{1}{2\sqrt{2\pi e}}
$$

□

6.7 Auxiliary Lemmas

Lemma 122. *(Azuma-Hoeffding inequality) We define $\{X_s\}_{s\in[T]}$ a super-martingale associated to the filtration $\mathcal{F}_t$. If it holds that for any $s \geq 1$, $|X_{s+1} - X_s| \leq c_{s+1}$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}(X_T - X_0 \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sum_{s=1}^{T} c_s^2}\right).$$

Lemma 123. *(Martingale Lemma [Abbasi-Yadkori et al., 2011b]) Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration, $(m_t)_{t\geq 1}$ be an $\mathbb{R}^d$-valued stochastic process such that m_t is $(\mathcal{F}'_{t-1})$-measurable, $(\epsilon_t)_{t\geq 1}$ be a real-valued martingale difference process such that ϵ_t is $(\mathcal{F}'_t)$-measurable. For $t \geq 0$, define $\zeta_t = \sum_{\tau=1}^{t} m_\tau \epsilon_\tau$ and $M_t = \mathrm{I}_d + \sum_{\tau=1}^{t} m_\tau m_\tau^\top$, where I_d is the d-dimensional identity matrix. Assume ϵ_t is conditionally R-sub-Gaussian. Then, for any $\delta' > 0$, $t \geq 0$, with probability at least $1 - \delta'$,*

$$\|\zeta_t\|_{M_t^{-1}} \leq R\sqrt{d\log\left(\frac{t+1}{\delta'}\right)}$$

where $\|\zeta_t\|_{M_t^{-1}} = \sqrt{\zeta_t^\top M_t^{-1} \zeta_t}$

Lemma 124. *(Gaussian concentration [Abramowitz and Stegun, 1964]) Suppose Z is a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$. For $0 \leq z \leq 1$, we have*

$$\mathbb{P}(Z > \mu + z\sigma) \geq \frac{1}{\sqrt{8\pi}} e^{-\frac{z^2}{2}}, \quad \mathbb{P}(Z < \mu - z\sigma) \geq \frac{1}{\sqrt{8\pi}} e^{-\frac{z^2}{2}} \tag{8.45}$$

And for $z \geq 1$, we have

$$\frac{e^{-z^2/2}}{2z\sqrt{\pi}} \leq \mathbb{P}(|Z - \mu| > z\sigma) \leq \frac{e^{-\frac{z^2}{2}}}{z\sqrt{\pi}}$$

7 Approximation of our algorithm and complexity

In this section, the objective is to approximate the inversion of the matrix $B_{t,k}$ of Algorithm 4. Ideed, Algorithm 4, requires to compute the inversion of a $d \times d$ matrix at each step t and k, which represents a complexity of $\mathcal{O}(d^3)$. In the approximated version of Algorithm 4, we consider both the sequence of square root covariance matrix $\{B_{t,k}\}_{k=1}^{K_t}$ and the sequence of their approximations $\{C_{t,k}\}_{k=1}^{K_t}$ such that, for any $t \in [T]$ and $k \in [K_t]$

$$C_{t,k} \approx B_{t,k}^{-1}.$$

Recall the recursive definition of $B_{t,k}$,

$$\begin{aligned} B_{t,k+1} &= \left\{\mathrm{I}_d - h_t A_{t,k}\right\} B_{t,k} + h_t (B_{t,k}^\top)^{-1} \\ &\approx \left\{\mathrm{I}_d - h_t A_{t,k}\right\} B_{t,k} + h_t C_{t,k}^\top, \end{aligned} \tag{8.46}$$

where $A_{t,k} = B_{t,k}^2(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t(\tilde{\theta}_{t,k})^\top$ if the hessian free algorithm is used or $A_{t,k} = \nabla^2 U(\tilde{\theta}_{t,k})$ otherwise. Recall that $\tilde{\theta}_{t,k} \sim \mathcal{N}(\tilde{\mu}_{t,k}, B_{t,k}B_{t,k}^\top)$. Furthermore, we can now focus on the definition of the sequence $\{C_{t,k}\}_{k=1}^{K_t}$. Firstly, we recall that

$$\begin{aligned} B_{t,k+1} &= \left\{\mathrm{I}_d - h_t A_{t,k}\right\} B_{t,k} + h_t (B_{t,k}^\top)^{-1} \\ &= \{\mathrm{I}_d + h_t((B_{t,k}^\top)^{-1}(B_{t,k})^{-1} - A_{t,k})\} B_{t,k}. \end{aligned}$$

Then, let's use a first order Taylor expansion of the previous equation in h_t, we obtain the approximated inverse square root covariance matrix:

$$C_{t,k+1} = C_{t,k}^{-1}\{\mathrm{I}_d - h_t(C_{t,k}^\top C_{t,k} - A_{t,k})\}. \tag{8.47}$$

Note that the lower is h_t, the better is the approximation and in our case the step size h_t is decreasing with t. The approximated recursive definition of the square root covariance matrix defined in equation (8.46) and its approximated inverse defined in equation (8.47) are used to defined our the approximated version of `VITS` called `VITS − II` and is presented in Algorithm 5. Moreover, note that the updating step of Algorithm 5 uses only matrix multiplication and sampling from independent Gaussian distribution $\mathcal{N}(0, \mathrm{I}_d)$. Therefore the global complexity of the overall algorithm is $\mathcal{O}(d^2)$.

8 Discussion on the difference between the algorithm of VTS and our algorithm VITS.

The main difference between our setting and the one of [Urteaga and Wiggins, 2018] (VTS) is the bandit modelisation. Indeed, given a context x and an action a, in our setting, the agent receives a reward $r \sim \mathrm{R}(\cdot|x, a)$. Consequently, a parametric model R_θ is used to approximate the reward distribution and it yields to a posterior distribution $\hat{\mathrm{p}}$. In the setting of [Urteaga and Wiggins, 2018], the agent receives a reward $r \sim \mathrm{R}_a(\cdot|x)$. Then, it considers a set of parametric models $\{\mathrm{R}_{\theta_a}\}_{a=1}^K$ and a set of posterior distributions: $\{\hat{\mathrm{p}}_a\}_{a=1}^K$. The setting we have used in this paper is richer as it consider the correlation between the arms distributions compared to [Urteaga and Wiggins, 2018] which consider that the arm distributions are independents. For example, if we consider the case of the Linear bandit. In this setting, the posterior distribution is Gaussian. With the modelisation of [Urteaga and Wiggins, 2018], we have for any $a \in [K]$, $\hat{\mathrm{p}}_a := \mathcal{N}(\mu_a, \Sigma_a)$, where $\mu_a \in \mathbb{R}^d$ and $\Sigma_a \in \mathcal{S}_+^d$. However, with our modelisation, $\hat{\mathrm{p}} := \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^{d\times K}$ and $\Sigma \in \mathcal{S}_+^{d\times K}$. We can see that the covariance matrix Σ encodes the correlations between the different arms, which is not the case of $\{\Sigma_a\}_{a=1}^K$. In addition, in our setting, we can consider any model for the mean of the reward distribution. For example we can choose $g(\theta, x, a)$ as a Neural Networks. This kind of model is unusable in the formulation of [Urteaga and Wiggins, 2018].

Moreover, the approximate families used in both papers are different. Indeed, we consider the family of non-degenerate Gaussian distributions, and [Urteaga and Wiggins, 2018] is focused on the family of mixture of mean-field Gaussian distribution. The mixture of Gaussian distribution is richer than the classic Gaussian distribution. However, the non mean-field hypothesis allow to keep the correlation between arms distributions.

Furthermore, `VTS` from [Urteaga and Wiggins, 2018] scales very poorly with the size of the problem. The variational parameters are very large: $\alpha \in \mathbb{R}^{K\times M}$, $\beta \in \mathbb{R}^{K\times M}$, $\gamma \in \mathbb{R}^{K\times M}$, $u \in \mathbb{R}^{K\times M\times d}$, $V \in \mathbb{R}^{K\times M\times d\times d}$ where K is the number of arms, M is the number of mixtures and d the parameter dimension. In addition, the parameter updating step is also very costly in term of memory and speed. We have re-implemented an efficient version of their algorithm in JAX in order to scale as much as possible but many memory problems occur.

Finally, our algorithm comes with theoretical guarantees in the Linear Bandit case and outperforms empirically the others approximate TS methods. `VTS` performs poorly in practice and has no theoretical guarantee, even in the Linear case.

9 Additional details about numerical settings

9.1 Hyper-parameters tuning

This Subsection summarizes the different grid-search used to compute all plots in this paper for the algorithms: LinTS, LMC-TS, `VITS − I`, `VITS − II` and `VITS − II Hessian-free`.

Parameter	Value
inverse temperature η	$10, 100, 500, 1000$
regularization λ	$0.1, 1, 10$

Tab. 8.1 LinTS hyperparameter grid-search

Parameter	Value
inverse temperature η	$10, 100, 500, 1000$
regularization λ	$0.1, 1, 10$
Nb gradient steps K_t	10, 50
learning rate h	$0.001, 0.01, 0.1$

Tab. 8.2 LMC-TS hyperparameter grid-search

Parameter	Value
inverse temperature η	$10, 100, 500, 1000$
regularization λ	$0.1, 1, 10$
Nb gradient steps K_t	10
learning rate h	$0.001/\eta, 0.01/\eta, 0.1/\eta$
Monte Carlo samples	1 (`Hessian`) and 20 (`Hessian-free`)

Tab. 8.3 `VITS` hyperparameter grid-search

9.2 Details about experiences in synthetic contextual bandits with synthetic data

In this subsection, we provide more details about the toy example derive in this paper. Firstly, we consider a fixed pool of arms denoted as $P = [\tilde{x}_1, \ldots, \tilde{x}_n]$ with $n = 50$, where each arm $\tilde{x}_i$ follows a normal distribution $\mathcal{N}(0_d, \mathrm{I}_d)$. Then, at each step $t \in [T]$, for every arm, we randomly sample a vector $\tilde{x}_i$ from the pool P, and the contextual vector associated with this arm is defined as $x = \tilde{x}_i + \zeta\epsilon$, where $\epsilon \sim \mathcal{N}(0_d, \mathrm{I}_d)$. The bandit environment is simulated using a random vector $\theta^\star$ sampled from a normal distribution $\mathcal{N}(0_d, \sigma^\star \mathrm{I}_d)$. We opted for $\sigma^\star = 1/d$ to ensure that the variance of the scalar product $x^\top\theta^\star$ remains independent of the dimension d. Indeed, both linear and quadratic settings, the reward only depends on the scalar product between the context and the true parameter. If we denote by $x[i]$ and $\theta^\star[i]$ the i^{th} coordinate of the vector x and $\theta^\star$ respectively, then the scalar product is defined by

$$x^\top\theta^\star = \sum_{i=1}^{d} x[i]\theta^\star[i],$$

and its variance is

$$\begin{aligned}\mathbb{V}[x^\top\theta^\star] &= \mathbb{V}[\sum_{i=1}^{d} x[i]\theta^\star[i]] \\ &= \sum_{i=1}^{d} \mathbb{V}[x[i]]\mathbb{V}[\theta^\star[i]] \\ &= d\sigma^\star\, \mathbb{V}[x[i]].\end{aligned}$$

In the previous equations we have used that all coordinates are independents identically distributed and centered. Therefore, taking $\sigma^\star = 1/d$ ensure that the variance of the scalar product remains independent of d. In the linear bandit setting, the reward depends linearly on the contextual vector x, more precisely,

$$r = x^\top\theta^\star + \alpha\epsilon\,,$$

where $\epsilon \sim \mathcal{N}(0_d, \mathrm{I}_d)$. However, to maintain problem complexity independent of ζ, we have set the signal-to-noise ratio to a fixed value of 1. This signal-to-noise ratio is the ratio between $\mathbb{E}[(x^\top\theta^\star)^2]$ and $\mathbb{E}[(\alpha\epsilon)^2]$. Firstly,

$$\begin{aligned}\mathbb{E}[(x^\top\theta^\star)^2] &= \mathbb{V}[x^\top\theta^\star] \\ &= \mathbb{V}[x[i]] \\ &= 1 + \zeta^2\,,\end{aligned}$$

where in the last equation we have used that $x = \tilde{x}_i + \zeta\epsilon$ and $\mathbb{V}[x[i]] = 1 + \zeta^2$. Moreover, the denominator of the signal-to-noise ratio is $\mathbb{E}[(\alpha\epsilon)^2] = \alpha^2$. Consequently, a signal-to-noise ratio equals to 1 implies that $\sqrt{1+\zeta^2} = \alpha$.

In the quadratic bandit setting, the reward depends quadratically on the contextual vector x, more precisely,

$$r = (x^\top \theta^\star)^2 + \alpha\epsilon\,,$$

where $\epsilon \sim \mathcal{N}(0, \mathrm{I}_d)$. In this setting, the reward also depends only on the scalar product between x and $\theta^\star$, thus, we also choose $\sigma^\star = 1/d$. We also ensure a signal-to-noise equal to 1, it implies a more sophisticated condition on the noise: $\alpha = (\zeta^2+1)\sqrt{3+6/d}$. More precisely, in the quadratic setting, the signal-to-noise ratio is defined as follow

$$\frac{\mathbb{E}[(x^\top\theta^\star)^4]}{\mathbb{E}[(\alpha\epsilon)^2]} = 1.$$

Firstly,

$$\begin{aligned}
\mathbb{E}[(x^\top\theta^\star)^4] &= \mathbb{E}[(\sum_{i=1}^d x[i]\theta^\star[i])^4] \\
&= \mathbb{E}\Bigg[\sum_{i=1}^d (x[i]\theta^\star[i])^4 + 4\sum_{i=1}^d\sum_{j\neq i}(x[i]\theta^\star[i])^3 x[j]\theta^\star[j] + 6\sum_{i=1}^d\sum_{j<i}(x[i]\theta^\star[i])^2(x[j]\theta^\star[j])^2 \\
&+ 12\sum_{i=1}^d\sum_{j\neq i}\sum_{k\neq i,k<j}(x[i]\theta^\star[i])^2 x[j]\theta^\star[j]x[k]\theta^\star[k] \\
&+ 24\sum_{i=1}^d\sum_{j<i}\sum_{k<j}\sum_{l<k} x[i]\theta^\star[i]x[j]\theta^\star[j]x[k]\theta^\star[k]x[l]\theta^\star[l]\Bigg] \\
&= \sum_{i=1}^d \mathbb{E}[x[i]^4]\mathbb{E}[\theta^\star[i]^4] + 6\sum_{i=1}^d\sum_{j<i}\mathbb{E}[x_i^2]\mathbb{E}[x_j^2]\mathbb{E}[\theta^\star[i]^2]\mathbb{E}[\theta^\star[j]^2] \\
&= \frac{9(\zeta^2+1)^2}{d} + 6\binom{d}{2}\frac{(\zeta^2+1)^2}{d^2} \\
&= (\zeta^2+1)^2(\frac{9}{d} + \frac{3(d-1)}{d}) \\
&= (\zeta^2+1)^2(\frac{6}{d}+3)
\end{aligned}$$

which gives that $\alpha = (\zeta^2+1)\sqrt{3+6/d}$

9.3 Computational Power

In this work, we use GPUs v100–16g or v100–32g for running our code with GPU Nvidia Tesla V100 SXM2 16 Go and CPUs with 192 Go per node.

10 Additional numerical experiments

10.1 Experimental comparison between Langevin Monte Carlo and VI

In this Subsection, we conduct an experimental comparison between Langevin Monte Carlo (LMC) and two variants of Variational Inference (VI), denoted as VI-I and VI-II, in approximating a specific target distribution. Our target distribution is a straightforward Gaussian distribution, represented as $\mathrm{p}_\star = \mathcal{N}(\mu_\star, \Sigma_\star)$. We perform LMC, VI-I, and VI-II for a designated number of iterations. In each iteration, we calculate the Kullback-Leibler distance between the approximated distribution and the target distribution. In this context, all distributions generated by LMC, VI-I, and VI-II take the form of Gaussians. To compute the mean and covariance matrix for LMC, we perform parameter

averaging over the results obtained after 1000 burn-in steps (which are excluded from the plotted data). Then, the training is stopped when

$$\mathrm{KL}(q_k, \mathrm{p}_\star) \leq \epsilon \,, \tag{8.48}$$

or if the number of steps exceeds 50000 steps.

Figure 8.4 illustrates the relationship between the condition number of $\Sigma_\star$ and the number of steps needed to achieve (8.48). We conducted these experiments with three different step sizes and repeated them across 100 different seeds. The red dashed line in the figure represents the maximum allowable number of iterations.

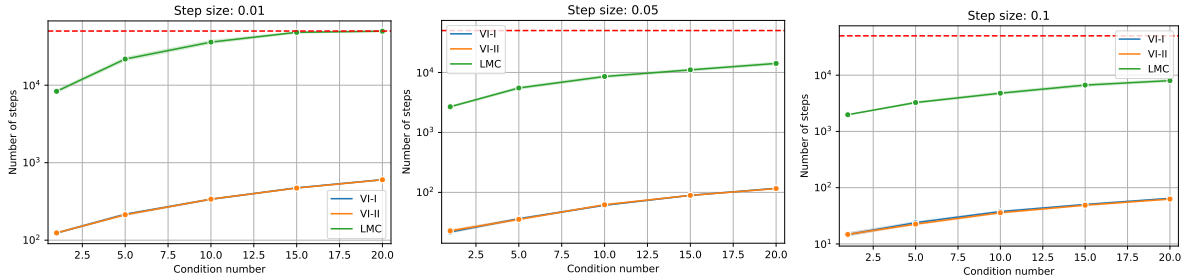


Fig. 8.4 Comparison Langevin Monte Carlo and Variational inference

The first observation drawn from these figures is that VI-I and VI-II exhibit identical behavior, even when using a relatively large step size of 0.1. The second finding suggests that both LMC and VI exhibit a linear dependency on the condition number. However, we cannot definitively conclude that one algorithm is more robust in the face of varying condition numbers. Lastly, the third conclusion highlights that VI consistently requires fewer iterations to achieve (8.48).

10.2 Additional Results on non-contextual bandits

In this subsection, we consider a contextual bandit setting with a parameter dimension $d = 10$ and a number of arms $K = 10$. The bandit environment is simulated by a random vector $\theta^\star \in \mathbb{R}^d$ sampled from a normal distribution $\mathcal{N}(0, \mathrm{I}_d)$ and subsequently scaled to unit norm. To create a complex environment that necessitates exploration, we define the set of contextual vectors as $\mathsf{X} := \{\theta^\star, \theta^\star_\epsilon, x_2, \ldots, x_K\}$. Here, $\theta^\star_\epsilon$ is defined as $\theta^\star_\epsilon = (\theta^\star + \epsilon)/\|(\theta^\star + \epsilon)\|_2$, where ϵ is sampled from a normal distribution with mean 0 and standard deviation 0.1. This contextual vector corresponds to a small modification of $\theta^\star$. The other contextual vectors are sampled from a normal distribution $\mathcal{N}(0, 1)$ and then scaled to unit norm.

Linear bandit scenario. Here, the true reward $\mathcal{R}(\cdot|x_a, a)$ associated to an action $a \in \{1, \ldots, K\}$ and an arm $x_a \in \mathbb{R}^d$ corresponds to the distribution of $r_a = x_a^\top \theta^\star + \xi$, where the noise ξ is sampled from $\mathcal{N}(0, \mathrm{I}_d)$. In this complex setting, we can calculate the expected reward for each arm as follows: $\mu_0 = \mathbb{E}[r_0] = 1$, $\mu_1 \approx 1 < \mu_0$, and for any $i > 1$, $\mu_i < \mu_1$. Intuitively, the first and second arms offered high rewards, while the remaining arms offered low rewards. On the other hand, finding the optimal arm is challenging and needs a significant amount of exploration.

Logistic bandit framework. We consider the same contextual set X, but the true reward $\mathcal{R}(\cdot|x_a, a)$ associated to an action $a \in \{1, \ldots, K\}$ and an arm x_a now corresponds to $r_a \sim \mathrm{Ber}(\sigma(\langle x_a, \theta^\star \rangle))$, where Ber is the Bernoulli distribution, and $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. Similarly to the linear bandit, the logistic framework introduces a complex environment where a significant amount of exploration is required to accurately distinguish between the first and second arm.

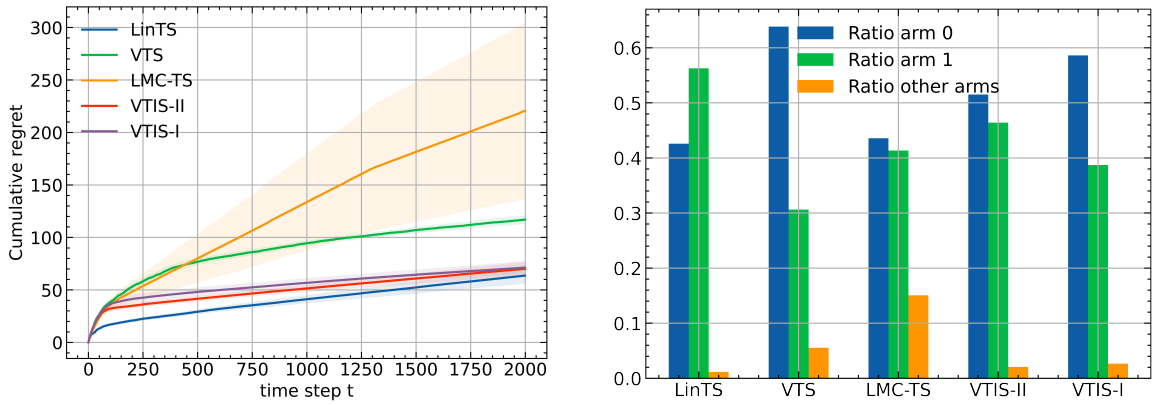


Fig. 8.5 Linear Bandits

Figures 8.5 and 8.6 display the cumulative regret (8.3) obtained by various TS algorithms, namely Linear TS (LinTS), Langevin Monte Carlo TS (LMC-TS), Variational TS (VTS), `VITS-I` and `VITS-II` in the linear and logistic bandit settings.

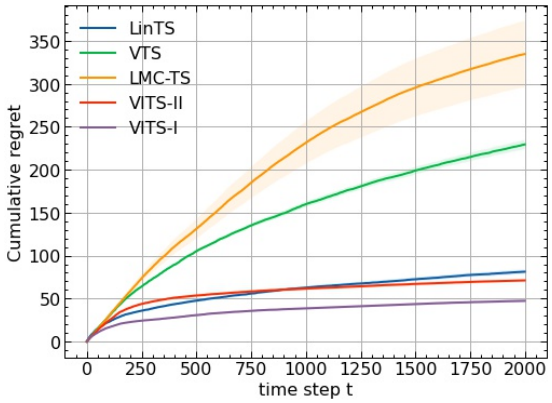


Fig. 8.6 Logistic Bandits

The figures show the mean and standard error of the cumulative regret over 20 samples. As depicted in Figure 8.5, `VITS-I` outperforms the other approximate TS algorithms in the linear bandit scenario. Note that the cumulative regret of `VITS-I` and `VITS-II` is comparable to that of Lin-TS, which uses the true posterior distribution. This observation highlights the efficiency of the variational TS algorithms in approximating the true posterior distribution and achieving similar performance to the Lin-TS algorithm. Figure 8.6 shows that `VITS` outperforms all other TS algorithms in the logistic setting too. This highlights the importance of employing approximation techniques in scenarios where the true posterior distribution cannot be sampled exactly. Moreover, both figures illustrate that `VITS-II` achieves a comparable regret to `VITS-I` while significantly reducing the computational complexity of the algorithm. Finally, as emphasized earlier, the settings we have chosen require a good tradeoff between exploration and exploitation that LMC-TS cannot achieve, as illustrated by the histogram in Figure 8.5.

10.3 Computation complexity

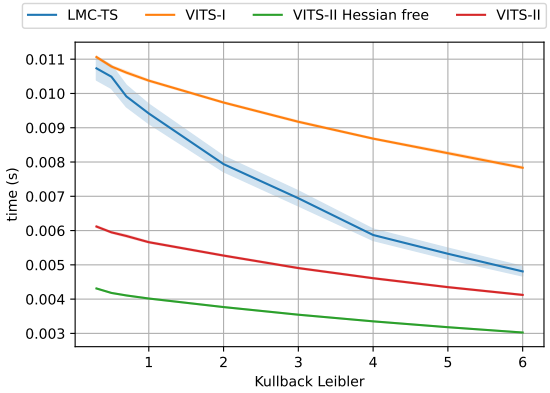
We conduct an experimental comparison between Langevin Monte Carlo (LMC) and three variants of Variational Inference, denoted as `VITS` $-$ `I`, `VITS` $-$ `II` and `VITS` $-$ `II Hessian-free`, in approximating a specific target distribution. Our target distribution is a straightforward Gaussian distribution, represented as $p^\star = \mathcal{N}(\mu^\star, \Sigma^\star)$. At each iteration, we calculate the Kullback-Leibler distance between the approximated distribution and the target distribution. In this context, all distributions generated by LMCTS, `VITS` $-$ `I`, `VITS` $-$ `II` and `VITS` $-$ `II Hessian-free` take the form of Gaussians. To compute the mean and covariance matrix for LMC, we perform parameter averaging. As

both the posterior and its approximation are Gaussians, the Kullback-Leibler divergence is easily tractable. Then, the training is stopped when

$$\mathrm{KL}(q_k, p^\star) \leq \epsilon$$

or if the number of steps exceeds 10000 steps.

The following Figure illustrates the relationship between the obtained Kullback-Leibler divergence and the computational time needed to achieve 10.3. The computational time is the total time (in second) required to run all updating steps of the algorithm. This experiment is repeated across 1000 different seeds to compute the confidence interval. We decide not to compare with LinTS or LinUCB algorithms as they do not allow to approximate complex posteriors compared to LMCTS and VITS algorithms.



This figure shows that `VITS-II` and `VITS-II Hessian-free` are faster (in term of computational time) than LMCTS to obtain a certain Kullback-Leibler divergence. Note that `VITS-I` is the slowest algorithm, this is due to the costly inverse matrix calculation.

CHAPTER 9

FEEL-GOOD THOMPSON SAMPLING VIA LANGEVIN MONTE CARLO

Chapter abstract: *In this chapter, we consider high dimensional contextual bandit problems. Within this setting, Thompson Sampling and its variants have been proposed and successfully applied to multiple machine learning problems. Existing theory on Thompson Sampling shows that it has suboptimal dimension dependency in contrast to upper confidence bound (UCB) algorithms. To circumvent this issue and obtain optimal regret bounds, [Zhang, 2022b] recently proposed to modify Thompson Sampling by enforcing more exploration and hence is able to attain optimal regret bounds. Nonetheless, this analysis does not permit tractable implementation in high dimensions. The main challenge therein is the simulation of the posterior samples at each step given the available observations. To overcome this, we propose and analyze the use of Markov Chain Monte Carlo methods. As a corollary, we show that for contextual linear bandits, using Langevin Monte Carlo (LMC) or Metropolis Adjusted Langevin Algorithm (MALA), our algorithm attains optimal regret bounds of $\tilde{\mathcal{O}}(d\sqrt{T})$. Furthermore, we show that this is obtained with $\tilde{\mathcal{O}}(dT^4)$, $\tilde{\mathcal{O}}(dT^2)$ data evaluations respectively for LMC and MALA. Finally, we validate our findings through numerical simulations and show that we outperform vanilla Thompson sampling in high dimensions.*

1 Introduction

Bandit models have proven to be one of the most successful paradigms for decision making in random environments [Robbins, 1952, Katehakis and Veinott, 1987, Berry and Fristedt, 1985, Auer et al., 2002, Lattimore and Szepesvári, 2020]. Formally, it models an agent which for some rounds has to choose between several potential actions. The agent selects each action according to its current policy and receives a reward once this action is made. In this paper, we are especially interested in the contextual bandit problem [Langford and Zhang, 2007b] which supposes that the set of actions at each round and the corresponding reward mean function depend on a context vector which is specified by the environment under consideration. This setting has been developed and studied intensively over the past decade [Langford and Zhang, 2007b, Filippi et al., 2010, Abbasi-Yadkori et al., 2011b, Chu et al., 2011, Agrawal and Goyal, 2013b, Li et al., 2017b, Lale et al., 2019, Kveton et al., 2020a] and has been successfully applied in various fields; see e.g. for applications in content recommendation, mobile health and finance [Li et al., 2010, Agarwal et al., 2016b, Tewari and Murphy, 2017, Bouneffouf et al., 2020]. To address this problem, bandits algorithms deal with the research and design of efficient algorithms that seek to optimize the cumulative reward. To this end, they recursively define a sequence of policies which is adjusted at each round given the previous historical state-action-reward tuples. The main challenge towards the adaptation and implementation of these policies is to

find a compromise between (1) exploitation of the arms with good empirical expected rewards and (2) exploration of the worse arms with under-sampled rewards.

The approaches to maximizing cumulative reward (alternatively, minimizing cumulative regret) can be broadly divided into two categories. Maximum likelihood methods with optimistic adjustment (UCB) follow the principle of optimism in the face of uncertainty and were adopted in [Auer et al., 2002, Ménard and Garivier, 2017, Chu et al., 2011, Abbasi-Yadkori et al., 2011b, Li et al., 2017b, Zhou et al., 2020, Zenati et al., 2022, Foster and Rakhlin, 2020]. The second approach is based on the Bayesian paradigm, and involves the sampling of a sequence of posterior distributions associated with a statistical model for the reward function; see e.g. , [Thompson, 1933, Agrawal and Goyal, 2012, Kaufmann et al., 2012b, Russo and Van Roy, 2016, 2014, Jin et al., 2021b]. Thompson Sampling (TS) is one of the most famous algorithms that fall into this latter category. Both of these aim to inject uncertainty into the model in order to encourage "exploration"-type behaviour, and have demonstrated their efficiency and robustness in a wide range of applications. In addition, they come with important theoretical guarantees, complementing each other while providing comparable results empirically; see [Chapelle and Li, 2011]. However, existing regret bounds for TS are often sub-optimal when compared to analogous rates for UCB. In particular, the bounds established in [Agrawal and Goyal, 2012] for Thompson Sampling (TS) applied to linear models are of order $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ where d is the dimension of the model considered and T the time horizon. These bounds are worse by a factor $\sqrt{d}$ than the ones proved in [Dani et al., 2008, Abbasi-Yadkori et al., 2011b] for Linear UCB type algortihms. In fact [Zhang, 2022b] showed that this discrepancy between usual TS and UCB cannot be reduced, providing an instance where regret bounds for usual TS can be lower bounded by $\widetilde{\mathcal{O}}(T)$ whereas results on UCB from [Foster and Rakhlin, 2020] achieve a cumulative regret of order $\widetilde{\mathcal{O}}(\sqrt{KT})$, where K is the number of possible actions. To circumvent this issue, [Zhang, 2022b] proposed to modify the likelihood function in TS by adding a penalty term to enforce more optimistic exploration. In addition, the author was able to show that this version of TS, coined Feel-Good Thompson sampling (FG-TS), comes with an upper bound for the cumulative regret which is of order $\widetilde{\mathcal{O}}(d\sqrt{T})$. This matches the minimax regret lower bound established in [Agarwal et al., 2012].

One defect in the methodology and the analysis of [Zhang, 2022b] is that they do not take into account that the sequence of posterior distributions associated with FG-TS is intractable to sample from in practice, even for linear contextual bandits. This is in contrast to the standard TS algorithm. The objective of the present paper is precisely to fill this gap. To address this problem, we propose the use of Markov Chain Monte Carlo methods at each round to obtain approximate samples from the target posterior distribution.

2 Contextual bandit and Thompson sampling methods

We describe the contextual Bandit framework below. Let X be a contextual set and $\mathcal{A} : \mathsf{X} \to 2^{\mathsf{A}}$ be a set-valued action map, where 2^{A} denotes the power set of the action space A. While we do not assume that A is finite, we suppose $\sup_{x\in\mathsf{X}} \mathrm{Card}(\mathcal{A}(x)) < \infty$. In the sequel, we consider policies $\pi : \mathsf{X} \to \mathsf{A}$ such that for any $x \in \mathsf{X}$, $\pi(x) \in \mathcal{A}(x)$, and π can be either deterministic or random. Given a horizon $T \in \mathbb{N}$, and the past observations $\mathrm{D}_{t-1} = \{(x_s, a_s, r_s)\}_{s<t}$ let the following procedure define the bandit framework:

Contextual bandit process. At each iteration $t \in [T]$ and given D_{t-1}:

- The agent observes a contextual vector $x_t \in \mathsf{X}$;
- The agent chooses a policy π_t from some conditional distribution $\mathbb{Q}_t(\cdot|\mathrm{D}_{t-1})$ and sets its action to $a_t = \pi_t(x_t)$;
- The agent receives a reward r_t with conditional distribution $\mathrm{R}(\cdot|x_t, a_t)$ given D_{t-1} (where R is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$, where R is some subset of $\mathbb{R}$).

Given a sequence of conditionals $\mathbb{Q}_{1:T} = \{\mathbb{Q}_t\}_{t\leq T}$, this process defines a distribution on the sequence of policies $\pi_{1:T} = \{\pi_t\}_{t\leq T}$ still denoted by $\mathbb{Q}_{1:T}$ by abuse of notation.
The bandit problem then consists in finding the conditional $\{\mathbb{Q}_t\}_{t\leq T}$ that minimizes the cumulative regret that we will define below. However, as the reward distribution R is unknown, the agent has to simultaneously learn this distribution and choose the best policy. This is a classical exploitation/exploration problem. First, define the

expected reward under the optimal action:

$$f_*(x) = \max_{a \in \mathcal{A}(x)} \int r \, \mathrm{R}(\mathrm{d}r|x,a) \,, \tag{9.1}$$

and the expected reward under any particular action as the following respectively:

$$f(x,a) = \int r \, \mathrm{R}(\mathrm{d}r|x,a) \,.$$

We define then the regret at time s with respect to a policy π_s and a context x_s as

$$\mathrm{REG}_s^{\pi_s} = f_\star(x_s) - f(x_s, \pi_s(x_s)) \,, \tag{9.2}$$

and finally, we seek to find $\mathbb{Q}_{1:T}$ such that the cumulative regret is minimized

$$\mathrm{CREG}(\mathbb{Q}_{1:T}) = \mathbb{E}_{\pi_{1:T} \sim \mathbb{Q}_{1:T}}[\textstyle\sum_{s \leq T} \mathrm{REG}_s^{\pi_s}] \,. \tag{9.3}$$

Thompson sampling (TS) algorithm is a well known algorithm which achieves this goal, with strong performance in practice. First we present the standard TS framework to highlight its limitations. Firstly, consider the Gaussian parametric model $\{\mathrm{R}_\theta^{(\mathrm{TS})} : \theta \in \mathbb{R}^d\}$ based on $g : \mathbb{R}^d \times \mathsf{X} \times \mathsf{A} \to \mathbb{R}$, where $\mathrm{R}_\theta^{(\mathrm{TS})}(\cdot|x,a)$ is the Gaussian distribution with mean $g(\theta,x,a)$ and variance $1/(2\eta)$ for some $\eta > 0$. For example, in linear contextual bandits [Chu et al., 2011, Abbasi-Yadkori et al., 2011b], $g(\theta,x,a) = \langle a, \theta \rangle$ assuming that $\mathcal{A}(x) \subset \mathbb{R}^d$ for any $x \in \mathsf{X}$. Under the same condition, generalized linear bandits [Filippi et al., 2010, Kveton et al., 2020a] consist of $g(\theta,x,a) = \sigma(\langle a, \theta \rangle)$ for some link function σ. Finally, in neural contextual bandits [Riquelme et al., 2018, Zhou et al., 2020, Xu et al., 2020], g is a neural network taking as input a pair (x,a) and θ stands for the weights of g.

Then, the likelihood function associated with the observations D_t at step t is given by

$$\mathrm{L}_t^{(\mathrm{TS})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t} \ell^{(\mathrm{TS})}(\theta|x_s,a_s,r_s) \right) \,, \tag{9.4}$$

where the negative log-likelihood $\ell^{(\mathrm{TS})}$ is given by

$$\ell^{(\mathrm{TS})}(\theta|x,a,r) = \eta(g(\theta,x,a) - r)^2 \,. \tag{9.5}$$

Then, at each iteration $t \in [T]$, TS considers the policy π_t determined, for any x, by

$$\pi_t^{(\mathrm{TS})}(x) = a^{\theta_t}(x) \tag{9.6}$$

where, for θ_t a sample from the posterior distribution $\mu_t^{(\mathrm{TS})}(\cdot|\mathrm{D}_{t-1})$, $a^{\theta_t}(x)$ is defined as follow

$$a^{\theta_t}(x) = \operatorname*{argmax}_{a \in \mathcal{A}(x)} g(\theta_t,x,a) \,.$$

Here $\mu_t^{(\mathrm{TS})}(\theta|\mathrm{D}_{t-1}) \propto \mathrm{L}_t^{(\mathrm{TS})}(\theta|\mathrm{D}_{t-1})\mathrm{p}_0(\theta)$, where p_0 is the prior on θ. However, as mentioned in [Zhang, 2022b], the classic TS algorithm may yield to sub-optimal cumulative regret. They described a simple example where the cumulative regret defined in (9.3) is linear ($\mathcal{O}(T)$), which is sub-optimal compared to the regret bound of $\mathcal{O}(\sqrt{T \log T})$ achieved in [Foster and Rakhlin, 2020] for UCB models. This behavior comes from the choice of Gaussians as the model, which leads to sub-exploration of the action space.

To overcome this difficulty, [Zhang, 2022b] proposes a new model where the classic negative log-likelihood is replaced by the Feel-Good negative log-likelihood, defined by

$$\ell^{(\mathrm{FG})}(\theta|x,a,r) = \eta(g(\theta,x,a) - r)^2 - \lambda \min(b, g_\star(\theta,x)) \,, \tag{9.7}$$

where λ, η and b are hyperparameters in $\mathbb{R}_+$ and $g_\star(\theta,x) = \max_{a \in \mathcal{A}(x)} g(\theta,x,a)$. Then the Feel-Good Thompson sampling algorithm analysed in [Zhang, 2022b] considers the resulting sequence of likelihoods $\{\mathrm{L}_t^{(\mathrm{FG})}\}_{t \leq T}$ and

sequence of posteriors $\{\mu_t^{(\mathrm{FG})}\}_{t\leq T}$ defined similarly to the classic TS method, and defines the sequence of policies $\{\pi_t^{(\mathrm{FG})}\}_{t\leq T}$ as in (9.6) where this time θ_t is a sample from $\mu_t^{(\mathrm{FG})}(\cdot|\mathrm{D}_{t-1})$.

However, exact sampling from $\mu_t^{(\mathrm{FG})}(\cdot|\mathrm{D}_{t-1})$ is usually not tractable, and MCMC algorithms have to be used in their place. This difficulty is not tackled in [Zhang, 2022b]. Consequently, the main objective and contribution of the present paper is to extend the analysis by considering the additional complexity from using approximate samples of the posteriors. More precisely, we consider a gradient-based MCMC schemes to generate these approximate samples. The non-smoothness of the prior definition raises a challenge to this end. While gradient-based MCMC has been developed to sample from such non-smooth densities, they do not enjoy the same theoretical guarantees as smooth densities. For that reason, we propose to consider a smoothed posterior (sFG-TS) with the negative log-likelihood

$$\ell^{(\mathrm{sFG})}(\theta|x,a,r) = \eta(g(\theta,x,a)-r)^2 - \lambda\left[b-\phi_\varsigma(b-g_\star(\theta,x))\right] , \tag{9.8}$$

with $\phi_\varsigma(u)=\log(1+\exp(\varsigma u))/\varsigma$ for $u\in\mathbb{R}$ and $\varsigma>0$ is a hyperparameter which controls the regularity of $\ell^{(\mathrm{sFG})}$. Through an application of the Bayes theorem, assuming that the prior distribution p_0 is correctly specified, then the posterior distribution at time $t\leq T$ can be defined as

$$\mu_t^{(\mathrm{sFG})}(\theta|\mathrm{D}_{t-1}) \propto \mathrm{e}^{-\sum_{s=1}^{t-1}\ell^{(\mathrm{sFG})}(\theta|x_s,a_s,r_s)}\mathrm{p}_0(\theta) . \tag{9.9}$$

For simplicity, we denote $\mu_{t-1}^{(\mathrm{sFG})}(\theta|\mathrm{D}_{t-1})$ by $\mu_{t-1}^{(\mathrm{sFG})}(\theta)$. With this notation, we present the MCMC-sFG-TS method in Algorithm 6. In this algorithm, the choice of the sequence of initial distributions $\{p_{t,0}\}_{t\geq T}$ and the sequence of Markov kernels $\{\mathrm{K}_t\}_{t\leq T}$ is left arbitrary. Indeed, we first extend the analysis provided in [Zhang, 2022b] to this setting and derive general bounds depending on quantities related to the convergence of Markov chains with Markov kernels $\{\mathrm{K}_t\}_{t\leq T}$ and initialized with $\{p_{t,0}\}_{t\geq T}$. We then illustrate our results by considering two examples of MCMC algorithms in particular, which we provide below.

(1) Langevin Monte Carlo: For a fixed step $t\in[T]$, given the target $\mu_t^{(\mathrm{sFG})}$ and an initial distribution $p_{t,0}$, Langevin Monte Carlo (LMC) follows the Markov chain $\{\theta_{t,k}^{\mathrm{L}}\}_{k=0}^{N_t}$ initialized $\theta_{t,0}^{\mathrm{L}}\sim p_{t,0}$, defined through the recursion:

$$\theta_{t,k+1}^{\mathrm{L}} = \theta_{t,k}^{\mathrm{L}} + \gamma_t \nabla \log \mu_t^{(\mathrm{sFG})}(\theta_{t,k}^{\mathrm{L}}) + \sqrt{2\gamma_t} Z_{t,k} , \tag{9.10}$$

where $N_t\in\mathbb{N}^*$ is a number of iterations, γ_t a step size, and $\{Z_{t,k}\}_{k\in[N_t]}$ are i.i.d. samples from the d-dimensional standard Gaussian. It amounts to choosing the Markov kernel $\mathrm{K}_t^{\mathrm{L}}$ with transition density given for $\theta_0,\theta_1\in\mathbb{R}^d$ by

$$k_t^{\mathrm{L}}(\theta_0,\theta_1) \propto \exp\left[-\|\theta_1-\theta_0+\gamma_t\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_0)\|^2/(4\gamma_t)\right]. \tag{9.11}$$

For better rates [Durmus et al., 2019], in our analysis we consider the final parameter to be the ergodic average after some burn-in time, i.e. $\theta_t^{\mathrm{L}} = 2/N_t \sum_{k=N_t/2}^{N_t}\theta_{t,k}$ for some even N_t.

LMC is the Euler discretization of the overdamped Langevin diffusion [Roberts and Tweedie, 1996b] and is a popular way to sample approximately from a smooth positive target density. The Langevin diffusion is a Markov process associated with solutions to the stochastic differential equation (SDE) $\mathrm{d}\boldsymbol{\theta}_{t,s} = \nabla\log\mu_t^{(\mathrm{sFG})}(\boldsymbol{\theta}_{t,s})\mathrm{d}s + \sqrt{2}\mathrm{d}\mathbf{B}_s$, where $(\mathbf{B}_s)_{s\geq 0}$ is a d-dimensional standard Brownian motion. However, while $\{\boldsymbol{\theta}_{t,s}\}_{s\geq 0}$ admits $\mu_t^{(\mathrm{sFG})}$ as its stationary distribution, this is not the case for the Markov kernel associated with (9.10). Therefore, LMC comes with a bias which is the same order as the stepsize γ_t under appropriate conditions [Talay and Tubaro, 1990, Durmus and Eberle, 2021].

(2) Metropolis Adjusted Langevin Algorithm: To correct the discretization bias of the Langevin SDE, a Metropolis filter can be applied at each iteration as suggested for example in [Roberts and Tweedie, 1996b]. This corresponds to the Metropolis Adjusted Langevin Algorithm (MALA). For technical reasons, we study the $1/2$-lazy version of this algorithm, which defines the Markov chain $\{\theta_{t,k}^{\mathrm{M}}\}_{k=0}^{N_t}$ initialized with $\theta_{t,0}^{\mathrm{M}}\sim p_{t,0}$ following the recursion:

- generate a proposal $\tilde{\theta}^{\mathrm{M}}_{t,k+1} \sim \mathrm{K}^{\mathrm{L}}_t(\theta^{\mathrm{M}}_{t,k}, \cdot)$;
- with probability $1/2\alpha^{\mathrm{M}}_t(\theta^{\mathrm{M}}_{t,k}, \tilde{\theta}^{\mathrm{M}}_{t,k+1})$: set $\theta^{\mathrm{M}}_{t,k+1} = \tilde{\theta}^{\mathrm{M}}_{t,k+1}$
- otherwise set $\theta^{\mathrm{M}}_{t,k+1} = \theta^{\mathrm{M}}_{t,k}$,

where

$$\alpha^{\mathrm{M}}_t(\theta_0, \theta_1) = 1 \wedge \frac{\mu^{(\mathrm{sFG})}_t(\theta_1) k^{\mathrm{L}}_t(\theta_1, \theta_0)}{\mu^{(\mathrm{sFG})}_t(\theta_0) k^{\mathrm{L}}_t(\theta_0, \theta_1)} \,. \tag{9.12}$$

For MALA, we take $\theta_t = \theta_{t,N_t}$ to be the last iterate. We refer to the resulting methods as LMC-sFG-TS (resp. MALA-sFG-TS) in the sequel.

Related Works Appoximate sampling in TS algorithms is in general based on Laplace approximation [Chapelle and Li, 2011], which fits the mean and the covariance matrix of a Gaussian distribution based on the target. This is then used to approximately sample from the posterior. However, high-dimensional Gaussian distribution with general covariance matrices may be expensive to compute. Further, in non-linear models such as generalized linear bandits and neural contextual bandits, the sequence of posteriors may be far from Gaussian distributions and Laplace approximation may fail in capturing their complex properties. Finally, Laplace approximation does not come with any theoretical guarantees on the quality of the resulting approximation.

The use of LMC or Stochastic Gradient Langevin Dynamics in Thompson Sampling for non-contextual bandits has been proposed in [Mazumdar et al., 2020b]. This idea has been recently extended to contextual bandits in [Xu et al., 2022], which introduced LMC-TS. Algorithm 6 extends this method in two ways: (1) by considering the more complex likelihood (9.4), (2) taking as an input the MCMC algorithms which are used to sample in sFG-TS. Finally, [Xu et al., 2022] is only applicable for linear bandits, where the TS posteriors are Gaussian distributions. In contrast, we are able to establish very generic bounds for MCMC-sFG-TS by adapting and extending the FG-TS theory in [Zhang, 2022b]. We specify these results in Section 3.3 to the particular instance of linear bandits, when the MCMC method used in MCMC-sFG-TS is LMC or MALA.

Algorithm 6 MCMC-sFG-TS

Initialize:
$\mathrm{D}_0 = \emptyset$
for $t = 1, \dots T$ **do**
receive $x_t \in \mathsf{X}$
initialize the Markov chain $\theta_{t,0} | \mathrm{D}_{t-1} \sim p_{t,0}$ where $p_{t,0}$ may depend on D_{t-1};
for $k = 0, \dots N_t - 1$ **do**
$\theta_{t,k+1} | \mathrm{D}_{t-1} \sim \mathrm{K}_t(\theta_{t,k}, \cdot)$ where K_t is a Markov kernel which targets $\mu^{(\mathrm{sFG})}_t(\cdot | \mathrm{D}_{t-1})$, e.g., LMC or MALA
end for
choose $\theta_t = F(\{\theta_{t,k}\}_{k \le N_t})$
choose $a_t = \mathrm{argmax}_{a \in \mathcal{A}(x_t)} g(\theta_t, x_t, a)$
receive the reward $r_t \sim \mathrm{R}(\cdot | x_t, a_t)$
end for

3 Main results

3.1 Analysis of MCMC-sFG-TS

We make these assumptions on the reward distribution.

Assumption 10. *(Sub-Gaussian Reward Distribution) There exists $c > 0$ such that for any $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$, $\rho > 0$,*

$$\log \int \exp\{\rho(r - f(x,a))\} \mathrm{R}(\mathrm{d}r | x, a) \le c\rho^2 \,, \tag{9.13}$$

where f is defined in (9.1)*. Furthermore, assume* $\sup_{x \in \mathsf{X}, a \in \mathcal{A}(x)} |f(x,a)| \le b_f$.

Note that Assumption 10 is automatically satisfied if the rewards are bounded almost surely, i.e, for any x and a, $\mathrm{R}(\cdot|x,a)$ has a bounded support.

We state our main result regarding the cumulative regret for MCMC-FG-TS. First recall that we have assumed a finite action set A and therefore we can define $K = \max_{x\in\mathsf{X}} \mathrm{Card}(\mathcal{A}(x))$. Second, we denote by $\tilde{\mathrm{q}}_t^{(\mathrm{sFG})}$ the distribution of θ_t given D_{t-1}, as defined in Algorithm 6, and define for $t \in [T]$, $\delta_t = ||\tilde{\mathrm{q}}_t^{(\mathrm{sFG})} - \mu_t^{(\mathrm{sFG})}||_{\mathrm{TV}}$. Note that the sequence $(x_t, a_t, r_t, \theta_t)_{t=0}^T$, defined in Algorithm 6, is a Markov chain, possibly inhomogeneous, and we define by $\mathbb{E}_{\nu_0}^T$ and $\mathbb{P}_{\nu_0}^T$ the canonical expectation and probability respectively associated with this process and with initial distribution ν_0. Define the filtration $(\mathcal{F}_t)_{t\in[T]}$ by $\mathcal{F}_t = \sigma\{\{x_s, a_s, r_s\}_{s\in[t]}\}$. With this notation, the cumulative regret associated with the distribution $\mathbb{Q}_{1:T}^{(\mathrm{sFG})}$ defined by Algorithm 6 can be written as $\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) = \mathbb{E}_{\nu_0}^T[\sum_{s\leq T} f_\star(x_s) - r_s]$.

Theorem 125. *Assume that Assumption 10 holds and let $\varsigma > 0$. If η is chosen according to* (9.27) *with $\epsilon \in (0,1)$, then there exists C_1, C_1, C_2 and C_3, independent of $\epsilon, \eta, \lambda, d, T, K$ such that*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \tfrac{\lambda}{\eta\epsilon} KT + C_1\lambda T - \tfrac{Z_T}{\lambda} + (C_2 + \tfrac{C_3}{\lambda}) \textstyle\sum_{t=0}^T \mathbb{E}_{\nu_0}^T[\delta_t] \,,$$

where

$$Z_T = \mathbb{E}_{\nu_0}^T \log \int \exp\Big(-\textstyle\sum_{s=1}^T \Delta\ell^{(\mathrm{sFG})}(\tilde{\theta}, x_s, a_s, r_s)\Big) \,\mathrm{dp}_0(\tilde{\theta}) \,,$$

and

$$\Delta\ell^{(\mathrm{sFG})}(\theta, x, a) = \eta\Big\{(g(\theta,x,a) - r)^2 - (f(x,a) - r)^2\Big\} - \lambda\{b - \phi_\varsigma(b - g_\star(\theta,x)) - f_\star(x)\} \,.$$

Proof. We provide here the main steps leading to Theorem 125 based on Lemmas which are stated and proved in Section 6.1 of the supplement.

(A) **Regret decomposition.** The first step of the proof is to decompose the expected regret at time s into two terms as follows

$$\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] = \mathbb{E}_{\nu_0}^T[\mathrm{B}_{x_s}(\theta_s, a^{\theta_s}(x_s))] - \mathbb{E}_{\nu_0}^T[\mathrm{FG}_{x_s}(\theta_s, a^{\theta_s}(x_s))] \tag{9.14}$$

where

- $\mathrm{B}_x : (\theta, a) \to g_b(\theta, x, a) - f(x, a)$,
- $g_b(\theta, x) = \max\{-b,\, \min(b,\, g_\star(\theta, x))\}$,
- $\mathrm{FG}_x \;:\; (\theta, a) \mapsto g_b(\theta, x, a) - f_*(x)$.

On the right hand side, the first term is referred to as the Bellman error in the reinforcement learning literature [Bellman, 1966], and the second one as the Feel-Good exploration term. The proof of the decomposition is provided in Lemma 130.

(B) **Bellman error.** By using Lemma 131 we can bound the Bellman error by

$$\mathbb{E}_{\nu_0}^T[\mathrm{B}_{x_s}(\theta_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \leq \inf_{\gamma>0}\Big(\tfrac{K}{4\gamma} + \gamma\mathbb{E}_{\nu_0}^T[\psi(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}]\Big) \,,$$

where $\psi(x_s, a) = \mathbb{E}_{\nu_0}^T[\mathrm{LS}_{x_s}^b(\theta_s, a)|x_s, \mathcal{F}_{s-1}]$, and

$$\mathrm{LS}_x^b \;:\; (\theta, a) \mapsto (g_b(\theta, x, a) - f(x, a))^2 \,. \tag{9.15}$$

This step allows us to decouple the contribution of the random parameter θ_s and its associated action $a^{\theta_s}(x_s)$ to the Bellman error. In the right hand side, we first take the expectation with respect to the parameter for a fixed action, and then with respect to the random action $a^{\theta_s}(x_s)$. This inequality holds for any $\gamma > 0$, in particular for $\gamma = 2C_\eta/(3\lambda)$, with

$$C_\eta = 1.5\eta(1 - 4c\eta)[1 \; - 0.75\eta(1 - 4c\eta)(b + b_f)^2] \,, \tag{9.16}$$

where c is the sub-Gaussian coefficient and b_f is the supremum of the true reward function, both defined in Assumption 10. Lemma 136 shows that $2C_\eta/(3\lambda)$ is strictly positive. Hence, the Bellman error bound becomes

$$\mathbb{E}^T_{\nu_0}[\mathrm{B}_{x_s}(\theta_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \leq \tfrac{3K\lambda}{8C_\eta} + \tfrac{2C_\eta}{3\lambda}\mathbb{E}^T_{\nu_0}[\psi(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \,. \tag{9.17}$$

In the next step of the proof, we focus on bounding the resulting error $\mathbb{E}^T_{\nu_0}[\psi(x_s, a^{\theta_s}_s(x_s))|x_s, \mathcal{F}_{s-1}]$. More precisely, given D_{s-1}, $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$, Lemma 132 with $\tau = 3\eta(1-4c\eta)/2$ (which is positive according to Lemma 136) gives

$$C_\eta \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}^{(\mathrm{sFG})}_s}[\mathrm{LS}^b_x(\theta, a)] \leq -\log \mathbb{E}_{\theta\sim\mu^{(\mathrm{sFG})}_s}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}] + C_\eta(b+b_f)^2\delta_s \,, \tag{9.18}$$

where LS_x is defined in (9.15), and

$$\mathrm{LS}_x \;:\; (\theta, a) \mapsto (g(\theta, x, a) - f(x, a))^2 \,. \tag{9.19}$$

Next, we will focus on the second term in the regret decomposition (9.14), the Feel-Good exploration term.

(C) **Feel Good exploration term.** Similarly, given D_{s-1}, for any $x \in \mathsf{X}$, Lemma 133 with $\tau = 3\lambda$ gives

$$-\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_s}[\mathrm{FG}_x(\theta, a^\theta(x))] \leq -\tfrac{1}{3\lambda}\log \mathbb{E}_{\theta\sim\mu^{(\mathrm{sFG})}_s}[\mathrm{e}^{3\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] + \tfrac{3\lambda(b+b_f)^2}{2} + (b+b_f)\delta_s \,. \tag{9.20}$$

Now, the Bellman error bound (9.17) and the Feel-Good bound (9.20) can be merged.

(D) **Combining the bounds.** The combination of (9.18) and (9.20) gives

$$\begin{aligned}\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}^{(\mathrm{sFG})}_s}\left[\tfrac{2C_\eta}{3\lambda}\mathrm{LS}^b_x(\theta, a) - \mathrm{FG}_x(\theta, a^\theta(x))\right] \leq &-\frac{2}{3\lambda}\log \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}^{(\mathrm{sFG})}_s}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}] \\ &- \frac{1}{3\lambda}\log \mathbb{E}_{\theta\sim\mu^{(\mathrm{sFG})}_s}[\mathrm{e}^{3\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \\ &+ \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\delta_s + \frac{3\lambda}{2}(b+b_f)^2 \,.\end{aligned}$$

Moreover, given D_{s-1}, we can use Lemma 134 to get for any $x \in \mathsf{X}$ and $a \in \mathcal{A}(x)$,

$$\begin{aligned}\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}^{(\mathrm{sFG})}_s}\left[\tfrac{2C_\eta}{3\lambda}\mathrm{LS}^b_x(\theta, a) - \mathrm{FG}_x(\theta, a^\theta(x))\right] \leq &-\frac{1}{\lambda}\log \mathbb{E}_{\theta\sim\mu^{(\mathrm{sFG})}_s}[\Gamma(a, x)] \\ &+ \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\delta_s + \frac{3\lambda}{2}(b+b_f)^2 \,,\end{aligned} \tag{9.21}$$

setting $\Gamma(a, x) = \mathbb{E}_{r\sim\mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)}]$. We now have all tools to bound the cumulative regret and conclude the proof.

(E) **Cumulative Regret Bound.** Using the regret decomposition (9.14) and the Bellman error bound (9.17), we have

$$\mathbb{E}^T_{\nu_0}[\mathrm{REG}^{\pi_s}_s] \leq \tfrac{3K\lambda}{8C_\eta} + \tfrac{2C_\eta}{3\lambda}\mathbb{E}^T_{\nu_0}\left[\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}^{(\mathrm{sFG})}_s}[\mathrm{LS}^b_{x_s}(\theta, a_s)]\right] - \mathbb{E}^T_{\nu_0}\left[\mathbb{E}_{\theta\sim\tilde{\mathrm{q}}^{(\mathrm{sFG})}_s}[\mathrm{FG}_{x_s}(\theta, a^\theta(x_s))]\right] \,.$$

Then Eq. (9.21) gives

$$\begin{aligned}\mathbb{E}^T_{\nu_0}[\mathrm{REG}^{\pi_s}_s] \leq &\frac{3\lambda K}{8C_\eta} - \frac{1}{\lambda}\mathbb{E}^T_{\nu_0}\left[\log \mathbb{E}_{\theta\sim\mu^{(\mathrm{sFG})}_s}[\Gamma(a_s, x_s)]\right] \\ &+ \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\mathbb{E}^T_{\nu_0}[\delta_s] + \frac{3\lambda}{2}(b+b_f)^2 \,.\end{aligned}$$

Finally, we can use Lemma 135 to get,

$$Z_t - Z_{t-1} \leq \mathbb{E}^T_{\nu_0}\left[\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\Gamma(a_s,x_s)]\right] . \tag{9.22}$$

We conclude the proof by summing over t to get,

$$\begin{aligned}
\mathrm{CREG}(\mathbb{Q}_{1:T}) &= \sum_{s\leq T} \mathbb{E}^T_{\nu_0}[\mathrm{REG}_s^{\pi_s}] \\
&\leq \left[\frac{3\lambda K}{8C_\eta} + \frac{3\lambda}{2}(b+b_f)^2\right] T - \frac{Z_T}{\lambda} + \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right] \sum_{s\leq T} \mathbb{E}^T_{\nu_0}[\delta_s] \\
&\leq \frac{\lambda K T}{\epsilon\eta} + C_1\lambda T + (C_2 + \frac{C_3}{\lambda}) \sum_{s\leq T} \mathbb{E}^T_{\eta_0}[\delta_s] - \frac{Z_T}{\lambda} ,
\end{aligned}$$

where $C_1 = 3(b+b_f)^2/2$, $C_2 = (b+b_f)$ and $C_3 = (b+b_f)^2/4$, these constants do not depend neither on η nor in λ. The last inequality uses Lemmas 137-138.

3.2 Regret Bounds for Bandits

We now specify the bounds provided by Theorem 125 assuming the following condition on the prior distribution p_0 and the family of models $\{(x,a) \mapsto g(\theta,x,a) : \theta \in \mathbb{R}^d\}$.

Assumption 11. *Assume that* $\log \mathrm{p}_0$ *is continuously differentiable,* L_0*-smooth and* m_0*-strongly concave for some* $L_0 \geq m_0 \geq 0$*. This implies that the following holds for all* $\theta_1, \theta_2 \in \mathbb{R}^d$*:*

$$\begin{aligned}
&\|\nabla \log \mathrm{p}_0(\theta_2) - \nabla \log \mathrm{p}_0(\theta_1)\| \leq L_0 \|\theta_1 - \theta_2\| \\
&\langle \nabla \log \mathrm{p}_0(\theta_2) - \nabla \log \mathrm{p}_0(\theta_1), \theta_1 - \theta_2\rangle \geq \frac{m_0}{2} \|\theta_1 - \theta_2\|^2 .
\end{aligned}$$

In addition, we assume that the family of models $\{(x,a) \mapsto g(\theta,x,a) : \theta \in \mathbb{R}^d\}$ is regular enough and close to the true model, in the following senses.

Assumption 12. *(Uniform Smoothness) Suppose that for all* $\theta_1, \theta_2 \in \mathbb{R}^d, x \in \mathsf{X}, a \in \mathcal{A}(x)$*, the following bound holds for some* $L_g \in \mathbb{R}_+$*:*

$$|g(\theta_1,x,a) - g(\theta_2,x,a)| \leq L_g \|\theta_1 - \theta_2\| .$$

Assumption 13. *(Well Specified Model) Suppose that there exist* $\theta_* \in \mathbb{R}^d$ *and* $\xi \in \mathbb{R}_+$ *such that for all* $x \in \mathsf{X}, a \in \mathcal{A}(x)$*:*

$$|g(\theta_*,x,a) - f(x,a)| \leq \xi .$$

Corollary 126. *Let Assumptions 10-13 hold and let* $\varsigma > 0$*. For* ω, η, λ *specified in* (9.28)*, and* T *large enough (specified in* (9.31)*), and for constants* C_4, C_5, C_6 *not dependent on* $\omega, \epsilon, d, K, T$

$$\begin{aligned}
\mathrm{CREG}(\mathbb{Q}^{(\mathrm{sFG})}_{1:T}) \leq{}& \frac{C_4}{\epsilon}\sqrt{\omega d K T \log(dT)} + \big(4\xi + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)\big)T \\
&+ C_5\sqrt{\frac{\omega K T}{d\log(dT)}}\big(-\log \mathrm{p}_0(\theta_*) + L_g + \xi T + \xi^2 T\big) \\
&+ C_6\left(1 + \sqrt{\frac{\omega K T}{d\log(dT)}}\right)\sum_{t=0}^{T} \mathbb{E}^T_{\nu_0}[\delta_t] + 4L_g.
\end{aligned}$$

Here θ_* *is the parameter defined in Assumption 13.*

The proof of this result along with explicit bounds are given in Section 6.2.

3.3 Linear Bandits

A concrete example where Assumptions 11-13 hold is the linear contextual bandits framework:

Example 127. *(Linear Gaussian Function Class) Consider the function class with $f(x,a) = \langle \varphi(x,a), \theta_* \rangle$, with $\theta_* \in \mathbb{R}^d$, $x \in \mathsf{X}, a \in \mathcal{A}(x)$, with $\varphi : \mathsf{X} \times \mathsf{A} \to \mathbb{R}^d$ being some feature map. Let the reward be absolutely bounded by some constant b_r almost surely, and let $\sup_{(x,a)\in \mathsf{X}\times\mathsf{A}(x)} \|\varphi(x,a)\| \le \sqrt{M}$ with $0 < M < \infty$. Finally, let $|\mathcal{A}(x)| \le d$ for all $x \in \mathsf{X}$.*

Remark: The absolute bound on the reward is only needed to guarantee the almost sure complexity bounds on the gradient descent step.

We now define an appropriate notion of complexity, which is different from the typical definition seen in bandit literature.

Definition 128. *(Data Complexity) The agent has access to both the value $g(\theta,x,a)$ and the gradient $\nabla g(\theta,x,a)$ for any $\theta \in \mathbb{R}^d, x \in \mathsf{X}, a \in \mathcal{A}(x)$. Then, if g is evaluated a_t times and ∇g is evaluated b_t times at any timestep t, then we define $G_t = \mathsf{a}_t + \mathsf{b}_t$ as the data complexity at time t, and $\mathrm{CG} = \sum_{t\le T} G_t$ be the cumulative data complexity.*

Theorem 129. *Consider Example 127 with the linear function class $g(\theta,x,a) = \langle \varphi(x,a), \theta \rangle$ and a Gaussian prior $\mathcal{N}(0, \mathtt{m}_0^{-1}\mathrm{I}_d)$, with $\mathtt{m}_0 > 0$. Assume Assumption 10 holds, let $\omega_{\mathsf{LG}}, \lambda, \eta$ be as specified in (9.32), and let T be large enough (specified in (9.33)). Assume in addition let there exist $\kappa > 0$ such that almost surely, for any $t \in [T]$ the Hessian matrix of $-\log \mu_t^{(\mathrm{sFG})}(\theta)$ (9.9) satisfies for some $\mathtt{m}_t, \mathtt{L}_t > 0$:*

$$\mathtt{L}_t \mathrm{I}_d \succeq -\nabla^2 \log \mu_t^{(\mathrm{sFG})}(\theta) \succeq \mathtt{m}_t \mathrm{I}_d \quad , \quad \mathtt{L}_t/\mathtt{m}_t \le \kappa \, . \tag{9.23}$$

(a) Then, starting from an initial point $\hat{\theta}_0^\star = \theta_0$, we can find at each round recursively $\hat{\theta}_t^\star$ satisfying $\|\hat{\theta}_t^\star - \theta_t^\star\| \le \sqrt{d/(2\mathtt{L}_t)}$ using the gradient descent algorithm to maximize $\log \mu_t^{(\mathrm{sFG})}(\theta)$ and initialized with $\hat{\theta}_{t-1}^*$. Here θ_t^* is the maximizer of $\log \mu_t^{(\mathrm{sFG})}(\theta)$. The cumulative data complexity of this procedure is of order $C_{\mathrm{GD}} \kappa T^2 \log\Big(b_r \mathtt{L}_t \sqrt{M} T/\mathtt{m}_0\Big)$, for some absolute constant C_{GD}, and the step size is $2/(\mathtt{L}_t + \mathtt{m}_t)$.

(b) In addition setting $p_{t,0} = \mathcal{N}(\hat{\theta}_t^\star, (\mathtt{L}_t)^{-1}\mathrm{I}_d)$, for any of the following standard choices of Markov kernel, we attain the regret bound for some constant C_7 not dependent on $\omega_{\mathsf{LG}}, \epsilon, d, K, T, M$

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \le C_7 \sqrt{\omega_{\mathsf{LG}} T \log^3(dT)} \Big(d(\epsilon \wedge \mathtt{m}_0)^{-1} + \sqrt{M}\mathtt{m}_0 \|\theta_*\|^2 \Big) \, ,$$

with the number of oracle calls stated below:

- K^{L} *(Langevin Monte Carlo): has* $\mathrm{CG}^{\mathsf{LMC}} \le C_{\mathrm{L}}\, C_\kappa d T^4 \log\Big(4\sqrt{d\kappa}/\mathtt{m}_0\Big)$ *cumulative data complexity, with step-size* $\gamma_t^{\mathrm{L}} = \mathrm{A}_{\mathrm{L}}/(\max(\kappa, \mathtt{L}_t) d T^2)$, $C_\kappa = \max(\mathtt{L}_T/\mathtt{m}_0^2, \mathtt{L}_T)$.
- K^{M} *(Metropolis Adjusted Langevin Monte Carlo): has* $\mathrm{CG}^{\mathsf{MALA}} \le C_{\mathrm{M}} \kappa d T^2 (1 \vee \sqrt{\kappa d^{-1}}) \log\big(dT^2\big)$ *cumulative data complexity, with step-size* $\gamma_t^{\mathrm{M}} = \mathrm{A}_{\mathrm{M}}/(\mathtt{L}_t d \max(1, \sqrt{\kappa d^{-1}}))$

Here C_{L}, C_{M}, $\mathrm{A}_{\mathrm{L}}, \mathrm{A}_{\mathrm{M}}$ are absolute constants depending on which MCMC algorithm was chosen.

The proof of this result along with explicit bounds are given in Section 6.3.

Remarks: We note that the Gaussian prior can be replaced with an arbitrary prior satisfying Assumption 11, so long as a good bound on $\mathrm{p}_0(\theta_*)$ exists.

Now, let's compare Theorem 129 with [Xu et al., 2022, Theorem 4.2]. [Xu et al., 2022, Theorem 4.2] has a bound on the cumulative data complexity for LMC-TS of order κT^2, which is used to obtain a cumulative regret of order $d^{3/2}T^{1/2}$. In contrast, for our results under MALA, we pay an extra factor of d in the cumulative data complexity in order to remove the suboptimal factor of $d^{1/2}$ in the resulting cumulative regret. We see this increased complexity as a necessary cost in order to obtain our tighter regret bounds. It may be possible to more finely balance this trade-off by e.g. annealing the Feel-Good parameter, but we defer this investigation to subsequent work.

4 Experiments

In this section, we illustrate the benefits of our methodology on several contextual bandit benchmarks associated with both synthetic and real data. In our comparisons, we first perform grid searches for the hyperparameters, and then fix the best ones. Additional details about experimental design are provided in Section 7.

4.1 Toy example

We first illustrate our approaches on a synthetic contextual bandit problem. At each round $t \in [T]$, the agent observes a contextual vector sampled from a 4 dimensional Gaussian distribution, i.e., $x_t \sim \mathcal{N}(0_4, \mathrm{I}_4)$. Then, the agent has to choose an action a_t between $K = 5$ arms, and finally, receives a reward $r_t = \varphi(x_t, a_t)^\top \theta^\star + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\theta^\star \in \mathbb{R}^{20}$ is the true parameter of the model, σ is the noise level of the problem. Here φ allows us to transform the context vector and the arm index into a vector v such as, $\varphi(x, 0) = (x, 0, \dots, 0)$, $\varphi(x, 1) = (0, x, 0, \dots)$ and $\varphi(x, d-1) = (0, \dots, 0, x)$. We consider the corresponding model defined as $g(\theta, x, a) = \varphi(x, a)^\top \theta$. Under these settings, note that posterior distributions associated with TS are Gaussian distributions and are therefore tractable.

In Figure 9.1, we compare our methodology MCMC-sFG-TS using LMC and MALA with Linear TS, along with LMC-TS. For completeness, we also consider TS where at each iteration, we approximate the TS posterior (9.4) with MALA. This simply corresponds to MCMC-sFG-TS but choosing $\lambda = 0$. For these results, we only display the best combination of hyperparameters for each algorithm. More details for the experiment settings are provided in Section 7. Note that for MALA-sFG-TS and MALA-TS, we initialize MALA with the output of a gradient descent scheme using full-batch gradient. Moreover, we also consider Linear UCB for which results can be found in Section 7. We observe that adding the Feel-Good framework allow us to converge to a better regret. Similarly, approximating the posterior using MALA seems to improve the algorithmic performance by converging faster to the target. Finally, by combining the Feel-Good adjustment with MALA, we obtain MALA-sFG-TS which provides the best cumulative regret.

Similar conclusions are drawn on different bandit settings, including logistic and quadratic bandits trained with benchmark algorithms; see Section 7.

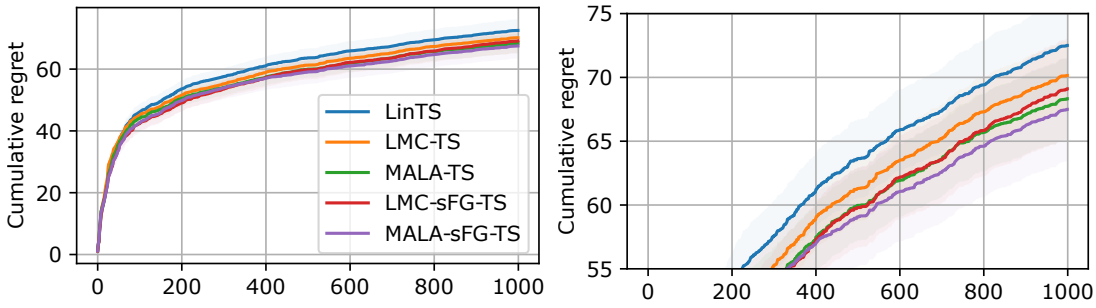


Fig. 9.1 Cumulative regret with respect to the time step t, for the toy example. Whole curve (left) and its zoomed version (right) are represented. Statistics are reported over 50 runs.

4.2 Real-World dataset

In this subsection, we compare the algorithms on the Yahoo! Front Page Today Module dataset, which is a standard benchmark for contextual bandits [Li et al., 2010, Mellor and Shapiro, 2013, Liu et al., 2018]. This seeks to model a user's interest in a specified news article using the contextual bandit framework. At each round, we consider a user and a pool of articles. Here, the context is composed by a user-features vector and user-article interaction information. In addition, the set of arms is the pool of articles. Then, given a current bandit model, we choose an article and check if it is clicked. If so, a reward of 1 is incurred; otherwise, the reward is 0. With this definition and our bandit formulation, we seek here to maximize the average expected cumulative reward $T^{-1}\mathbb{E}_{\Pi \sim \mathbb{Q}_{1:T}}[\sum_{t=1}^{T} f(x_t, \pi_s(x_t))]$, which is precisely the click-through rate (CTR) in [Li et al., 2010]. A more detailed description on the implementation can be found in [Li et al., 2010].

In our experiments, we consider just a subset of 500 thousand recommendations made the 3^{th} of May 2009, with the statistics reported over 10 trials. For each run the dataset is shuffled.

In Figure 9.2 we compare the different approaches using their relative CTR, which is the algorithm's CTR divided by that of a baseline random policy. It can be seen that LMC-sFG-TS and MALA-sFG-TS deliver the best recommendations amongst their competitors.

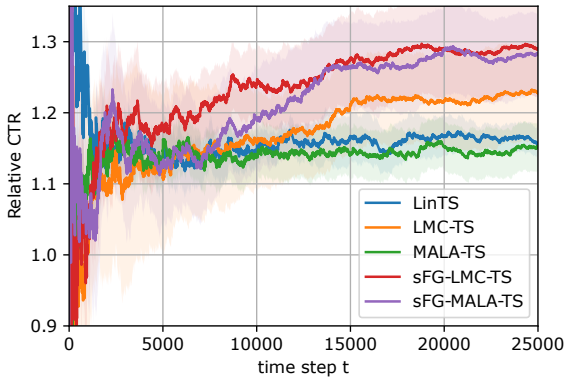


Fig. 9.2 Relative CTR for the Yahoo recommendation task

5 Conclusion

In this work we proposed and analyzed the MCMC-sFG-TS algorithm for contextual bandits, which is a tractable implementation of Thompson sampling with an optimistic Feel-Good adjustment term. We showed that this obtains the optimal regret bound of $\tilde{\mathcal{O}}(d\sqrt{T})$ in high dimensions, in contrast to the $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ that was previously known for MCMC algorithms in the Thompson sampling setting. We also validated the superior performance of this algorithm in practice, relative to the standard Thompson sampling.

Further extensions to our approach include non-quadratic log-likelihoods, which would extend our results to classes such as logistic bandits and bandits with generalized linear models. Finally, applying our framework to some classes of reinforcement learning problems would be an important step towards a general understanding of Thompson sampling algorithms in that setting.

6 Main Proofs

6.1 Proof of Theorem 125

Lemma 130. *(Regret decomposition) The regret at time s can be decomposed into two terms as follows*

$$\mathbb{E}^T_{\nu_0}[\mathrm{REG}^{\pi_s}_s] = \mathbb{E}^T_{\nu_0}[g_b(\theta, x_s, a^\theta_s(x_s)) - f_\star(x_s, a^\theta_s(x_s))] - \mathbb{E}^T_{\nu_0}[g_b(\theta, x_s, a^\theta_s(x_s)) - f_\star(x_s)]\,.$$

Proof. Using the definition of $\mathrm{REG}^{\pi_s}_s$ in (9.2) and the definition of policy π_s in (9.6), we have

$$\begin{aligned}\mathbb{E}^T_{\nu_0}[\mathrm{REG}^{\pi_s}_s] &= \mathbb{E}^T_{\nu_0}[f_\star(x_s) - f(x_s, \pi_s(x_s))] \\ &= \mathbb{E}^T_{\nu_0}[f_\star(x_s) - f(x_s, a^\theta_s(x_s))] \\ &= \mathbb{E}^T_{\nu_0}[g_b(\theta, x_s, a^\theta_s(x_s)) - f(x_s, a^\theta_s(x_s))] - \mathbb{E}^T_{\nu_0}[g_b(\theta, x_s, a^\theta_s(x_s)) - f_\star(x_s)]\,.\end{aligned}$$

□

Lemma 131. *Let $b > 0$. Then, we have the following decoupling bound*

$$\mathbb{E}^T_{\nu_0}[g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \le \inf_{\gamma>0}\left(K/(4\gamma) + \gamma\mathbb{E}^T_{\nu_0}[\psi_{x_s}(a^{\theta_s}_s(x_s))|x_s, \mathcal{F}_{s-1}]\right)\,,$$

where $\psi_{x_s}(a) = \mathbb{E}^T_{\nu_0}[\mathrm{LS}^b_{x_s}(\theta_s, a)|x_s, \mathcal{F}_{s-1}]$.

Proof. Note first that

$$\begin{aligned}
&\mathbb{E}^T_{\nu_0}[g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \\
&\quad \leq \mathbb{E}^T_{\nu_0}[|g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))||\mathcal{F}_{s-1}, x_s] \\
&\quad = \sum_{a \in \mathcal{A}(x_s)} \mathbb{E}^T_{\nu_0}[\mathbb{1}\{a^{\theta_s}(x_s) = a\}|g_b(\theta_s, x_s, a) - f(x_s, a)||\mathcal{F}_{s-1}, x_s] \,. \qquad (9.24)
\end{aligned}$$

Consider for any $\tilde{a} \in \mathcal{A}(x_s)$, $\mathbf{q}(\tilde{a}|x_s) = \mathbb{E}^T_{\nu_0}[\mathbb{1}\{a^{\theta_s}(x_s) = \tilde{a}\}|\mathcal{F}_{s-1}, x_s]$. Then for any $\gamma > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{\nu_0}[\mathbb{1}\{a^{\theta_s}(x_s) = a\}|g_b(\theta_s, x_s, a) - f(x_s, a)||\mathcal{F}_{s-1}, x_s] \\
&\qquad \leq \mathbb{E}^T_{\nu_0}[\frac{\mathbb{1}\{a^{\theta_s}(x_s) = a\}}{4\gamma \mathbf{q}(a|x_s)} + \gamma \mathbf{q}(a|x_s)(g_b(\theta_s, x_s, a) - f(x_s, a))^2|\mathcal{F}_{s-1}, x_s] \\
&\qquad = 1/(4\gamma) + \gamma \mathbf{q}(a|x_s)\mathbb{E}^T_{\nu_0}[(g_b(\theta_s, x_s, a) - f(x_s, a))^2|\mathcal{F}_{s-1}, x_s]
\end{aligned}$$

where the inequality comes from the algebraic inequality $z_1 \cdot z_2 \leq z_1^2/2 + z_2^2/2$ and the last equality from the definition of the distribution $\mathbf{q}$. Plugging the previous inequality in (9.24), and using that for any $x \in \mathsf{X}, \mathrm{Card}(\mathcal{A}(x)) \leq K$, then we have

$$\begin{aligned}
&\mathbb{E}^T_{\nu_0}[g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))|\mathcal{F}_{s-1}, x_s] \\
&\qquad \leq K/(4\gamma) + \gamma \sum_{a \in \mathcal{A}(x_s)} \mathbf{q}(a|x_s)\mathbb{E}^T_{\nu_0}[(g_b(\theta_s, x_s, a) - f(x_s, a_s))^2|\mathcal{F}_{s-1}, x_s] \\
&\qquad = K/(4\gamma) + \gamma \mathbb{E}^T_{\nu_0}[\psi(x_s, a_s^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \,.
\end{aligned}$$

□

Lemma 132. *Assume Assumption 10. Given* D_{s-1}*, for any* $x \in \mathsf{X}$*,* $a \in \mathcal{A}(x)$ *and* $\tau > 0$*, it holds*

$$C_\tau \mathbb{E}_{\theta \sim \tilde{\mathrm{q}}_s}[\mathrm{LS}^b_x(\theta, a)] \leq -\log \mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta, a)\}] + C_\tau (b + b_f)^2 \delta_s \,,$$

where

$$C_\tau = \tau[1 - \tau(b + b_f)^2/2] \,.$$

Proof. Since for any $z \leq 0$, we have $\exp z \leq z^2/2 + z + 1$, we obtain for any $\tau > 0$,

$$\begin{aligned}
\mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta, a)\}] &\leq \mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}^b_x(\theta, a)\}] \\
&\leq -\tau \mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\mathrm{LS}^b_x(\theta, a)] + \frac{\tau^2}{2}\mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\mathrm{LS}^b_x(\theta, a)^2] + 1 \\
&\leq -\tau[1 - \frac{\tau(b + b_f)^2}{2}]\mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\mathrm{LS}^b_x(\theta, a)] + 1 \\
&\leq -C_\tau \mathbb{E}_{\theta \sim \tilde{\mathrm{q}}_s}[\mathrm{LS}^b_x(\theta, a)] + 1 + C_\tau (b + b_f)^2 \delta_s \,,
\end{aligned}$$

where the first inequality uses $\mathrm{LS}_x(\theta, a) \geq \mathrm{LS}^b_x(\theta, a)$, third inequality $\mathrm{LS}^b_x \leq (b + b_f)^2$ and the last inequality the definition of the total variation distance. Moreover, using $\log z \leq z - 1$ for $z \leq 1$, we have,

$$\begin{aligned}
\log \mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta, a)\}] &\leq \mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta, a)\}] - 1 \\
&\leq -C_\tau \mathbb{E}_{\theta \sim \tilde{\mathrm{q}}_s}[\mathrm{LS}^b_x(\theta, a)] + C_\tau (b + b_f)^2 \delta_s \,.
\end{aligned}$$

□

Lemma 133. *Assume Assumption 10. Given* D_{s-1}*, for any* $x \in \mathsf{X}$*,* $a \in \mathcal{A}(x)$ *and* $\tau > 0$*, the Feel-Good exploration term is bounded as follows*

$$-\mathbb{E}_{\theta \sim \tilde{\mathrm{q}}_s}[\mathrm{FG}_x(\theta, a^\theta(x))] \leq -\frac{1}{\tau} \log \mathbb{E}_{\theta \sim \mu_s^{(\mathrm{sFG})}}[\exp\Big(\tau \mathrm{FG}_x(\theta, a^\theta(x))\Big)] + \frac{\tau}{2}(b + b_f)^2 + (b + b_f)\delta_s \,.$$

Proof. Using Hoeffding's lemma since $\mathrm{FG}_x(\theta, a^\theta(x)) \in [-(b+b_f), (b+b_f)]$, for any $\tau > 0$, we have

$$\begin{aligned}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{\tau \mathrm{FG}_x(\theta, a^\theta(x))\}] &\leq \tau \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{FG}_x(\theta, a^\theta(x))] + \frac{\tau^2}{2}(b+b_f)^2 \\ &\leq \tau \mathbb{E}_{\theta\sim\tilde{\mathrm{q}}_s}[\mathrm{FG}_x(\theta, a^\theta(x))] + \frac{\tau^2}{2}(b+b_f)^2 + \tau(b+b_f)\delta_s \,,\end{aligned}$$

where the second line uses the definition of the total variation distance. □

Lemma 134. *Assume Assumption 10. Given* D_{s-1}*, for any* $x \in \mathsf{X}$*, and* $a \in \mathcal{A}(x)$*,*

$$\begin{aligned}-\frac{2}{3}\log\left(\mathbb{E}_{\theta\sim\mu_s(\mathrm{sFG})}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}]\right) - \frac{1}{3}\log\left(\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda\mathrm{FG}_x(\theta,a^\theta(x))}]\right) \\ \leq -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})},\, r\sim\mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)}] \,,\end{aligned}$$

where $\Delta\ell^{(\mathrm{sFG})}(\theta, x, a, r)$ *is defined in* (125)*.*

Proof. Firstly, we can apply the Hölder's inequality with $p = 3/2$ and $q = 3$:

$$\begin{aligned}&\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)+\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \\ &\quad\leq \frac{2}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}] + \frac{1}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \,.\end{aligned} \tag{9.25}$$

Subsequently, by Assumption 10 with $\rho = 2\eta(f(x,a) - g(\theta,x,a))$, if we denote $\epsilon = r - f(x,a)$, we find that: $\exists\, c > 0$ such that

$$\begin{aligned}\int \exp\{-2\eta(f(x,a) - g(\theta,x,a))\epsilon\}\mathrm{R}(\mathrm{d}r|x,a) &\leq \exp\{4c\eta^2(f(x,a) - g(\theta,x,a))^2\} \\ &= \exp\{4c\eta^2 \mathrm{LS}_x(\theta,a)\} \,.\end{aligned}$$

Recall the definition of $\Delta\ell^{(\mathrm{sFG})}$ in (125). Then,

$$\begin{aligned}-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r) &= -\eta(\epsilon + f(x,a) - g(\theta,x,a))^2 + \eta\epsilon^2 + \lambda(b - \phi_\varsigma(b, g_\star(\theta,x)) - f_\star(x)) \\ &= -2\eta\epsilon(f(x,a) - g(\theta,x,a)) - \eta(f(x,a) - g(\theta,x,a))^2 + \lambda(b - \phi_\varsigma(b, g_\star(\theta,x)) - f_\star(x)) \\ &\leq -2\eta\epsilon(f(x,a) - g(\theta,x,a)) - \eta(f(x,a_s) - g(\theta,x,a))^2 + \lambda(g_b(\theta,x) - f_\star(x)) \\ &= -2\eta\epsilon(f(x,a) - g(\theta,x,a)) - \eta\mathrm{LS}_x(\theta,a) + \lambda\mathrm{FG}_x(\theta,a^\theta(x)) \,.\end{aligned}$$

Combining the sub-Gaussian equation with (9.25) and the bound of $-\Delta\ell^{(\mathrm{sFG})}$, we find

$$\begin{aligned}-\Delta\ell^{(\mathrm{sFG})} &\leq -\frac{2}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a_s)/2}] - \frac{1}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \\ &\leq -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})},\, r\sim\mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-2\eta(f(x,a)-g(\theta,x,a))\epsilon - \eta\mathrm{LS}_x(\theta,a) + \lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \\ &\leq -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})},\, r\sim\mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)}] \,.\end{aligned}$$

□

Lemma 135.

$$Z_t - Z_{t-1} \leq \mathbb{E}_{\nu_0}^T\left[\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathbb{E}_{r\sim R(\cdot|x_s,a_s)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x_s,a_s,r)}]]\right]$$

where

$$Z_t = \mathbb{E}_{\nu_0}^T \log \int \exp\left(-\sum_{s=1}^t \Delta\ell^{(\mathrm{sFG})}(\tilde{\theta}, x_s, a_s, r_s)\right) \mathrm{d}\mathrm{p}_0(\tilde{\theta}) \,, \tag{9.26}$$

Proof. The proof is provided in [Zhang, 2022b] but has been rewritten for completeness.

For ease of notation, let define $K_t(\theta|\mathrm{D}_t) = \exp\{-\sum_{s=1}^{t}\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)\}$ such that $Z_t = \mathbb{E}_{\nu_0}^{T}[\log \mathbb{E}_{\theta\sim \mathrm{p}_0}[K_t(\theta|\mathrm{D}_t)]]$. Then we have

$$
\begin{aligned}
Z_t - Z_{t-1} &= \mathbb{E}_{\nu_0}^{T} \log \frac{\mathbb{E}_{\theta\sim \mathrm{p}_0}[K_t(\theta|\mathrm{D}_t)]}{\mathbb{E}_{\theta\sim \mathrm{p}_0}[K_{t-1}(\theta|\mathrm{D}_{t-1})]} \\
&= \mathbb{E}_{\nu_0}^{T} \log \mathbb{E}_{\theta\sim \mathrm{p}_0}\left[\frac{K_t(\theta|\mathrm{D}_t)}{\mathbb{E}_{\tilde{\theta}\sim \mathrm{p}_0}[K_{t-1}(\tilde{\theta}|\mathrm{D}_{t-1})]}\right] \\
&= \mathbb{E}_{\nu_0}^{T} \log \mathbb{E}_{\theta\sim \mathrm{p}_0}\left[\frac{K_{t-1}(\theta|\mathrm{D}_{t-1})\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta, x_t, a_t, r_t)}}{\mathbb{E}_{\tilde{\theta}\sim \mathrm{p}_0}[K_{t-1}(\tilde{\theta}|\mathrm{D}_{t-1})]}\right] \\
&= \mathbb{E}_{\nu_0}^{T} \log \mathbb{E}_{\theta\sim \mu_t^{(\mathrm{sFG})}}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta, x_t, a_t, r_t)}] \\
&\leq \mathbb{E}_{\nu_0}^{T}\left[\log \mathbb{E}_{\theta\sim \mu_s^{(\mathrm{sFG})}}[\mathbb{E}_{r\sim R(\cdot|x_s,a_s)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r)}]]\right],
\end{aligned}
$$

where the last line uses Jensen's inequality.

□

6.1.1 Technical Lemmas

Lemma 136. *Let $c > 0$ be given in H 10. If η is chosen according to the following strategy, for any $\epsilon \in (0, 1)$,*

$$
0 < \eta \leq \begin{cases} 3/(16c) & \text{if } \frac{1}{16c^2} \leq \frac{1-\epsilon}{3c(b+b_f)^2} \\ \min\left(\frac{3}{16c}, \frac{1}{8c} - \sqrt{\frac{1}{64c^2} - \frac{1-\epsilon}{3c(b+b_f)^2}}\right) & \text{otherwise.} \end{cases} \tag{9.27}
$$

Then we have these useful properties

(i) $\eta > 0$,

(ii) $\eta \leq 3/(16c) < 1/(4c)$,

(iii) $1 - (3\eta(1 - 4c\eta)(b + b_f)^2)/4 \geq \epsilon$,

(iv) $C_\eta > 0$ *where* C_η *is defined in* (9.16).

Proof. The results for (i) and (ii) are obvious regarding the definition of η in (9.27). Moreover, $\mathrm{P}(\eta) = \eta^2 - \eta/(4c) + (1-\epsilon)/(3c(b+b_f)^2 s)$ is a second order polynomial with determinant $\Delta_\mathrm{P} = 1/(16c^2) - 4(1-\epsilon)/(3c(b+b_f)^2)$. If $\Delta_\mathrm{P} \leq 0 \iff (b+b_f)^2 \leq 64(1-\epsilon)c/3$, then P is always positive on its domain. However, if $\Delta_\mathrm{P} > 0 \iff (b+b_f)^2 > 64(1-\epsilon)c/3$ then P admits two zeros:

$$
\begin{aligned}
x_1 &= \frac{1}{8c} - \sqrt{\frac{1}{64c^2} - \frac{1-\epsilon}{3c(b+b_f)^2}} \geq 0 \\
x_2 &= \frac{1}{8c} + \sqrt{\frac{1}{64c^2} - \frac{1-\epsilon}{3c(b+b_f)^2}} \geq 0
\end{aligned}
$$

As x_1 is obviously positive, by taking $\eta \leq x_1$ we have $\mathrm{P}(\eta)$ positive and then (iii) is true.
Finally, given (i), (ii) and (iii), C_η is obvisouly strictly positive. □

Lemma 137. *If η is chosen according to 9.27, then we have,*

$$\frac{3\lambda KT}{8C_\eta} \leq \frac{\lambda KT}{\epsilon\eta} .$$

Proof. By definition of C_η in (9.16) and using the property (iii) of Lemma 136, then we have

$$\begin{aligned} C_\eta &= 1.5\eta(1-4c\eta)[1-3\eta(1-4c\eta)(b+b_f)^2/4] \\ &\geq 1.5\eta(1-4c\eta)\epsilon \end{aligned}$$

Moreover $\eta \leq 3/(16c)$ we have $1-4c\eta \geq 1/4$. Hence,

$$C_\eta \geq \frac{3\epsilon}{8}\eta .$$

This last inequality concludes the proof. □

Lemma 138. *If η is chosen according to* (9.27)*, then*

$$\frac{2C_\eta(b+b_f)^2}{3\lambda} \leq \frac{(b+b_f)^2}{4\lambda} ,$$

Proof. By definition of C_η in (9.16),

$$\begin{aligned} C_\eta &= 1.5\eta(1-4c\eta)[1-3\eta(1-4c\eta)(b+b_f)^2/4] \\ &\leq 1.5\eta \leq \frac{3}{8} , \end{aligned}$$

where the last inequality comes from (9.27). □

6.2 Proof of Corollary 126

Proof. Hereafter we specify the choice of

$$\omega = D_\eta^{-1} \vee L_g \vee 1, \qquad \eta = \frac{1}{\omega}, \qquad \lambda = \sqrt{\frac{d\log(dT)}{\omega KT}}, \tag{9.28}$$

where D_η is the RHS of equation (9.27). Consider the compact set $\mathrm{B}_\gamma = \{\theta \in \mathbb{R}^d : \|\theta - \theta_*\| \leq \frac{1}{\gamma}\}$ for some $\gamma \geq 1$. By Assumption 11, we know that for any $\theta \in \mathrm{B}_\gamma$, if $\tilde{\theta}_s = (1-s)\theta + s\theta_*$,

$$\begin{aligned} \log \mathrm{p}_0(\theta) - \log \mathrm{p}_0(\theta_*) &\geq -\int_0^1 \left\langle \nabla \log \mathrm{p}_0(\tilde{\theta}_s), \theta_* - \theta \right\rangle \mathrm{d}s \\ &\geq -\int_0^1 \langle \nabla \log \mathrm{p}_0(\theta_*), \theta_* - \theta \rangle \, \mathrm{d}s - L_0 \int_0^1 \|\theta - \theta_*\|^2 \, \mathrm{d}s \\ &\geq -\frac{\|\nabla \log \mathrm{p}_0(\theta_*)\|}{\gamma} - \frac{L_0}{2\gamma^2} . \end{aligned}$$

From Assumptions 12-13, we get for any $\theta \in \mathrm{B}_\gamma$,

$$\sup_{x \in \mathsf{X}, a \in \mathcal{A}(x)} |g(\theta, x, a) - f(x, a)| \leq \frac{L_g}{\gamma} + \xi . \tag{9.29}$$

Consequently for $\theta \in \mathrm{B}_\gamma$, if we let $a_*(x) = \mathrm{argmax}_{a \in \mathcal{A}(x)} f(x, a)$, then we have

$$\begin{aligned} -\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s) \geq & -\eta(g(\theta, x_s, a_s) - f(x_s, a_s))^2 - 2\eta|g(\theta, x_s, a_s) - f(x_s, a_s)||r_s - f(x_s, a_s)| \\ & - \lambda\left(f_*(x_s) - b + \phi_\varsigma(b - g_*(\theta, x_s))\right) \\ \geq & -\Big(\frac{\eta L_g}{\gamma} + \eta\xi + 2\eta|r_s - f(x_s, a_s)|\Big)(\xi + \frac{L_g}{\gamma}) \\ & - \lambda\left(f_*(x_s) - b + \phi_\varsigma(b - g(\theta, x_s, a^\theta(x_s)))\right) \end{aligned}$$

In the last line, we used (9.29). Now, let's focus on the last term of the previous inequality

$$\begin{aligned} f_\star(x_s) - b + \phi_\varsigma\big(b - g(\theta, x_s, a^\theta(x_s))\big) &= f_\star(x_s) - g(\theta, x_s, a^\theta(x_s)) + \phi_\varsigma\big(g(\theta, x_s, a^\theta(x_s)) - b\big) \\ &\leq \frac{L_g}{\gamma} + \xi + \phi_\varsigma(g(\theta, x_s, a^\theta(x_s)) - b)\,, \end{aligned}$$

In the first line, we used that $\phi_\varsigma(x) = x + \phi_\varsigma(-x)$. The second line comes from 9.29 and that for any $a \in \mathcal{A}(x)$, $f_\star(x) - f_\star(x,a) \leq 0$. Moreover, as ϕ_ς is a growing function, we just have to found an upper bound of $g(\theta, x_s, a^\theta(x_s)) - b$ to bound the previous term.

$$\begin{aligned} g(\theta, x_s, a^\theta(x_s)) - b &= g(\theta, x_s, a^\theta(x_s)) - f_\star(x_s, a^\theta(x_s)) + f_\star(x_s, a^\theta(x_s)) - b \\ &\leq \frac{L_g}{\gamma} + \xi + b_f - b\,. \end{aligned}$$

Consequently,

$$-\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s) \geq -\Big(\frac{\eta L_g}{\gamma} + \eta\xi + \lambda + 2\eta|r_s - f(x_s, a_s)|\Big)(\xi + \frac{L_g}{\gamma}) - \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b)\,.$$

Then, taking expectation and using Assumption 10 to control $\mathbb{E}_{\nu_0}[|r_s - f(x_s, a_s)|] \leq \sqrt{2c}$ (see e.g. [Wainwright, 2019], Theorem 2.6),

$$\begin{aligned} \mathbb{E}[\inf_{\theta \in \mathrm{B}_\gamma} -\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)] &\geq -\left(\eta\left(\frac{L_g}{\gamma} + \xi\right) + \lambda + 2\sqrt{2c}\eta\right)(\xi + \frac{L_g}{\gamma}) - \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b) \\ &\geq -4(1 + \xi + \lambda)(\xi + \frac{L_g}{\gamma}) - \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b). \end{aligned}$$

The last line follows from our choice of η, and $\gamma \geq 1$. Finally, noting that the volume of a d-dimensional ball can be lower bounded by $\exp(-10d\log d)$, we can estimate the probability of B_γ under p_0 with the following

$$\begin{aligned} \log \mathrm{p}_0(\mathrm{B}_\gamma) &\geq \inf_{\theta \in \mathrm{B}_\gamma} \log \mathrm{p}_0(\theta) - 10d\log\gamma d \\ &\geq \log \mathrm{p}_0(\theta_*) - \frac{L_0}{2\gamma^2} - 10d\log(\gamma d) \end{aligned}$$

Then we can bound as follows:

$$\begin{aligned} Z_T &= \mathbb{E}\left[\log \mathbb{E}_{\theta \sim \mathrm{p}_0}\left[\exp\left(-\sum_{s=1}^T \Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)\right)\right]\right] \\ &\geq \mathbb{E}\log\left(\mathrm{p}_0(\mathrm{B}_\gamma) \inf_{\theta \in \mathrm{B}_\gamma} \exp\left(-\sum_{s=1}^T \Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)\right)\right) \\ &\geq \log \mathrm{p}_0(\theta_*) - \frac{L_0}{2\gamma^2} - 10d\log(\gamma d) - \Big(4(1 + \xi + \lambda)(\xi + \frac{L_g}{\gamma}) + \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b)\Big)T, \end{aligned}$$

where in the last step we used our bound on $\mathrm{p}_0(\mathrm{B}_\gamma)$.

Finally, substituting $Z_T, \lambda, \eta, \gamma = T$ into Theorem 125, and expanding the product:

$$\begin{aligned} \mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) &\leq \frac{\lambda}{\eta\epsilon} KT + C_1\lambda T - \frac{Z_T}{\lambda} + \left(C_2 + \frac{C_3}{\lambda}\right)\sum_{t=0}^T \mathbb{E}_{\nu_0}^T[\delta_t] \\ &\leq \frac{\sqrt{\omega dKT\log(dT)}}{\epsilon} + C_1\sqrt{\frac{dT\log(dT)}{\omega K}} + \left(C_2 + C_3\sqrt{\frac{\omega KT}{d\log(dT)}}\right)\sum_{t=0}^T \mathbb{E}_{\nu_0}^T[\delta_t] \\ &\quad + \sqrt{\frac{\omega KT}{d\log(dT)}}\left(-\log \mathrm{p}_0(\theta_*) + \frac{L_0}{2T^2} + 10d\log(dT) + 4L_g\right) \\ &\quad + 4\xi\sqrt{\frac{\omega KT}{d\log(dT)}}\,(T + \xi T + L_g) + 4(\xi T + L_g) + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)T\,. \end{aligned} \tag{9.30}$$

When T satisfies

$$T \geq \sqrt{\frac{L_0}{2d}} \vee L_g \vee e, \tag{9.31}$$

then the following inequalities hold:

$$\frac{L_0}{2T^2} \leq d, \qquad L_g \leq T, \qquad \log T \geq 1.$$

This is a mild assumption and does not impact the viability of the result; the second term is only needed to absorb ξL_g into ξT, and is not necessary when ξ is small.

Consequently, we can make some simplifications to find

$$\begin{aligned}\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \frac{C_4\sqrt{\omega dKT\log(dT)}}{\epsilon} + C_6\left(1+\sqrt{\frac{\omega KT}{d\log(dT)}}\right)\sum_{t=0}^{T}\mathbb{E}_{\nu_0}^{T}[\delta_t] + 4L_g \\ + C_5\sqrt{\frac{\omega KT}{d\log(dT)}}\left(-\log \mathrm{p}_0(\theta_*) + L_g + \xi T + \xi^2 T\right) + \left(4\xi + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)\right)T,\end{aligned}$$

where here we define $C_4 = 1 + 11\epsilon + \epsilon\, C_1/(\omega K) \leq 14 + C_1$, $C_6 = C_2 + C_3$, $C_5 = 8$, such that they can be loosely upper bounded by constants not depending on $\epsilon, \omega, d, K, T$. Note that this restriction on T is dimension-free and quite mild.

□

6.3 Proof of Theorem 129

Let D_η again be the RHS of (9.27). Hereafter we specify the choice of

$$\omega_{\mathsf{LG}} = D_\eta^{-1} \vee \sqrt{M} \vee 1, \qquad \eta = \frac{1}{\omega_{\mathsf{LG}}}, \qquad \lambda = \sqrt{\frac{\log(dT)}{\omega_{\mathsf{LG}} T}}, \qquad \varsigma = \sqrt{T}, \qquad b \geq b_f. \tag{9.32}$$

Secondly, the condition on T is now

$$T \geq e \vee \sqrt{\frac{\mathtt{m}_0}{2d}}, \tag{9.33}$$

since as ξ is zero, the second condition in (9.31) is not necessary. Note that this assumption is not very restrictive on T, especially when the dimension is large.

Lemma 139. *If the MCMC method can output p_{t,N_t} such that $\delta_t \leq \frac{1}{T}$, then we obtain the bound for $C_7 = (C_4 + C_5) \vee C_6$ when the parameters satisfy* (9.32)*,* (9.33)*:*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq C_7\sqrt{\omega_{\mathsf{LG}} T\log^3(dT)}\left(d(\frac{1}{\epsilon} + \frac{1}{\mathtt{m}_0}) + \sqrt{M} + \mathtt{m}_0\left\|\theta_*\right\|^2\right). \tag{9.34}$$

Proof.

The setting of Theorem 129 satisfies all the assumptions of Proposition 126 with $\xi = 0$, $L_g = \sqrt{M}$, $\omega = \omega_{\mathsf{LG}}$. Let us first examine the term $\phi_\varsigma(\sqrt{M}/T + b_f - b)T$ for our choice of ς, b. In this case,

$$\begin{aligned}\phi_\varsigma(\frac{L_g}{T} + b_f - b)T &= \frac{\log\left(1+\exp\left(\sqrt{M}/\sqrt{T} + \sqrt{T}(b_f - b)\right)\right)}{\sqrt{T}} \times T \\ &\leq \sqrt{T}\log\left(1+\exp\left(\sqrt{\frac{M}{T}}\right)\right) \\ &\leq \sqrt{M} + \sqrt{T}.\end{aligned}$$

In the second line we use that $b \geq b_f$, and in the third line we use that $\log(1+\exp(x)) \leq 1+x$ for $x \geq 0$. Subsequently, we get the following bound immediately, using that $K/d \leq 1$:

$$\begin{aligned}\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) &\leq \sqrt{\omega_{\mathsf{LG}} T \log(dT)} \left(2\, C_4 \frac{d}{\epsilon} + 3\, C_5 \sqrt{M} - \, C_5 \log \mathrm{p}_0(\theta_*) + \, C_6 \sum_{t=0}^{T} \mathbb{E}_{\nu_0}^{T}[\delta_t] \right) \\ &\leq \sqrt{\omega_{\mathsf{LG}} T \log(dT)} \left(2\, C_4 \frac{d}{\epsilon} + 3\, C_5 \sqrt{M} + \frac{C_5 \mathtt{m}_0 \left\|\theta_*\right\|^2}{2} + \frac{C_5 d \log 2\pi}{2\mathtt{m}_0} + \, C_6 \sum_{t=0}^{T} \mathbb{E}_{\nu_0}^{T}[\delta_t] \right),\end{aligned}$$

where in the second line we substitute the density of the Gaussian prior. We absorb the $\sqrt{M}$, $\sqrt{T}$ tersm from ϕ_ς into C_4, C_5. Since $4 \leq 2\sqrt{wT\log(dT)}C_5$, the $4L_g$ term in Corollary 126 is absorbed into the $C_5\sqrt{M}$ seen above. If we substitute $\delta_t \leq 1/T$, this last part of the sum can be absorbed as a factor of $\log(T) \leq \log(dT)$, and then we choose $C_7 = (2\, C_4 + 3\, C_5) \vee \, C_6$ to complete the proof.

□

Remark: We can assume instead $K \leq \, C_K d$ for some absolute constant C_K, with this constant subsequently appearing at multiple places in the proof. For ease of presentation, we do not do this.

Consequently, this allows us to use gradient descent to estimate the modes of the successive posteriors with negligible cost (with the previous mode for bootstrapping). We state a theorem for gradient descent which makes this rate rigorous:

Lemma 140 (Adapted from [Nesterov et al., 2018], Theorem 2.1.15)**.** *Given a μ-strongly convex, λ-smooth function g with condition number κ and an initial point θ_0, gradient descent with step-size $2/(\mu+\lambda)$ can find the mode $\theta^* = \mathrm{argmin}_\theta\, g(\theta)$ with rate*

$$N \geq 2\kappa \log\left(\frac{\|\theta_0 - \theta\|}{\epsilon} \right) \implies \|\theta_N - \theta^*\| \leq \epsilon.$$

We will not discuss this result extensively as it is only necessary to furnish a modal estimate for MCMC methods. The use of gradient descent is standard and has been well-studied, e.g. in the aforementioned [Nesterov et al., 2018].

We show a polynomial in time bound on the norms of the iterates, which is crude but sufficient for our purposes.

Lemma 141. *Let θ_t^* be the mode of the posterior $\mu_t^{(\mathrm{sFG})}$. Then the following holds, where b_r is the a.s. bound on the reward:*

$$\|\theta_t^*\| \leq \frac{2t\sqrt{Mt}}{\mathtt{m}_0}\left(\frac{b_r}{\omega_{\mathsf{LG}}} + \lambda\right) \tag{9.35}$$

In particular, we immediately get the crude bound

$$\left\|\theta_t^* - \theta_{t-1}^*\right\| \leq \frac{4t\sqrt{Mt}}{\mathtt{m}_0}\left(\frac{b_r}{\omega_{\mathsf{LG}}} + \lambda\right). \tag{9.36}$$

Proof. First consider the minimizer of the posterior for Thompson sampling without Feel-good adjustment ($\mu_t^{(\mathrm{TS})}$), and denote it by ζ_t^*. Then, since ζ_t^* is just the solution of a regularized least squares problem, we know the following bound on ζ_t^*:

$$\zeta_t^* = \left(\mathbf{\Phi}_t^\top \mathbf{\Phi}_t + \frac{\mathtt{m}_0}{\eta} I_d\right)^{-1} \mathbf{\Phi}_t^\top \mathbf{r}_t, \qquad \mathbf{\Phi}_t = \begin{bmatrix} \varphi(x_1, a_1) \\ \varphi(x_2, a_2) \\ \cdots \\ \varphi(x_t, a_t) \end{bmatrix}, \qquad \mathbf{r}_t = \begin{bmatrix} r_1 \\ r_2 \\ \cdots \\ r_t \end{bmatrix}.$$

Here $\mathbf{\Phi}_t$ is the data matrix which has $\varphi(x_i, a_i)$ in its i-th row. In particular, since the matrix $\mathbf{\Phi}_t^\top \mathbf{\Phi}_t + \frac{\mathtt{m}_0}{\eta}\mathrm{I}_d \succeq \omega_{\mathsf{LG}}\mathtt{m}_0 \mathrm{I}_d$, $\|\mathbf{\Phi}_t\|_2 \leq t\sqrt{M}$ and $\|\mathbf{r_t}\|_2 \leq b_r\sqrt{t}$, we obtain

$$\|\zeta_t^*\| \leq \frac{1}{\omega_{\mathsf{LG}}\mathtt{m}_0} \|\mathbf{\Phi_t}\|_2 \|\mathbf{r_t}\|_2 \leq \frac{b_r t\sqrt{Mt}}{\omega_{\mathsf{LG}}\mathtt{m}_0}. \tag{9.37}$$

Secondly, writing the difference in negative log-likelihoods as:

$$-\log\mu_t^{(\mathrm{TS})}(\theta) = -\log\mu_t^{(\mathrm{sFG})}(\theta) + \underbrace{\lambda\sum_{s=1}^{t}\Big[b - \phi_\varsigma(b - \Big\langle\theta, \varphi(x_s, a^\theta(x_s))\Big\rangle)\Big]}_{\mathsf{J}_t(\theta)}.$$

We now seek to estimate $\|\theta_t^* - \zeta_t^*\|$, using that ζ_t^*, θ_t^* minimize their respective posteriors:

$$\begin{aligned}
0 &= \Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{TS})}(\zeta_t^*)\Big\|^2 \\
&= \Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*) + \nabla\mathsf{J}_t(\zeta_t^*)\Big\|^2 \\
&= \Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + \|\nabla\mathsf{J}_t(\zeta_t^*)\|^2 + 2\Big\langle\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) + \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*), \nabla\mathsf{J}_t(\zeta_t^*)\Big\rangle \\
&\geq \Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + \|\nabla\mathsf{J}_t(\zeta_t^*)\|^2 \\
&\quad - 2\Big|\Big\langle\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*), \nabla\mathsf{J}_t(\zeta_t^*)\Big\rangle\Big|.
\end{aligned}$$

Let us proceed to use Young's inequality $|\langle a, b\rangle| \leq 1/4\,\|a\|^2 + \|b\|^2$, to find

$$\begin{aligned}
&\Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + \|\nabla\mathsf{J}_t(\zeta_t^*)\|^2 \\
&\qquad \leq 2\Big|\Big\langle\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*), \nabla\mathsf{J}_t(\zeta_t^*)\Big\rangle\Big| \\
&\qquad \leq \frac{1}{2}\Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + 2\,\|\nabla\mathsf{J}_t(\zeta_t^*)\|^2.
\end{aligned}$$

After some rearranging, we get

$$\Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 \leq 2\,\|\nabla\mathsf{J}_t(\zeta_t^*)\|^2.$$

We use triangle inequality and the boundedness of φ to get for all $\theta \in \mathbb{R}^d$

$$\begin{aligned}
\|\nabla\mathsf{J}_t(\theta)\| &= \left\|\lambda\sum_{s=1}^{t}\frac{\exp\big(\varsigma(b - \langle\theta, \varphi(x_s, a^\theta(x_s))\rangle)\big)}{\exp\big(\varsigma(b - \langle\theta, \varphi(x_s, a^\theta(x_s))\rangle)\big) + 1}\varphi(x_s, a^\theta(x_s))\right\| \\
&\leq \lambda t\sqrt{M}.
\end{aligned}$$

From the strong convexity of $-\log\mu_t^{(\mathrm{sFG})}$, we get $\Big\|\nabla\log\mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla\log\mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 \geq \mathtt{m}_t^2\,\|\theta_t^* - \zeta_t^*\|^2 \geq \mathtt{m}_0^2\,\|\theta_t^* - \zeta_t^*\|^2$. Finally, this implies

$$\|\theta_t^* - \zeta_t^*\| \leq \frac{\sqrt{2M}\lambda t}{\mathtt{m}_0}.$$

Substituting (9.37) completes the proof. □

Remarks: Much better bounds are possible through more careful analysis, but since it is only necessary to provide very rough bounds (as gradient descent is a fast algorithm), this will suffice for our purposes.

First we formally state the warm-start condition:

Definition 142. *(Warm-Start Condition) Let μ, ν be two distributions on $\mathbb{R}^d$. We say that a distribution μ is a $c_W(\mu, \nu)$ warm-start for another distribution ν if*

$$\sup_{\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)}\frac{\mu(\mathsf{A})}{\nu(\mathsf{A})} \leq c_W(\mu, \nu)\,, \tag{9.38}$$

where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field of $\mathbb{R}^d$.

Finally, we state the consequences of gradient descent for finding appropriate warm-starts for our MCMC methods.

Corollary 143. *Using gradient descent methods, at time t, we can find an approximate mode $\hat{\theta}_t^*$, so that when we construct the prior $p_{t,0} = \mathcal{N}(\hat{\theta}_t^*, (2\mathtt{L}_t^{(\mathrm{sFG})})^{-1}\mathrm{I}_d)$ (with I_d the d-dimensional identity matrix), then*

$$\log c_W(p_{t,0}, \mu_t^{(\mathrm{sFG})}) \leq d\log 2\kappa, \qquad \mathsf{KL}(p_{t,0} \,\|\, \mu_t^{(\mathrm{sFG})}) \leq d\log 2\kappa.$$

is satisfied with only $2\kappa\log^2(8b\mathtt{L}_t^{(\mathrm{sFG})}\sqrt{M}T/\mathtt{m}_0)$ iterations of gradient descent.

Proof. For each time t, we can first estimate $\hat{\theta}_t^*$ using gradient descent from $\hat{\theta}_{t-1}^*$. We choose the desired accuracy to be $\epsilon = \sqrt{d/(2\mathtt{L}_t^{(\mathrm{sFG})})}$ at each time t. Using Lemmas 140 and (9.36), this can be done with number of iterations $4\kappa\log\Big(8b\mathtt{L}_t^{(\mathrm{sFG})}\sqrt{M}T/\mathtt{m}_0\Big)$.

Then, Section 3.2.1 of [Dwivedi et al., 2018] shows that $p_{t,0}$ chosen here attains a warm-start with $c_W(p_{t,0}, \mu_t^{(\mathrm{sFG})}) \leq \exp(d\log(2\kappa))$. Finally, for the KL bound, we need only note that

$$\mathsf{KL}(p \,\|\, q) = \int \log\frac{p}{q}\,\mathrm{d}p \leq \log c_W(p \,\|\, q).$$

□

Remark: Summing the number of iterations over $t \in [T]$, and noting that each iteration of gradient descent is equal to a full pass through the data, this yields $4\kappa T^2\log\Big(8b\mathtt{L}_t^{(\mathrm{sFG})}\sqrt{M}T/\mathtt{m}_0\Big)$ data complexity. This is dominated by the data complexity due to sampling in all cases.

6.3.1 Langevin Monte Carlo

For the result under LMC, we can give the following state-of-the-art rate, following the result of [Durmus et al., 2019].

Lemma 144 (Adapted from [Durmus et al., 2019], Corollary 11)**.** *For targets with condition number $\kappa = \mathtt{L}/\mathtt{m}$, ambient dimension d and error tolerance ϵ, if we take the ergodic distribution of the $N/2$ to N LMC iterates, for some N even, $2/N\sum_{k=N/2}^{N}\theta_k$ with the law of θ_k denoted p_k and the stationary distribution μ, we get*

$$N^{\mathrm{L}} = \frac{C_{\mathrm{L}}\,\tilde{\mathrm{C}}_\kappa d}{\delta^2}\log\frac{2\mathcal{W}_2(p_0 \,\|\, \mu)}{\delta^2} \implies ||\frac{2}{N}\sum_{k=N/2}^{N} p_k - \mu||_{\mathrm{TV}} \leq \delta, \tag{9.39}$$

for some absolute constant C_{L}, with $\tilde{\mathrm{C}}_\kappa = \max(\mathtt{L}/\mathtt{m}^2, \mathtt{L})$ and $\mathcal{W}_2$ is the 2-Wasserstein distance between measures. Here the step size is chosen as

$$\gamma^{\mathrm{L}} = \mathrm{A}_{\mathrm{L}}\frac{\delta_t^2}{(\kappa \vee \mathtt{L})d}. \tag{9.40}$$

where $\mathrm{A}_{\mathrm{L}} > 0$ is an absolute constant.

Secondly, we state a lemma:

Lemma 145 (Talagrand's Inequality, [Bakry et al., 2014] Corollary 9.3.2)**.** *If p is strongly convex with constant α, then $\mathcal{W}_2^2(q \,\|\, p) \leq 2/\alpha\mathsf{KL}(q \,\|\, p)$.*

Finally, we are ready to show the complexity for LMC.

Proof of Proposition 129, LMC : To show the MCMC complexity, it remains only to combine Lemma 145 with Corollary 143. This shows that the Wasserstein term can be bounded

$$\log\Big(2\mathcal{W}_2(p_{t,0} \,\|\, \mu_t^{(\mathrm{sFG})})\Big) \leq \log\left(\frac{2}{\mathtt{m}_t}\sqrt{d\log 2\kappa}\right).$$

Consequently, we apply with the choice $\delta_t \leq \frac{1}{T}$, which yields $N_t^{\mathrm{L}} \leq C_{\mathrm{L}}\, C_\kappa dT^2 \log\Big(4\sqrt{d\kappa}/\mathtt{m}_0\Big)$, and $\gamma_t^{\mathrm{L}} = \mathrm{A}_{\mathrm{L}}/((\kappa \vee \mathtt{L})dT^2)$. This implies that at time t, the data complexity is $G_t \leq C_{\mathrm{L}}\, C_\kappa dT^3 \log\Big(4\sqrt{d\kappa}/\mathtt{m}_0\Big)$, and that cumulative data complexity is

$$\sum_{t=1}^{T} G_t \leq C_{\mathrm{L}}\, C_\kappa dT^4 \log\Big(4\sqrt{d\kappa}/\mathtt{m}_0\Big)$$

6.3.2 Metropolis Algorithm

Let us state the conditions required for MALA to obtain a fast rate, seen e.g. in [Chen et al., 2020b].

Proposition 146. *(One-Step Convergence of Bandit MALA [Chen et al., 2020b], Theorem 5) Assume that the initial distribution p_0 satisfies Definition 142 with $\log c_W(p_0, \mu) \leq d\log(2\kappa)$, where μ is the stationary distribution of the chain. Assume further that the potential has condition number κ. Then the MALA algorithm converges to the true posterior with the following rate:*

$$N \geq C_{\mathrm{M}}\kappa d\log\left(\frac{d}{\delta^2}\right)\left(1 \vee \sqrt{\frac{\kappa}{d}}\right) \implies ||p_N - \mu||_{\mathrm{TV}} \leq \delta,$$

when we take the step size to be

$$\gamma^{\mathrm{M}} = \frac{\mathrm{A}_{\mathrm{M}}}{\mathtt{L}d\max\left(1, \sqrt{\kappa/d}\right)},$$

with A_{M} again an absolute constant.

Immediately, we can see that the critically dependency on the error tolerance ϵ are significantly better when contrasted with the unadjusted Langevin algorithm.

Proof of Proposition 129, MALA: The warm-start condition for all $t \leq T$ is immediately implied by Corollary 143. Consequently, recalling that we pick $\delta_t \leq \frac{1}{T}$ at each iteration t, we only need to perform $N_t = C_{\mathrm{M}}\kappa d\log\big(dT^2\big)(1 \vee \kappa/d)$ MALA iterations at each time t. Since each MALA iteration contains t gradients, this has data complexity $G_t \leq C_{\mathrm{M}}\kappa dT\log\big(dT^2\big)(1 \vee \kappa/d)$. Finally,

$$\sum_{t=1}^{T} G_t \leq C_{\mathrm{M}}\kappa dT^2\log\big(dT^2\big)(1 \vee \kappa/d)$$

7 Additional numerical experiments

7.1 Toy Example

In this section, we give additional details about the Toy example settings. As presented in the Section 4.1, the reward distribution considered in this toy example is Gaussian and all parameters used to describe the problem are provided in Table 9.1.

For each algorithm, we studied a pool of hyperparameters, and Figure 9.1 represents the best combination of hyperparameter for each approach. Table 9.2 summarizes the pool of hyperparameters studied during the experiment. Notice that the step size, parameter λ, and the standard deviation of the prior depend on the parameter η. This choice

Parameter dimension (d)	20
Context dimension (d_x)	4
Number of arms (K)	5
Noise level (σ)	1
Time horizon (T)	1000

Tab. 9.1 Environment hyperparameters

η	[1, 5, 10, 50, 100, 500, 1000]
Step Size	$[1/(t\eta), 0.5/(\eta t), 0.1/(\eta t), 0.05/(\eta t), 0.01/(\eta t)]$
λ	$[0.5\eta, 0.1\eta, 0.05\eta]$
Gaussian Prior Std	0.01η
Number of gradient updates	[25, 50, 100]
b	1000
Gradient descent steps for MALA / FG-MALA	20

Tab. 9.2 Algorithm hyperparameters

is subjective but seems to be quite logical. The step size is also depending on the time step t. For MALA-TS and FG-MALA-TS, we initialize MALA with the output of a full-batch gradient descent during 20 steps.

The baseline algorithm LinUCB has been studied for different values of α. However, for clarity, figure 9.1 shows only the performance of LinTS, LMC-TS, MALA-TS, FG-LMC-TS and FG-MALA-TS. The study of LinUCb is provided in figure 9.3. Notice that the best α among the pool studied is 0.1 and with this setting LinUCB outperforms all algorithms except FG-MALA-TS.

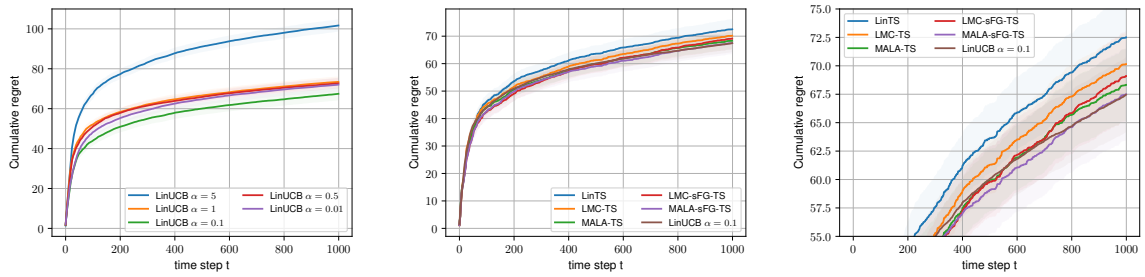


Fig. 9.3 Linear UCB study

7.2 Real World Dataset

Table 9.3 summarizes the main parameters used for the Yahoo! Front Page Today Module Dataset. A more detailed description of the problem can be found in [Li et al., 2010]. Our implementation of this task is based on the git repository: `https://github.com/antonismand/Personalized-News-Recommendation`.

Parameter dimension (d)	12
Context dimension (d_x)	12
Number of arms (K)	22
Time horizon (T)	25000

Tab. 9.3 Environment hyperparameters

Similarly, Table 9.4 describes the pool of hyperparameters studied during this experiment. Therefore, Figure 9.2 shows only the best comparison among this pool.

η	[1, 3, 5, 10, 20, 30, 40, 50]
Step size	$0.1/(t\eta)$
λ	$[0.1\eta, 0.3\eta, 0.5\eta]$
Gaussian Prior std	0.01η
Number of gradient updates	100
b	1000
Gradient Descent steps for MALA/FG-MALA	20

Tab. 9.4 Algorithm hyperparameters

7.3 Logistic bandit

In this section we investigate the behavior of Feel-Good Thompson Sampling on a more complex setting: the logistic bandit. We follow the setting of [Kveton et al., 2020b] and [Xu et al., 2022]. We consider a contextual vector $x \in \mathbb{R}^{20}$ sampled from $\mathcal{N}(0_{20}, \mathrm{I}_{20})$ and scaled to unit norm. A fixed set of 50 arms. And a Bernoulli reward distribution such that $r \sim \mathrm{B}(\phi(\theta^{\star T} x))$ where $\theta^\star$ is the true parameter, sampled from $\mathcal{N}(0_{20}, \mathrm{I}_{20})$ and scaled to unit norm. The function $\phi(u) = 1/(1+e^{-u})$ is the logistic function.

Figure 9.4 shows the cumulative regret, ie, $\mathbb{E}_{\Pi \sim \mathbb{Q}_{1:T}}[\sum_{t=1}^{T} 1 - f(x_t, \pi_s(x_t))]$ for LMC-TS and FG-LMC-TS. For the later, we consider four different values of λ. We observe that for small λ (≤ 0.01) FG-LMC-TS outperforms LMC-TS. However, when λ is too high, FG-LMC-TS becomes unstable and linear. It means that in this setting, the parameter λ has to be carefully determined. The implementation is based on the repository git: `https://github.com/devzhk/LMCTS`. The hyperparameters used for this experiment are provided in Table 9.5

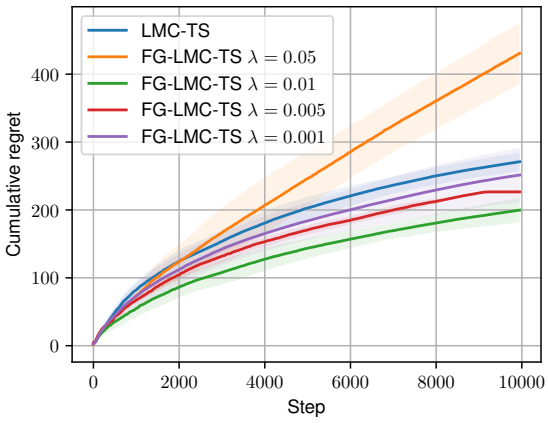


Fig. 9.4 Cumulative regret for logistic bandit over 10 runs

Time horizon (T)	10000
Number of LMC steps	500
Step size	0.001
Inverse temperature (β^{-1})	0.001

Tab. 9.5 Hyperparameters for logistic bandit

CHAPTER 10

Conclusion, limitations and perspectives

The frequentist approach to machine learning is often seen as simple and effective, but it is limited and cannot incorporate any prior knowledge or properly quantify uncertainty in predictions. In contrast, the Bayesian approach, known as Bayesian Machine Learning, naturally addresses these issues. However, it comes with an algorithmic challenge: how to sample from the posterior distribution. Variational Inference (VI) is a knwon method used to solve this issue. The main idea behind VI is to transform the sampling problem into an optimization problem. It results in a computationally efficient, simple, and highly parallelizable algorithm. By choosing an appropriate variational family, VI offers high flexibility, allowing us to adapt to the target distribution. The choice of variational family is crucial, as it directly impacts the approximation's quality, but it requires a profound understanding and expertise of the specific problem.

In this thesis, we have explored the potential of VI for Bayesian Neural Networks (BNNs). These models integrate Bayesian inference with deep learning, offering a promising path for combining flexibility with uncertainty quantification, a vital aspect in many real-world applications. In this setting, the large model size and non-convexity of the method make the MCMC approach impractically slow and unscalable. In contrast, Variational Inference, which comes with all the efficient optimization tools, appears far more suitable. For instance, in this scenario, we operate in the mean-field regime (large model size), meaning the model weights can be assumed to be independent. As a result, we can represent the variational distribution with a diagonal covariance matrix, enabling the use of a simpler and faster optimization algorithm. For these reasons, I strongly believe that Variational Inference is particularly promising for BNNs. However, in the current state of research, this approach lacks sufficient theoretical guarantees, both in terms of optimization error and approximation error, making it a promising area for future investigation.

Furthermore, we have explored the use of Variational Inference in a specific sequential decision-making problem known as contextual bandits. Our work demonstrates that VI can efficiently estimate the posterior distribution of the reward function, enabling informed decision-making at each step of the process. In this context, I believe the Gaussian variational approach is particularly well-suited. For instance, in the contextual bandit setting, the agent receives a new context at each step, leading to a large-data regime where the Bernstein-von Mises theorem suggests that the posterior distribution converges to a Gaussian. However, this domain also faces a lack of theoretical guarantees. An intriguing area for future research would be to investigate the behavior of VI in contextual bandits, particularly in more complex scenarios, such as non-linear reward distributions.

In conclusion, I believe that Variational Inference is a powerful tool with the potential to drive significant advancements in Bayesian Machine Learning. However, there are still several challenges that need to be addressed.

First of all, despite its empirical success, the theoretical properties of VI have only received attention recently, and mostly when the parametric family is the one of Gaussians. To unlock the full potential of VI, it is crucial to extend these theoretical guarantees to more complex and flexible variational families. This will enable us to better harness the flexibility and robustness that VI offers, paving the way for broader and more effective applications.

Bibliography

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 41–50, 2019.

Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. In *European Conference on Computer Vision*, pages 623–640. Springer, 2022.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. *Advances in neural information processing systems*, 29, 2016.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. 1996a.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. 2017.

Adam D Cobb and Brian Jalaian. Scaling Hamiltonian monte carlo inference for Bayesian neural networks with symmetric splitting. In *Uncertainty in Artificial Intelligence*, 2021.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are Bayesian neural network posteriors really like? *International Conference on Machine Learning*, 2021.

Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pages 5–13. ACM Press, 1993.

David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

David JC MacKay et al. Ensemble learning and evidence maximization. In *Proc. Nips*, page 4083. Citeseer, 1995.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Marc Lambert, Silvere Bonnabel, and Francis Bach. The recursive variational gaussian approximation (r-vga). *Statistics and Computing*, 32(1):10, 2022a.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR, 2023.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Samuel Gershman, Matt Hoffman, and David Blei. Nonparametric variational inference. *International Conference of Machine Learning*, 2012.

Oleg Arenz, Gerhard Neumann, and Mingjun Zhong. Efficient gradient-free variational inference using policy search. In *International conference on machine learning*, pages 234–243. PMLR, 2018.

Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning*, pages 3992–4002. PMLR, 2019.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 2015.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902*, 2022b.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

A. Graves. Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, 2011.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Michael N. Katehakis and Arthur F. Veinott. The multi-armed bandit problem: Decomposition and computation. *Math. Oper. Res.*, 12:262–268, 1987.

Donald A Berry and Bert Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5(71-87):7–7, 1985.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2066–2076. PMLR, 26–28 Aug 2020a. URL `https://proceedings.mlr.press/v108/kveton20a.html`.

Clémence Réda. *Combination of gene regulatory networks and sequential machine learning for drug repurposing*. Theses, Université Paris Cité, September 2022. URL `https://hal.science/tel-03846072`.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013a.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt, 2016a. URL https://arxiv.org/abs/1606.03966.

Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, 2017.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011b. URL https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR, 8 2017a. URL https://proceedings.mlr.press/v70/li17c.html.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre Gaillard. Efficient kernelized ucb for contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 5689–5720. PMLR, 2022.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012a.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyYe6k-CW.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021a.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 2018.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *International Conference on Machine Learning*, 2019.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent Gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. *Advances in Neural Information Processing Systems*, 2020a.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022a.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, 2008.

Alain Durmus, Éric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: When langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018. doi: 10.1137/16M1108340. URL `https://doi.org/10.1137/16M1108340`.

Alekh Agarwal, Miroslav Dudik, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: review and debates. *arXiv preprint arXiv:2309.16314*, 2023.

Rhiannon Michelmore, Matthew Wicker, Luca Laurenti, Luca Cardelli, Yarin Gal, and Marta Kwiatkowska. Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control. In *IEEE International Conference on Robotics and Automation*, 2020.

Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *IJCAI*, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *4th workshop on Bayesian Deep Learning (NeurIPS)*, 2019.

Abdullah A Abdullah, Masoud M Hassan, and Yaseen T Mustafa. A review on bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*, 2022.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, Aliaksandr Hubin, Alexander Immer, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Agustinus Kristiadi, Yingzhen Li, Stephan Mandt, Christopher Nemeth, Michael A Osborne, Tim GJ Rudner, David Rügamer, Yee Whye Teh, Max Welling, Andrew Gordon Wilson, and Ruqi Zhang. Position paper: Bayesian deep learning in the age of large-scale ai. *arXiv preprint arXiv:2402.00809*, 2024.

Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on bayesian neural networks–a tutorial for deep learning users. *arXiv preprint arXiv:2007.06823*, 2020.

Radford M Neal. MCMC using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2011.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, pages 681–688. Citeseer, 2011.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, 2014.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, 2018.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, 2017.

Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in adam. In *International Conference on Machine Learning*, 2018.

Mohammad Emtiyaz Khan and Håvard Rue. The bayesian learning rule. *Journal of Machine Learning Research*, 2023.

Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pages 5852–5861. PMLR, 2018a.

Kazuki Osawa, Siddharth Swaroop, Anirudh Jain, Runa Eschenhagen, Richard E Turner, Rio Yokota, and Mohammad Emtiyaz Khan. Practical deep learning with Bayesian principles. *Advances in Neural Information Processing Systems*, 2019.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *International Conference of Learning Representations*, 2020.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? *International Conference on Machine Learning*, 2020.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 2020.

Peter Grünwald. The safe bayesian. In *International Conference on Algorithmic Learning Theory*, 2012.

Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 2017.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 2019.

Rianne Heide, Alisa Kirichenko, Peter Grunwald, and Nishant Mehta. Safe-bayesian generalized linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2623–2633. PMLR, 2020.

Peter Grunwald, Thomas Steinke, and Lydia Zakynthinou. PAC-Bayes, MAC-bayes and conditional mutual information: Fast rate bounds that handle general VC classes. In *Proceedings of Thirty Fourth Conference on Learning Theory*, 2021.

Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 1991.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 2006.

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. *International Conference of Learning Representations*, 2021.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in Bayesian neural networks and the cold posterior effect. In *Conference on Uncertainty in Artificial Intelligence*, 2022.

Sebastian Farquhar, Michael Osborne, and Yarin Gal. Radial bayesian neural networks: Robust variational inference in big models. *Training*, 2019.

Vincent Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 2022.

Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All you need is a good functional prior for bayesian deep learning. *Journal of Machine Learning Research*, 2022.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *International Conference on Learning Representations*, 2022.

Sebastian W Ober and Laurence Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *International Conference on Machine Learning*, 2021.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Beau Coker, Wessel P Bruinsma, David R Burt, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field Bayesian neural networks ignore the data. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Yuri Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Arnaud Descours, Tom Huix, Arnaud Guillin, Manon Michel, Éric Moulines, and Boris Nectoux. Law of large numbers for bayesian two-layer neural network trained with variational inference. In *Proceedings of Thirty Sixth Conference on Learning Theory*. PMLR, 2023a.

Arnaud Descours, Tom Huix, Arnaud Guillin, Manon Michel, Éric Moulines, and Boris Nectoux. Central limit theorem for bayesian neural network trained with variational inference. *arXiv preprint*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 2020a.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *International Conference of Learning Representations*, 2018.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. *International Conference of Learning Representations*, 2019.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *International Conference of Learning Representations*, 2019.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pages 4376–4386. PMLR, 2020a.

Jiri Hron, Yasaman Bahri, Roman Novak, Jeffrey Pennington, and Jascha Sohl-Dickstein. Exact posterior distributions of wide Bayesian neural networks. *Workshop on Uncertainty and Robustness in Deep Learning (ICML)*, 2020b.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference of Learning Representations*, 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015.

Lorenzo Noci, Kevin Roth, Gregor Bachmann, Sebastian Nowozin, and Thomas Hofmann. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. *Advances in Neural Information Processing Systems*, 2021.

Max-Heinrich Laves, Malte Tölle, Alexander Schlaefer, and Sandy Engelhardt. Posterior temperature optimization in variational inference for inverse problems. *arXiv preprint arXiv:2106.07533*, 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, 2015.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

M. Krzywinski and N. Altman. Importance of being uncertain. *Nature methods*, 10(9):809–811, 2013.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

T. Huix, S. Majewski, A. Durmus, E. Moulines, and A. Korba. Variational inference of overparameterized bayesian neural networks: a theoretical and empirical study, 2022. URL https://arxiv.org/abs/2207.03859.

G.M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *Preprint arXiv:1805.00915, to appear in Comm. Pure App. Math.*, 2018.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020b.

A. Descours, A. Guillin, M. Michel, and B. Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case. *arXiv preprint arXiv:2207.12734*, 2022a.

L. Chizat. Mean-field langevin dynamics: Exponential convergence and annealing, 2022. URL https://arxiv.org/abs/2202.01009.

L. Chizat, M. Colombo, X. Fernandez-Real, and A. Figalli. Infinite-width limit of deep linear neural networks, 2022. URL https://arxiv.org/abs/2211.16980.

M. Hitsuda and I. Mitoma. Tightness problem and stochastic evolution equation arising from fluctuation phenomena for interacting diffusions. *Journal of Multivariate Analysis*, 19(2):311–328, 1986.

A-S. Sznitman. Topics in propagation of chaos. In *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251. Springer, 1991. ISBN 978-3-540-46319-1.

B. Fernandez and S. Méléard. A Hilbertian approach for fluctuations on the Mckean-Vlasov model. *Stochastic Processes and their Applications*, 71(1):33–53, 1997.

B. Jourdain and S. Méléard. Propagation of chaos and fluctuations for a moderate model with smooth initial data. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 34(6):727–766, 1998.

F. Delarue, D. Lacker, and K. Ramanan. From the master equation to mean field game limit theory: a central limit theorem. *Electronic Journal of Probability*, 24:1–54, 2019.

P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *The Annals of Applied Probability*, 9(2):275–297, 1999.

T. Kurtz and J. Xiong. A stochastic evolution equation arising from the fluctuations of a class of interacting particle systems. *Communications in Mathematical Sciences*, 2(3):325–358, 2004.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 278–288. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/02e74f10e0327ad868d138f2b4fdd6f0-Paper.pdf.

A. Jakubowski. On the skorokhod topology. In *Annales de l'IHP Probabilités et statistiques*, pages 263–285, 1986.

B. Piccoli, F. Rossi, and E. Trélat. Control to flocking of the kinetic Cucker–Smale model. *SIAM Journal on Mathematical Analysis*, 47(6):4685–4719, 2015.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

B. Piccoli and F. Rossi. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365, 2016.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition, 1999.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

J. Jacod and A. Shiryaev. *Skorokhod Topology and Convergence of Processes*. Springer, 1987.

A. Descours, T. Huix, A. Guillin, M. Michel, É. Moulines, and B. Nectoux. Law of large numbers for bayesian two-layer neural network trained with variational inference. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4657–4695. PMLR, 12–15 Jul 2023b. URL `https://proceedings.mlr.press/v195/descours23a.html`.

A. Descours, A. Guillin, M. Michel, and B. Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case. *To appear in Journal of Machine Learning Research*, 2022b.

Z. Chen, G.M. Rotskoff, J. Bruna, and E. Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22217–22230. Curran Associates, Inc., 2020a. URL `https://proceedings.neurips.cc/paper/2020/file/fc5b3186f1cf0daece964f78259b7ba0-Paper.pdf`.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020c.

O. Kallenberg. *Foundations of modern probability*. Springer, 2nd edition, 2002.

Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

Théo Moins, Julyan Arbel, Anne Dutfoy, and Stéphane Girard. On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic. *Bayesian Analysis*, 1(1):1–26, 2023.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018b.

Anya Katsevich and Philippe Rigollet. On the approximation accuracy of gaussian variational inference. *arXiv preprint arXiv:2301.02168*, 2023.

Tapio Helin and Remo Kretschmann. Non-asymptotic error estimates for the laplace approximation in bayesian inverse problems. *Numerische Mathematik*, 150(2):521–549, 2022.

Justin Domke, Guillaume Garrigos, and Robert Gower. Provable convergence guarantees for black-box variational inference. *Advances in neural information processing systems*, 2023.

Mingxuan Yi and Song Liu. Bridging the gap between variational inference and wasserstein gradient flows. *arXiv preprint arXiv:2310.20090*, 2023.

Kurt Otto Friedrichs. The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society*, 55:132–151, 1944.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Taylor & Francis*, 2001.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR, 2021.

A Duncan, N Nüsken, and L Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 2023.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in neural information processing systems*, 2019.

Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

Jonathan Li and Andrew Barron. Mixture density estimation. *Advances in neural information processing systems*, 12, 1999.

Krzysztof Łatuszyński, Błażej Miasojedow, and Wojciech Niemiro. Nonasymptotic bounds on the estimation error of mcmc algorithms. *Bernoulli*, 19(5A):2033–2066, 2013.

Rentian Yao and Yun Yang. Mean field variational inference via Wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*, 2022.

Daniel Lacker. Independent projections of diffusions: Gradient flows for variational inference and optimal mean field approximations. *arXiv preprint arXiv:2309.13332*, 2023.

Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *arXiv preprint arXiv:2312.02849*, 2023.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019.

Katy Craig, Karthik Elamvazhuthi, Matt Haberland, and Olga Turanova. A blob method for inhomogeneous diffusion with applications to multi-agent control and sampling. *Mathematics of Computation*, 2023a.

Katy Craig, Matt Jacobs, and Olga Turanova. Nonlocal approximation of slow and fast diffusion. *arXiv preprint arXiv:2312.11438*, 2023b.

José Antonio Carrillo, Antonio Esposito, and Jeremy Sheung-Him Wu. Nonlocal approximation of nonlinear diffusion equations. *Calculus of Variations and Partial Differential Equations*, 63(4):100, 2024.

Lingxiao Li, Qiang Liu, Anna Korba, Mikhail Yurochkin, and Justin Solomon. Sampling with mollified interaction energy descent. *arXiv preprint arXiv:2210.13400*, 2022.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *Advances in Neural Information Processing Systems*, 35: 17263–17275, 2022.

Frederik Kunstner, Raunak Kumar, and Mark Schmidt. Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages 3295–3303. PMLR, 2021.

Carles Domingo-Enrich and Aram-Alexandre Pooladian. An explicit expansion of the Kullback-Leibler divergence along its Fisher-Rao gradient flow. *Transactions of Machine Learning Research*, 2023.

Nicolas Chopin, Francesca R Crucinio, and Anna Korba. A connection between tempering and entropic mirror descent. *International Conference of Machine Learning*, 2024.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

FLEMMING Topsøe. Some bounds for the logarithmic function. *Inequality theory and applications*, 4:137, 2007.

Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR, 2022.

Eric Mazumdar, Aldo Pacchiano, Yi-an Ma, Peter L Bartlett, and Michael I Jordan. On thompson sampling with langevin algorithms. *arXiv preprint arXiv:2002.10002*, 2020a.

Pan Xu, Hongkai Zheng, Eric Mazumdar, Kamyar Azizzadenesheli, and Anima Anandkumar. Langevin monte carlo for contextual bandits. *arXiv preprint arXiv:2206.11254*, 2022.

Tom Huix, Matthew Zhang, and Alain Durmus. Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8749–8770. PMLR, 4 2023. URL `https://proceedings.mlr.press/v206/huix23a.html`.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. *arXiv preprint arXiv:2010.00827*, 2020.

Iñigo Urteaga and Chris Wiggins. Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 698–706. PMLR, 2018.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL `https://proceedings.neurips.cc/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf`.

Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration. *arXiv preprint arXiv:2012.01780*, 2020.

Nima Hamidi and Mohsen Bayati. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Weiqiang Wu, Jing Yang, and Cong Shen. Stochastic linear contextual bandits with diverse contexts. In *International Conference on Artificial Intelligence and Statistics*, pages 2392–2401. PMLR, 2020.

Imad Aouali, Branislav Kveton, and Sumeet Katariya. Generalizing hierarchical bayesian bandits. *arXiv preprint arXiv:2205.15124*, 2022.

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022b.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007b. URL https://proceedings.neurips.cc/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, number 3 in Proceedings of Machine Learning Research, pages 127–135, Atlanta, Georgia, USA, 6 2013b. PMLR. URL https://proceedings.mlr.press/v28/agrawal13.html.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR, 8 2017b. URL https://proceedings.mlr.press/v70/li17c.html.

Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic linear bandits with hidden low rank structure. *CoRR*, abs/1901.09490, 2019. URL http://arxiv.org/abs/1901.09490.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt, 2016b. URL https://arxiv.org/abs/1606.03966.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In Steve Hanneke and Lev Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 223–237. PMLR, 15–17 Oct 2017. URL https://proceedings.mlr.press/v76/mC3A9nard17a.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR, 2012b.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5074–5083. PMLR, 7 2021b. URL https://proceedings.mlr.press/v139/jin21d.html.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996b.

Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic analysis and applications*, 8(4):483–509, 1990.

Alain Durmus and Andreas Eberle. Asymptotic bias of inexact markov chain monte carlo methods in high dimension. *arXiv preprint arXiv:2108.00682*, 2021.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On approximate thompson sampling with Langevin algorithms. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6797–6807. PMLR, 7 2020b. URL https://proceedings.mlr.press/v119/mazumdar20a.html.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change detection. In *Artificial intelligence and statistics*, pages 442–450. PMLR, 2013.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21:92–1, 2020b.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020b.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Inférence Variationnelle : théorie et applications en grande dimension.

Mots clés : Inférence Variationnelle, Machine Learning Bayesien, Bandit Contextuel.

Résumé : Cette thèse développe des méthodes d'Inférence Variationnelle pour l'apprentissage bayésien en grande dimension. L'approche bayésienne en machine learning permet de gérer l'incertitude épistémique des modèles et ainsi de mieux quantifier l'incertitude de ces modèles, ce qui est nécessaire dans de nombreuses applications de machine learning. Cependant, l'Inférence Bayésienne n'est souvent pas réalisable car la distribution à posteriori des paramètres du modèle n'est pas calculable en général. L'Inférence Variationnelle (VI) est une approche qui permet de contourner ce problème en approximant la distribution à posteriori par une distribution plus simple appelée distribution Variationnelle.
Dans la première partie de cette thèse, nous avons travaillé sur les garanties théoriques de l'Inférence Variationnelle. Dans un premier temps, nous avons étudié cette approche lorsque la distribution Variationnelle est une Gaussienne et dans le régime surparamétré, c'est-à-dire lorsque les modèles sont en très grande dimension. Puis, nous nous sommes intéressés aux distributions Variationnelles plus expressives que sont les mélanges de Gaussiennes et nous avons étudié à la fois l'erreur d'optimisation et l'erreur d'approximation de cette méthode.
Dans la deuxième partie de la thèse, nous avons étudié les garanties théoriques des problèmes de bandit contextuels en utilisant une approche Bayésienne appelée Thompson Sampling. Dans un premier temps, nous avons exploré l'utilisation d'Inférence Variationnelle pour l'algorithme Thompson Sampling. Nous avons notament démontré que dans le cadre linéaire, cette approche permet d'obtenir les mêmes garanties théoriques que si la distribution à posteriori était connue. Dans un deuxième temps, nous avons étudié une variante de Thompson Sampling appelée Feel-Good Thompson Sampling (FG-TS). Cette méthode permet d'obtenir de meilleures garanties théoriques que l'algorithme classique. Nous avons alors étudié l'utilisation d'une méthode de Monte Carlo Markov Chain pour approximer la distribution à posteriori. Plus spécifiquement, nous avons ajouté à FG-TS un algorithme de Langevin Monte Carlo et de Metropolized Langevin Monte Carlo. De plus, nous avons obtenu les mêmes garanties théoriques que pour FG-TS lorsque la distribution à posteriori est connue.

Title : Variational Inference : theory and large scale applications.

Keywords : Variational Inference, Bayesian Machine Learning, Contextual Bandit

Abstract : This thesis explores Variational Inference methods for high-dimensional Bayesian learning. In Machine Learning, the Bayesian approach allows one to deal with epistemic uncertainty and provides a better uncertainty quantification, which is necessary in many machine learning applications. However, Bayesian Inference is often not feasible because the posterior distribution of the model parameters is generally untractable. Variational Inference (VI) allows to overcome this problem by approximating the posterior distribution with a simpler distribution called the Variational distribution.
In the first part of this thesis, we worked on the theoretical guarantees of Variational Inference. First, we studied VI when the Variational distribution is a Gaussian and in the overparameterized regime, i.e., when the models are high dimensionals. Finally, we explore the Gaussian mixtures Variational distributions, as it is a more expressive distribution. We studied both the optimization error and the approximation error of this method.
In the second part of the thesis, we studied the theoretical guarantees for contextual bandit problems using a Bayesian approach called Thompson Sampling. First, we explored the use of Variational Inference for Thompson Sampling algorithm. We notably showed that in the linear framework, this approach allows us to obtain the same theoretical guarantees as if we had access to the true posterior distribution. Finally, we consider a variant of Thompson Sampling called Feel-Good Thompson Sampling (FG-TS). This method allows to provide better theoretical guarantees than the classical algorithm. We then studied the use of a Monte Carlo Markov Chain method to approximate the posterior distribution. Specifically, we incorporated into FG-TS a Langevin Monte Carlo algorithm and a Metropolized Langevin Monte Carlo algorithm. Moreover, we obtained the same theoretical guarantees as for FG-TS when the posterior distribution is known.

Institut Polytechnique de Paris
91120 Palaiseau, France