On the tradeoffs of statistical learning with privacy

Clément Lalanne

THÈSE

en vue de l'obtention du grade de Docteur, délivré par

l'École Normale Supérieure de Lyon

École Doctorale N°512

Discipline : Informatique

Soutenue publiquement le 4 octobre 2023, par :
Clément Lalanne

On the tradeoffs of statistical learning with privacy

Sur les compromis liés à l'apprentissage statistique sous contraintes de confidentialité

Devant le jury composé de :

Aurélien Bellet, Directeur de Recherche, Inria - Univ. de Montpellier	Rapporteur
Béatrice Laurent-Bonneau, Professeur des universités, INSA Toulouse	Rapporteuse
Élisa Fromont, Professeur des universités, Univ. de Rennes	Examinatrice
Aurélien Garivier, Professeur des universités, ENS de Lyon	Directeur de thèse
Rémi Gribonval, Directeur de Recherche, Inria - ENS de Lyon	Co-directeur de thèse

Remerciements - Acknowledgements

Je tiens à remercier mes directeurs de thèse, Aurélien Garivier et Rémi Gribonval, pour leur encadrement assidu durant ces trois années. Merci d'abord d'avoir accepté d'encadrer un étudiant qui vous était alors inconnu, et merci pour la confiance que vous m'avez accordée. Vous avez su rendre ces trois années agréables, tout en m'enseignant votre goût pour le travail rigoureux. En outre, j'ai beaucoup appris à vos côtés, tant sur un plan purement scientifique que sur tous les autres aspects de la recherche. Aurélien, merci en particulier d'avoir su me donner la confiance nécessaire pour mener à bien cette aventure, y compris dans les moments de doute. Rémi, merci pour tous tes efforts dans la création et l'animation de l'équipe de recherche Ockham, qui m'a offert tellement de bons moments et d'expériences enrichissantes. Grâce à vous, j'ai pu évoluer d'un étudiant animé d'une volonté naïve d'impacter positivement la confidentialité dans l'ère du numérique, à un chercheur capable d'identifier des problèmes concrets pour y arriver et de les résoudre.

Je tiens ensuite à remercier Aurélien Bellet et Béatrice Laurent-Bonneau pour avoir rapporté cette thèse, et Élisa Fromont pour avoir présidé mon jury de soutenance. J'espère de tout cœur que la lecture de ce document, ainsi que la présentation que j'en ai donné le 4 octobre 2023 lors de ma soutenance, vous auront été agréables. J'ai intégré au mieux (dans la limite des modifications tolérées) les différentes remarques que vous m'aviez faites au sujet de la rédaction de ce document dans cette version finale.

Cette thèse a également été rendue possible grâce à l'intervention de deux personnes importantes : Francis Bach et Nicolas Grislain. Francis, merci de m'avoir mis en relation avec Aurélien lorsqu'en dernière année à l'ENS, je t'avais exprimé mon souhait de travailler sur la confidentialité des données. Tes conseils auront été excellents et ont directement conduit à la production de cette thèse. Nicolas, merci d'avoir participé à mon encadrement non-officiel durant mon doctorat. Ta vision pratique des problèmes liés à la confidentialité m'a beaucoup apporté scientifiquement, et il ne fait aucun doute qu'elle influencera ma carrière pendant longtemps.

I would then like to thank anyone that contributed to the scientific content of this thesis without being explicitely credited for it. This includes, for instance, the anonymous reviewers and the action editors of the different articles that we published, and that end up being reused here.

Furthermore, I would like to thank anyone that contributed to my training as a researcher, and in particular Florian Simatos, Nicolas Vayatis, Laurent Oudre, Thomas Moreau and Volkan Cevher for their supervision during various intership.

Cette thèse clôturant mes études, j'aimerais prendre le temps de remercier des professeurs qui ont eu une grande influence sur ma scolarité. En particulier, mes professeurs dans les matières scientifiques en classes préparatoires aux grandes écoles, Adrien Joseph, Bernard Randé, Emmanuel Ranz, François Pantigny, Guillaume Pagot, Hélène Briand, Malek Kious, Philippe Fortin et Roger Mansuy. Durant ces années, j'ai pu développer mon goût pour l'étude des sciences, et j'ai alors décidé de poursuivre dans des études menant au doctorat. Merci de m'avoir communiqué cette passion. Ensuite, j'aimerais remercier Jean-Claude Blaye-Félice et Olivier Delord pour leurs conseils d'orientation en classe de terminale. Sans vous, je n'aurais probablement pas envisagé de poursuivre en classes préparatoires aux grandes écoles, et il ne fait aucun doute que mes études supérieures auraient-été bien différentes. Enfin, j'aimerais remercier ma professeure de mathématiques en classe de sixième, Ghislaine Budon. C'est avec vous que j'ai pour la première fois appris les bases du raisonnement en mathématiques (je me souviens encore de "On sait que ... Or ... Donc ..."), et je me souviens avoir vite pris goût à ce nouveau jeu, dans lequel le but est de démontrer. C'est probablement la compétence la plus élémentaire ayant conduit à la production de cette thèse.

Sur un plan plus personnel, j'aimerais remercier toutes les personnes qui m'ont entouré durant ces trois années à l'ENS de Lyon. Je n'aurais pas pu rêver mieux que de me retrouver dans un environnement comme celui-ci. Alors que mes productions scientifiques font (soi-disant) de moi un expert dans un domaine très spécifique qui sera développé dans la suite de ce document, vous m'avez permis d'élargir mes compétences à plein d'autres sujets. Je peux grâce à vous, par exemple, comparer le niveau d'épice de n'importe quel plat à celui du piment pakistanais, construire plein de contre-exemples rigolos nécessitant l'axiome du choix, reconnaitre les différents types de Spritz, et prendre une décision éclairée entre faire mon footing en montagne, ou sur un tapis de course incliné. Du fond du cœur, merci pour les repas au self ou au restaurant, les pauses-café, et tous les autres moments de convivialité.

Enfin, j'aimerais remercier ma famille et mes amis sans qui rien de tout ça n'aurait été possible. En particulier, merci à mes parents, Magali Ricarde et Vincent Lalanne, de m'avoir donné le goût de l'apprentissage et de m'avoir donné les conditions matérielles nécessaires pour pouvoir mener à bien mes études. Merci à mes grand-parents, Claudine

Lavie, Jean-Léon Ricarde et Marie Lavie, d'avoir toujours veillé à mon assiduité scolaire. Merci enfin à Astrid De Oliveira de partager ma vie et d'avoir été un soutien constant durant ces trois années.

Résumé

Cette thèse étudie les compromis entre l'apprentissage statistique et la protection de la vie privée. D'une part, l'apprentissage, qui se définit comme l'estimation de quantités ou de tendances significatives à l'échelle d'une population en n'ayant accès qu'à des observations échantillonnées de cette population, sera plus facile si l'on accorde un accès illimité aux données d'apprentissage. D'un autre côté, les données d'apprentissage peuvent être sensibles et leur utilisation sans restriction pourrait entraîner des problèmes de confidentialité.

Le spectre des problèmes de sécurité et de confidentialité pouvant être très large, il est nécessaire de préciser le champ d'application de cette thèse. Dans la configuration considérée, les données sont agrégées par un seul acteur qui les utilise pour entraîner un modèle statistique (procédure d'estimation, réseau neuronal, ...). Ce modèle est ensuite partagé avec le monde entier. Le problème considéré est celui de l'inversion : est-il possible de briser la confidentialité des échantillons des données d'entraînement individuels par la seule observation du modèle entraîné ?

Après une introduction, cette thèse est composée de six chapitres et d'une conclusion. Le premier chapitre présente une étude de cas pratique. Il établit empiriquement le compromis entre l'utilité en classification d'images et la protection contre les attaques par inférence d'appartenance en tirant parti de la parcimonie du modèle.

Le deuxième chapitre est consacré à la présentation des principaux résultats de la théorie de la confidentialité différentielle. Cette définition mathématique de la confidentialité permet de se défendre contre n'importe quel adversaire, avec de fortes garanties de confidentialité. Le chapitre illustre des résultats connus et importants de la littérature en les intégrant dans un cadre statistique, introduisant ainsi le lecteur à des concepts clés pour le reste de la thèse.

Le troisième chapitre se concentre sur les bornes inférieures sur l'utilité statistique des algorithmes d'apprentissage lorsqu'ils sont soumis à des contraintes de confidentialité différentielle. En particulier, il présente un cadre de preuve qui s'appuie sur une formalisation en tant que problème de transport, où les ensembles de données sont comparés en utilisant des fonctions de similarité qui capturent l'essence de la confidentialité. Ce cadre permet de retrouver les résultats de l'état de l'art des dernières années sur le sujet, tout en unifiant la théorie qui les sous-tend. Ce cadre de preuve est également prêt à l'emploi, ce qui signifie qu'il est facile de l'utiliser pour élargir la théorie, notamment pour de nouvelles définitions de la confidentialité ou de nouvelles structures probabilistes de l'espace de données.

Le quatrième chapitre explique comment appliquer les techniques du chapitre précédent, en regardant des exemples paramétriques. En particulier, il étudie le modèle de Bernoulli, le modèle uniforme, le modèle gaussien et l'estimation de familles exponentielles à domaine compact. Il donne également des références bibliographiques pour de nombreux problèmes similaires intéressants.

Le cinquième chapitre étudie l'estimation non paramétrique privée des densités. Il présente plusieurs procédures d'estimation optimales ou quasi-optimales pour des densités appartenant à des espaces fonctionnels de Lipschitz et de Sobolev. Le problème est étudié dans le cadre de la confidentialité différentielle régulière et de la confidentialité différentielle concentrée.

Le sixième (et dernier) chapitre traite du problème de l'estimation de la fonction quantile. Il s'appuie tout d'abord sur l'idée que les quantiles empiriques d'un ensemble de données sont de bons estimateurs des quantiles de la distribution sous-jacente. À partir de là, les propriétés de concentration des algorithmes de pointe pour l'estimation des quantiles empiriques privés sont dérivées pour le problème de l'estimation statistique. Le chapitre présente les limites de ces estimateurs, notamment en soulignant leur sous-optimalité possible sur des instances spécifiques du problème.

Abstract

This thesis studies the tradeoffs between statistical learning and privacy. On the one hand, learning, which is defined as estimating meaningful quantities or trends at the scale of a population with only access to sampled observations of that population, will be easier if granted unrestricted access to its training data. On the other hand, the training data can be sensitive, and its unrestricted use can lead to privacy issues.

Since the spectrum of security and privacy issues can be very large, it is necessary to specify the range of this thesis. In the considered setup, the data is aggregated by a single actor that uses it to train a statistical model (estimation procedure, neural network, ...). This model is then shared to the entire world. The problem that is considered is the inversion one : is it possible to break the privacy of the individual samples of the training data by the sole observation of the trained model ?

After an introduction, this thesis is composed of six chapters and of a conclusion. The first chapter presents a practical case study. It empirically draws the tradeoff between utility in image classification and privacy against membership inference attacks by leveraging the sparsity of the model.

The second chapter is devoted to the presentation of key results of the theory of differential privacy. This mathematical definition of privacy allows defending against any adversary, with strong privacy guarantees. The chapter illustrates known and important results from the literature by embedding them in a statistical framework, thus introducing the reader to key concepts for the rest of the thesis.

The third chapter focuses on lower-bounds on the statistical utility of learning algorithms when constrained by differential privacy. In particular, it presents a proof framework that builds on a formalization as a transport problem, where datasets are compared using similarity functions that capture the essence of privacy. This framework allows recovering

the state-of-the-art results of the last few years on the subject, while unifying the theory behind them. It is also plug-and-play, meaning that it is easy to build on, notably for new definitions of privacy, or new probabilistic structures of the data space.

The fourth chapter details how to apply the techniques of the previous one on parametric examples. In particular, it studies the Bernoulli model, the uniform model, the Gaussian model, and the estimation of exponential families with compact domain. It also gives bibliographic pointers for many interesting similar problems.

The fifth chapter studies the private nonparametric estimation of densities. It presents multiple optimal or near-optimal estimation procedures for densities that belong to Lipschitz and Sobolev functional spaces. The problem is studied under regular differential privacy and under concentrated differential privacy.

The sixth (and last) chapter considers the quantile function estimation problem. First, it builds on the idea that empirical quantiles of a dataset are good proxies for the quantiles of the underlying distribution. From that, the concentration properties of state-of-the-art algorithms for the private empirical quantile estimation are derived for the statistical estimation problem. The chapter presents the limits of these estimators, notably by pointing-out their possible sub-optimality on specific instances of the problem.

Notations

$\mathbb{N}$	Set of *natural numbers*.
$\mathbb{Z}$	Set of *relative integers*.
$\mathbb{Q}$	Set of *rational numbers*.
$\mathbb{R}$	Set of *real numbers*.
$\mathbb{C}$	Set of *complex numbers*.
$\{m, \dots, n\}$	Set of i*ntegers from m (included) to n (included)*.
$\mathcal{P}_k$	Simplex of $\mathbb{R}^k$ of vectors with *positive entries that sum to* 1.
$\mathbb{X}_*$	*Non-null* elements from $\mathbb{X}$.
$\mathbb{X}_+$ (resp. $\mathbb{X}_-$)	*Non-negative* (resp. *non-positive*) elements from $\mathbb{X}$.
$\mathbb{X}^k$	Set *k-tuples* from $\mathbb{X}$.
$\mathbb{X}_\cdot^k$	Should be interpreted as $(\mathbb{X}_\cdot)^k$.
$\mathbb{X}^{k\nearrow}$	Set *k-tuples* from $\mathbb{X}$ sorted by non-decreasing order.
$\#(S)$	*Cardinality* of set S.
$\cdot \sqcup \cdot$	*Disjoint union* of sets.
$\mathbf{X}$	*Vector* $\mathbf{X} = (X_1, \dots, X_n)$, or *dataset*.
$\cdot \otimes \cdot$	*Kronecker product*.
$\cdot \sim \cdot$	*Neighboring relation*.
$\ln$	*Natural logarithm*.
$\log_k$	*Logarithm in basis k* (i.e. $\ln(\cdot)/\ln(k)$).
$\mathrm{dom}\,(\mathfrak{M})$	*Domain* (set of admissible inputs) of the mechanism $\mathfrak{M}$.
$\mathrm{codom}\,(\mathfrak{M})$	*Codomain* (set of admissible outputs) of the mechanism $\mathfrak{M}$.
∇f	*Gradient* of the function f.
$\nabla^2 f$	*Hessian* of the function f.
Δf	*Sensitivity* (l_1) of f.
$\Delta_k f$	*l_k sensitivity* of f.
$\mathcal{C}^k(E, F)$	Functions from E to F that are *k-times continuously differentiable*.
$\mathcal{C}^\infty(E, F)$	$\cap_{k\geq 0}\mathcal{C}^k(E, F)$.
Θ_L^{Lip}	Set of L-Lipschitz functions on $[0, 1]$.
$\Theta_{L,\beta}^{\mathrm{Sob}}$	Set of *β-Sobolev* functions on $[0, 1]$.
$\Theta_{L,\beta}^{\mathrm{PSob}}$	Set of *β-Periodic Sobolev* functions on $[0, 1]$.

$\mathbb{P}_X$	*Probability distribution.* X may be used to specify the randomness.
$\mathbb{E}_X$	*Expectation.* X may be used to specify the randomness.
$\mathbb{V}_X$	*Variance.* X may be used to specify the randomness.
Cov_X	*Covariance matrix.* X may be used to specify the randomness.
$\mathcal{B}(p)$	*Bernoulli distribution* of probability of success p.
$\mathcal{B}(n, p)$	*Binomial distribution* of probability of success p and n trials.
$\mathcal{U}(S)$	*Uniform distribution* on S.
$\mathcal{E}(\lambda)$	*Exponential distribution* of p.d.f. $p(x) = \lambda e^{-\lambda x}$.
$\mathcal{L}(b)$	*Centered Laplace distribution* of p.d.f. $p(x) = \frac{1}{2b} e^{-\frac{\lvert x \rvert}{b}}$.
$\mathcal{L}(bI_d)$	Distribution on $\mathbb{R}^d$ with independent components of distribution $\mathcal{L}(b)$.
$\mathcal{N}(\mu, \Sigma)$	*Multivariate normal distribution* with mean μ and covariance matrix Σ.
$\mathbb{P} \ll \mathbb{Q}$	$\mathbb{P}$ is *absolutely continuous* w.r.t. $\mathbb{Q}$.
$\mu^{\otimes n}$	*Product measure* with n times μ as marginal measure.
$\xrightarrow{\mathcal{L}}$	*Convergence in distribution.*
$\mathrm{TV}(\mathbb{P}, \mathbb{Q})$	*Total variation distance* between $\mathbb{P}$ and $\mathbb{Q}$.
$\mathrm{KL}(\mathbb{P} \Vert \mathbb{Q})$	*Kullback–Leibler divergence* of $\mathbb{P}$ from $\mathbb{Q}$.
$\mathrm{D}_\alpha(\mathbb{P} \Vert \mathbb{Q})$	*Rényi divergence of level* α of $\mathbb{P}$ from $\mathbb{Q}$.
$\Pi(\mathbb{P}_1, \ldots, \mathbb{P}_N)$	Set of *couplings* with marginal distributions $\mathbb{P}_1, \ldots, \mathbb{P}_N$.

Contents

Introduction

With the generalization of large-scale data collection, the ever-increasing computational power of modern computers, and the ingenious contributions of the scientific community, statistical learning (which is often referred to as machine learning or simply as artificial intelligence) has revolutionized many aspects of our modern lives. This revolution is not only quantitative, but it is also a qualitative one in the sense that it changes the way knowledge is built. Traditionally, sensible science (e.g. physics, chemistry, social sciences, ...) is built by proposing a model that has later to be confirmed by the experiments. In contrast, statistical learning builds a model *from* the experiments. For various reasons, one could want to share this learned model with the world (e.g. to help with the diagnosis of certain diseases). However, when this model is trained on sensitive data (e.g. medical data [Dubost et al., 2020, Jung et al., 2021, Truong & Oudre, 2022, Bargiotas et al., 2022, Sbidian et al., 2020, la Tour et al., 2018, Czernichow et al., 2020, Sebia et al., 2023, Brat et al., 2020, Lalanne et al., 2020]), this task is challenging, and extra caution measures should be taken.

Let me start with the following, extremely famous story [Kearns & Roth, 2019], that illustrates well the catastrophic consequences of poor data management. Netflix[1][2] is an American media company. It produces movies and TV shows, but they mostly serve as a way to promote its main product : an over-the-top[3], on-demand[4], paid-subscription-based, video platform. On this platform, users can watch movies and TV shows produced either by Netflix directly, but also by many others to who Netflix gives in turn some money. For such a platform, the first necessary condition to succeed is to offer a large collection of high quality movies and TV shows. However, this is certainly not the only one. At the core of what makes the Netflix experience is its recommendation system. It is the algorithm that recommends the Netflix users new programs to watch based on what the platform thinks the user's tastes are. Good suggestions lead to users that spend more time on the platform, and in turn to users that are more satisfied. In practice, the platform

[1] https://about.netflix.com/en
[2] https://en.wikipedia.org/wiki/Netflix
[3] Offered directly to the consumer via internet.
[4] Not constrained by a strict schedule.

recommends its users shows that were appreciated by users that the platform believes to have similar tastes.

So, in 2006, Netflix started a data challenge with the hope that, by releasing an *anonymized* dataset of ratings given by a subset of its users to a subset of its shows, the machine learning community would find a better recommendation algorithm than their home-brewed version. The metrics for the evaluation, and the continuous improvements on the challenge, are not relevant here. However, the hot topic for this thesis is the dataset that was released by Netflix.

The dataset that Netflix released is a collection of more than 100 million [Bennett et al., 2007] records of the form (`anonymized_user_id`, `show_name`, `rating`), where

- `anonymized_user_id` refers to a field that allows to uniquely identify the user that gives the review, without further detail about the user's identity,
- `show_name` is the name of the show to which the review is given,
- and `rating` is the actual value of the review, which is an integer between 1 and 5 (think about it as "stars").

Did the release of this dataset respect Netflix's users' privacy ? There is no clear answer to that question. Someone could indeed argue that "yes", Netflix has made a sufficient effort at hiding its users' identity. Indeed, this dataset does not contain any information that allows to directly identify the users (such as names, zip codes, ...). However, no one can guarantee that such dataset won't ever be deanonymized.

In fact, and this is the reason why this story is so popular, it *was* partially deanonymized. During the same year (2006), a young researcher (at the time PhD student of Vitaly Shmatikov[5]) named Arvind Narayanan[6] uploaded an article on arXiv[7] [Narayanan & Shmatikov, 2006], claiming to have recovered sensitive information about the users that were part of the "anonymized" dataset.

How did they proceed ? Without diving too much into technicalities, they used the

[5] https://www.cs.cornell.edu/~shmat/
[6] https://www.cs.princeton.edu/~arvindn/
[7] https://arxiv.org/

dataset of *public* reviews of the website IMDb[8] in order to recover the *hidden* reviews of the *common* users with the Netflix dataset. The idea is the following : if a user belongs to both datasets, he should give similar reviews to the shows that he rated both on Netflix and on IMDb. Those similarities can be leveraged in order to find matchings. Once a matching between a user in the IMDb database and in the Netflix dataset has been found, it is possible to look at the shows that were rated by the user on Netflix (thinking that he was anonymous), but not on IMDb (knowing that the reviews were public). By doing so, [Narayanan & Shmatikov, 2006] claims that it is possible to find the political orientation or even the sexual preferences of a subset of common users.

This revelation ultimately led to Netflix canceling its challenge[9], and to a class action lawsuit against Netflix, that was ultimately dismissed after a settlement with the plaintiffs was found[10].

So, what is the take-home message from that story ? For me, it is that the privacy of a pipeline that involves sensitive data can be more complex than it seems, and sometimes the intuition can be fooled. In the case of the Netflix challenge, even though the identifiable labels (name, zip code, ...) were removed, people were still identifiable because of the relative *uniqueness of their tastes*.

Let me extrapolate a bit from this example. What would have happened if, instead of publicly releasing the dataset, Netflix only gave it to a restricted list of whitelisted researchers ? Maybe one of the researchers could have been compromised and would have then leaked the dataset, but for the sake of reasoning, let us suppose that it is not the case. The researchers develop their new recommendation algorithm, no data leaks, and Netflix implements it on its platform. Would it be possible that, through its own recommendations that were generated by this algorithm, a user infers the private tastes of other specific users ?

Even though this question may sound at first as if it was borrowed from a conspiracy theory, giving a scientific answer is not an easy task. It raises two important questions : "How would it even be possible ?" and "How can we guarantee that it would be impossible, or at least hard ?".

In this thesis, I will try to give answers to both questions (not necessarily in the case of the Netflix example), and to precisely characterize the frontier between what is doable or not while guaranteeing a certain level of privacy.

[8] https://www.imdb.com
[9] https://www.nytimes.com/2010/03/13/technology/13netflix.html
[10] https://www.wired.com/2009/12/netflix-privacy-lawsuit/

Note that this example is not isolated, and that the literature on privacy attacks is massive [Narayanan & Shmatikov, 2006, Backstrom et al., 2007, Fredrikson et al., 2015, Dinur & Nissim, 2003, Homer et al., 2008, Loukides et al., 2010, Narayanan & Shmatikov, 2008, Sweeney, 2000, Gonon et al., 2023b, Wagner & Eckhoff, 2018, Sweeney, 2002, Voyez et al., 2022b].

0.1 Some context and definitions

I chose to name my thesis "On the tradeoffs of statistical learning with privacy" for reasons that will hopefully be clear by the end of the Introduction. Here, I define what the terms *statistical learning* and *privacy* mean in this title, and I intuitively introduce the *tradeoff* between them that will be at the core of this thesis.

0.1.1 A definition of statistical learning

The term *statistical learning* [Shalev-Shwartz & Ben-David, 2014, Bach, 2021] can be defined as *a way to "learn" quantities or behaviors that are meaningful at the scale of a population, with only access to observations of that population (samples).* For instance, estimating the proportion of people from the general population that like a movie based on the reviews of a restricted set of reviewers.

In particular, in this thesis, all the problems that are considered fit in this framework : we will have access to a dataset $\mathbf{X} = (X_1, \dots, X_n)$ corresponding to the observations of the data of n individuals. We will suppose that this dataset was generated from a distribution $\mathbb{P}$, and the objective will be to construct an estimator $\hat{\theta}(\mathbf{X})$ such that $\hat{\theta}(\mathbf{X}) \approx \theta(\mathbb{P})$, $\theta(\mathbb{P})$ being the quantity of interest from $\mathbb{P}$ (e.g. parametric or non-parametric estimation, more abstract behavior like regression error, comparative behavior to the empirical data). For instance, in the previous example, the likings of the reviewers can be modelled as $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{B}(p)$ (independently and identically distributed according a Bernoulli distribution of probability of success p). The problem would be to build an estimator from $X_1, \dots, X_n$ that approximates p.

0.1.2 A definition of privacy

The notion of privacy is harder to define. We will see down the line that it is possible to define it mathematically via the property of *differential privacy* [Dwork et al., 2006b, Dwork et al., 2006a]. However, a mathematical model is only as good as it reflects what we want it to. At first, I will thus try to informally draw the boundaries of the concept of privacy. It will later help us to understand why differential privacy works.

According to the Cambridge Dictionary[11], privacy is about *"not revealing somebody's information"*. This definition is a bit imprecise as to how to interpret it. Indeed, it

[11] https://dictionary.cambridge.org/

doesn't specify what "revealing" means. For instance, if I reveal somebody's age +1 year, did I reveal his age ? Someone could argue that "no", the age wasn't revealed. However, any good willing people would probably argue that "yes" the age was leaked. Indeed, the inverse transformation is extremely easy to do.

From the last example, we can try to modify this definition and say that privacy is about *not revealing something that could be used to recover somebody's information.* However, this definition is extremely strong. It implies that the only things that can be communicated should be independent of the data. This definition completely blocks learning. The challenge is to find a definition that allows both learning and privacy.

Instead, let me propose (inspired by differential privacy [Dwork et al., 2006a, Dwork et al., 2006b]) the following definition of privacy : privacy is about *only revealing things that make discriminating if one's information was used hard.* With this definition, it is still possible to have correlation between the data and the quantities that are communicated, but the recovering of one's information must be hard. "How hard ?" is a question that will later be mathematically characterized. Another question that must be answered is "hard for who ?". We will call *weak* privacy the scenario in which the adversaries are known in advance, and *strong* privacy the one in which the adversaries can be anything. Besides, the property of differential privacy that will be used later guarantees a certain level of privacy against any adversary. It is thus a *strong* definition of privacy.

0.1.3 An example of a non-private learning model

Before properly defining privacy, let us start by looking at a famous learning algorithm, and let us show that it has big privacy issues, in the sense that it is easily reverted. The dataset consists of n pairs (data, value) $((x_i, y_i))_{i=1,\dots,n}$ where the x's live in a metric space $\mathcal{X}$, equipped with a distance function d. The learning algorithm that we consider is the nearest neighbor predictor. Given a new data x, it predicts its associated value $\hat{y}$ with

$$\hat{y} := y_{\hat{i}} \ ,$$

where

$$\hat{i} := \operatorname*{arg\,min}_{i=1,\dots,n} d(x, x_i) \ .$$

With only a black box access to the nearest neighbor predictor (that is the ability to query its output for any x), it can be very easy to recover the full training set $((x_i, y_i))_{i=1,\dots,n}$. For instance, in a regression setup, it is not unreasonable to suppose that the marginal distribution of the y's is continuous. Then, almost surely, all the y's from the training set are distinct. Furthermore, by querying a grid of arbitrary precision of the space $\mathcal{X}$, it is possible to (i) obtain the values of all the y's of the training set, and to (ii) obtain the Voronoi diagram[12] or Dirichlet tessellations induced by the predictor on $\mathcal{X}$ with an arbitrary precision. Reconstructing the x's from the training set then boils down to the

[12] https://en.wikipedia.org/wiki/Voronoi_diagram

inversion of this Dirichlet tessellations, which is a well studied problem [Ash & Bolker, 1985, Aurenhammer, 1987, Hartvigsen, 1992, Schoenberg et al., 2003, Yeganova et al., 2001, Aloupis et al., 2013]. For instance, when d is Euclidean, the solution is often either unique, or parametrized by a few parameters only [Ash & Bolker, 1985]. In the first case, the inversion is thus possible. In the second one, if the information of only a few of the x's leak, it is possible to reconstruct all the other ones.

0.1.4 The nature of the tradeoff

The tradeoff between utility and privacy will later be investigated mathematically. However, at this point, we can already feel its nature. Privacy acts as a *constraint* on the data pipeline. Furthermore, by taking the last definition of privacy, the harder we want the discrimination process to be, the more restrictive privacy is on the data pipeline. This filtration of the usable pipelines means that possibly, we will exclude all the pipelines that obtained good utility for the task at hand. This forms the basis of the fundamental tradeoff implying privacy : the *tradeoff between utility and privacy.*

This tradeoff can be strong or weak in nature. For instance, the tradeoff is weak if it only measures the degradation of performance of *a given* data pipeline by the addition of privacy, compared to the same unrestricted pipeline, or to the best unrestricted pipeline. In contrast, the tradeoff is strong if it characterizes the degradation of performance of *any* private data pipeline compared to the *best* unrestricted data pipeline for the same task. By the end of the introduction, we will already have seen an example of *strong* tradeoff.

0.2 Attack surface, and the boundaries of the thesis

The topic of security and privacy is so large that entire journals and conferences are devoted to the subject[13] [14] [15]. This subsection of the introduction presents the boundaries of the subject of this thesis, and gives bibliographic pointers to what is immediately outside this boundary.

Let us enumerate the most common steps involved in a data pipeline. Data is collected, communicated, aggregated and processed. Furthermore, this list is not necessarily sequential, and in particular, it is possible to communicate on various quantities that appear at different stages of the pipeline.

In this thesis, we will always fall into the following scenario : we suppose that after collection, the data is centralized by a common aggregator that processes it. This aggregator then communicates to the world a quantity that is built from the data that is collected.

[13] https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=8013
[14] https://www.ieee-security.org/TC/SP2022/
[15] https://onlinelibrary.wiley.com/journal/24756725

The privacy problem is the following : Is it possible to recover information about the data from the sole observation of the outputted quantity.

In particular, we consider that the aggregator is trustworthy, and that the communication channels hide all information to everyone, except to the two communicating parties. This scenario is rather restrictive, but still extremely rich. The rest of this section presents techniques that allow relaxing those strong hypotheses. They will not be explored further in the rest of the thesis.

0.2.1 Can an actor be trusted ?

The first important hypothesis is that actors can be trusted, and that communication channels are secured. Thanks to symmetric and asymmetric encryption [Simmons, 1979], the second part of this hypothesis is reasonable. For the first part, on the other hand, as long as the data is stored somewhere and is not encrypted, data breaches can happen. Furthermore, the task of learning from the data often requires being able to perform arithmetic operations on the representations of the data (such as additions and multiplications), which are not compatible with classical encryption schemes.

In order to solve this problem, a possible solution is the use of homomorphic encryption (see [Acar et al., 2018] for a comprehensive survey). The term homomorphic encryption refers to encryption schemes that preserve the morphisms (such as multiplications and additions). It thus makes the task of learning (at least some) possible, with only access to encrypted data. It often comes however at the cost of extra computational complexity, and possible error terms.

0.2.2 Data collection and centralization

The second hypothesis about the setup of the thesis is that the data is centralized by a common aggregator. From a security point of view, this is often seen as a problem since it introduces a single point of failure. However, with our hypothesis that the actors can be trusted, this is not a problem. It also poses another problem with data sovereignty. In order to relax the hypothesis, both federated learning and local differential privacy are great options.

Federated learning [McMahan et al., 2017, Konečný et al., 2016, McMahan et al., 2018a, Bonawitz et al., 2019, Vanhaesebrouck et al., 2017, Bellet et al., 2018, Marfoq et al., 2021] has been proposed as a way to solve the problem of data centralization. Without digging into details, it can be defined as *a set of techniques allowing the training of common machine learning models, by aggregating the information of decentralized datasets or data holders.* A simple example would be the following : a model is trained via SGD. An agent

that holds a dataset and the current iterate of the model runs an EPOCH[16] of SGD[17] on the data that he owns. He then gives the updated model to the next agent, and the process continues. For more details, the reader may refer to the surveys [Zhang et al., 2021b, Kairouz et al., 2021].

Local differential privacy [Duchi et al., 2013] blurs the data with noise as soon as it exits its original data holder (as opposed to regular differential privacy that blurs the information when it exits the aggregator). At first, it seems like a more appealing notion of privacy (since the privacy guarantees are stronger), however, it comes at the cost of utility since local differentially private mechanisms typically end up with a lot more noise than their non-local alternatives. The reader may refer to the survey [Yang et al., 2020].

0.3 The vocabulary of statistical learning

Linking back to the subject of the thesis, an important question is whether an estimator (private or not) estimates well the quantity of interest (that is defined at the scale of the population).

In the following, we consider that we have access to a dataset $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}^n$ generated from some distribution $\mathbb{P}$. In many applications, the independence and identical distribution assumption is made, namely that $\mathbb{P} = \mathbb{p}^{\otimes n}$ for some distribution $\mathbb{p}$ on $\mathcal{X}$.

0.3.1 The formalization for estimation

First, for the estimation problem (which will be the main topic of this thesis), we suppose that the data distribution can be fully described by a parameter $\theta \in \Theta$. We note it $(X_1, \dots, X_n) \sim \mathbb{P}_\theta$. Say we have built an estimator $\hat{\theta}$ from the dataset $(X_1, \dots, X_n)$, how to measure its utility as an estimator of the true parameter θ ?

This is usually done by taking a cost function $c : \Theta \times \Theta \to \mathbb{R}_+$ that measures how close its two arguments are, with the convention that the lower, the better. For instance, when Θ is a subspace of some Euclidean space, it is common practice to take $c(x, y) = \|x - y\|^\alpha$ for $\alpha \geq 1$. Hence, the utility of our estimator $\hat{\theta}$ may be measured as

$$c(\hat{\theta}, \theta) \ . \tag{1}$$

However, this quantity is a random variable. A common practice is thus either to control it in probability (i.e. that it is small with high probability), or to take its expectation (notice that c is positive). The performance can hence be measured as

$$\mathcal{R}\left(\hat{\theta}\right) := \mathbb{E}_{(X_1,\dots,X_n)\sim\mathbb{P}_\theta,\hat{\theta}}\left(c(\hat{\theta}, \theta)\right) \ , \tag{2}$$

[16] i.e. a complete pass over the dataset
[17] Stochastic Gradient Descent

which is commonly called the *risk* of $\hat{\theta}$, and where the subscript $\hat{\theta}$ in the expectation is used to refer to all the sources of randomness that might be in $\hat{\theta}$ and that are independent of $(X_1, \dots, X_n)$. For instance, in the Euclidean case, when $c(x, y) = \|x - y\|^2$ the quantity Equation (2) is usually referred to as the quadratic risk of the estimator $\hat{\theta}$.

0.3.2 The formalization for a specific task

Even if it will not be the main subject of this thesis, it is not possible to talk about learning without presenting the case in point of the measure of performance for supervised learning. Here, we suppose that $((a_1, b_1), \dots, (a_n, b_n)) \sim \mathbb{p}_\theta^{\otimes n}$. For instance, it handles the famous regression model where $b_i = f(a_i) + \epsilon_i$ where f is the regression function to learn, and ϵ_i is some noise.

A first approach would be to build an estimator $\hat{f}$ of the function f and to measure its utility as previously by leveraging a norm or a semi-norm in a functional space. Another approach is to measure the utility of $\hat{f}$ by measuring its predictive utility on unseen data with the same distribution. For instance, the utility could be measured with

$$\mathcal{R}\left(\hat{f}\right) := \mathbb{E}_{((a_1,b_1),\dots,(a_n,b_n),(a,b))\sim\mathbb{p}_\theta^{\otimes(n+1)},\hat{f}}\left(c'(\hat{f}(a), b)\right) ,$$

where c' is a cost function defined on the output space of f instead of the parameter space.

This measure of performance measures the relevance of $\hat{f}$ not for how close it is to f but for how well it behaves like f on a specific task and against the true data distribution. In particular, this formulation makes a lot of sense when the mapping $\theta \mapsto \mathbb{p}_\theta$ is not "injective" (i.e. when the model is not identifiable), which is for instance the case with neural networks which are often invariant by permutation and rescaling of the weights.

Notice that this new measure of performance perfectly fits in the previous scenario, however, this formulation is of key interest in predictive scenarios. In particular, the measure of performance can be approximated by Monte-Carlo methods without having to know the ground-truth parameter θ. The empirical risk of an estimator $\hat{f}$ is hence defined as

$$\mathcal{R}_n\left(\hat{f}\right) := \frac{1}{n}\sum_{i=1}^{n} c'(\hat{f}(a_i), b_i) .$$

In particular, this formulation of the problem often allows controlling the *generalization error* [Musavi et al., 1994, Vapni, 1995, Mohri et al., 2012, Gonon et al., 2023a], which is equal to $\mathcal{R}_n\left(\hat{f}\right) - \mathcal{R}\left(\hat{f}\right)$ for a given estimator $\hat{f}$ of f. Such bounds are usually derived using the theory of VC dimension [Blumer et al., 1989, Vapnik, 2006], the Rademacher complexity [Koltchinskii & Panchenko, 2000, Koltchinskii, 2001, Bartlett & Mendelson,

2002], covering number arguments [Dudley, 1967, Haussler, 1995], Pac-Bayesian methods [McAllester, 1999, Haddouche et al., 2020, Haddouche et al., 2021, Haddouche & Guedj, 2022, Haddouche & Guedj, 2023, Haddouche et al., 2023], or the algorithmic stability [Rogers & Wagner, 1978, Bousquet & Elisseeff, 2002, Xu et al., 2012] of the learning algorithm.

Recently, this last proof framework has been used to prove that with differential privacy (wait for the next section for a proper definition), the generalization gap can be upper-bounded [Dwork et al., 2015, Oneto et al., 2017, Nissim & Stemmer, 2015, He et al., 2021] by the sole effect of privacy. In other word, it shows that differential privacy is a *sufficient condition* to have a small generalization gap. That being said, the empirical risk of the produced private estimator $\mathcal{R}_n\left(\hat{f}\right)$ may be high, and in this case, a small generalization gap just says $\hat{f}$ behaves as poorly on the data that he has not seen as on the data that he has seen. A small generalization gap is a desirable property to have, but it is not sufficient to characterize the effect of privacy on the estimation difficulty. We may now close the parenthesis about the generalization gap under differential privacy. The rest of the Thesis fits in the general setup of Section 0.3.1, and Chapter 1 will be the only part where the specific formalism of Section 0.3.2 is more suited.

0.4 How to formalize privacy ?

The gold standard in privacy protection is the definition of differential privacy. It gives strong privacy guarantees while still being extremely handy to use in many situations. It is notably used by the US Census Bureau [Abowd, 2018], Google [Erlingsson et al., 2014], Apple [Thakurta et al., 2017] and Microsoft [Ding et al., 2017], among many others.

0.4.1 Differential privacy

Given $n \in \mathbb{N}_*$ and a feature space $\mathcal{X}$, $\mathcal{X}^n$ may be viewed as a set of datasets containing n elements from $\mathcal{X}$. Given $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}^n$ and $i \in \{1, \dots, n\}$, X_i is the data record of the individual i from the database.

On $(\mathcal{X}^n)^2$, the *Hamming*[18] distance is defined as

$$d_{\text{ham}}\left((X_1, \dots X_n), (Y_1, \dots Y_n)\right) := \sum_{i=1}^{n} \mathbb{1}_{X_i \neq Y_i} \ .$$

In particular, for $\mathbf{X}, \mathbf{Y} \in \mathcal{X}^n$ and $k \in \mathbb{N}$, $d_{\text{ham}}(\mathbf{X}, \mathbf{Y}) \leq k$ when the datasets $\mathbf{X}$ and $\mathbf{Y}$ differ by the records of at most k individuals.

The core idea of *differential privacy* is to say that if a mechanism $\mathfrak{M}$ was to work on a dataset $\mathbf{X} \in \mathcal{X}^n$ and to output $\mathfrak{M}(\mathbf{X})$ in some output space $\text{codom}(\mathfrak{M})$, then it should

[18] from its inventor Richard Hamming

have outputted a *similar* output if applied to any other dataset that differed from $\mathbf{X}$ only on few records. Formally, this "similarity" is characterized in terms of distributions. Hence, a differentially private mechanism that is not constant is *necessarily stochastic*. The following definition is due Cynthia Dwork[19], Frank McSherry[20], Kobbi Nissim[21], and Adam D. Smith[22] [Dwork et al., 2006b], and it owed them the Gödel prize in 2017. Given $\epsilon \in \mathbb{R}_{+*}$, a randomized mechanism $\mathfrak{M} : \mathcal{X}^n \to \text{codom}(\mathfrak{M})$ is ϵ-differentially private (or ϵ-DP) if for any $\mathbf{X}, \mathbf{Y} \in \mathcal{X}^n$ and any measurable S of the output space $\text{codom}(\mathfrak{M})$, we have

$$d_{\text{ham}}(\mathbf{X}, \mathbf{Y}) \leq 1 \implies \mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\mathbf{X}) \in S) \leq e^{\epsilon} \mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\mathbf{Y}) \in S) \ . \tag{3}$$

When $d_{\text{ham}}(\mathbf{X}, \mathbf{Y}) \leq 1$, we say that $\mathbf{X}$ and $\mathbf{Y}$ are *neighbors*. An interpretation of this definition is that if the datasets $\mathbf{X}$ and $\mathbf{Y}$ vary on the records of at most one individual, then the output distributions should be close.

The parameter ϵ is usually referred to as the *privacy budget*. The bigger it is, the looser the constraint of Equation (3) becomes. On the opposite side, if it is very small, it forces the distributions of $\mathfrak{M}(\mathbf{X})$ and of $\mathfrak{M}(\mathbf{Y})$ to be extremely close.

As a point of reference, in 2020, the data anonymized by the US Census Bureau was released with a ϵ of around 20[23].

Personally, I like to view differential privacy as an analogous notion to the one of quotient in algebra. Indeed, the famous isomorphism theorem states that given a morphism, its image is isomorphic to the original structure quotiented by its kernel. In the construction of the measure theory and of the L^p spaces, it is frequent to identify objects that differ by negligible aspects. With differential privacy, the output distributions of neighboring datasets are not equal (contrary to quotient structures), but they are close. The answer to the question "how close ?" can be tuned by varying ϵ. An alternative abstract view of differential privacy could be as a form of Lipschitz condition for the mechanism, that is of probabilistic nature.

This notion of neighboring (i.e. differing on the data record of at most one individual) makes differential privacy all the more compatible with statistical learning. Indeed, by its nature, statistical estimation does not care about the data of a single individual. It tries to find patterns that are meaningful at the scale of the population.

[19] https://en.wikipedia.org/wiki/Cynthia_Dwork
[20] https://en.wikipedia.org/wiki/Frank_McSherry
[21] https://people.cs.georgetown.edu/~kobbi/
[22] https://cs-people.bu.edu/ads22/
[23] https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html

0.4.2 Privacy guarantees

In order to understand the guarantees provided by differential privacy, we can play the role of an attacker. There is a ϵ-DP mechanism $\mathfrak{M}$ and two datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ that differ on the records of at most one individual. We have access to $\mathfrak{M}(\mathbf{X}_?)$, and we want to determine if $\mathbf{X}_? = \mathbf{X}_1$ or if $\mathbf{X}_? = \mathbf{X}_2$. That is, we have access to the records of all the individuals except one, and we want to test between two possibilities for this remaining record.

We set up a decision rule (or statistical test). We decide of a $S \subset \mathrm{codom}(\mathfrak{M})$. If $\mathfrak{M}(\mathbf{X}_?) \in S$, we say that $\mathbf{X}_? = \mathbf{X}_1$. Conversely, if $\mathfrak{M}(\mathbf{X}_?) \notin S$, we say that $\mathbf{X}_? = \mathbf{X}_2$. What is the error of this test ?

The type 1 error α is defined as

$$\alpha := \mathbb{P}_{\mathbf{X}_?=\mathbf{X}_1}(\mathfrak{M}(\mathbf{X}_?) \notin S) \ .$$

It measures the probability of $\mathbf{X}_1$ being falsely rejected. Likewise, the type 2 error β is defined as

$$\beta := \mathbb{P}_{\mathbf{X}_?=\mathbf{X}_2}(\mathfrak{M}(\mathbf{X}_?) \in S) \ .$$

It measures the probability of $\mathbf{X}_1$ being falsely selected.

Since $\mathfrak{M}$ is ϵ-DP, it follows that

$$\beta = \mathbb{P}_{\mathbf{X}_?=\mathbf{X}_2}(\mathfrak{M}(\mathbf{X}_?) \in S) \geq e^{-\epsilon}\mathbb{P}_{\mathbf{X}_?=\mathbf{X}_1}(\mathfrak{M}(\mathbf{X}_?) \in S) = e^{-\epsilon}(1-\alpha) \ ,$$

and likewise that

$$\alpha \geq e^{-\epsilon}(1-\beta) \ .$$

In other words, α and β cannot be arbitrarily small at the same time. Any test will either falsely reject or falsely select $\mathbf{X}_1$ frequently.

This proves the strong nature of the guarantees provided by differential privacy. No adversary can do better than a certain efficiency fixed by ϵ.

0.4.3 Statistical implications and Bernoulli example

Differential privacy acts as a constraint on the set of usable estimators. As for other constraints (e.g. restricted bandwidth, ...) [Barnes et al., 2019, Barnes et al., 2020b, Acharya et al., 2021a, Acharya et al., 2021c, Acharya et al., 2021d, Acharya et al., 2021b], it is interesting to study its consequences on learning and statistical estimation. This question will be the central question of this thesis.

We start by looking at the simple example of Bernoulli parameter estimation. We are given $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{B}(p)$ where $p \in [0,1]$, and the task is to estimate p from $\mathbf{X} := (X_1, \dots, X_n)$.

If privacy was not an issue, the most natural estimator would probably be the moment estimator

$$\hat{p}(X_1, \dots, X_n) := \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) . \tag{4}$$

Its quadratic risk (the measure of performance) may be computed as

$$\begin{aligned} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}}\left((\hat{p}(\mathbf{X}) - p)^2\right) &= \left(\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}}\left(\hat{p}(\mathbf{X})\right) - p\right)^2 + \mathbb{V}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}}\left(\hat{p}(\mathbf{X})\right) \\ &= (p-p)^2 + \frac{p(1-p)}{n} = \frac{p(1-p)}{n} . \end{aligned} \tag{5}$$

However, this estimator is not differentially private. Indeed, it is deterministic and not constant.

In contrast, the following estimator is ϵ-DP :

$$\mathfrak{M}(X_1, \dots, X_n) := \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{n\epsilon}\mathcal{L}(1) ,$$

where $\mathcal{L}(1)$ should be interpreted as a random variable (independent of $\mathbf{X}$) following the Laplace distribution $\mathcal{L}(1)$. This claim is in fact a simple application of the so-called Laplace mechanism [Dwork et al., 2006b, Dwork et al., 2006a] that will be presented in Chapter 2. For the completeness on this introduction, we give a brief proof.

Let $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^n$ such that $d_{\text{ham}}(\mathbf{X}, \mathbf{X}') \leq 1$. By noting $\mathbb{P}_{\mathfrak{M}(\mathbf{X})}$ (rest. $\mathbb{P}_{\mathfrak{M}(\mathbf{X}')}$) the output distribution of $\mathfrak{M}$ when applied to $\mathbf{X}$ (resp. $\mathbf{X}'$), we can first notice that both of them are absolutely continuous w.r.t. Lebesgue's measure on $\mathbb{R}$. We can thus use $p_{\mathfrak{M}(\mathbf{X})}$ and $p_{\mathfrak{M}(\mathbf{X}')}$ to refer to their respective densities. We have (almost surely in x) that

$$\frac{p_{\mathfrak{M}(\mathbf{X})}(x)}{p_{\mathfrak{M}(\mathbf{X}')}(x)} = \frac{\frac{1}{2}e^{-\left|x-\epsilon\left(\sum_{i=1}^n X_i\right)\right|}}{\frac{1}{2}e^{-\left|x-\epsilon\left(\sum_{i=1}^n X'_i\right)\right|}} \overset{\text{triangular inequality}}{\leq} e^{\epsilon\left|\left(\sum_{i=1}^n X_i\right)-\left(\sum_{i=1}^n X'_i\right)\right|} ,$$

and since $d_{\text{ham}}(\mathbf{X}, \mathbf{X}') \leq 1$, it follows that almost surely in x,

$$\frac{p_{\mathfrak{M}(\mathbf{X})}(x)}{p_{\mathfrak{M}(\mathbf{X}')}(x)} \leq e^{\epsilon} .$$

Hence, for any Borel set S,

$$\int_S p_{\mathfrak{M}(\mathbf{X})} \leq e^{\epsilon} \int_S p_{\mathfrak{M}(\mathbf{X}')} ,$$

which translates to

$$\mathbb{P}\left(\mathfrak{M}(\mathbf{X}) \in S\right) \leq e^{\epsilon}\mathbb{P}\left(\mathfrak{M}(\mathbf{X}') \in S\right) .$$

This proves that $\mathfrak{M}$ is ϵ-DP.

We may now measure the performance of $\mathfrak{M}$ by looking at its quadratic risk as

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}}\left((\mathfrak{M}(\mathbf{X})-p)^2\right) &= \left(\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n},\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X})\right)-p\right)^2 + \mathbb{V}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n},\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X})\right) \\ &= (p-p)^2 + \frac{p(1-p)}{n} + \frac{2}{n^2\epsilon^2} = \frac{p(1-p)}{n} + \frac{2}{n^2\epsilon^2}\,. \end{aligned} \tag{6}$$

0.5 What is the cost of privacy ?

What was the cost of privacy in the estimation of p, the parameter of the Bernoulli distribution, of the last section ? The comparison of Equation (5) and of Equation (6) shows that the performance with privacy is degraded additively by $\frac{2}{n^2\epsilon^2}$. But why do we have to compare those two estimators ? For instance, if we consider the estimator that is constant, equal to p, we can see that it has null quadratic risk under $\mathcal{B}(p)^{\otimes n}$. Furthermore, it is constant, and is hence ϵ-DP for any $\epsilon > 0$! Should we conclude that the best estimator achieves perfect estimation, and that this conclusion is not changed when the estimator is restricted to be private ?

If the last example feels puzzling, it is normal. The estimator constant to p works well under $\mathcal{B}(p)^{\otimes n}$, but knowing p in advance is cheating of course. In contrast, it performs poorly under any $\mathcal{B}(p')^{\otimes n}$ whenever $p' \neq p$. Instead, we see that the performance of an estimator should not only be measured against what *is*, but also against what *could have been*. With this in mind, we can see that for any constant estimator, we can always find a Bernoulli distribution such that it has a quadratic risk bigger than $\frac{1}{4}$. In comparison, $\hat{p}(\cdot)$ (see Section 0.4.3) has a quadratic risk that is smaller than $\frac{1/4}{n}$ on any Bernoulli distribution.

This idea of good performance under any possible outcome has led to the notion of *minimax* optimality. In this theory, it is only meaningful to compare the performance of a given estimator to the one *of the best estimator on its worst outcome*. Formally, the minimax risk of estimation of the model $(\mathcal{B}(p)^{\otimes n})_{p\in[0,1]}$ is the quantity

$$\inf_{\hat{p}' \text{ estimator}} \sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}}\left((\hat{p}'(\mathbf{X})-p)^2\right)\,. \tag{7}$$

For the Bernoulli model, it is well known that this minimax risk of estimation is $\Omega\left(\frac{1}{n}\right)$ [Rigollet & Hütter, 2015] (asymptotically lower-bounded by a positive constant times $\frac{1}{n}$). The proof can be found in Chapter 4, where it is used as an illustration. In particular, in comparison with the upper-bound of the non-private estimator $\hat{p}(\cdot)$, it means that $\hat{p}(\cdot)$ has *minimax-optimal convergence rate of estimation*, which means that the uniform upper

bound on the quadratic risk of $\hat{p}(\cdot)$ is comparable to a lower bound on the minimax risk for the estimation problem, up to multiplicative constants. Without privacy, the optimal rate of estimation is thus $\Theta\left(\frac{1}{n}\right)$ ($O\left(\frac{1}{n}\right)$ and $\Omega\left(\frac{1}{n}\right)$).

An important question that will be central in most of this thesis is whether this minimax rate of estimation is modified by privacy. Formally, the question is whether the quantity

$$\inf_{\mathfrak{M}:\epsilon-\mathrm{DP}} \sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n},\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X})-p)^2\right) \ . \tag{8}$$

is significantly bigger than the quantity of Equation (7) or not.

Equation (6) tells us that this quantity is upper-bounded by $O\left(\max\left(\frac{1}{n}, \frac{1}{(n\epsilon)^2}\right)\right)$, but we still need a matching lower-bound. For the Bernoulli example, we give a concise proof of such lower-bound that is a direct application of the techniques presented in Chapter 3.

Let $p_1 < p_2$ be two parameters in $(0,1)$ and let $U_1, \ldots, U_n$, be n independent and identically distributed uniform random variables on $[0,1]$. The random variables $Z_i := (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$, $1 \leq i \leq n$, defined by

$$(X_i^{(1)}, X_i^{(2)}) = (\mathbb{1}_{[0,p_1)}(U_i), \mathbb{1}_{[0,p_2)}(U_i))$$

are independent and identically distributed with marginal distributions Bernoulli $\mathcal{B}(p_1)$ and $\mathcal{B}(p_2)$. In the sequel we note $\mathbf{X}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})$, $j = 1, 2$, $\mathbf{U} = (U_1, \ldots, U_n)$, $S_1 := [0, (p_1+p_2)/2)$. and $S_2 := [(p_1+p_2)/2, 1]$. Given any $(\epsilon, 0)$-DP mechanism $\mathfrak{M} : [0,1]^n \to [0,1]$ (where $\epsilon > 0$) to estimate the Bernoulli parameter, the risk satisfies

$$\begin{aligned}
&\sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}}\left((\mathfrak{M}(\mathbf{X})-p)^2\right) \\
&\qquad \geq \left(\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p_1)^{\otimes n},\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X})-p_1)^2\right) + \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p_2)^{\otimes n},\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X})-p_2)^2\right)\right)/2 \\
&\qquad \overset{\text{Coupling}}{=} \left(\mathbb{E}_{\mathbf{U},\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X}^{(1)})-p_1)^2\right) + \mathbb{E}_{\mathbf{U},\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X}^{(2)})-p_2)^2\right)\right)/2 \\
&\qquad \overset{\text{Conditioning}}{=} \mathbb{E}_{\mathbf{U}}\left(\mathbb{E}_{\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X}^{(1)})-p_1)^2\right) + \mathbb{E}_{\mathfrak{M}}\left((\mathfrak{M}(\mathbf{X}^{(2)})-p_2)^2\right)\right)/2 \\
&\qquad \geq \left(\tfrac{p_2-p_1}{2}\right)^2 \mathbb{E}_{\mathbf{U}}\left(\mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}^{(1)}) \in S_2\right) + \mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_1\right)\right)/2.
\end{aligned} \tag{9}$$

This is where the DP property yields a lower bound on the second factor as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{U}}\Big(e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})}\mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_2\right) + \mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_1\right)\Big) \\
&\quad \overset{d_{\mathrm{ham}}(\cdot,\cdot)\geq 0}{\geq} \mathbb{E}_{\mathbf{U}}\Big(e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})}\Big(\mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_2\right) + \mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_1\right)\Big)\Big) \\
&\quad = \mathbb{E}_{\mathbf{U}}\Big(e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})}\Big) \overset{\text{Jensen}}{\geq} e^{-n\epsilon|p_2-p_1|} \ ,
\end{aligned} \tag{10}$$

which overall yields the lower bound $\frac{(p_2-p_1)^2}{8}e^{-n\epsilon|p_2-p_1|}$.

A good lower bound on the minimax risk is then provided by optimizing over p_1 and p_2. For instance, when $n \geq \frac{2}{\epsilon}$, $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{2} + \frac{1}{n\epsilon}$ lead to

$$\sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n},\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}) - p)^2\right) \geq \frac{1}{8}\frac{1}{(n\epsilon)^2} \ .$$

Since this is true for any ϵ-DP $\mathfrak{M}$, and since any ϵ-DP estimator is also in particular an estimator, it is possible to write that

$$\inf_{\mathfrak{M}:\epsilon-\mathrm{DP}} \sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}} \left((\mathfrak{M}(\mathbf{X}) - p)^2\right) \geq \Omega\left(\max\left(\frac{1}{n}, \frac{1}{(n\epsilon)^2}\right)\right) \ . \tag{11}$$

Together, the upper-bound of Equation (6) and the lower-bound of Equation (11) prove that the minimax rate of estimation of the parameter of a Bernoulli distribution under ϵ-differential privacy is $\Theta\left(\max\left(\frac{1}{n}, \frac{1}{(n\epsilon)^2}\right)\right)$, which is to compare to the non-private minimax rate of estimation of $\Theta\left(\frac{1}{n}\right)$. In particular, two regimes arise.

Low privacy regime. When $\epsilon = \Omega\left(\frac{1}{\sqrt{n}}\right)$, the minimax rate of estimation is *unchanged* by privacy. We call this regime the *low privacy* regime, in which privacy basically comes for free on a statistical point of view.

High privacy regime. In contrast, the regime $\epsilon \ll \frac{1}{\sqrt{n}}$ comes with a degradation of the minimax risk of estimation. Privacy in this regime comes with a *necessary cost* on the estimation complexity. In other words, any private estimator performs significantly worse than the best non-private estimator.

Through this example, we have seen our first example of strong privacy-utility tradeoff. Many more will be investigated in the rest of this thesis.

0.6 An overview of the thesis

This thesis is composed of six chapters. They are based on five research articles [Lalanne et al., 2023d, Lalanne et al., 2023b, Lalanne et al., 2023c, Gonon et al., 2023b], but the structure does not necessarily reflect a clear one to one mapping between the chapters and the articles (with one extra chapter). At the beginning of each chapter, a small clarification is made as to what articles were used for the chapter, and to who contributed to them.

0.6.1 Overview of Chapter 1

Deep neural networks are state-of-the-art for many learning problems. In practice, it is possible to tune the parameters of a given network in order to perfectly interpolate the available data [Zhang et al., 2021a]. This overfitting regime is of practical interest since good performance can be obtained this way [Belkin et al., 2019]. However, it comes with an increased risk in terms of privacy [Rigaki & Garcia, 2020], since the network memorizes information about training data, up to the point of interpolating them. Among these information, some might be confidential. This raises the question of what information can be inferred given a black-box access to the model.

To detect an overfitting situation, an indicator is given by the ratio of the number of parameters by the number of data points available: the more parameters there are, the more the model is likely to be able to interpolate the data. In order to hinder the capacity of the model to overfit, and thus to store confidential information, this work studies the role of the number of nonzero parameters used. *Can we find a good trade-off between model accuracy and privacy by tuning the sparsity (number of nonzero parameters) of neural networks?*

Attacks such as "Membership Inference Attack" (MIA) [Hu et al., 2022, Shokri et al., 2017, Truex et al., 2021, Rezaei & Liu, 2021, Hui et al., 2021, Long et al., 2018, Yeom et al., 2018, Salem et al., 2018, Sablayrolles et al., 2019, Voyez et al., 2022a] can infer whether a data point was a member of the training set [Shokri et al., 2017], using only a black-box (or white-box is other cases) access to the targeted model. This can be problematic in case of sensitive data (medical data, etc.). Given a network, how could one reduce the risk of such attacks, while preserving its performances as much as possible?

This chapter leverages sparsity in neural networks as a defense against MIA's. This is an empirical study and introduces the reader to real-world privacy attacks.

0.6.2 Overview of Chapter 2

This chapter presents key concepts and technical results about differential privacy that are necessary for the rest of the thesis. It puts an emphasis on embedding those results in a statistical framework in order to link them with the main theme of the thesis.

0.6.3 Overview of Chapter 3

Similarly to the small example for the Bernoulli estimation, this chapter studies minimax lower bounds for classes of differentially private estimators. In particular, it shows how to characterize the power of a statistical test under differential privacy in a plug-and-play fashion by solving an appropriate transport problem. With specific coupling constructions, this observation allows deriving Le Cam-type and Fano-type inequalities not only for regular definitions of differential privacy but also for those based on Rényi divergence.

This is a core chapter for the thesis that introduces theoretical tools that are used in the rest of the thesis.

0.6.4 Overview of Chapter 4

This chapter illustrates the results of the last chapter on three simple, fully worked out parametric examples. In particular, it shows that the problem class has a huge importance on the provable degradation of utility due to privacy. In certain scenarios, it shows that maintaining privacy results in a noticeable reduction in performance only when the level of privacy protection is very high. Conversely, for other problems, even a modest level of privacy protection can lead to a significant decrease in performance.

It also observes that the DP-SGLD algorithm, a private convex solver, can be employed for maximum likelihood estimation with a high degree of confidence, as it provides near-optimal results with respect to both the size of the sample and the level of privacy protection. This algorithm is applicable to a broad range of parametric estimation procedures, including exponential families.

Finally, it gives bibliographical pointers to many recent research articles studying similar problems of private parametric estimation problems.

0.6.5 Overview of Chapter 5

Given $\mathbf{X} := (X_1, \ldots, X_n) \sim \mathbb{P}_\pi^{\otimes n}$, where $\mathbb{P}_\pi$ refers to a distribution of probability that has a density π that is absolutely continuous with respect to Lebesgue measure on $[0, 1]$, this chapter studies the private estimation of π.

In terms of upper-bounds, this chapter analyzes histogram and so-called projection estimators at a resolution that captures the impact of the privacy and smoothness parameters. Furthermore, it proves new lower bounds by using classical packing method combined with new tools that characterize the testing difficulty under global privacy from [Acharya et al., 2021e, Kamath et al., 2022, Lalanne et al., 2023b].

In particular, for Lipschitz densities and under pure differential privacy, it recovers known results from [Barber & Duchi, 2014] with a few complements. It then extends the estimation on this class of distributions to the context of concentrated differential privacy [Bun & Steinke, 2016], a more modern definition of privacy that is compatible with stochastic processes relying on Gaussian noise. It finally investigates higher degrees of smoothness by looking at periodic Sobolev distributions.

0.6.6 Overview of Chapter 6

Any probability distribution $\mathbb{P}$ on $[0,1]$ is fully characterized by its cumulative distribution function (CDF) defined by

$$F_{\mathbb{P}}(t) := \mathbb{P}\big((-\infty, t]\big), \quad \forall t \in \mathbb{R}\,.$$

The central topic of this chapter is the quantile function (QF), $F_{\mathbb{P}}^{-1}$, defined as the generalized inverse of $F_{\mathbb{P}}$:

$$F_{\mathbb{P}}^{-1}(p) = \inf\left\{t \in \mathbb{R} \mid p \leq F_{\mathbb{P}}(t)\right\}, \quad \forall p \in [0,1]\,,$$

with the convention $\inf \emptyset = +\infty$. When $\mathbb{P}$ is absolutely continuous w.r.t. Lebesgue's measure with a density that is bounded away from 0, $F_{\mathbb{P}}$ and $F_{\mathbb{P}}^{-1}$ are bijective and are inverse from one another.

A well-known result is that, under mild hypotheses on $\mathbb{P}$, if $U \sim \mathcal{U}([0,1])$ (U follows a uniform distribution on $[0,1]$), then $F_{\mathbb{P}}^{-1}(U) \sim \mathbb{P}$ [Devroye, 1986]. In other words, knowing $F_{\mathbb{P}}^{-1}$ allows to generate data with distribution $\mathbb{P}$. It makes the estimation of $F_{\mathbb{P}}^{-1}$ a key component in data generation. Indeed, privately learning the quantile function would then allow generating infinitely many new coherent samples at no extra cost on privacy.

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, this chapter studies the *private* estimation of $F_{\mathbb{P}}^{-1}(p_j)$ from these samples at prescribed values $\{p_1, \ldots, p_m\} \subset (0,1)$.

Without privacy and under mild hypotheses on the distribution, it is well-known [Van der Vaart, 1998] that for each $p \in (0,1)$, the quantity $X_{(E(np))}$ is a good estimator of $F_{\mathbb{P}}^{-1}(p)$, where $X_{(1)}, \ldots, X_{(n)}$ are the order statistic of $X_1, \ldots, X_n$ (i.e. a permutation of the observations such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$) and $E(x)$ denotes the largest integer smaller or equal to x. The quantity $X_{(E(np))}$ is called the empirical (as opposed to statistical) quantile of the dataset $(X_1, \ldots, X_n)$ (as opposed to the distribution $\mathbb{P}$) of order p.

This chapter studies the properties of private *empirical* quantiles procedures, when applied for the corresponding statistical task. It started as a project in collaboration with Clément Gastaud and Nicolas Grislain from Sarus Technologies[24] with whom we proved the quasi-equivalence between the JointExp mechanism [Gillenwater et al., 2021] and the inverse sensitivity mechanism [Asi & Duchi, 2020b, Asi & Duchi, 2020a], and with whom we studied the statistical properties of those two estimators on continuous and atomic distributions. For them, quantiles are interesting for private data generation. Later during the preparation of my thesis, [Kaplan et al., 2022] proposed a new state-of-the-art mechanism for the empirical quantile problem. We chose to investigate the statistical properties

[24] https://www.sarus.tech/

of this new mechanism, and came up with nice concentration inequalities, proving a poly-logarithmic degradation of the utility when the number of quantiles increases.

Chapter 1

A practical case study : membership inference attacks and sparsity in neural networks

The origin of this chapter, and the use of the first person. This chapter is based on the article [Gonon et al., 2023b], written by my colleagues and friends Antoine Gonon[1], Can Pouliquen[2], Guillaume Lauga[3], Léon Zheng[4], and Quoc-Tung Le[5], and by myself. The genesis of the project was to find and investigate a common theme among the PhD students that were in the same research team. In this chapter, I will try to respect the following rule : the use of the first person of the plural (we, our, ...) represents all the above-mentioned people, while the use of the first person of the singular (I, my, ...) represents myself.

Deep neural networks are state-of-the-art for many learning problems. In practice, it is possible to tune the parameters of a given network in order to perfectly interpolate the available data [Zhang et al., 2021a]. This overfitting regime is of practical interest since good performance can be obtained this way [Belkin et al., 2019]. However, it comes with an

[1] https://agonon.github.io/
[2] https://perceptronium.github.io/
[3] https://laugaguillaume.github.io/
[4] https://leonzheng2.github.io/
[5] https://tung-qle.github.io/

increased risk in terms of privacy [Rigaki & Garcia, 2020], since the network memorizes information about training data, up to the point of interpolating them. Among these information, some might be confidential. This raises the question of what information can be inferred given a black-box access to the model.

To detect an overfitting situation, an indicator is given by the ratio of the number of parameters by the number of data points available: the more parameters there are, the more the model is likely to be able to interpolate the data. In order to hinder the capacity of the model to overfit, and thus to store confidential information, this work studies the role of the number of nonzero parameters used. *Can we find a good trade-off between model accuracy and privacy by tuning the sparsity (number of nonzero parameters) of neural networks?*

Attacks such as "Membership Inference Attack" (MIA) [Hu et al., 2022, Shokri et al., 2017, Truex et al., 2021, Rezaei & Liu, 2021, Hui et al., 2021, Long et al., 2018, Yeom et al., 2018, Salem et al., 2018, Sablayrolles et al., 2019, Voyez et al., 2022a] can infer the membership of a data point to the training set [Shokri et al., 2017], using only a black-box (or white-box is other cases) access to the targeted model. This can be problematic in case of sensitive data (medical data, etc.). Given a network, how could one reduce the risk of such attacks, while preserving its performance as much as possible?

Numerous procedures have been proposed to defend against MIAs [Hu et al., 2022]. In this work, the studied approach consists in decreasing the number of nonzero parameters used by the network in order to reduce its memorization capacity, while preserving as much as possible its accuracy.

Related works. The links between neural network sparsity and privacy have already been partially explored, but, to the best of our knowledge, it has not yet been shown that sparsity improves privacy *without further adjustment* of the training algorithm. A comparison with literature is done in section 1.3.

Contributions and results. The results of the experiments in section 1.3 support the hypothesis that sparsity improves the defense against MIAs while maintaining comparable performance on the learning task. However, the standard deviations reported in the experiments suggest that larger scale experiments are needed before confirming this trend. Figure 1.1 shows that the trade-off between robustness to MIA and network accuracy is similar between unstructured sparsity, obtained by an Iterative Magnitude Pruning (IMP) [Frankle & Carbin, 2019] of the weights, and structured "butterfly" sparsity, where the weights matrices are constrained to admit some structured sparse factorization [Lin et al., 2021, Dao et al., 2022]. To the best of our knowledge, the "butterfly" structure has not been studied before in this context. This structure achieves similar trade-offs as IMP, which is remarkable, as the structure is fixed beforehand, independently of the data. Moreover, software and hardware optimizations can be envisioned to leverage butterfly

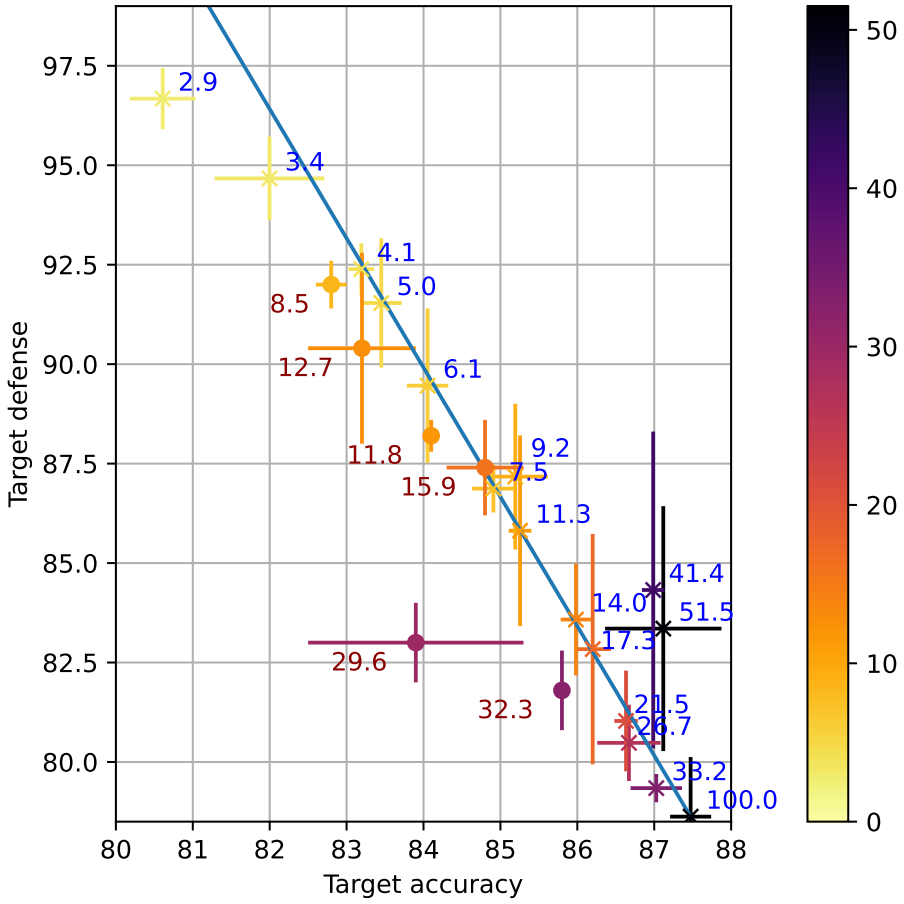


Figure 1.1: Means and standard deviations of the accuracy and defense level of various sparse networks. The percentage of nonzero weights is given in blue for IMP ($*$ p%), and in red for Butterfly ($\bullet$ p%). The color (as represented on the heat scale) emphasizes the sparsity level (in % of non-zero weights). The line has a slope of -3.25.

sparsity in order to implement matrix-vector multiplications in a more efficient way than it is without sparsity or with unstructured sparsity.

Experiments on CIFAR-10 show that when the percentage of nonzero weights in ResNet-20 is between 3.4% and 17.3%, a relative loss of p% in accuracy, compared to the trained dense network [6], leads to a relative gain of $3.6 \times p$% in defense against MIA, see Figure 1.1.

Section 1.1 introduces the MIAs used for the experiments. Section 1.2 describes the types of sparsity used to defend against MIAs. The results of the experiments are presented in section 1.3, with a comparison to literature.

[6] The dense network is the original network, with 100% of the nonzero weights.

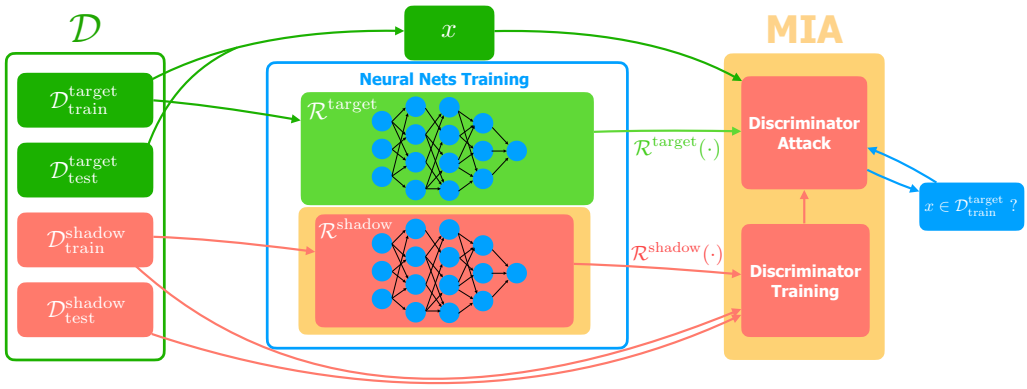


Figure 1.2: Experiments obey to the following pipeline: two networks are trained in the same fashion on $\mathcal{D}^{\text{target}}_{\text{train}}$ and $\mathcal{D}^{\text{shadow}}_{\text{train}}$ respectively. $\mathcal{R}^{\text{shadow}}$, $\mathcal{D}^{\text{shadow}}_{\text{train}}$ and $\mathcal{D}^{\text{shadow}}_{\text{test}}$ are then used to train a discriminator that will attack $\mathcal{R}^{\text{target}}$ by trying to infer the membership of x in $\mathcal{D}^{\text{target}}_{\text{train}}$.

1.1 MIA with a shadow model

Let $\mathcal{D}$ be a dataset and $\mathcal{D}_{\text{train}} \subset \mathcal{D}$ be a training subset. The associated *membership function* $m_{\mathcal{D}_{\text{train}},\mathcal{D}}$ is defined by:

$$m_{\mathcal{D}_{\text{train}},\mathcal{D}} : x \in \mathcal{D} \mapsto \begin{cases} 1 & \text{if } x \in \mathcal{D}_{\text{train}}, \\ 0 & \text{otherwise.} \end{cases}$$

Given a dataset $\mathcal{D}^{\text{target}}$, and a target network $\mathcal{R}^{\text{target}}$ trained on a subset $\mathcal{D}^{\text{target}}_{\text{train}}$ of $\mathcal{D}^{\text{target}}$, a MIA consists in retrieving the associated membership function $m_{\text{target}} := m_{\mathcal{D}^{\text{target}}_{\text{train}},\mathcal{D}^{\text{target}}}$, with only a *black-box* access to the function $x \mapsto \mathcal{R}^{\text{target}}(x)$. Most of the known attacks are based on an observation of the output of the $\mathcal{R}^{\text{target}}$ model, locally around x [Hu et al., 2022]. In general, these attacks seek to measure the confidence of the model in its predictions made locally around x. If the measured confidence is high enough, then the attacker answers positively to the membership question.

In practice, the most efficient attacks consist in training a discriminator model that makes a decision based on local information of $\mathcal{R}^{\text{target}}$ around x. This discriminator is trained from a *shadow* network [Hu et al., 2022], as explained below (see also Figure 1.2).

Suppose that the attacker has access to a dataset $\mathcal{D}^{\text{shadow}}$ from the same distribution as $\mathcal{D}^{\text{target}}$. It then trains its own shadow network $\mathcal{R}^{\text{shadow}}$ on a subset $\mathcal{D}^{\text{shadow}}_{\text{train}}$ of the data it owns. Ideally, $\mathcal{R}^{\text{shadow}}$ is trained under the same conditions as $\mathcal{R}^{\text{target}}$ (same architecture and same optimization algorithm). The attacker then has a tuple $(\mathcal{R}^{\text{shadow}}, \mathcal{D}^{\text{shadow}}, \mathcal{D}^{\text{shadow}}_{\text{train}})$ which is similar to $(\mathcal{R}^{\text{target}}, \mathcal{D}^{\text{target}}, \mathcal{D}^{\text{target}}_{\text{train}})$, and he knows the shadow membership function $m_{\text{shadow}} := m_{\mathcal{D}^{\text{shadow}}_{\text{train}},\mathcal{D}^{\text{shadow}}}$.

Discriminator. The attacker can then train a discriminator to approximate m_{shadow}, given a black box access to $\mathcal{R}^{\text{shadow}}$. This discriminator can then be used to approximate m_{target} given a black box access to $\mathcal{R}^{\text{target}}$. The model for the discriminator can be any classical classifier (logistic regression, neural network, etc.) [Hu et al., 2022].

1.2 Defense and neural network pruning

Training sparse neural networks is first motivated by needs for frugality in resources (memory, inference time, training time, etc.).

Here, the following hypothesis is investigated: sparsity can limit the model's ability to store private information about the data it has been trained on. A perfectly confidential network has not learned anything from its data and has no practical interest. A trade-off between confidentiality and accuracy must be made according to the task at hand. In what follows, two types of sparsity are considered.

1.2.1 Unstructured sparsity via IMP

In the first case, no specific structure is imposed on the set of nonzero weights. The weights that are set to zero (pruned) are selected by an iterative magnitude pruning process (IMP) [Frankle & Carbin, 2019]:

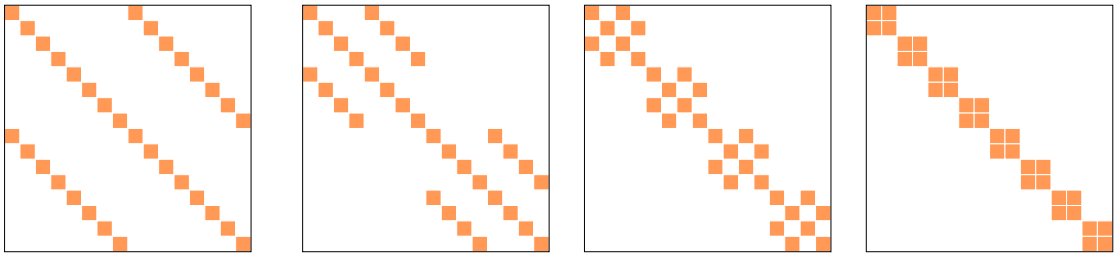
- train a network the usual way,
- prune $p\%$ of the weights having the smallest magnitude,
- adjust the remaining weights by re-training the network (weights that have been pruned are masked and are no longer updated), then go back to the second point until the desired level of sparsity is reached.

This procedure allows to find sparse networks with empirical good statistical properties [Frankle & Carbin, 2019, Frankle et al., 2021, Malach et al., 2020, Orseau et al., 2020, Paul et al., 2022].

1.2.2 Structured butterfly sparsity

In the second case, the sparsity is structured: the weight matrices of the neural network are constrained to admit a "butterfly" factorization [Zheng et al., 2022, Le et al., 2022, Dao et al., 2021, Dao et al., 2020], for which the associated matrix-vector multiplication can be efficiently implemented [Dao et al., 2022]. A square matrix $\mathbf{W}$ of size $N := 2^L$ has a butterfly factorization if it can be written as an exact product $\mathbf{W} = \mathbf{X}^{(1)} \dots \mathbf{X}^{(L)}$ of L

square factors of size N, where each factor satisfies the support constraint[7] $\mathrm{supp}(\mathbf{X}^{(\ell)}) \subseteq \mathrm{supp}(\mathbf{S}_{\mathrm{bf}}^{(\ell)})$, with $\mathbf{S}_{\mathrm{bf}}^{(\ell)} := \mathbf{I}_{2^{\ell-1}} \otimes \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right] \otimes \mathbf{I}_{N/2^{\ell}}$. See Figure 1.3 for an illustration. The factors have at most two nonzero entries per row and per column. Leveraging this factorization, matrix-vector multiplication has a complexity of $\mathcal{O}(N \log N)$, against $\mathcal{O}(N^2)$ in general.



(a) $\mathbf{S}_{\mathrm{bf}}^{(1)}$ (b) $\mathbf{S}_{\mathrm{bf}}^{(2)}$ (c) $\mathbf{S}_{\mathrm{bf}}^{(3)}$ (d) $\mathbf{S}_{\mathrm{bf}}^{(4)}$

Figure 1.3: Supports in a butterfly factorization of size $N = 16$.

To enforce the butterfly structure in a neural network, the weight matrices $\mathbf{W}$ are parameterized as $\mathbf{W} = \mathbf{X}^{(1)} \dots \mathbf{X}^{(L)}$, and only the nonzero coefficients of $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(L)}$ are initialized and then optimized by stochastic gradient descent.

In general, for a matrix $\mathbf{W}$ of arbitrary size, it is also possible to impose a similar structure but the definitions are more involved. We refer the reader to [Lin et al., 2021]. In the case of a convolution layer, the matrix $\mathbf{W}$ for which we impose such a structure corresponds to the concatenation of convolution kernels [Lin et al., 2021]. In our experiments, for a fixed size of $\mathbf{W}$ and a fixed number of factors L, the rectangular butterfly factorization is parameterized according to a so-called *monotone* chain following [Lin et al., 2021]. Among all possible chains, the one with the minimal number of parameters is selected.

Butterfly networks can reach empirical performance comparable to a dense network on image classification tasks [Dao et al., 2022, Lin et al., 2021].

1.3 Experimental results

All hyperparameters (including the discriminator architecture) have been determined following a grid search, averaged on three experiments to take into account randomness.

Dataset. Experiments are performed on the CIFAR-10 dataset (60000 images $32 \times 32 \times 3$, 10 classes). The dataset is randomly (uniformly) partitioned into 4 subsets $\mathcal{D}_{\mathrm{train}}^{\mathrm{target}}, \mathcal{D}_{\mathrm{test}}^{\mathrm{target}}, \mathcal{D}_{\mathrm{train}}^{\mathrm{shadow}}, \mathcal{D}_{\mathrm{test}}^{\mathrm{shadow}}$ of 15000 images, respectively used to train and test the target and shadow networks. The membership functions are defined as in section 1.1, with

[7] $\mathrm{supp}(\cdot)$ is the set of nonzero entries of a matrix, $\mathbf{I}_N$ is the identity matrix of size $N \times N$, and $\otimes$ is the Kronecker product.

$\mathcal{D}^{\text{target}} := \mathcal{D}^{\text{target}}_{\text{train}} \cup \mathcal{D}^{\text{target}}_{\text{test}}$ and $\mathcal{D}^{\text{shadow}} := \mathcal{D}^{\text{shadow}}_{\text{train}} \cup \mathcal{D}^{\text{shadow}}_{\text{test}}$. For the target and shadow network, among their 15000 training data points, 1000 are randomly chosen and fixed for all our experiments as a validation set (used to tune the hyper-parameters, and for the stopping criterion).

Training of the target and shadow models. The target and shadow networks have a ResNet-20 architecture [He et al., 2016] (272474 parameters). They are trained to minimize the cross-entropy loss by stochastic gradient descent (with 0.9 momentum and no Nesterov acceleration) on their respective training sets for 300 epochs, with a batch size of 256. The dataset is augmented with random horizontal flipping and random cropping. The initial learning rate is divided by 10 after 150 and after 225 epochs. The weights of the neural networks are initialized with the standard method on Pytorch, following a uniform distribution on $(-1/\sqrt{n}, 1/\sqrt{n})$ where n is the input dimension for a linear layer, and n is input dimension $\times$ kernel width $\times$ kernel height for a convolution.

Values of initial learning rate and weight decay are reported in table 1.1. Note that the chosen hyperparameters allow to reproduce results of [He et al., 2016] when using the whole 50000 training images of CIFAR-10 instead of 15000 of them as it is done for the target and shadow networks.

For IMP, 24 prunings and readjustments of the parameters are performed. Each readjustment is done with the same training procedure as above (300 epochs, etc.). Before each pruning, the weights are rewound to the values they had at the end of the epoch of maximum validation accuracy in the last 300 epochs.

For training ResNet-20 with the butterfly structure, the original weight matrices of some convolution layers are substituted by matrices admitting a butterfly factorization, with a number $L = 2$ or 3 of factors, following a monotonic chain minimizing the number of parameters in the factorization, as described in section 1.2.2. The substituted layers are those of the $S = 1, 2$ or 3 last segments[8] of ResNet-20.

Discriminator training. A discriminator takes as inputs the class i of x, the prediction $\mathcal{R}(x)$ made by a network $\mathcal{R}$ (target or shadow), as well as $\frac{1}{\epsilon}\mathbb{E}\left(|\mathcal{R}(x) - \mathcal{R}(x + \epsilon\mathcal{N})|\right)$ ($\epsilon = 0.001$ and $\mathcal{N}$, an independent centered and reduced Gaussian vector) that encodes local first order information of $\mathcal{R}$ around x. The expectation is estimated by averaging over 5 samples. For each pair of networks ($\mathcal{R}^{\text{target}}$,$\mathcal{R}^{\text{shadow}}$), three discriminators (perceptrons) are trained, with respectively 1, 2, 3 hidden layer(s) and 30, 30, 100 neurons on each hidden layer. The binary cross entropy is minimized with Adam for 80 epochs, without weight decay and for three different learning rates $\{0.01, 0.001, 0.0001\}$.

[8] A segment is three consecutive basic blocks with the same number of filters. A basic block is two convolutional layers surrounded by a residual connection.

Table 1.1: Hyperparameters for the training of the target and the shadow neural networks.

Network	% of nonzero params	Initial learning rate	Weight decay
ResNet-20 dense	100 %	0.03	0.005
Butterfly $(S = 1, L = 2)$	32.3 %	0.3	0.0005
Butterfly $(S = 1, L = 3)$	29.6 %	0.3	0.0001
Butterfly $(S = 2, L = 2)$	15.9 %	0.3	0.0005
Butterfly $(S = 2, L = 3)$	12.9 %	0.1	0.001
Butterfly $(S = 3, L = 2)$	11.8 %	0.3	0.0005
Butterfly $(S = 3, L = 3)$	8.5 %	0.1	0.001
IMP with k prunings	$\simeq 100 \times (0.8)^k\%$	0.03	0.005

Accuracy and defense The *accuracy* of a network is the percentage of data whose class is the one predicted with the highest probability by the network. The *defense* D of a network against a discriminator is defined as $D = 200 - 2A$ where A is the accuracy of the discriminator on the membership classification task associated with the training and test data of the considered network. For example, if a discriminator has an attack accuracy $A = 50 + x$, then the defense is $D = 100 - 2x$. In our case, there are as much training and testing data points for the network (target or shadow). Ideally, the discriminator should not do better than guessing randomly, having then an accuracy of 50%.

Results Dense target and shadow networks achieve on average 87.5% accuracy on the test set. This accuracy decreases with sparsity, see Figure 1.1. A gain (or loss) in defense is significant if the interval with upper (resp. lower) bound being the mean plus (resp. minus) the standard deviation is disjoint from the interval corresponding to the trained dense network. A significant gain (or loss) in defense is only observed for a proportion of nonzero weight between 0% and 17.3%, and for 41.4% and 51.5%. Between 3.4% and 17.3%, a relative loss of p% in accuracy, compared to the trained dense network, leads to a relative gain of $3.6 \times p$%: $3.6 \simeq \frac{|\text{defense} - \text{defense dense}|}{\text{defense dense}} \frac{\text{accuracy dense}}{|\text{accuracy} - \text{accuracy dense}|}$.

Related work on sparsity as a defense mechanism. Experimental results from [Yuan & Zhang, 2022] suggest on the contrary that training a network with sparse regularization from IMP *degrades* privacy. But these results were not averaged over multiple experiments to reduce variability due to randomness. The experiments of [Yuan & Zhang, 2022] are also performed on CIFAR-10 but with a model with 40 times as many weights as ResNet-20, and for a proportion of nonzero weights above 50%. Given the standard deviations observed in Figure 1.1 for sparsity levels above 20% on ResNet-20, one should remain cautious about the interpretation of the results of [Yuan & Zhang, 2022].

[Tan et al., 2023] also showed recently that decreasing the number of parameters of a model can improve defense to MIAs. This is complementary to this chapter. Note however that the way the number of parameters are reduced are fundamentally different since [Tan et al., 2023] consider smaller *dense networks* while, here, *sparse subnetworks* are consider. These types of networks may not have the same privacy-accuracy trade-off.

Given a sparsity level, [Wang et al., 2021] looks for the parameters that minimize the loss function of the learning problem, penalized by the highest MIA attack accuracy achievable against these parameters. Note that this penalty term is in general not explicitly computable, and difficult to minimize. Moreover, this requires to know in advance the type of attack that targets the network, e.g., the architecture of the attacker, etc. No comparison with the non-penalized case has been proposed in [Wang et al., 2021], which makes it unclear whether this penalization is necessary to improve privacy or if sparsity *without additional penalization* is sufficient. In contrast, our experiments do suggest the latter. Moreover, [Wang et al., 2021] only displays the defense achieved at the sparsity level with the smallest penalized loss function. In comparison, Figure 1.1 shows the robustness to MIAs for a whole range of different sparsity levels.

Finally, it has been observed that enforcing sparsity during the training of neural networks with DP-SGD ("Differentially Private Stochastic Gradient Descent") [Abadi et al., 2016, Adamczewski & Park, 2023] improves the accuracy, compared to the dense network, while keeping the same guarantees of Differential Privacy (giving strong privacy guarantees) [Huang et al., 2020, Adamczewski & Park, 2023]. However, compared to SGD, DP-SGD suffers from a performance drop and a high computational demand that is prohibitive for large-scale experiments [Sander et al., 2022, Lalanne et al., 2023b]. In contrast, the privacy enhancement investigated in this work comes at a lower cost (in both accuracy and resources) but does not provide any theoretical differential privacy guarantee.

1.4 Take home message

The results obtained support the following conjecture: sparsity is a defense mechanism against membership inference attacks, as it reduces the effectiveness of attacks with a relatively low cost on network accuracy. This is in particular the case for structured butterfly sparsity, which had not yet been investigated in this context to the best of our knowledge.

Extending the experiments to a richer class of models, datasets and attacks would support the interest of sparsity as a defense mechanism. In the future, sparsity could serve as a baseline to decrease privacy threats since it comes at a lower computational cost than methods providing strong theoretical guarantees such as DP-SGD, does not require to know the kind of attack in advance, allows for fast matrix-vector multiplication when using structured sparsity such as the butterfly structure, and, compared to penalized loss where the attacker could infer the typical behavior of the model on training data [Song et al.,], it may not lead to bias easily exploitable by an attacker.

Chapter 2

A survival guide to differential privacy

The origin of this chapter, and the use of the first person. This chapter presents important results that the reader must know about differential privacy. In particular, it does not present direct contributions, other than the reformulation effort. Additionally, this chapter is in part inspired by Rachel Cumming's lecture on differential privacy that was given at a summer school at the CIRM in May 2022 in Marseille, France. Hence, in this chapter, I will try to respect the following rule : the use of the first person of the plural (we, our, ...) will be used as a generic inclusion formula to include the reader. I will refrain from using the first person of the singular (I, my, ...), except for editorial clarifications about this thesis.

This chapter presents key concepts and technical results about differential privacy that are necessary for the rest of the thesis. It puts an emphasis on embedding those results in a statistical framework in order to link them with the main theme of the thesis.

2.1 Formal definitions of privacy

All the formal definitions of differential privacy are based on neighboring relations.

Definition 2.1.1 (Neighboring relation)**.** Let $\mathcal{D}$ be the set of possible datasets (for the

problem of interest). A neighboring relation is a *symmetric* relation on $\mathcal{D}$, which means that it is a subset $\mathcal{R}$ of $\mathcal{D}^2$ such that $(x, y) \in \mathcal{R}$ iff $(y, x) \in \mathcal{R}$. We will note $x \sim y$ iff $(x, y) \in \mathcal{R}$.

Intuitively, when two datasets are neighbors, one wants the output of any private learner to have similar outputs on them.

When the objective is to hide the *individual values* of a dataset, a few natural neighboring relations arise.

Example 2.1.2 (Addition / deletion)**.** When $\mathcal{D} = \cup_{k \geq 0} \mathcal{X}^k$, representing the collection of all finite datasets that are collections of elements of $\mathcal{X}$, the addition/replacement neighboring relation is defined as $x \sim y$ iff $x = y$ (up to a permutation) or x can be obtained from y by addition or deletion of a single element (up to a permutation).

Example 2.1.3 (Substitution - permutation dependent)**.** When $\mathcal{D} = \mathcal{X}^n$, representing the collection of all datasets that are collections of n elements of $\mathcal{X}$, the permutation dependent neighboring relation is defined as $x \sim y$ iff $d_{\text{ham}}(x, y) \leq 1$ where $d_{\text{ham}}(\cdot, \cdot)$ refers to the Hamming distance.

Example 2.1.4 (Substitution - permutation invariant)**.** When $\mathcal{D} = \mathcal{X}^n$, representing the collection of all datasets that are collections of n elements of $\mathcal{X}$, the permutation independent neighboring relation is defined as $x \sim y$ iff there exists σ a permutation of the indices such that $d_{\text{ham}}(\sigma(x), y) \leq 1$.

In particular, for the substitution neighboring relations, n (the sample size) is a constant of the problem (even if it is possible to consider a series of problems of different sizes). In contrast, for the addition/deletion neighboring relation, the sample size is not fixed and datasets can be of any size.

It can be interesting to design your own dataset spaces and neighboring relations depending on the problem at hand. However, some of the properties that follow depend on the fact that $\mathcal{D}$ is connex for the neighboring relation $\sim$, that is that for any pair of datasets in $\mathcal{D}$, there exist a path of neighboring (for $\sim$) datasets in $\mathcal{D}$ linking them. Except when specified, all the results of this chapter are in this general setup. In the rest of the thesis, without further specifications, the setup is the one of the permutation dependent substitution neighboring relation.

2.1.1 The historic definition

The first definition of differential privacy [Dwork et al., 2006b, Dwork et al., 2006a] bounds the output distributions of any pair of neighboring datasets on any measurable element of the output space.

Definition 2.1.5 (Differential privacy)**.** Let $\mathcal{O}$ be a set (that will be our output space) endowed with a σ-algebra $\sigma(\mathcal{O})$. Let $\epsilon \geq 0$ (called the privacy budget) and $\delta \geq 0$ (the relaxation parameter). A randomized mechanism $\mathfrak{M} : \mathcal{D} \to \mathcal{O}$ is (ϵ, δ)-differentially private (or simply (ϵ, δ)-DP) if for any $S \in \sigma(\mathcal{O})$, for any $\mathbf{X}, \mathbf{X}' \in \mathcal{D}$,

$$\mathbf{X} \sim \mathbf{X}' \implies \mathbb{P}\left(\mathfrak{M}(\mathbf{X}) \in S\right) \leq e^{\epsilon}\mathbb{P}\left(\mathfrak{M}(\mathbf{X}') \in S\right) + \delta \, .$$

Furthermore, a mechanism that is $(\epsilon, 0)$-DP is said to satisfy ϵ-pure differential privacy, or simply ϵ-DP. This definition needs a small clarification on what a randomized mechanism $\mathfrak{M} : \mathcal{D} \to \mathcal{O}$ means. It means that to each $\mathbf{X} \in \mathcal{D}$ is associated a distribution $\mathbb{P}_{\mathfrak{M}(\mathbf{X})}$ on $(\mathcal{O}, \sigma(\mathcal{O}))$. In the definition, $\mathbb{P}\left(\mathfrak{M}(\mathbf{X}) \in S\right)$ is a proxy for $\mathbb{P}_{\mathfrak{M}(\mathbf{X})}(S)$.

The role of ϵ. If a mechanism is (ϵ, δ)-DP, it is also (ϵ', δ)-DP if $\epsilon' > \epsilon$. As a result, the smaller ϵ, the stronger the constraint on privacy. The two following limit behaviors arise:

- $\epsilon = 0$: Perfect privacy, where the result cannot depend at all on the data. As a result, no learning is possible.

- $\epsilon = +\infty$: No privacy since the constraint vanishes. Privacy is no longer implied by the definition.

We want to be somewhere in the middle, and the "correct "choice of ϵ depends on the level of privacy that we want to guarantee.

The role of δ. Similarly, we can observe that the smaller δ, the stronger the privacy guarantees. δ differs from ϵ because:

- It gives a small additive slack in the privacy guarantee (relaxation).

- It allows for a family of output distributions that are not all absolutely continuous with respect to each other.

- Even with uniform support, it allows for an easier mechanism design.

In order to tune δ, we can fall back on the following observations and interpretations of this parameter:

- δ may be viewed as the probability under which the output mechanism does not respect the ϵ-DP guarantee [Dwork & Roth, 2014]. In fact, this rule of thumb is not always true (despite being a good guideline), but it is often true (indeed it is a common proof technique to conclude to the (ϵ, δ)-DP [Dwork & Roth, 2014])

- If $\delta = 1$ then we're back to no privacy, even for $\epsilon = 0$.

- Usually, δ is considered to be acceptable if $\delta \ll \frac{1}{n}$ [Dwork & Roth, 2014].

Remark 2.1.6. One might think that the definition of differential privacy is arbitrary, and it is. However, it is becoming increasingly adopted because this is the best that has been proposed to this date. Indeed, it ensures strong privacy guarantees (see [Kairouz et al., 2015, Dong et al., 2019]) while allowing for a nice algebra of private mechanisms (as we will see later). As a consequence, it is both conceptually powerful and handy, in a way that wasn't matched by previous definitions (such as k-anonymity [Sweeney, 2002]).

Example 2.1.7 (Randomized data leak)**.** For the replacement neighboring relation, the mechanism that select an element in the dataset uniformly at random and shares it to the world is $(0, 1/n)$-DP (n being the sample size). Since this mechanism couldn't realistically be considered as private, this observation strengthens the guideline that one must have $\delta \ll \frac{1}{n}$.

2.1.2 Definitions based on Rényi divergences

Working under pure differential privacy is often preferable compared to working under (ϵ, δ)-DP. However, it can easily be shown that a pure differentially private mechanisms cannot be-non trivial (i.e. not having the same output distribution on any pair of neighboring datasets), and have Gaussian output distributions. This observation is problematic since the Gaussian structure is extremely handy and is broadly used in data science (in has strong tail bounds and allows the exact computation of otherwise intractable terms). It is possible to encapsulate those mechanisms with the larger definition of (ϵ, δ)-DP (approximate differential privacy, but choosing the relaxation parameter δ is always a haste, and often lead to suboptimal errors.

In comparison, modern definitions of privacy that are specifically tailored to handle mechanisms with a Gaussian structure are based on the Rényi divergence. For the rest of this thesis, for any $\alpha > 1$, $\mathrm{D}_\alpha(\cdot\|\cdot)$ denotes the Renyi divergence of level α, which defined for two distributions of probability $\mathbb{P}$ and $\mathbb{Q}$ as

$$\mathrm{D}_\alpha(\mathbb{P}\|\mathbb{Q}) := \frac{1}{\alpha-1}\ln\int\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^{\alpha-1} d\mathbb{Q}\,.$$

For more details, we recommend referring to the excellent article [van Erven & Harremoës, 2014]. With this new divergence between probability distributions, it is possible to define the Rényi differential privacy, and the more restrictive zero-concentrated differential privacy as :

Definition 2.1.8 (Rényi differential privacy)**.** Let $\mathcal{O}$ be a set (that will be our output space) endowed with a σ-algebra $\sigma(\mathcal{O})$. Let $\epsilon \geq 0$ (called the privacy budget), and $\alpha > 1$ (called the level). A randomized mechanism $\mathfrak{M} : \mathcal{D} \to \mathcal{O}$ is (α, ϵ)-Rényi differentially private (or simply (α, ϵ)-RDP) if for any $S \in \sigma(\mathcal{O})$, for any $\mathbf{X}, \mathbf{X}' \in \mathcal{D}$,

$$\mathbf{X} \sim \mathbf{X}' \implies \mathrm{D}_\alpha\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X})}\middle\|\,\mathbb{P}_{\mathfrak{M}(\mathbf{X}')}\right) \leq \epsilon\,. \tag{2.1}$$

For more details on Rényi differential privacy, please refer to [Mironov, 2017].

Definition 2.1.9 (Concentrated differential privacy)**.** Let $\mathcal{O}$ be a set (that will be our output space) endowed with a σ-algebra $\sigma(\mathcal{O})$. Let $\rho \geq 0$ (called the privacy budget). A randomized mechanism $\mathfrak{M} : \mathcal{D} \to \mathcal{O}$ is ρ-zero-concentrated differentially private (or simply ρ-zCDP) if for any $S \in \sigma(\mathcal{O})$, for any $\mathbf{X}, \mathbf{X}' \in \mathcal{D}$,

$$\mathbf{X} \sim \mathbf{X}' \implies \forall 1 < \alpha < +\infty, \mathrm{D}_\alpha\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X})}\middle\|\,\mathbb{P}_{\mathfrak{M}(\mathbf{X}')}\right) \leq \rho\alpha\,. \tag{2.2}$$

For more details on concentrated differential privacy, please refer to the original paper [Dwork & Rothblum, 2016], or to the updated version [Bun & Steinke, 2016].

2.1.3 Other interesting attempts

As we have seen in the introduction, from differential privacy, it is possible to deduce strong lower bounds on the testing difficulty between two neighboring datasets. In fact, it is an equivalence [Kairouz et al., 2015]. A mechanism is (ϵ, δ)-DP iff for any pair of neighboring datasets $\mathbf{X}$ and $\mathbf{X}'$, if an adversary tries to discriminate $\mathbf{X}$ from $\mathbf{X}'$ with a type 1 error α, he should have a type 2 error $\beta \geq f_{\epsilon,\delta}(\alpha)$. $f_{\epsilon,\delta}$ is called a *tradeoff* function.

The expression of $f_{\epsilon,\delta}$ is fully characterized by ϵ and δ, but it is possible to take any (as long as it satisfies a few hypotheses) tradeoff function to define a new type of privacy.

This is what is done in the excellent article [Dong et al., 2019], which also provides a lot of conceptual insights on the effect of differential privacy.

2.2 The algebra of private mechanisms

Differential privacy offers strong privacy guarantees, but it is not the only reason why this definition of privacy became so popular. It is also extremely handy to use, and it is possible to talk about the *algebra* of private mechanisms. Indeed, it is possible to control the privacy of any procedure building on private mechanisms by the so-called post-processing, composition, and group privacy properties. Furthermore, providing privacy guarantees for any of the above-mentioned definitions of privacy usually allows providing guarantees for all the other ones. This section presents key building blocks that are used almost all the time.

2.2.1 Post processing

The first important property is the post-processing property. Informally, it states that any quantity that is build from a private observation of the dataset, and without further information about the dataset, is still private.

Fact 2.2.1 (Post processing)**.** *If $\mathfrak{M}$ is a mechanism satisfies one of the above-mentioned definitions of privacy, and if f is a deterministic function, then $f(\mathfrak{M})$ satisfies the same definition of privacy as $\mathfrak{M}$, and with the same privacy parameters.*

It is possible to deal with stochastic functions by integrating the last result w.r.t. the extra source of randomness in f. In order to obtain the same conclusion when f is randomized, we typically need the extra source of randomness to be independent of the ones on which $\mathfrak{M}$ builds.

2.2.2 Various conversions, and corresponding fees

It is possible to convert guarantees between (ϵ, δ)-DP, (α, ϵ)-RDP, and ρ-zCDP. Between (ϵ, δ)-DP and (α, ϵ)-RDP, [Bun & Steinke, 2016] states that any $(\epsilon, 0)$-DP mechanism is also $(\alpha, \alpha\frac{\epsilon^2}{2})$-RDP for any α. Furthermore, [Mironov, 2017] states that any (α, ϵ)-RDP mechanism is also $\left(\epsilon + \frac{\ln \frac{1}{\delta}}{\alpha - 1}, \delta\right)$-DP for any $\delta > 0$. Between RDP and zCDP, the conversion is only possible from zCDP to RDP, and is given by the definition. Between zCDP and DP, [Bun & Steinke, 2016] states that any $(\epsilon, 0)$-DP mechanism is also $\frac{\epsilon^2}{2}$-zCDP. Furthermore, it also states that any ρ-zCDP mechanism is also $\left(\rho + 2\sqrt{\rho \ln \frac{1}{\delta}}, \delta\right)$-DP for any $\delta > 0$.

2.2.3 Comparing any pair of datasets

The definitions of differential privacy characterize the testing difficulty between pairs of neighboring datasets. However, under the connexity assumption, it is possible to charac-

terize the testing difficulty between any pair of datasets depending on their distance k on the neighboring relation $\sim$ (i.e. the minimal length of a path on $\sim$ linking them). Such property is usually called the *group privacy* property.

Indeed, by inductively applying the definition of differential privacy, it directly follows that

Fact 2.2.2 (Group privacy)**.** *If a randomized mechanism $\mathfrak{M}$ is (ϵ, δ)-differentially private, then, for any pair of datasets $\mathbf{X}, \mathbf{X}' \in \mathcal{D}$ and any measurable $S \subseteq \mathrm{codom}(\mathfrak{M})$, if $\mathbf{X}$ and $\mathbf{X}'$ are at distance at most k on $\sim$, then*

$$\mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\mathbf{X}) \in S) \leq e^{\epsilon k}\mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\mathbf{Y}) \in S) + \delta k e^{\epsilon(k-1)} .$$

For concentrated differential privacy, [Bun & Steinke, 2016] states that

Fact 2.2.3 (Group privacy (zCDP case))**.** *If a randomized mechanism $\mathfrak{M}$ is ρ-zCDP, then, for any pair of datasets $\mathbf{X}, \mathbf{X}' \in \mathcal{D}$, if $\mathbf{X}$ and $\mathbf{X}'$ are at distance at most k on $\sim$, then*

$$\forall 1 < \alpha < +\infty, D_\alpha\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X})} \middle\| \mathbb{P}_{\mathfrak{M}(\mathbf{X}')}\right) \leq \rho k^2 \alpha .$$

2.2.4 The case in point of composition

The most important property of differential privacy is probably the so-called *composition* property. Informally, it states that if each access to the dataset during a complex data pipeline is done with a certain privacy budget, then the whole procedure is differentially private with privacy budget the sum of the individual privacy budgets.

Let $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$ be randomized mechanisms from a common dataset space $\mathcal{D}$ (and eventually taking auxiliary inputs) to their respective output spaces. An adaptive composition of $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$ is a pipeline (or computational acyclic graph) in which each of the mechanisms appears at most once. In particular, those mechanisms are allowed to take as auxiliary input the result of the previous ones. In this scenario, we say that $\mathfrak{M}_i$ satisfies a certain property of differential privacy if its restriction to the dataset variable satisfies it for any value of its auxiliary inputs. Composition theorems allow characterizing the privacy of the whole pipeline, depending on the privacy budgets of the building blocks $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$.

Under (ϵ, δ)-DP, the following result [Kairouz et al., 2015] sharply characterizes the composition of private mechanisms. Its expression can be a bit terrifying at first, but it is easily derived in simpler (but suboptimal) applicable forms.

Fact 2.2.4 (Advanced composition)**.** *If $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$ are respectively $(\epsilon_1, \delta_1), \ldots, (\epsilon_k, \delta_k)$-DP, then for any $\tilde{\delta} \in [0, 1]$, any adaptive composition of those mechanisms is*

$$\left(\tilde{\epsilon}, 1 - (1 - \tilde{\delta})\Pi_{i=1}^{k}(1 - \delta_i)\right) \text{ -DP} ,$$

for

$$\tilde{\epsilon} := \min\left(A, B, B\right) ,$$

where

$$A := \sum_{i=1}^{k} \epsilon_i ,$$

$$B := \sum_{i=1}^{k} \frac{(e^{\epsilon_i} - 1)\epsilon_i}{e^{\epsilon_i} + 1} + \sqrt{\sum_{i=1}^{k} 2\epsilon_i^2 \ln\left(e + \frac{\sqrt{\sum_{i=1}^{k} \epsilon_i^2}}{\tilde{\delta}}\right)} ,$$

and

$$C := \sum_{i=1}^{k} \frac{(e^{\epsilon_i} - 1)\epsilon_i}{e^{\epsilon_i} + 1} + \sqrt{\sum_{i=1}^{k} 2\epsilon_i^2 \ln\left(\frac{1}{\tilde{\delta}}\right)} .$$

Two more applicable corollaries of this theorem (and that were known earlier) are the so called *simple composition* theorem [Dwork et al., 2006b, Dwork et al., 2006a] which states that if $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$ are respectively $(\epsilon_1, \delta_1), \ldots, (\epsilon_k, \delta_k)$-DP, then any adaptive composition of those mechanisms is $\left(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i\right)$-DP, and the first form of the so-called *advanced composition* theorem [Dwork et al., 2010], stating that if $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$ are all (ϵ, δ)-DP, then for any $\tilde{\delta} \in [0, 1]$, any adaptive composition of those mechanisms is $(\tilde{\epsilon}, k\delta + \tilde{\delta})$-DP, where

$$\tilde{\epsilon} := k\epsilon\left(e^{\epsilon} - 1\right) + \epsilon\sqrt{2k \ln\left(\frac{1}{\tilde{\delta}}\right)} .$$

In particular, the small extra slack given in δ allows for a $\tilde{\epsilon}$ that scales in $O(k\epsilon^2 + \sqrt{k}\epsilon)$, which is often a lot better than the scaling in $k\epsilon$ given by simple composition.

Remark 2.2.5 (From addition/deletion to replacement)**.** By noticing that a replacement in a dataset can be decomposed as an addition and then a deletion, thanks to the simple composition property, if an algorithm is (ϵ, δ)-DP for the addition/deletion neighboring relation, it is $(2\epsilon, 2\delta)$-DP for the permutation invariant replacement neighboring relation.

With RDP and its variants, composition is a lot simpler [Mironov, 2017]. Indeed, we have the following composition theorem :

Fact 2.2.6 (Composition with Rényi divergence)**.** *If $\mathfrak{M}_1, \dots, \mathfrak{M}_k$ are respectively $(\alpha, \epsilon_1), \dots, (\alpha, \epsilon_k)$-RDP, then any adaptive composition of those mechanisms is $\left(\alpha, \sum_{i=1}^k \epsilon_i\right)$-RDP.*

Note that a direct consequence on this result is the composition of zCDP mechanisms, stating that if $\mathfrak{M}_1, \dots, \mathfrak{M}_k$ are respectively $\rho_1, \dots, \rho_k$-zCDP, then any adaptive composition of those mechanisms is $\sum_{i=1}^k \rho_i$-zCDP.

2.2.5 Privacy amplification and subsampling

A last property that is interesting (but not exploited directly in this thesis) is the privacy amplification by subsampling. Namely, the idea that the privacy of a randomized mechanism is amplified by previously subsampling its dataset. I recommend [Balle et al., 2018a] and [Wang et al., 2020] for recent results.

2.3 The private jungle

We saw that private mechanisms form a nice algebra, making them appealing for many data pipelines. However, we still have to present elementary building blocks that are versatile enough to adapt to many problems. We do so in this section by presenting the ubiquitous mechanisms that are the Laplace mechanism, the Gaussian mechanism, the exponential mechanism, and some results about private optimization.

2.3.1 Laplace mechanism

The Laplace mechanism [Dwork et al., 2006b, Dwork et al., 2006a] was the first example of private mechanism. It is based on the following simple idea. Let us say that we have a *deterministic* function f defined on a set of datasets $\mathcal{D}$ and taking values in $\mathbb{R}^k$ that we want to make private. The idea is to replace f by the randomized mechanism

$$\mathbf{X} \in \mathcal{D} \mapsto f(\mathbf{X}) + \alpha \mathcal{L}(I_k) \ ,$$

where the notation $\mathcal{L}(I_k)$ refers to a random vector with independent components following Laplace distributions of parameter 1.

The amount of noise α to add in order to make this mechanism (called the Laplace mechanism) depends on the *sensitivity* of the function f.

Definition 2.3.1 (l_k-sensitivity)**.** For $f : \mathcal{D} \to \mathbb{R}^k$ defined on a set of datasets $\mathcal{D}$ equipped with a neighboring relation $\sim$, the l_k-sensitivity of f ($\Delta_k f$) is defined as

$$\Delta_k f := \sup_{\mathbf{X} \sim \mathbf{X}'} \|f(\mathbf{X}) - f(\mathbf{X}')\|_k \ .$$

For brevity, Δf is often used to refer to $\Delta_1 f$.

Example 2.3.2 (Mean estimation, sensitivity)**.** Let us give a small example that highlights the importance of the choice of the neighboring relation $\sim$, and the dataset space $\mathcal{D}$. Let us consider the example where we observe n samples $X_1, \dots, X_n$ living in $[a, b]$, and where the objective is to privately estimate their mean $\frac{1}{n}\sum_{i=1}^{n} X_i$. For the addition/deletion relationship in the setup where the sample-size n is not fixed, the sensitivity of this query is $|b-a|$. On the other hand, in the replacement setup where n is fixed, the sensitivity of the same query is $\frac{|b-a|}{n}$. Changing the setup allows to greatly reduce the sensitivity.

The privacy of the Laplace mechanism is given by the following result :

Fact 2.3.3 (Privacy guarantees)**.** *If $\alpha \geq \frac{\Delta f}{\epsilon}$, then the Laplace mechanism is ϵ-DP.*

Furthermore, the tail bounds on the Laplace distribution allow deriving the following utility guarantees of the Laplace mechanism :

Fact 2.3.4 (Utility guarantees)**.** *Let us note y the output of the Laplace mechanism with noise magnitude α. For any $\gamma > 0$,*

$$\mathbb{P}\left(\|f(x) - y\|_1 > \alpha \ln\left(\frac{k}{\gamma}\right)\right) < \gamma \,.$$

Example 2.3.5 (Learning finite distributions)**.** Let $\mathcal{S} = \{s_1, \dots, s_k\}$ be a finite set. We call $\mathcal{P}_k$ the simplex of $\mathbb{R}^k$ of vectors with positive entries that sum to 1. To any distribution of probability $\mathbb{p}$ on $\mathcal{S}$ canonically corresponds a vector $\mathbf{p} = (p_1, \dots, p_k) \in \mathcal{P}_k$ such that $\mathbf{p}(\{s_i\}) = p_i$ for any i. Hence, we will simply use a vector in $\mathcal{P}_k$ to refer to distributions. Finally, for $\mathbf{p} \in \mathcal{P}_k$, $X \sim \mathbf{p}$ means that the random variable X follows the distribution associated to $\mathbf{p}$ on $\mathcal{S}$.

The problem is the following : let $\mathbf{p} \in \mathcal{P}_k$, and given $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbf{p}$, can we build a ϵ-DP estimator of $\mathbf{p}$? Guided by the simple moment estimator, we can look at the performances of the histogram estimator.

For any i, we define $h_i := \sum_{j=1}^{n} \mathbb{1}_{X_j = s_i}$. The non-private histogram estimator is defined as $\hat{\mathbf{p}} := \left(\dots, \frac{1}{n}h_i, \dots\right)$. Can we make it private ?

This estimator builds on the deterministic function of the data

$$f := (X_1, \dots, X_n) \mapsto \left(\dots, \sum_{j=1}^{n} \mathbb{1}_{X_j = s_i}, \dots \right) . \tag{2.3}$$

What is its l_1 sensitivity ? Adding or removing a single element to the dataset will change the value of at most one coordinate of the output vector by at most $+1$ or -1. Consequently, the l_1 sensitivity for the addition/removal relation is 1. By replacing an element of the dataset, the counts of at most two coordinates of the output vector can change by at most $+1$ or -1 each. Consequently, the l_1 sensitivity of f for the replacement neighboring relation is 2. In order to be conservative, let us keep 2 as an upper-bound on both sensitivities.

Applying Fact 2.3.3, if $\mathcal{L}_1, \dots, \mathcal{L}_n \overset{\text{i.i.d.}}{\sim} \mathcal{L}(1)$, then the mechanism

$$\mathfrak{M} := (X_1, \dots, X_n) \mapsto \left(\dots, \sum_{j=1}^{n} \mathbb{1}_{X_j = s_i} + \frac{2}{\epsilon} \mathcal{L}_i, \dots \right) \tag{2.4}$$

is ϵ-DP. Finally, by post-processing (Fact 2.2.1), the following estimator is ϵ-DP :

$$\hat{\mathbf{p}}_\epsilon := (X_1, \dots, X_n) \mapsto \frac{1}{n} \left(\dots, \sum_{j=1}^{n} \mathbb{1}_{X_j = s_i} + \frac{2}{\epsilon} \mathcal{L}_i, \dots \right) . \tag{2.5}$$

Let us analyze its performances as an estimator of the true distribution. First, we can notice that for any i, $h_i \sim \mathcal{B}(n, p_i)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{p} - \hat{\mathbf{p}}_\epsilon\|_2^2 \right) &= \frac{1}{n^2} \mathbb{E}\left(\sum_{i=1}^{k} \left(np_i - h_i - \frac{2}{\epsilon} \mathcal{L}_i \right)^2 \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{k} \mathbb{E}\left(\left(np_i - h_i - \frac{2}{\epsilon} \mathcal{L}_i \right)^2 \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{k} \left(\mathbb{E}\left(np_i - h_i - \frac{2}{\epsilon} \mathcal{L}_i \right)^2 + \mathbb{V}\left(h_i + \frac{2}{\epsilon} \mathcal{L}_i \right) \right) \\
&\overset{\text{indep.}}{=} \frac{1}{n^2} \sum_{i=1}^{k} \left(\mathbb{E}\left(np_i - h_i - \frac{2}{\epsilon} \mathcal{L}_i \right)^2 + \mathbb{V}\left(h_i \right) + \mathbb{V}\left(\frac{2}{\epsilon} \mathcal{L}_i \right) \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{k} \left(\mathbb{E}\left(0 \right)^2 + np_i(1 - p_i) + \frac{8}{\epsilon^2} \right) \\
&= \frac{\sum_{i=1}^{k} p_i(1 - p_i)}{n} + \frac{8k}{n^2 \epsilon^2} .
\end{aligned}$$

Finding, or upper-bounding the quantity $\sup_{\forall i, p_i \geq 0, \sum_i p_i = 1} \sum_{i=1}^k p_i(1-p_i)$ will allow to conclude. First, we notice that $\sup_{\forall i, p_i \geq 0, \sum_i p_i = 1} \sum_{i=1}^k p_i(1-p_i) \leq \sup_{\sum_i p_i = 1} \sum_{i=1}^k p_i(1-p_i)$. We dropped the positivity constraints. Then, we see that the gradient of the only remaining constraint is equal to $(1, \dots, 1)$ uniformly. In particular, it is never 0 and thus the KKT conditions apply. They tell us that at the optimum, there exist a $\lambda^* \in \mathbb{R}$ such that

$$(\dots, 1 - 2p_i^*, \dots) = \lambda^*(\dots, 1, \dots).$$

In other words, at the optimum, all the components are equal. Since the vector is a vector of probability, the only possibility is to have $p_i^* = \frac{1}{k}$ for all i.

Finally, we get

$$\mathbb{E}\left(\|\mathbf{p} - \hat{\mathbf{p}}_\epsilon\|_2^2\right) \leq \frac{1}{n} + \frac{8k}{n^2\epsilon^2} \,. \tag{2.6}$$

It can be proven, by using techniques presented in Chapter 3, that when the estimator is not private, the optimal rate of estimation is $\Theta(1/n)$. In particular, when $\epsilon = \Omega(\sqrt{k}/\sqrt{n})$, we can see that the rate of estimation provided by Equation (2.6) is not degraded. On the other hand, when $\epsilon \ll \sqrt{k}/\sqrt{n}$, the guarantees obtained via Equation (2.6) start to degrade.

Furthermore, since Equation (2.6) is an upper bound, it only says that the *guarantees* start to degrade. However, by looking at the special case of the uniform distribution, we have

$$\mathbb{E}\left(\|\mathbf{p}_{\text{unif}} - \hat{\mathbf{p}}_\epsilon\|_2^2\right) = \frac{\left(1 - \frac{1}{k}\right)}{n} + \frac{8k}{n^2\epsilon^2} \,. \tag{2.7}$$

On this example, the utility is effectively degraded. The tools necessary to study the optimality of such estimation will be presented in Chapter 3. In this case, the estimation is not optimal, and a projection step (convex projection on the set of probability distributions) must be added. It then becomes optimal up to polylog factors (see [Acharya et al., 2021e]).

2.3.2 Gaussian mechanism

Using the same formalism as with the Laplace mechanism (deterministic f to make private), the idea of the Gaussian mechanism is to replace f by the randomized mechanism

$$\mathbf{X} \in \mathcal{D} \mapsto f(\mathbf{X}) + \alpha\mathcal{N}(0, I_k) \,,$$

where the notation $\mathcal{N}(0, I_k)$ refers to a random vector with independent components following centered normal distributions of variance 1.

It can easily be shown that if f is not constant on a given pair of neighboring databases, then this mechanism has no chance to be ϵ-DP for any finite ϵ. This is where concentrated differential privacy comes handy : this mechanism is ρ-zCDP for a certain $\rho > 0$. It is

then possible to give results for (ϵ, δ)-DP for strictly positive δ by leveraging the conversion from zCDP to DP.

Fact 2.3.6 (Privacy guarantees)**.** *If* $\alpha \geq \frac{\Delta_2 f}{\sqrt{2\rho}}$*, then the Gaussian mechanism is* ρ*-zCDP.*

Furthermore, the utility of this mechanism is controlled by classical tail bounds on the chi-squared distribution [Laurent & Massart, 2000].

Fact 2.3.7 (Utility guarantees)**.** *Let us note* y *the output of the Gaussian mechanism with noise magnitude* α*. For any* $\gamma > 0$*,*

$$\mathbb{P}\left(\|f(x) - y\|_2^2 \geq \alpha^2 \left(k + 2\sqrt{k \ln\left(\frac{1}{\gamma}\right)} + 2\ln\left(\frac{1}{\gamma}\right)\right)\right) \leq \gamma \, .$$

From global to local sensitivity. The Laplace and the Gaussian mechanisms are defined with a *uniform* bound on the sensitivity (i.e. that it holds for any pair of neighboring datasets), one might be tempted to use the local sensitivity (i.e. it holds for any pair of neighboring datasets with one fixed extremity). With the Laplace mechanism for instance, it can however be shown that it is not possible. It is possible however to use the so-called *smoothed* sensitivity [Nissim et al., 2007] at the cost of an extra slack in the δ. This thesis does not exploit such techniques directly, but it is important to know their existence.

2.3.3 Exponential mechanism

Another extremely important example of private mechanism building block is the so-called *exponential* mechanism [McSherry & Talwar, 2007]. Let us present the setup : The dataset space $\mathcal{D}$ is equipped with a neighboring relation $\sim$, and we want to build a private mechanism taking its output in some output space $\mathcal{O}$ equipped with a reference σ-finite measure μ.

For a dataset $\mathbf{X} \in \mathcal{D}$ and an output candidate $o \in \mathcal{O}$, the utility of o relatively to the dataset $\mathbf{X}$ (i.e. how good o would be if returned by the mechanism applied to $\mathbf{X}$) is measured by a utility function $u : \mathcal{D} \times \mathcal{O} \to \mathbb{R}$. Usually, the convention that "the higher, the better" is taken.

The idea of the exponential mechanism is, given a dataset $\mathbf{X}$, to return a random variable on $\mathcal{O}$ that has a density p w.r.t. μ that almost-surely satisfies

$$p(o) \propto e^{\frac{u(\mathbf{X}, o)}{\alpha}} \, ,$$

where $\propto$ means proportionality, and $\alpha > 0$. In particular, computing the normalization factor is often a problem for sampling from this mechanism, and may require smart sampling algorithms [Gillenwater et al., 2021].

Similarly as with the Laplace and Gaussian mechanisms, the privacy guarantees of the exponential mechanism depend on the *sensitivity* of u.

Definition 2.3.8 (Sensitivity)**.** The sensitivity of u (Δu) is defined as

$$\sup_{\mathbf{X} \sim \mathbf{X}'} \sup_{o \in O} \left| u(\mathbf{X}, o) - u(\mathbf{X}', o) \right| .$$

Fact 2.3.9 (Privacy guarantees)**.** *If $\alpha \geq \frac{2\Delta u}{\epsilon}$, then the exponential mechanism is ϵ-DP. Furthermore, when the normalization factor is independent of the dataset, $\alpha \geq \frac{\Delta u}{\epsilon}$ is enough to arrive to the same conclusions.*

When the output space $\mathcal{O}$ is finite, utility guarantees of the exponential mechanism are easily derived.

Fact 2.3.10 (Utility guarantees, finite case)**.** *Denoting by o the output of the exponential mechanism on $\mathbf{X}$, for any $\gamma > 0$,*

$$\mathbb{P}\left(\sup_{o' \in \mathcal{O}} u(\mathbf{X}, o') - u(\mathbf{X}, o) > \alpha \ln\left(\frac{\#(O)}{\gamma} \right) \right) < \gamma .$$

When the output space is not finite, it can be harder to give utility results for the exponential mechanism. It often requires to control the normalization factor of the exponential mechanism. For instance, in the theory of the inverse sensitivity [Asi & Duchi, 2020b, Asi & Duchi, 2020a], this is done by a technique called *smoothing*. [Kaplan et al., 2022] does it by imposing a minimal gap condition for the quantile problem. This approach is built on in Chapter 6. Finally, Chapter 6 also presents an ad-hoc technique for the multiquantile problem that is based on neutralizing the normalization term by working on probability ratios on continuous domains.

Example 2.3.11 (Inverse sensitivity)**.** An important mechanism building tool that is exploited in Chapter 6 is the *inverse sensitivity* mechanism [Asi & Duchi, 2020b, Asi & Duchi, 2020a]. Recall that since $\mathcal{D}$ is connex for the neighboring relation $\sim$, it is possible to define a distance on $\mathcal{D}$ by stating that the distance between two datasets is the minimum length of the path linking them. Let us note d this distance.

When the target of the private procedure is a deterministic function $f : \mathcal{D} \to \mathcal{O}$, the inverse sensitivity mechanism corresponds to the exponential mechanism with utility function

$$u(\mathbf{X}, o) = -\min \left\{ d(\mathbf{X}, \mathbf{X}') | \mathbf{X}' \in \mathcal{D}, f(\mathbf{X}') = o \right\} .$$

This mechanism first appeared in [McSherry, 2010] for the mean and the median estimation. It was later generalized to arbitrary queries in [Asi & Duchi, 2020b, Asi & Duchi, 2020a]. Very recently, this mechanism has been used to prove the strong equivalence between private mechanisms and robust mechanisms [Asi et al., 2023].

2.3.4 Private optimization

Many problems can be formalized as an optimization problem, i.e. to find the global (or local) extremum(a) of a well-chosen function. In particular, in a learning setup, this function is built from a dataset (e.g. empirical risk of an estimator, least squares, ...). In this scenario, the question of how to find a differential private point that is close to the optimum is crucial.

The results of the literature are numerous and often too verbose to be presented in this overview chapter. However, we can give a few pointers to interesting approaches of this ongoing body of literature [Song et al., 2013, McMahan et al., 2018b, Abadi et al., 2016, Smith et al., 2017, Wu et al., 2017, Iyengar et al., 2019, Song et al., 2020, Song et al., 2021, Mangold et al., 2022, Ganesh et al., 2022, Gopi et al., 2022, Bassily et al., 2019, Bassily et al., 2014, Avella-Medina et al., 2021, Ganesh et al., 2023], and leave the technical details aside.

DP-SGD. Probably the most well-known private optimization algorithm, DP-SGD (for Differentially Private Stochastic Gradient Descent) [Abadi et al., 2016] mimics the behavior of the classical stochastic gradient descent algorithm, with the difference that it clips each sample's relative gradient to a ball, and add noise. Furthermore, it leverages a specific stochastic batch structure in order to increase privacy by subsampling.

In the setup where the problem can be expressed as

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} l((x_i, y_i), \theta)$$

for a loss function l, and where $((x_1, y_1), \ldots, (x_n, y_n))$ is the dataset (considered with the replacement relationship), DP-SGD considers the sequence

$$\theta_{t+1} = \theta_t - \eta_t g_t .$$

The sequence of positive real numbers $(\eta_t)_t$ is the sequence of learning rates. The quantity g_t is a private estimate of the gradient defined as

$$g_t = \frac{1}{\#(B)} \sum_{i \in B} \text{clip}_C \left(\nabla_\theta l((x_i, y_i), \theta) \right) + \mathcal{N}\left(0, \frac{C^2 \sigma^2}{B^2}\right) .$$

Here, B is a batch obtained by i.i.d. selection of each element in the dataset with probability p, $\sigma > 0$ tunes the amount of noise, and clip_C is the function that project onto the Euclidean ball of radius C and centered in 0.

The mechanism generating the noisy gradient is $(\alpha, g(\alpha, \sigma, p))$-Rényi differentially private for any $\alpha > 0$ where

$$g(\alpha, \sigma, p) = \mathrm{D}_\alpha \left((1-p)\mathcal{N}\left(0, \sigma^2\right) + p\mathcal{N}\left(1, \sigma^2\right) \middle\| \mathcal{N}\left(0, \sigma^2\right) \right) .$$

The privacy of a full trajectory can then be characterized in terms of Rényi differential privacy by composition theorems.

Its limitations are that in many application scenarios (e.g. Deep Learning), it comes with little to no utility guarantees. It is much more computationally demanding than its non-private counterpart (for instance requiring very large batch sizes). And, it makes hyperparameter tuning a hassle. Luckily for this last point, rules of thumbs have been designed for interpolating the results of DP-SGD based on a few observations [Sander et al., 2022].

Langevin diffusion. Langevin diffusion refers to the continuous-time stochastic process of a gradient-flow perturbed by a standard Brownian motion. This Brownian motion can be leveraged to obtain privacy, playing the role of the Gaussian noise in DP-SGD. In the convex case, this observation leads to the state of the art first order private solver for convex problems [Ganesh et al., 2022]. It was later adapted to handle stochastic gradients in [Ryffel et al., 2022].

Second order optimization. Recently, second order optimization methods have been investigated in order to reduce the high number of steps that first order private solvers take to converge. In non-private optimization, this is done by adapting the learning rate and the direction of the gradient based on a surrogate of the Hessian matrix. Adaptations for the private setup appear in [Avella-Medina et al., 2021, Ganesh et al., 2023].

Fixed-point methods. Many optimization problems can be reduced to fixed-point equations [Bauschke & Combettes, 2011] of the form

$$x = f(x) .$$

Under suited hypotheses on f and on the set in which x lives, it is often possible to converge to a solution (often unique) of the fixed-point equation via a iterative series $x_{n+1} := f(x_n)$

(e.g. Fixed-point theorem of Banach-Picard in Banach spaces, fixed-point theorem in compact spaces). The choice of f is often equivalent to the choice of an optimization algorithm (e.g. gradient descent).

A very recent piece of work [Cyffers et al., 2023] modified this framework by adding noise to the iterates, in order to obtain differential privacy. The resulting method is general enough to recovers algorithms such as DP-SGD, but also allows directly adapting methods such as ADMM to differential privacy [Boyd et al., 2011].

Chapter 3

Lower-bounds on the statistical risk : a unified framework

The origin of this chapter, and the use of the first person. This chapter is based on the article [Lalanne et al., 2023b], written by Aurélien Garivier[1], Rémi Gribonval[2], and by myself. In this chapter, I will try to respect the following rule : the use of the first person of the plural (we, our, ...) represents all the above-mentioned people, while the use of the first person of the singular (I, my, ...) represents myself.

This chapter studies minimax lower bounds for classes of differentially private estimators. In particular, it shows how to characterize the power of a statistical test under differential privacy in a plug-and-play fashion by solving an appropriate transport problem. With specific coupling constructions, this observation allows deriving Le Cam-type and Fano-type inequalities not only for regular definitions of differential privacy but also for those based on Renyi divergence. This is a core chapter for the thesis that introduces theoretical tools that are used in the rest of the thesis.

3.1 Context on minimax lower-bounds

The lower-bounds and the optimality will be investigated in a *minimax* sense.

[1] https://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/index.html/

[2] https://people.irisa.fr/Remi.Gribonval/

3.1.1 The Minimax Risk and Private Estimators

We start by defining the minimax risk. Given $n \in \mathbb{N}_*$ and a feature space $\mathcal{X}$, $\mathcal{X}^n$ may be viewed as a set of datasets containing n elements from $\mathcal{X}$. We consider a family of probability distributions $(\mathbb{P}_\theta)_{\theta\in\Theta}$ on $\mathcal{X}^n$ where Θ is equipped with a semi-metric[3] $d_\Theta : \Theta^2 \to \mathbb{R}_+$. Often, for all $\theta \in \Theta$, $\mathbb{P}_\theta = \mathbb{p}_\theta^{\otimes n}$ where $(\mathbb{p}_\theta)_{\theta\in\Theta}$ is a family of probability distributions on $\mathcal{X}$. This corresponds to the classical statistical setup where we observe n i.i.d. random variables. The general setup allows capturing phenomena that are not i.i.d., for instance Markov processes. Given an estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$ one might look at its uniform risk of estimation over Θ for a loss function $\Phi : [0, +\infty) \to [0, +\infty)$ that is non-decreasing and such that $\Phi(0) = 0$ which is

$$\sup_{\theta\in\Theta} \int_{\mathcal{X}^n} \Phi(d_\Theta(\hat{\theta}(\mathbf{X}), \theta)) d\mathbb{P}_\theta(\mathbf{X}) \ .$$

The best achievable uniform risk defines what is called the *minimax* risk

$$\mathfrak{M}_n := \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \int_{\mathcal{X}^n} \Phi(d_\Theta(\hat{\theta}(\mathbf{X}), \theta)) d\mathbb{P}_\theta(\mathbf{X}) \ . \tag{3.1}$$

Here, the infimum over $\hat{\theta}$ is taken among all possible measurable functions of the samples.

In order to factorize the results, we will use the abstract formulation that a randomized mechanism $\mathfrak{M} : \mathcal{X}^n \to \Theta$ satisfies a certain condition $\mathcal{C}$ rather than fixing the class in which it belongs. We define the *private minimax* risk as the best achievable uniform risk with mechanisms that satisfy the privacy condition $\mathcal{C}$

$$\mathfrak{M}_n^{(\mathcal{C})} := \inf_{\mathfrak{M} \text{ s.t. } \mathcal{C}} \sup_{\theta\in\Theta} \int_{\mathcal{X}^n} \mathbb{E}_{\mathbb{P}_\mathfrak{M}} \left(\Phi(d_\Theta(\mathfrak{M}(\mathbf{X}), \theta))\right) d\mathbb{P}_\theta(\mathbf{X}) \ . \tag{3.2}$$

Note that with both notations $\mathfrak{M}_n$ and $\mathfrak{M}_n^{(\mathcal{C})}$, there is a lot of implicit (the semi-metric, ...). This is a choice in order to simplify the notations, and the context will fix the ambiguities.

3.1.2 Introducing example

As a warmup we discuss here the simplest possible example on which we can present the questions that this chapter addresses and the flavor of the developed approaches. Let $p_1 < p_2$ be two parameters in $(0, 1)$ and let $U_1, \ldots, U_n$, n be independent and identically distributed uniform random variables on $[0, 1]$. The random variables $Z_i := (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$, $1 \leq i \leq n$, defined by

$$(X_i^{(1)}, X_i^{(2)}) = (\mathbb{1}_{[0,p_1)}(U_i), \mathbb{1}_{[0,p_2)}(U_i))$$

are independent and identically distributed with marginal distributions Bernoulli $\mathcal{B}(p_1)$ and $\mathcal{B}(p_2)$. In the sequel we note $\mathbf{X}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})$, $j = 1, 2$, $\mathbf{U} = (U_1, \ldots, U_n)$,

[3] i.e. that is positive, symmetric, that satisfies the triangular inequality and $d_\Theta(\theta, \theta) = 0, \forall\theta \in \Theta$

$S_1 := [0, (p_1+p_2)/2)$. and $S_2 := [(p_1+p_2)/2, 1]$. Given any $(\epsilon, 0)$-DP mechanism $\mathfrak{M} : [0,1]^n \to [0,1]$ (where $\epsilon > 0$) to estimate the Bernoulli parameter, the risk satisfies

$$\begin{aligned}
&\sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}} \left((\mathfrak{M}(\mathbf{X}) - p)^2\right) \\
&\qquad \geq \left(\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p_1)^{\otimes n},\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}) - p_1)^2\right) + \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p_2)^{\otimes n},\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}) - p_2)^2\right)\right)/2 \\
&\qquad \overset{\text{Coupling}}{=} \left(\mathbb{E}_{\mathbf{U},\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}^{(1)}) - p_1)^2\right) + \mathbb{E}_{\mathbf{U},\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}^{(2)}) - p_2)^2\right)\right)/2 \\
&\qquad \overset{\text{Conditioning}}{=} \mathbb{E}_{\mathbf{U}} \left(\mathbb{E}_{\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}^{(1)}) - p_1)^2\right) + \mathbb{E}_{\mathfrak{M}} \left((\mathfrak{M}(\mathbf{X}^{(2)}) - p_2)^2\right)\right)/2 \\
&\qquad \geq \left(\tfrac{p_2-p_1}{2}\right)^2 \mathbb{E}_{\mathbf{U}} \left(\mathbb{P}_{\mathfrak{M}} \left(\mathfrak{M}(\mathbf{X}^{(1)}) \in S_2\right) + \mathbb{P}_{\mathfrak{M}} \left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_1\right)\right)/2.
\end{aligned} \tag{3.3}$$

This is where the DP property yields a lower bound on the second factor as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{U}}\Big(e^{-\epsilon d_{\text{ham}}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})} \mathbb{P}_{\mathfrak{M}} \left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_2\right) + \mathbb{P}_{\mathfrak{M}} \left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_1\right)\Big) \\
&\quad \overset{d_{\text{ham}}(\cdot,\cdot)\geq 0}{\geq} \mathbb{E}_{\mathbf{U}}\Big(e^{-\epsilon d_{\text{ham}}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})} \Big(\mathbb{P}_{\mathfrak{M}} \left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_2\right) + \mathbb{P}_{\mathfrak{M}} \left(\mathfrak{M}(\mathbf{X}^{(2)}) \in S_1\right)\Big)\Big) \\
&\quad = \mathbb{E}_{\mathbf{U}}\Big(e^{-\epsilon d_{\text{ham}}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})}\Big) \overset{\text{Jensen}}{\geq} e^{-n\epsilon|p_2-p_1|},
\end{aligned} \tag{3.4}$$

which overall yields the lower bound $\frac{(p_2-p_1)^2}{8} e^{-n\epsilon|p_2-p_1|}$. A good lower bound on the minimax risk is then provided by optimizing over p_1 and p_2. For instance, when $n \geq \frac{2}{\epsilon}$, $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{2} + \frac{1}{n\epsilon}$ leads to

$$\sup_{p\in[0,1]} \mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^{\otimes n}} \left((\mathfrak{M}(\mathbf{X}) - p)^2\right) \geq \frac{1}{8} \frac{1}{(n\epsilon)^2}.$$

The idea behind the first inequality in (3.3) is classical in the minimax literature and is recalled in Section 3.1.3 using the notion of *packing*. The *coupling construction* can be generalized and taylored to other settings and has a critical impact on the deduced lower bounds, as we present in Section 3.4. The minoration involving differential privacy is a special case of the techniques that we formalize under the notion of *admissible similarity functions* in Section 3.3, which are adapted to various types of privacy constraints.

3.1.3 From Minimax Lower Bounds to Hypothesis Testing

A classical technique (see [Duchi et al., 2013]) for finding lower bounds on $\mathfrak{M}_n\left((\mathbb{P}_\theta)_{\theta\in\Theta}, d_\Theta, \Phi\right)$ is to replace the parameter set Θ by a much "simpler" set $\Theta' \subseteq \Theta$ and to use the trivial lower bound

$$\mathfrak{M}_n \geq \inf_{\hat{\theta}} \sup_{\theta\in\Theta'} \int_{\mathcal{X}^n} \Phi(d_\Theta(\hat{\theta}(\mathbf{X}), \theta)) d\mathbb{P}_\theta(\mathbf{X}).$$

Usually Θ' is chosen as an Ω-*packing* of Θ, for some real number $\Omega > 0$: it is a countable family $\Theta' := \{\theta_i, i \in \mathbb{N}_*\}$ $(\theta_i)_{i\in\mathbb{N}_*}$ (and most of the time, including in this chapter, it is

taken to be finite) such that: a) $\theta_i \in \Theta$ for all i; b) $i \neq j \implies d_\Theta(\theta_i, \theta_j) \geq 2\Omega$; and c) there is a well-defined function $\Psi_{\Theta'}$ satisfying

$$\Psi_{\Theta'}(\theta) \in \arg\min_{i \geq 1} d_\Theta(\theta_i, \theta)$$

for each $\theta \in \Theta$. Under such hypotheses, any estimator $\hat{\theta}$ satisfies [Duchi et al., 2013]

$$\sup_{\theta \in \Theta'} \int_{\mathcal{X}^n} \Phi(d_\Theta(\hat{\theta}(\mathbf{X}), \theta)) d\mathbb{P}_\theta(\mathbf{X}) \geq \Phi(\Omega) \sup_{i \in \{1,\dots,\#(\Theta')\}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\theta_i}} \left(\Psi_{\Theta'} \left(\hat{\theta}(\mathbf{X}) \right) \neq i \right) . \tag{3.5}$$

The mapping $\hat{\Psi} := \Psi_{\Theta'} \circ \hat{\theta} : \mathcal{X}^n \to \{1, \dots, \#(\Theta')\}$ may be viewed as a *test function* (that selects the model number) and thus

$$\mathfrak{M}_n \geq \Phi(\Omega) \inf_{\Psi : \mathcal{X}^n \to \{1,\dots,\#(\Theta')\}} \sup_{i \in \{1,\dots,\#(\Theta')\}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\theta_i}} (\Psi(\mathbf{X}) \neq i) . \tag{3.6}$$

Finding minimax lower bounds is thus done by finding a suitable Ω-packing of the parameter space and then by providing lower bounds on

$$\inf_{\Psi : \mathcal{X}^n \to \{1,\dots,\#(\Theta')\}} \sup_{i \in \{1,\dots,\#(\Theta')\}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\theta_i}} (\Psi(\mathbf{X}) \neq i) . \tag{3.7}$$

Two powerful tools to find such lower bounds come from information theory: Le Cam's lemma (see Fact 3.1.1) can be used when Θ' only contains two elements, while Fano's lemma (see Fact 3.1.2) is applicable when Θ' contains $N \geq 2$ elements.

Fact 3.1.1 (Neyman-Pearson & Le Cam's lemma [Rigollet & Hütter, 2015, Lemma 5.3])**.** *Let $\mathbb{P}_1, \mathbb{P}_2$ be two probability distributions on a measure space $\mathcal{E}$, then*

$$\begin{aligned} \inf_{\Psi : \mathcal{E} \to \{1,2\}} \max_{i \in \{1,2\}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i} (\Psi(\mathbf{X}) \neq i) &\geq \frac{1}{2} \inf_{\Psi : \mathcal{E} \to \{1,2\}} \sum_{i=1}^{2} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i} (\Psi(\mathbf{X}) \neq i) \\ &= \frac{1}{2} (1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)) . \end{aligned} \tag{3.8}$$

Let us highlight that along this thesis, the term "Fact" refers to results directly borrowed from existing literature, independently of the supposed technicality of the result and/or of its proof. It is simply used to easily emphasize whether a result is a contribution or not.

Fact 3.1.2 (Fano's lemma [Giraud, 2021, Theorem 3.1])**.** *Let $(\mathbb{P}_i)_{i \in \{1,\dots,N\}}$ be a family of probability distributions on a measure space $\mathcal{E}$. For any probability distribution $\mathbb{Q}$ on $\mathcal{E}$ such that $\mathbb{P}_i \ll \mathbb{Q}$ for all i, and for any test function $\Psi : \mathcal{X}^n \to \{1, \dots, N\}$,*

$$\begin{aligned} \max_{i \in \{1,\dots,N\}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i} (\Psi(\mathbf{X}) \neq i) &\geq \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i} (\Psi(\mathbf{X}) \neq i) \\ &\geq 1 - \frac{1 + \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i \| \mathbb{Q})}{\ln(N)} . \end{aligned} \tag{3.9}$$

Often $\mathbb{Q}$ *is set to* $\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i$.

With the same reasoning used [Duchi et al., 2013] to establish (3.5), with $\Theta' = (\theta_i)_{i\in\{1,\dots,\#(\Theta')\}}$ an Ω-packing of Θ, we can lower-bound the *private* minimax risk:

$$\mathfrak{M}_n^{(\mathcal{C})} \geq \Phi(\Omega) \inf_{\mathfrak{M} \text{ s.t. } \mathcal{C}} \inf_{\Psi:\text{codom}(\mathfrak{M})\to\{1,\dots,\#(\Theta')\}} \sup_{i\in\{1,\dots,\#(\Theta')\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\theta_i},\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \ . \tag{3.10}$$

Consequently, finding private minimax lower bounds is done analogously to the non-private setting by finding an appropriate Ω-packing and a lower bound on

$$\inf_{\Psi:\text{codom}(\mathfrak{M})\to\{1,\dots,\#(\Theta')\}} \sup_{i\in\{1,\dots,\#(\Theta')\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\theta_i},\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \tag{3.11}$$

that is independent on the mechanism $\mathfrak{M}$ but only depends on the privacy condition $\mathcal{C}$.

3.2 Quantitative results : constraint-specific lower-bounds

The main contribution of this work, presented in Section 3.3, is to propose a generic framework for the derivation of lower bounds on the minimax risk under various privacy conditions. Technically, the techniques of Le Cam and Fano are extended to the private context, reducing the distributional test problem (3.11) to a Kantorovich problem [Santambrogio, 2016, Peyré & Cuturi, 2019, Villani et al., 2009] of the form

$$\sup_{\mathbb{Q}\in\Pi(\mathbb{P}_1,\dots,\mathbb{P}_N)} \int_{(\mathcal{X}^n)^N} s_{\mathcal{C}}\left(\mathbf{X}_1,\dots,\mathbf{X}_N\right) d\mathbb{Q}\left(\mathbf{X}_1,\dots,\mathbf{X}_N\right) \ . \tag{3.12}$$

Here, $\Pi\left(\mathbb{P}_1,\dots,\mathbb{P}_N\right)$ is the set of *couplings* between the considered distributions and $s_{\mathcal{C}}$ is an *admissible similarity function* depending on the nature of the constraint $\mathcal{C}$ and the number of hypotheses (Theorem 3.3.3 and Theorem 3.3.4). For instance, regarding (ϵ,δ)-differential privacy, similarity functions are obtained by comparing datasets to a common *anchor*. This result is summarized in Theorem 3.3.2.

Unlike the prior work of [Acharya et al., 2021e], the proposed framework allow us to consider joint couplings across all instances rather than just pairwise couplings. Additionally, the level of generality of our proofs leaves room for subsequent work to build upon this framework.

The general idea behind the proofs is as follows. In classical Fano's, one considers the decoding error probability: on average over a family of instances, what is the probability that the estimator, given samples from a given instance, fails to identify that the samples came from that instance. In place of Fano's inequality, the present work lower bounds this by noting that, given datasets $\mathbf{X}_1,...,\mathbf{X}_N$ coming from each instance, as well as an "anchor" dataset Λ (or alternatively an anchor distribution), differential privacy implies

that the probability that the estimator decides $\mathbf{X}_i$ comes from instance i cannot differ by much from the probability it decides Λ comes from instance i, provided Λ and $\mathbf{X}_i$ are similar. The decoding error probability can thus be lower bounded in terms of the maximum distance between Λ and any of $\mathbf{X}_1, ..., \mathbf{X}_N$, averaged over the randomness of $\mathbf{X}_1, ..., \mathbf{X}_N$, where there is freedom in choosing how to couple this randomness.

Section 3.4 includes various coupling constructions yielding quantitative lower bounds for the Kantorovich formulation (3.12). These constructions only depend on the number of hypotheses N, the sample size n, the privacy parameters ϵ, δ, ρ, and information theoretic quantities such as the pairwise total variations or KL divergences between the distributions. Those results will be presented in Section 3.3 and in Section 3.4.

3.2.1 Differential privacy with two hypotheses

We now showcase useful consequences, starting with the case $N = 2$: similarly to [Acharya et al., 2021e], we extend Le Cam's lemma to the (ϵ, δ)-differentially private setting:

Theorem 3.2.1 (Le Cam for (ϵ, δ)-DP)**.** *If a randomized mechanism $\mathfrak{M}$ satisfies (ϵ, δ)-DP, then for any test function $\Psi : \mathrm{codom}(\mathfrak{M}) \to \{1, 2\}$ and any probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathcal{X}^n$ we have*

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right) \neq i\right) \geq \frac{1}{2}\max\Big\{1 - \mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right) \ , \\ 1 - \left(1 - e^{-n\epsilon} + 2ne^{-\epsilon}\delta\right)\mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right)\Big\} \ .$$

Furthermore, when $\mathbb{P}_1 = \mathbb{p}_1^{\otimes n}$ and $\mathbb{P}_2 = \mathbb{p}_2^{\otimes n}$ are product *distributions,*

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right) \neq i\right) \\ \geq \frac{1}{2}\left(\left(1 - \left(1 - e^{-\epsilon}\right)\mathrm{TV}\left(\mathbb{p}_1, \mathbb{p}_2\right)\right)^n - 2ne^{-\epsilon}\delta\mathrm{TV}\left(\mathbb{p}_1, \mathbb{p}_2\right)\right) \ .$$

The proof can be found in Section 3.4.2. The classical lower bound of Le Cam (3.8) allows for a tunable testing difficulty depending on $\mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right)$. However, in the regime $\epsilon, \delta = o(1/n)$, the private lower bound is $\Omega(1)$: it becomes arbitrarily hard to distinguish between any pair of distributions.

For i.i.d. observations ($\mathbb{P}_i = \mathbb{p}_i^{\otimes n}$), it follows by convexity that for any (ϵ, δ)-DP mechanism $\mathfrak{M}$ and test function $\Psi : \mathrm{codom}(\mathfrak{M}) \to \{1, 2\}$

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right) \neq i\right) \geq \frac{1}{2}\left(e^{-\epsilon n\mathrm{TV}\left(\mathbb{p}_1,\mathbb{p}_2\right)} - 2e^{-\epsilon}\delta n\mathrm{TV}\left(\mathbb{p}_1, \mathbb{p}_2\right)\right) \ .$$

This is to be compared to the state of the art lower bound of [Acharya et al., 2021e, Theorem 1]:

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right)\neq i\right) \geq \frac{1}{2}\left(0.9e^{-10\epsilon n\mathrm{TV}\left(\mathbb{p}_1,\mathbb{p}_2\right)} - 10\delta n\mathrm{TV}\left(\mathbb{p}_1,\mathbb{p}_2\right)\right) .$$

Theorem 3.2.1 gives tighter results with better constants, notably thanks to a different proof technique avoiding some convexity and concentration inequalities. While this difference does not change the obtained rates, such an improvement is significant on the resulting sample complexities by a factor 10 in the exponential. As an illustration, imagine that a statistician has to discriminate between the two hypotheses $H_0 : \mathcal{B}\left(\frac{50}{100}\right)$ and $H_1 : \mathcal{B}\left(\frac{51}{100}\right)$, two Bernoulli distributions. Under $\epsilon = 0.1$, if we want H_0 and H_1 to be falsely rejected with probability at most 1%, [Acharya et al., 2021e] says that the experiment will have to be calibrated with at least 381 participants while our Theorem 3.2.1 says that in fact, at least 4109 participants will be necessary, leading to a less over-optimistic estimation by a large factor.

3.2.2 Concentrated differential privacy with two hypotheses

We also prove an equivalent for so-called ρ-zero concentrated differential privacy (or in short ρ-zCDP), which is, to the best of our knowledge, the first successful attempt at doing so.

Theorem 3.2.2 (Le Cam for ρ-zCDP)**.** *If a randomized mechanism $\mathfrak{M}$ satisfies ρ-zCDP, then for any test function $\Psi : \mathrm{codom}\left(\mathfrak{M}\right) \to \{1,\dots,N\}$ and any probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathcal{X}^n$,*

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right)\neq i\right) \geq \frac{1}{2}\max\Big\{1-\mathrm{TV}\left(\mathbb{P}_1,\mathbb{P}_2\right) , \\ 1-n\sqrt{\rho/2}\mathrm{TV}\left(\mathbb{P}_1,\mathbb{P}_2\right)\Big\} .$$

Furthermore, when $\mathbb{P}_1 = \mathbb{p}_1^{\otimes n}$ and $\mathbb{P}_2 = \mathbb{p}_2^{\otimes n}$ are product *distributions,*

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right)\neq i\right) \geq \frac{1}{2}\left(1-n\sqrt{\rho/2}\mathrm{TV}\left(\mathbb{p}_1,\mathbb{p}_2\right)\right) .$$

The proof of this result can be found in Section 3.4.2. As above, any two distributions can no longer be distinguished in the regime $\rho \ll 1/n^2$.

3.2.3 Differential privacy with many hypotheses

For more than two hypotheses and (ϵ,δ)-DP, we also get a private version of Fano's lemma.

Theorem 3.2.3 (Multiple Distributional Tests for (ϵ,δ)-DP)**.** *If a randomized mechanism $\mathfrak{M}$ satisfies (ϵ,δ)-DP, then for any test function $\Psi : \mathrm{codom}\left(\mathfrak{M}\right) \to \{1,\dots,N\}$, any family*

of probability distributions $(\mathbb{P}_i)_{i\in\{1,\dots,N\}}$ *on* $\mathcal{X}^n$ *and any* $\mathbb{Q}$ *such that* $\mathbb{P}_i \ll \mathbb{Q}$ *for all* i,

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right) \geq \max\Bigg\{1-\frac{1+\frac{1}{N}\sum_{i=1}^N \mathrm{KL}\left(\mathbb{P}_i\|\,\mathbb{Q}\right)}{\ln(N)}\ ,\\ \frac{1}{2}-\frac{1-e^{-n\epsilon}+2ne^{-\epsilon}\delta}{2N^2}\sum_{i,j}\frac{2\mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right)}{1+\mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right)}\ ,\\ \mathbb{1}_{\delta=0}\times\left(1-\frac{1+\frac{n\epsilon}{N^2}\sum_{i,j}\frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}}{\ln(N)}\right)\Bigg\}\ .$$

Furthermore,when $\mathbb{P}_1=\mathbb{p}_1^{\otimes n}$, $\dots$, $\mathbb{P}_N=\mathbb{p}_N^{\otimes n}$ *are* product *distributions,*

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right) \geq \max\Bigg\{\frac{1}{2N^2}\sum_{i,j}\Bigg(\left(1-(1-e^{-\epsilon})\frac{2\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}{1+\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}\right)^n\\ -2ne^{-\epsilon}\delta\frac{2\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}{1+\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}\Bigg)\ ,\\ \mathbb{1}_{\delta=0}\times\left(1-\frac{1+\frac{n\epsilon}{N^2}\sum_{i,j}\frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}}{\ln(N)}\right)\Bigg\}\ .$$

The proof is given in Section 3.4.2. When dealing with product distributions, the quantity

$$D:=\frac{n}{N^2}\sum_{i,j}\frac{2\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}{1+\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}$$

can roughly be seen as an averaged hamming distance between pairs of marginals. An implication of Theorem 3.2.3 is then that

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right) \geq \ \mathbb{1}_{\delta=0}\times\left(1-\frac{1+\epsilon D}{\ln(N)}\right)\ .$$

As the bound of [Acharya et al., 2021e, Theorem 2]

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right) \geq \ \mathbb{1}_{\delta=0}\times 0.9\times\min\left\{1,\frac{N}{e^{10\epsilon D}}\right\}\ , \tag{3.13}$$

the lower bound is $\Omega(1)$ in the regime $D=o(\ln(N)/\epsilon)$. In particular, both inequalities are expected to yield similar qualitative results for a broad range of applications. However, the quantitative consequences of Theorem 3.2.3 can again be orders of magnitude better. Another improvement of our result is that, contrary to previous work, our bound allows to handle asymmetric hypotheses. Indeed, prior work is based on a uniform upper-bound on the family $\left(\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)\right)_{i,j}$ whereas our work uses only their mean value. As an illustration, if a statistician was to discriminate between a set of N distributions with for instance $N-1$ distributions close to each other in total variation distance and one outlier far from all

the others, the results of [Acharya et al., 2021e] only tell that the problem will be at least as hard as discriminating distributions that are far from one another (which is easy). In contrast, our Theorem 3.2.3 shows that the true testing difficulty lies in discriminating the distributions that are similar (the outlier vanishes), thus resulting in lower bounds that are less over-optimistic.

3.2.4 Concentrated differential privacy with many hypotheses

Similarly, we obtain results for multiple hypotheses under ρ-zCDP.

Theorem 3.2.4 (Multiple Distributional Tests for ρ-zCDP)**.** *If a randomized mechanism $\mathfrak{M}$ satisfies ρ-zCDP, then for any test function $\Psi : \mathrm{codom}(\mathfrak{M}) \to \{1, \dots, N\}$, any family of probability distributions $(\mathbb{P}_i)_{i\in\{1,\dots,N\}}$ on $\mathcal{X}^n$ and any $\mathbb{Q}$ such that $\mathbb{P}_i \ll \mathbb{Q}$ for all i,*

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \geq \max\left\{1 - \frac{1 + \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}\left(\mathbb{P}_i \| \mathbb{Q}\right)}{\ln(N)}\ ,\right.$$
$$\left. 1 - \frac{1 + \frac{n^2\rho}{N^2}\sum_{i,j} \frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}}{\ln(N)}\right\}.$$

Furthermore, when $\mathbb{P}_1 = \mathbb{p}_1^{\otimes n}$, …, $\mathbb{P}_N = \mathbb{p}_N^{\otimes n}$ are product *distributions,*

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \geq 1 - \frac{1 + \frac{n^2\rho}{N^2}\sum_{i,j} \frac{1}{n}\frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)} + \left(\frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}\right)^2}{\ln(N)}.$$

The proof is to be found in Section 3.4.2. This result recovers a recently published result in [Kamath et al., 2022], with the advantage again of better handling asymmetrical hypotheses (i.e. with possible outliers). Another interesting observation is that our framework unifies the proofs of lower bounds under a general technique based on multiple marginals coupling and similarity functions.

3.3 From Testing to a Transport Problem

This section presents our main theorem, which states that finding lower bounds on (3.11) can be done by solving a transport problem [Santambrogio, 2016, Peyré & Cuturi, 2019]. In some sense, this view is close to the coupling of [Acharya et al., 2021e] which considers couplings between pairs of marginals and controls the variations of the hamming distance compared to its expected value with Markov's inequality. However, the high level view that our result allows to obtain numerically sharper results because it allows to skip approximations such as those involving Jensen or Markov inequalities and more importantly, it allows handling divergence-based definitions of privacy which do not fit in the framework of [Acharya et al., 2021e]. Furthermore, a key difference is that the theory of [Acharya

et al., 2021e] only requires to build couplings between pairs of marginals, whereas our theory requires building couplings between all the marginals at the same time. This is both a drawback because it requires to use more complex coupling constructions, and an advantage because it allows to give results that are easier to use when there are more than two hypotheses.

Our analysis is based on the notion of *similarity* functions.

Definition 3.3.1. Given a condition $\mathcal{C}$, we say that a similarity function $s_{\mathcal{C}} : (\mathcal{X}^n)^N \to \mathbb{R}$ is admissible for $\mathcal{C}$ if for any mechanism $\mathfrak{M} : \mathcal{X}^n \to \mathrm{codom}(\mathfrak{M})$ that satisfies $\mathcal{C}$, for any test function $\Psi : \mathrm{codom}(\mathfrak{M}) \to \{1, \dots, N\}$, and for any $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathcal{X}^n$, the following inequality holds:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) \geq s_{\mathcal{C}} \left(\mathbf{X}_1, \dots, \mathbf{X}_N \right) \ .$$

Theorem 3.3.2. *If a randomized mechanism $\mathfrak{M} : \mathcal{X}^n \to \mathrm{codom}(\mathfrak{M})$ satisfies the privacy condition $\mathcal{C}$, for any $N \geq 2$, if $s_{\mathcal{C}} : (\mathcal{X}^n)^N \to \mathbb{R}$ is an admissible similarity function for $\mathcal{C}$, for any distributions $\mathbb{P}_1, \dots, \mathbb{P}_N$ over $\mathcal{X}^n$ we have*

$$\begin{aligned} \inf_{\Psi : \mathrm{codom}(\mathfrak{M}) \to \{1,\dots,N\}} \max_{i \in \{1,\dots,N\}} & \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i, \mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X} \right) \right) \neq i \right) \\ & \geq \sup_{\mathbb{Q} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)} \int_{(\mathcal{X}^n)^N} s_{\mathcal{C}} \left(\mathbf{X}_1, \dots, \mathbf{X}_N \right) d\mathbb{Q} \left(\mathbf{X}_1, \dots, \mathbf{X}_N \right) \ . \end{aligned} \tag{3.14}$$

Proof. Given a test function $\Psi : \mathrm{codom}(\mathfrak{M}) \to \{1, \dots, N\}$ and a coupling $\mathbb{Q} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$,

$$\begin{aligned} \max_{i \in \{1,\dots,N\}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i, \mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X} \right) \right) \neq i \right) &\geq \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_i, \mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X} \right) \right) \neq i \right) \\ &= \int_{(\mathcal{X}^n)^N} \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) d\mathbb{Q} \left(\mathbf{X}_1, \dots, \mathbf{X}_N \right) \\ &\geq \int_{(\mathcal{X}^n)^N} s_{\mathcal{C}} \left(\mathbf{X}_1, \dots, \mathbf{X}_N \right) d\mathbb{Q} \left(\mathbf{X}_1, \dots, \mathbf{X}_N \right) \ . \end{aligned}$$

□

In particular, under (ϵ, δ)-DP, similarity functions are built using a technique that we call *anchoring* which will be introduced in Section 3.3.1, where the proof of the following theorem is given.

Theorem 3.3.3 (Admissible similarity functions for (ϵ, δ)-DP)**.** *When $\mathcal{C}$ is (ϵ, δ)-differential privacy, the following approaches yield admissible similarity functions.*

- **Global anchoring.** *Consider any* anchor function $\Lambda : (\mathcal{X}^n)^N \to \mathcal{X}^n$*, and define the admissible similarity function as*

$$s_{\mathcal{C}}(\mathbf{X}_1, \ldots, \mathbf{X}_N) := \frac{N-1}{N} e^{-\epsilon \max_i (d_{\mathrm{ham}}(\mathbf{X}_i, \Lambda(\mathbf{X}_1, \ldots, \mathbf{X}_N)))} - e^{-\epsilon} \delta \max_i \left(d_{\mathrm{ham}}\left(\mathbf{X}_i, \Lambda\left(\mathbf{X}_1, \ldots, \mathbf{X}_N\right)\right)\right) .$$

- **Projection anchoring.** *In particular, for any $j \in \{1, \ldots, N\}$, consider the* projection anchor $\Lambda_j(\mathbf{X}_1, \ldots, \mathbf{X}_N) := \mathbf{X}_j$*, and define the corresponding admissible similarity function*

$$s_{\mathcal{C}}(\mathbf{X}_1, \ldots, \mathbf{X}_N) := \frac{N-1}{N} e^{-\epsilon \max_i (d_{\mathrm{ham}}(\mathbf{X}_i, \mathbf{X}_j))} - e^{-\epsilon} \delta \max_i \left(d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)\right)$$

- (ϵ, δ)**-DP Le Cam matching.** *When $N = 2$, there is a global anchor function yielding the admissible similarity function*

$$s_{\mathcal{C}}(\mathbf{X}_1, \mathbf{X}_2) := \frac{1}{2} e^{-\epsilon \lceil d_{\mathrm{ham}}(\mathbf{X}_1, \mathbf{X}_2)/2 \rceil} - e^{-\epsilon} \delta \left\lceil d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)/2 \right\rceil .$$

- **Pairwise anchoring.** *An admissible similarity function is*

$$s_{\mathcal{C}}(\mathbf{X}_1, \ldots, \mathbf{X}_N) := \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} e^{-\epsilon \lceil d_{\mathrm{ham}}(\mathbf{X}_i, \mathbf{X}_j)/2 \rceil} - 2e^{-\epsilon} \delta \left\lceil d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)/2 \right\rceil .$$

- $(\epsilon, 0)$**-DP Fano matching.** *When $\delta = 0$, an admissible similarity function is*

$$s_{\mathcal{C}}(\mathbf{X}_1, \ldots, \mathbf{X}_N) := 1 - \frac{1 + \frac{\epsilon}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)}{\ln(N)} .$$

When working under ρ-zCDP, admissible similarity functions are built using classical information theoretic inequalities directly. It can be seen as a form of anchoring on the distributions rather than on the observed random variables (i.e. all the distributions are compared to a common distribution directly that is not necessarily a pushforward by $\mathfrak{M}$). The following result is proved in Section 3.3.2.

Theorem 3.3.4 (Admissible similarity functions for ρ-zCDP)**.** *When $\mathcal{C}$ is the ρ-zero concentrated-differential privacy, the two following quantities are admissible similarity functions:*

- **ρ-zCDP Fano matching**

$$s_{\mathcal{C}}\left(\mathbf{X}_1, \ldots, \mathbf{X}_N\right) := 1 - \frac{1 + \frac{\rho}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)^2}{\ln(N)} \, .$$

- **ρ-zCDP Le Cam matching** *When $N = 2$*

$$s_{\mathcal{C}}\left(\mathbf{X}_1, \mathbf{X}_2\right) := \frac{1}{2}\left(1 - \sqrt{\rho/2} d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)\right) \, .$$

Note that similarity functions can also be easily built for the more general notion of (ξ, ρ) - concentrated differential privacy by swapping the group privacy property for its correct variant (see [Bun & Steinke, 2016]). We do not include the results about (ξ, ρ)-concentrated differential in this section because our objective is more to illustrate the versatility of our framework rather than to build a complete catalogue.

3.3.1 The case of (ϵ, δ)-differential privacy

(ϵ, δ)-differential privacy allows to compare conditional distributions for datasets depending on their Hamming distance. In particular, characterizing the pushforward of a distribution by a private mechanism in not an easy task. We overtake that difficulty with a technique that we call *anchoring*. Informally, an anchor is a function that, given multiple datasets, decides a common dataset to exploit so called *group privacy* of (ϵ, δ)-DP mechanisms and to give numerically tractable results.

Fact 3.3.5 ((ϵ, δ)-DP Group Privacy)**.** *Given $\epsilon \in \mathbb{R}_{+*}$ and $\delta \in [0, 1)$, if a randomized mechanism $\mathfrak{M} : \mathcal{X}^n \to \mathrm{codom}\left(\mathfrak{M}\right)$ is (ϵ, δ)-differentially private, then, for all $\mathbf{X}, \mathbf{Y} \in \mathcal{X}^n$ and all measurable $S \subseteq \mathrm{codom}\left(\mathfrak{M}\right)$, we have*

$$\mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{X}) \in S\right) \leq e^{\epsilon d_{\mathrm{ham}}(\mathbf{X}, \mathbf{Y})} \mathbb{P}_{\mathfrak{M}}\left(\mathfrak{M}(\mathbf{Y}) \in S\right) + \delta d_{\mathrm{ham}}\left(\mathbf{X}, \mathbf{Y}\right) e^{\epsilon\left(d_{\mathrm{ham}}(\mathbf{X}, \mathbf{Y}) - 1\right)} \, .$$

Global Anchoring

The first type of anchor is a global anchor, where all the marginal datasets are compared to the same one.

Lemma 3.3.6 (Global Anchoring)**.** *Consider an* (ϵ, δ)*-DP mechanism* $\mathfrak{M}$ *, a test function* $\Psi : \operatorname{codom}(\mathfrak{M}) \to \{1, \dots, N\}$*, and datasets* $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathcal{X}^n$*. For any* anchor *function* $\Lambda : (\mathcal{X}^n)^N \to \mathcal{X}^n$*, we have*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right) \neq i\right) \quad \geq \frac{N-1}{N} e^{-\epsilon \max_i d_{\mathrm{ham}}(\mathbf{X}_i, \Lambda)} - e^{-\epsilon}\delta \max_i d_{\mathrm{ham}}\left(\mathbf{X}_i, \Lambda\right)$$

where Λ *is a shorthand for* $\Lambda\left(\mathbf{X}_1, \dots, \mathbf{X}_N\right)$*.*

Proof. By the group privacy property (see Fact 3.3.5), we have for each $i \in \{1, \dots, N\}$

$$\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right) \neq i\right) \geq e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_i, \Lambda)} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\Lambda\right)\right) \neq i\right) - e^{-\epsilon}\delta d_{\mathrm{ham}}\left(\mathbf{X}_i, \Lambda\right) \ .$$

As a result,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right) \neq i\right) \\
&\qquad \geq \frac{1}{N}\left(\sum_{i=1}^{N} e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_i, \Lambda)} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\Lambda\right)\right) \neq i\right) - e^{-\epsilon}\delta d_{\mathrm{ham}}\left(\mathbf{X}_i, \Lambda\right)\right) \\
&\qquad \geq \frac{1}{N}\left(e^{-\epsilon \max_i d_{\mathrm{ham}}(\mathbf{X}_i, \Lambda)} \sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\Lambda\right)\right) \neq i\right)\right. \\
&\qquad\qquad \left. - N e^{-\epsilon}\delta \max_i d_{\mathrm{ham}}\left(\mathbf{X}_i, \Lambda\right)\right) \\
&\qquad = \frac{N-1}{N} e^{-\epsilon \max_i d_{\mathrm{ham}}(\mathbf{X}_i, \Lambda)} - e^{-\epsilon}\delta \max_i d_{\mathrm{ham}}\left(\mathbf{X}_i, \Lambda\right) \ ,
\end{aligned}$$

where we used $\sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\Lambda\right)\right) \neq i\right) = \sum_{i=1}^{N}(1 - \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\Lambda\right)\right) = i\right)) = N - 1$ to get the last equality. □

Remark 3.3.7 ((ϵ, δ)-DP Le Cam Matching)**.** When we have to find an anchor between only two datasets, we can design it optimally. Considering any $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}^n$, by definition these datasets disagree on exactly $d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)$ entries. The projection anchor $\Lambda = \Lambda_j$, $j \in \{1, 2\}$ consists in anchoring both $\mathbf{X}_1$ and $\mathbf{X}_2$ to $\mathbf{X}_j$. Consequently, we have $\max\left\{d_{\mathrm{ham}}\left(\mathbf{X}_1, \Lambda\right), d_{\mathrm{ham}}\left(\mathbf{X}_2, \Lambda\right)\right\} = d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)$. If instead we allocate in the anchor Λ half of the disagreeing components to $\mathbf{X}_1$ and the other half to $\mathbf{X}_2$, we get an anchor that satisfies

$$\max\left\{d_{\mathrm{ham}}\left(\mathbf{X}_1, \Lambda\right), d_{\mathrm{ham}}\left(\mathbf{X}_2, \Lambda\right)\right\} = \lceil d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)/2 \rceil \, .$$

Furthermore, one can check that no anchor can achieve a better bound. With this new anchor, the direct application of Lemma 3.3.6 yields

$$\begin{aligned}
&\frac{1}{2}\left(\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_1\right)\right) \neq 1\right) + \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_2\right)\right) \neq 2\right)\right) \\
&\qquad \geq \frac{1}{2} e^{-\epsilon \lceil d_{\mathrm{ham}}(\mathbf{X}_1, \mathbf{X}_2)/2 \rceil} - e^{-\epsilon}\delta \left\lceil d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)/2 \right\rceil \ .
\end{aligned} \tag{3.15}$$

Pairwise Anchoring

The fact that one needs to control the maximum of the hamming distances between a single anchor and the marginals might be prohibitive. We give here a symmetrized version that only requires to control the hamming distances between the pairs of marginals.

Lemma 3.3.8 (Pairwise Anchoring)**.** *Under the assumptions of Lemma 3.3.6 we have*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right)\neq i\right) \geq \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(e^{-\epsilon\lceil d_{\mathrm{ham}}(\mathbf{X}_i,\mathbf{X}_j)/2\rceil}-2e^{-\epsilon}\delta\left\lceil d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)/2\right\rceil\right)\ .$$

Proof. First we observe that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right)\neq i\right) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right)\neq i\right)+\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_j\right)\right)\neq j\right)\right)\ .$$

We then consider the two-point anchor defined in Remark 3.3.7 and get using (3.15) that for every $1\leq i,j\leq N$,

$$\begin{aligned}&\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right)\neq i\right)+\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_j\right)\right)\neq j\right)\\ &\quad\geq e^{-\epsilon\lceil d_{\mathrm{ham}}(\mathbf{X}_i,\mathbf{X}_j)/2\rceil}-2e^{-\epsilon}\delta\left\lceil d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)/2\right\rceil.\end{aligned}$$

□

The special case of $(\epsilon,0)$-DP

The following lemma yields a bound on the KL divergence between the output distributions of an $(\epsilon,0)$-DP mechanism applied to different datasets.

Lemma 3.3.9. *If a randomized mechanism $\mathfrak{M}:\mathcal{X}^n\to\mathrm{codom}\left(\mathfrak{M}\right)$ is $(\epsilon,0)$-DP, then*

$$\forall\mathbf{X},\mathbf{Y}\in\mathcal{X}^n,\quad \frac{d\mathbb{P}_{\mathfrak{M}(\mathbf{X})}}{d\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}}\leq e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon},\quad \mathbb{P}_{\mathfrak{M}(\mathbf{X})}-\ \textit{almost surely}\,,$$

where $\frac{d\mathbb{P}_{\mathfrak{M}(\mathbf{X})}}{d\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}}$ is the Radon-Nikodym density of the distribution of the output of the mechanism with input $\mathbf{X}$, with respect to the distribution of the output of the mechanism with input $\mathbf{Y}$. As a consequence,

$$\forall\mathbf{X},\mathbf{Y}\in\mathcal{X}^n,\quad \mathrm{KL}\left(\mathfrak{M}(\mathbf{X})\|\,\mathfrak{M}(\mathbf{Y})\right)\leq\epsilon d_{\mathrm{ham}}\left(\mathbf{X},\mathbf{Y}\right)\ .$$

Proof. By the group privacy property (see Fact 3.3.5), it is clear that the measurable sets of null measure for $\mathbb{P}_{\mathfrak{M}(\mathbf{X})}$ are exactly the measurable sets of null measure for $\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}$. In particular, $\mathbb{P}_{\mathfrak{M}(\mathbf{X})} \ll \mathbb{P}_{\mathfrak{M}(\mathbf{Y})}$ and hence $p := \frac{d\mathbb{P}_{\mathfrak{M}(\mathbf{X})}}{d\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}}$ exists. By group privacy property again for each measurable set $S \subseteq \operatorname{codom}(\mathfrak{M})$ we have

$$\begin{aligned}
\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}(S) &\geq e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}\mathbb{P}_{\mathfrak{M}(\mathbf{X})}(S) \\
&= e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})} \int_S p d\mathbb{P}_{\mathfrak{M}(\mathbf{Y})} \\
&\geq e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})} \left(\inf_S p\right) \mathbb{P}_{\mathfrak{M}(\mathbf{Y})}(S) \,.
\end{aligned}$$

So, for each measurable set S,

$$\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}(S) > 0 \implies \inf_S p \leq e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon} \,.$$

Furthermore, p is measurable for the Borel σ-algebra of $\mathbb{R}$. In particular, for any $n \in \mathbb{N}_*$, $p^{-1}\left([e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon} + \frac{1}{n}, +\infty)\right)$ is measurable. As a consequence,

$$\forall n \in \mathbb{N}_*, \quad \mathbb{P}_{\mathfrak{M}(\mathbf{Y})}\left(p^{-1}\left(\left[e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon} + \tfrac{1}{n}, +\infty\right)\right)\right) = 0$$

and then

$$\begin{aligned}
\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}\left(p^{-1}\left(\left(e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon}, +\infty\right)\right)\right) &= \mathbb{P}_{\mathfrak{M}(\mathbf{Y})}\left(p^{-1}\left(\cup_{n\in\mathbb{N}_*}\left[e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon} + \tfrac{1}{n}, +\infty\right)\right)\right) \\
&= \mathbb{P}_{\mathfrak{M}(\mathbf{Y})}\left(\cup_{n\in\mathbb{N}_*} p^{-1}\left(\left[e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon} + \tfrac{1}{n}, +\infty\right)\right)\right) \\
&\leq \sum_{n\in\mathbb{N}_*} \mathbb{P}_{\mathfrak{M}(\mathbf{Y})}\left(p^{-1}\left(\left[e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon} + \tfrac{1}{n}, +\infty\right)\right)\right) \\
&= 0
\end{aligned}$$

which proves that $\frac{d\mathbb{P}_{\mathfrak{M}(\mathbf{X})}}{d\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}} \leq e^{d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\epsilon}$, $\mathbb{P}_{\mathfrak{M}(\mathbf{Y})}$-almost surely, which is also the case $\mathbb{P}_{\mathfrak{M}(\mathbf{X})}$-almost surely, thanks to the first remark of the proof. The result about the KL divergence is a direct consequence of this inequality. □

In particular, this result allows us to apply Fano's lemma in order to obtain a similarity function that is based on anchoring the conditional distributions rather than the marginals. I.e., given, $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathcal{X}^n$, $\mathbb{P}_{\mathfrak{M}(\mathbf{X}_1)}, \ldots, \mathbb{P}_{\mathfrak{M}(\mathbf{X}_N)}$ are anchored to $\frac{1}{N}\sum_{j=1}^N \mathbb{P}_{\mathfrak{M}(\mathbf{X}_j)}$.

Lemma 3.3.10 ($(\epsilon, 0)$-DP Fano Matching)**.** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathcal{X}^n$ *and* $\Psi : \operatorname{codom}(\mathfrak{M}) \to \{1, \ldots, N\}$,

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right) \neq i\right) \geq 1 - \frac{1 + \frac{\epsilon}{N^2}\sum_{i=1}^N\sum_{j=1}^N d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)}{\ln(N)} \,.$$

Proof. By Fano's lemma (see Fact 3.1.2),

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right)\neq i\right)\geq 1-\frac{1+\frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X}_i)}\middle\|\,\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}_{\mathfrak{M}(\mathbf{X}_j)}\right)}{\ln(N)}\,.$$

By convexity of the KL divergence with respect to its second argument (see [van Erven & Harremoës, 2014, Theorem 12]), it follows that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right)\neq i\right)\geq 1-\frac{1+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{KL}\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X}_i)}\middle\|\,\mathbb{P}_{\mathfrak{M}(\mathbf{X}_j)}\right)}{\ln(N)}\,. \tag{3.16}$$

An application of Lemma 3.3.9 concludes the proof. □

The bound of Lemma 3.3.9 on the KL divergence between the output distributions works well because the product $\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})$ is typically high in the chosen applications. When it is small, better control on the KL divergence is possible. For instance, [Dwork et al., 2010] proves the bound

$$\mathrm{KL}\left(\mathfrak{M}(\mathbf{X})\|\,\mathfrak{M}(\mathbf{Y})\right)\leq\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\left(e^{\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}-1\right)\,,$$

which was later improved in [Dwork & Rothblum, 2016] to

$$\mathrm{KL}\left(\mathfrak{M}(\mathbf{X})\|\,\mathfrak{M}(\mathbf{Y})\right)\leq\frac{1}{2}\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\left(e^{\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}-1\right)\,.$$

Those two bounds are problematic when the product $\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})$ is too high. This was later improved in [Bun & Steinke, 2016] to

$$\mathrm{KL}\left(\mathfrak{M}(\mathbf{X})\|\,\mathfrak{M}(\mathbf{Y})\right)\leq\frac{1}{2}\left(\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\right)^2\,,$$

but it is still worse than the version that we use for high values of $\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})$. The best of both worlds is achieved in [He et al., 2021] with

$$\mathrm{KL}\left(\mathfrak{M}(\mathbf{X})\|\,\mathfrak{M}(\mathbf{Y})\right)\leq\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})\,\frac{e^{\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}-1}{e^{\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}+1}\,.$$

Again, when the value $\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})$ is typically high, there is no need to come to this degree of precision. However, in some settings, for instance when ϵ is very small or when the distributions to test are very close, this last expression can lead to better results than the one that we used.

3.3.2 The case of ρ-zero concentrated differential privacy

For ρ-zero concentrated differential privacy, the fact that the definition uses information theoretic quantities makes things easier than with the traditional definition of privacy. In particular, the anchoring technique happens implicitly on the distributions rather than on the marginals (similarly as with the $(\epsilon,0)$-DP case). Again, the notion of *group privacy* is central in our proofs.

Fact 3.3.11 (ρ-zCDP Group Privacy [Bun & Steinke, 2016, Proposition 27])**.** *Let $\rho \in \mathbb{R}_{+*}$, if a randomized mechanism $\mathfrak{M} : \mathcal{X}^n \to \operatorname{codom}(\mathfrak{M})$ is ρ-zero concentrated differentially private, then, for any $\mathbf{X}, \mathbf{Y} \in \mathcal{X}^n$ and for any $\alpha \in (1, \infty)$,*

$$D_\alpha \left(\mathfrak{M}(\mathbf{X}) \| \, \mathfrak{M}(\mathbf{Y}) \right) \leq \rho d_{\mathrm{ham}} \left(\mathbf{X}, \mathbf{Y} \right)^2 \alpha \, .$$

Lemma 3.3.12 (ρ-zCDP Le Cam Matching)**.** *Consider a ρ-zCDP mechanism $\mathfrak{M}$, a test function $\Psi : \operatorname{codom}(\mathfrak{M}) \to \{1, 2\}$, and two datasets $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}^n$. We have*

$$\frac{1}{2} \sum_{i=1}^{2} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) \geq \frac{1}{2} \left(1 - \sqrt{\rho/2} d_{\mathrm{ham}} \left(\mathbf{X}_1, \mathbf{X}_2 \right) \right) \, .$$

Proof. By the Neyman-Pearson lemma (see Fact 3.1.1),

$$\frac{1}{2} \sum_{i=1}^{2} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) \geq \frac{1}{2} \left(1 - \mathrm{TV} \left(\mathfrak{M}(\mathbf{X}_1), \mathfrak{M}(\mathbf{X}_2) \right) \right) \, .$$

By Pinsker's lemma (see [Tsybakov, 2009, Lemma 2.5]), $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\mathrm{KL}(\mathbb{P} \| \, \mathbb{Q}) / 2}$, and hence

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^{2} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) &\geq \frac{1}{2} \left(1 - \sqrt{\mathrm{KL} \left(\mathfrak{M}(\mathbf{X}_1) \| \, \mathfrak{M}(\mathbf{X}_2) \right) / 2} \right) \\ &= \frac{1}{2} \left(1 - \sqrt{\mathrm{D}_1 \left(\mathfrak{M}(\mathbf{X}_1) \| \, \mathfrak{M}(\mathbf{X}_2) \right) / 2} \right) \, . \end{aligned}$$

Since the Renyi divergence between a given pair of distributions $\mathrm{D}_\alpha(. \| \, .)$ is non-decreasing in α (see [van Erven & Harremoës, 2014, Theorem 3]), we obtain for any $\alpha \in (1, +\infty)$, s

$$\frac{1}{2} \sum_{i=1}^{2} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) \geq \frac{1}{2} \left(1 - \sqrt{\mathrm{D}_\alpha \left(\mathfrak{M}(\mathbf{X}_1) \| \, \mathfrak{M}(\mathbf{X}_2) \right) / 2} \right) \, .$$

Eventually, we obtain using group privacy (see Fact 3.3.11) that

$$\frac{1}{2} \sum_{i=1}^{2} \mathbb{P}_{\mathfrak{M}} \left(\Psi \left(\mathfrak{M} \left(\mathbf{X}_i \right) \right) \neq i \right) \geq \frac{1}{2} \left(1 - \sqrt{\rho \alpha / 2} d_{\mathrm{ham}} \left(\mathbf{X_1}, \mathbf{X_2} \right) \right) \, .$$

The supremum of the right hand side over $\alpha \in (1, +\infty)$ yields the result. □

We also obtain a zero concentrated DP version of the Fano matching method that we introduced previously for $(\epsilon, 0)$-DP.

Lemma 3.3.13 (ρ-zCDP Fano Matching)**.** *Consider a ρ-zCDP mechanism $\mathfrak{M}$, a test function $\Psi := \mathrm{codom}(\mathfrak{M}) \to \{1, \dots, N\}$, and datasets $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathcal{X}^n$. We have*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right) \neq i\right) \geq 1 - \frac{1 + \frac{\rho}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)^2}{\ln(N)} \,.$$

Proof. By the inequality (3.16) established in the proof of Lemma 3.3.10, and using again the fact that $\mathrm{D}_\alpha\left(.\middle\|.\right)$ is non-decreasing in α (see [van Erven & Harremoës, 2014, Theorem 3]), as well as the group privacy property (see Fact 3.3.11), we obtain that for any $\alpha \in (1, +\infty)$,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}_i\right)\right) \neq i\right) &\geq 1 - \frac{1 + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{KL}\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X}_i)}\middle\|\,\mathbb{P}_{\mathfrak{M}(\mathbf{X}_j)}\right)}{\ln(N)} \\
&\geq 1 - \frac{1 + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{D}_\alpha\left(\mathbb{P}_{\mathfrak{M}(\mathbf{X}_i)}\middle\|\,\mathbb{P}_{\mathfrak{M}(\mathbf{X}_j)}\right)}{\ln(N)} \,. \\
&\geq 1 - \frac{1 + \frac{\rho\alpha}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)^2}{\ln(N)} \,.
\end{aligned}$$

The supremum of the right-hand side over $\alpha \in (1, +\infty)$ yields the result. □

3.4 Lower-bounds via Couplings

The transport problem (3.12) can be studied either theoretically [Santambrogio, 2016] or numerically [Peyré & Cuturi, 2019] in order to give the best lower bounds that our technique permits. However, identifying an optimal coupling is out of the scope of this section. We here provide coupling constructions that are sufficient to exhibit useful lower bounds.

3.4.1 Near optimal couplings

Most of the similarity functions expressed in Theorem 3.3.2 yield lower bounds that are based on or further lower-bounded by expressions involving the quantities

$$\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{Q}}\left(g\left(d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right)\right)\right)$$

for a coupling $\mathbb{Q} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$ where g is a fixed non-increasing function. Hence, finding reasonably good lower bounds can be achieved by finding a coupling that *minimizes* the expected pairwise Hamming distance between the marginals.

As a proxy, we first aim at maximizing the probabilities of pairwise equalities between all the marginals simultaneously. We then control the Hamming distance by observing that when $\mathbf{X}_i = \mathbf{X}_j$, $d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right) = 0$ and otherwise, $d_{\mathrm{ham}}\left(\mathbf{X}_i, \mathbf{X}_j\right) \leq n$. It is known

[Kallenberg, 1993] that if $\mathbb{Q} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$, the disagreement probabilities (i.e. the probability that two marginal random variables are not equal) between the marginals satisfy

$$\forall i, j, \quad \mathrm{TV}\left(\mathbb{P}_i, \mathbb{P}_j\right) \leq \mathbb{P}_{(\mathbf{X}_1, \dots, \mathbf{X}_N) \sim \mathbb{Q}}\left(\mathbf{X}_i \neq \mathbf{X}_j\right) \ . \tag{3.17}$$

A natural question is whether this lower bound is achievable by a coupling simultaneously for all pairs of marginals. When there are only two marginals (i.e. $N = 2$), a classical construction (see [Kallenberg, 1993]) answers this question positively:

Fact 3.4.1 (Maximal coupling)**.** *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability distributions on $\mathcal{X}^n$ that share the same σ-algebra. There exists a coupling $\pi^\infty(\mathbb{P}_1, \mathbb{P}_2) \in \Pi(\mathbb{P}_1, \mathbb{P}_2)$ (which is a distribution on $(\mathcal{X}^n)^2$), called a* maximal *coupling, such that*

$$\begin{aligned} & \mathbb{P}_{(X_1, X_2) \sim \pi^\infty(\mathbb{P}_1, \mathbb{P}_2)}(X_1 \neq X_2) = \mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right) \ , \\ \forall S \text{ measurable} , \quad & \mathbb{P}_{(X_1, X_2) \sim \pi^\infty(\mathbb{P}_1, \mathbb{P}_2)}(X_1 \in S) = \mathbb{P}_1(X_1 \in S) \ , \\ \forall S \text{ measurable} , \quad & \mathbb{P}_{(X_1, X_2) \sim \pi^\infty(\mathbb{P}_1, \mathbb{P}_2)}(X_2 \in S) = \mathbb{P}_2(X_2 \in S) \ . \end{aligned}$$

This construction unfortunately does not generically scale to more than two marginals, even though on simple examples, couplings can be built that still match the lower bound (3.17) for any pair of marginals.

Example 3.4.2 (Bernoulli optimal coupling)**.** Given $\mathbb{P}_i = \mathcal{B}(p_i)$, $1 \leq i \leq N$ a family of Bernoulli distributions and $U \sim \mathcal{U}([0,1])$ a uniformly distributed variable on $[0,1]$, the random vector $(X_1, \dots, X_N)$ defined by $X_i := \mathbb{1}_{[0,p_i)}(U)$ is distributed according to a coupling $Q \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$, and for every i, j

$$\mathbb{P}(X_i \neq X_j) = |p_i - p_j| = \mathrm{TV}\left(\mathcal{B}(p_i), \mathcal{B}(p_j)\right) \ .$$

There are however examples for which it is provably impossible to build couplings that match the lower bound (3.17) for any pair of marginals.

Example 3.4.3 (A counterexample)**.** Let $X_1 \sim \mathcal{U}(\{-1, 0\})$, $X_2 \sim \mathcal{U}(\{0, 1\})$ and $X_3 \sim \mathcal{U}(\{1, -1\})$, and let $\mathbb{P}$ be a coupling between X_1, X_2 and X_3. We have that,

$$\mathbb{1}_{X_1 \neq X_2} + \mathbb{1}_{X_2 \neq X_3} + \mathbb{1}_{X_3 \neq X_1} \geq 2$$

and as a consequence on $\mathbb{P}$,

$$\begin{aligned} \mathbb{P}(X_1 \neq X_2) + \mathbb{P}(X_2 \neq X_3) + \mathbb{P}(X_3 \neq X_1) &\geq 2 \\ &> \mathrm{TV}\left(X_1, X_2\right) + \mathrm{TV}\left(X_2, X_3\right) + \mathrm{TV}\left(X_3, X_1\right) \ , \end{aligned}$$

which proves that at least one of the disagreement probabilities is strictly bigger than the corresponding total variation.

Recent constructions based on Poisson point processes allow in general, for any number of marginals N, to match the lower bound (3.17) up to a factor 2.

Fact 3.4.4 (Near-optimal coupling of multiple distributions [Angel & Spinka, 2021])**.** *Let $\mathbb{P}_1, \dots, \mathbb{P}_N$ be N distributions on the same measurable set. There exists a coupling $\mathbb{Q} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$ such that*

$$\forall i, j \in \{1, \dots, N\}, \quad \mathbb{P}_{(X_1,\dots,X_N)\sim\mathbb{Q}}(X_i \neq X_j) \leq \frac{2\mathrm{TV}(\mathbb{P}_i, \mathbb{P}_j)}{1 + \mathrm{TV}(\mathbb{P}_i, \mathbb{P}_j)} .$$

In the rest of this section, the notation $\pi^\infty(\mathbb{P}_1, \dots, \mathbb{P}_N)$ refers to a coupling that satisfies this condition. When there are only two distributions, it refers to the construction of Fact 3.4.1. This factor 2 is not a problem for minimax theory, since it is a common practice to overlook the constants by looking at rates of convergence. However, for some more precise applications, working on more specific couplings may improve our results. With either coupling constructions, the lower bounds of Theorem 3.3.2 can be controlled with the following straighforward lemma:

Lemma 3.4.5. *Let $\mathbb{P}_1, \dots, \mathbb{P}_N$ be N distributions on $\mathcal{X}^n$ and $\mathbb{Q} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$. Consider $1 \leq i, j \leq N$ and denote $\Delta_{i,j} := \mathbb{P}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{Q}}(\mathbf{X}_i \neq \mathbf{X}_j)$. We have*

$$\begin{aligned}
\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{Q}}\left(d_{\mathrm{ham}}(\mathbf{X}_i, \mathbf{X}_j)\right) &\leq n\Delta_{i,j} \\
\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{Q}}\left(d_{\mathrm{ham}}(\mathbf{X}_i, \mathbf{X}_j)^2\right) &\leq n^2\Delta_{i,j} \\
\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{Q}}\left(e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_i,\mathbf{X}_j)}\right) &\geq 1 - (1 - e^{-n\epsilon})\Delta_{i,j} .
\end{aligned}$$

Note that $\Delta_{i,j}$ directly depends on the coupling construction, but that with any of the ones presented above, we always have $\forall i, j,\ \Delta_{i,j} \leq 2\mathrm{TV}(\mathbb{P}_i, \mathbb{P}_j)$.

When the distributions that we are trying to couple are product distributions (i.e. $\mathbb{P}_1 = \mathbb{p}_1^{\otimes n}, \dots, \mathbb{P}_N = \mathbb{p}_N^{\otimes n}$), we can notice that any coupling $\mathbb{q} \in \Pi(\mathbb{p}_1, \dots, \mathbb{p}_N)$ induces a coupling $\mathbb{q}^{\otimes n} \in \Pi(\mathbb{P}_1, \dots, \mathbb{P}_N)$. Under this coupling, the Hamming distances between the pairs of marginals follow binomial distributions. For the rest of this section, we define the product (near) optimal coupling

$$\pi^\otimes(\mathbb{p}_1^{\otimes n}, \dots, \mathbb{p}_N^{\otimes n}) := \pi^\infty(\mathbb{p}_1, \dots, \mathbb{p}_N)^{\otimes n} .$$

Straightforward computations yield the following lemma.

Lemma 3.4.6. *Let* $\mathbb{P}_1 = \mathbb{p}_1^{\otimes n}, \dots, \mathbb{P}_N = \mathbb{p}_N^{\otimes n}$ *be* N product *distributions on* $\mathcal{X}^n$ *and* $\mathbb{q} \in \Pi(\mathbb{p}_1, \dots,$
$\mathbb{p}_N)$*. Consider any* $1 \leq i, j \leq N$ *and denote*[4] $\delta_{i,j} := \mathbb{P}_{(X_1,\dots,X_N)\sim\mathbb{q}}(X_i \neq X_j)$*. We have:*

$$\begin{aligned}
\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{q}^{\otimes n}}\left(d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)\right) &= n\delta_{i,j} \\
\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{q}^{\otimes n}}\left(d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)^2\right) &= n^2\delta_{i,j}^2 + n\delta_{i,j}(1-\delta_{i,j}) \leq n^2\delta_{i,j}^2 + n\delta_{i,j} \\
\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\mathbb{q}^{\otimes n}}\left(e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_i,\mathbf{X}_j)}\right) &= \left(1-(1-e^{-\epsilon})\delta_{i,j}\right)^n \geq e^{-n\epsilon\delta_{i,j}} \ .
\end{aligned}$$

Note that $\delta_{i,j}$ directly depends on the coupling construction, but that with any of the ones presented above (applied to $\mathbb{p}_1, \dots, \mathbb{p}_N$), we always have $\forall i, j, \quad \delta_{i,j} \leq 2\mathrm{TV}\left(\mathbb{p}_i, \mathbb{p}_j\right)$.

Each of the coupling constructions presented above has its own merits. They will all prove to be useful in the sequel.

3.4.2 Quantitative lower bounds

In this subsection, we finally put the pieces together in order to obtain quantitative lower bounds on (3.11). This subsection serves as a joint proof for Theorem 3.2.1, Theorem 3.2.2, Theorem 3.2.3 and Theorem 3.2.4.

Immediate results on the private minimax risk. A usual estimator (i.e. a measurable function of the samples) $\hat{\theta}$ may be viewed as randomized and almost surely constant to $\hat{\theta}$ (i.e. $\forall \mathbf{X}, \mathfrak{M}(\mathbf{X}) := \hat{\theta}(\mathbf{X})$ almost surely). As a result, it is clear that the private minimax risk is always bigger than the non-private one. For distributional tests, the result is not so obvious, and we give the following general purpose lemma that ensures that Fano's and Le Cam's regular inequalities still hold.

Lemma 3.4.7. *Let* $(\mathbb{P}_i)_{i\in\{1,\dots,N\}}$ *be a family of probability distributions on* $\mathcal{X}^n$ *and let* $\mathfrak{M} : \mathcal{X}^n \to \mathrm{codom}\left(\mathfrak{M}\right)$ *be a randomized mechanism,*

$$\inf_{\Psi:\mathrm{codom}(\mathfrak{M})\to\{1,\dots,N\}} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \geq \inf_{\Psi:\mathcal{X}^N\to\{1,\dots,N\}} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i}\left(\Psi\left(\mathbf{X}\right) \neq i\right) \ .$$

In particular, the inequalities in Le Cam's lemma (see Fact 3.1.1) or Fano's lemma (see Fact 3.1.2) still hold when the test function Ψ *is fed with an input* $\mathfrak{M}(\mathbf{X}) \in \mathrm{codom}\left(\mathfrak{M}\right)$ *instead of an input* $\mathbf{X} \in \mathcal{X}^n$*.*

[4] not to be confused with the Kronecker symbol.

Proof. Let $\Psi : \mathrm{codom}\left(\mathfrak{M}\right) \to \{1, \dots, N\}$ be a test function. Then,

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) &= \sum_{i=1}^{N} \int \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) d\mathbb{P}_{\mathfrak{M}} \\
&= \int \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i}\left(\left(\Psi \circ \mathfrak{M}\right)(\mathbf{X}) \neq i\right) d\mathbb{P}_{\mathfrak{M}} \\
&\geq \int \inf_{\Psi' : \mathcal{X}^N \to \{1,\dots,N\}} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i}\left(\Psi'\left(\mathbf{X}\right) \neq i\right) d\mathbb{P}_{\mathfrak{M}} \\
&= \inf_{\Psi' : \mathcal{X}^N \to \{1,\dots,N\}} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i}\left(\Psi'\left(\mathbf{X}\right) \neq i\right) \ .
\end{aligned}$$

Taking the infimum over $\Psi : \mathrm{codom}\left(\mathfrak{M}\right) \to \{1, \dots, N\}$ concludes the proof. □

The case of two hypotheses ($N = 2$). At first, we look at the implications of couplings between pairs of distributions. Given $\mathbb{P}_1$ and $\mathbb{P}_2$ distributions on $\mathcal{X}^n$, a direct implication of Lemma 3.4.7 and of Le Cam's lemma (see Fact 3.1.1) is that independently on the privacy condition $\mathcal{C}$ imposed on $\mathfrak{M}$,

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right) \neq i\right) \geq \frac{1}{2}\left(1 - \mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right)\right) \ .$$

This is the first ingredient in the proof of Theorem 3.2.1 and Theorem 3.2.2 that we now detail.

Proof of Theorem 3.2.1. When $\mathfrak{M}$ is (ϵ, δ)-DP, the generic bound of Theorem 3.3.2 applied with the Le Cam matching technique described in Theorem 3.3.3 and the coupling $\pi^{\infty}(\mathbb{P}_1, \mathbb{P}_2)$ leads to

$$\begin{aligned}
\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right) \neq i\right) &\geq \frac{1}{2}\mathbb{E}_{(\mathbf{X}_1,\mathbf{X}_2)\sim\pi^{\infty}(\mathbb{P}_1,\mathbb{P}_2)}\left(e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_1,\mathbf{X}_2)}\right) \\
&\quad - e^{-\epsilon}\delta\mathbb{E}_{(\mathbf{X}_1,\mathbf{X}_2)\sim\pi^{\infty}(\mathbb{P}_1,\mathbb{P}_2)}\left(d_{\mathrm{ham}}\left(\mathbf{X}_1, \mathbf{X}_2\right)\right) \\
&\overset{\text{Lemma 3.4.5}}{\geq} \frac{1}{2}\left(1 - (1 - e^{-n\epsilon})\Delta_{1,2}\right) - e^{-\epsilon} n \Delta_{1,2} \\
&= \frac{1}{2}\left(1 - \left(1 - e^{-n\epsilon} + 2ne^{-\epsilon}\delta\right)\mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right)\right) \ .
\end{aligned}$$

where in the second line we denote $\Delta_{1,2} := \mathbb{P}_{(\mathbf{X}_1,\mathbf{X}_2)\sim\pi^{\infty}(\mathbb{P}_1,\mathbb{P}_2)}\left(\mathbf{X}_1 \neq \mathbf{X}_2\right)$ and in the last line we use that $\Delta_{1,2} = \mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right)$ with the chosen coupling. Similarly, in the case of product distributions, with the same matching but $\pi^{\otimes}(\mathbb{p}_1^{\otimes n}, \mathbb{p}_2^{\otimes n})$ we obtain as a consequence of Lemma 3.4.6

$$\begin{aligned}
&\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right) \neq i\right) \\
&\qquad \geq \frac{1}{2}\left(\left(1 - \left(1 - e^{-\epsilon}\right)\mathrm{TV}\left(\mathbb{p}_1, \mathbb{p}_2\right)\right)^n - 2ne^{-\epsilon}\delta\mathrm{TV}\left(\mathbb{p}_1, \mathbb{p}_2\right)\right) \ .
\end{aligned}$$

□

Proof of Theorem 3.2.2. When $\mathfrak{M}$ is ρ-DP, the generic bound of Theorem 3.3.2 applied with the Le Cam matching technique described in Theorem 3.3.4 and the coupling $\pi^{\infty}(\mathbb{P}_1, \mathbb{P}_2)$ leads to

$$\begin{aligned}\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right)\neq i\right) &\geq \frac{1}{2}\left(1-\sqrt{\rho/2}\mathbb{E}_{(\mathbf{X}_1,\mathbf{X}_2)\sim\pi^{\infty}(\mathbb{P}_1,\mathbb{P}_2)}\left(d_{\mathrm{ham}}\left(\mathbf{X}_1,\mathbf{X}_2\right)\right)\right)\\ &\overset{\text{Lemma 3.4.5}}{\geq} \frac{1}{2}\left(1-\sqrt{\rho/2}\delta n\Delta_{1,2}\right)\\ &= \frac{1}{2}\left(1-n\sqrt{\rho/2}\mathrm{TV}\left(\mathbb{P}_1,\mathbb{P}_2\right)\right)\ .\end{aligned}$$

where in the second line we denote $\Delta_{1,2} := \mathbb{P}_{(\mathbf{X}_1,\mathbf{X}_2)\sim\pi^{\infty}(\mathbb{P}_1,\mathbb{P}_2)}\left(\mathbf{X}_1\neq\mathbf{X}_2\right)$ and in the last line we use that $\Delta_{1,2} = \mathrm{TV}\left(\mathbb{P}_1,\mathbb{P}_2\right)$ with the chosen coupling. Similarly, in the case of product distributions, with the same matching but $\pi^{\otimes}(\mathbb{p}_1^{\otimes n},\mathbb{p}_2^{\otimes n})$ we obtain as a consequence of Lemma 3.4.6

$$\max_{i\in\{1,2\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}\left(\mathbf{X}\right)\right)\neq i\right) \geq \frac{1}{2}\left(1-n\sqrt{\rho/2}\mathrm{TV}\left(\mathbb{p}_1,\mathbb{p}_2\right)\right)\ .$$

□

The case of arbitrary many hypotheses ($N \geq 2$). Given $\mathbb{P}_1,\ldots,\mathbb{P}_N$ distributions on $\mathcal{X}^n$, a direct implication of Lemma 3.4.7 and of Fano's lemma (see Fact 3.1.2) is that independently on the privacy condition $\mathcal{C}$ imposed on $\mathfrak{M}$, for any $\mathbb{Q}$ such that $\mathbb{P}_i \ll \mathbb{Q}$ for all i,

$$\max_{i\in\{1,\ldots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right) \geq 1-\frac{1+\frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\left(\mathbb{P}_i\|\,\mathbb{Q}\right)}{\ln(N)}\ .$$

Again, this serves as the first ingredient of the proof of Theorem 3.2.3 and Theorem 3.2.4 that we now detail.

Proof of Theorem 3.2.3. When $\mathfrak{M}$ is (ϵ,δ)-DP, the generic bound of Theorem 3.3.2 applied with the pairwise anchoring technique described in Theorem 3.3.3 and the coupling

$\pi^\infty(\mathbb{P}_1,\dots,\mathbb{P}_N)$ leads to (since $\lceil n/2\rceil \le n$ for each integer $n \ge 0$)

$$\begin{aligned}
&\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right)\\
&\qquad\ge \frac{1}{2N^2}\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^\infty(\mathbb{P}_1,\dots,\mathbb{P}_N)}\Big(\sum_{i=1}^N\sum_{j=1}^N e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_i,\mathbf{X}_j)}\\
&\qquad\qquad - 2e^{-\epsilon}\delta d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)\Big)\\
&\qquad\overset{\text{Lemma 3.4.5}}{\ge} \frac{1}{2N^2}\Big(\sum_{i=1}^N\sum_{j=1}^N\left(1-\left(1-e^{-n\epsilon}\right)\Delta_{i,j}\right)-2e^{-\epsilon}\delta n\Delta_{i,j}\Big)\\
&\qquad\ge \frac{1}{2}-\frac{1-e^{-n\epsilon}+2ne^{-\epsilon}\delta}{2N^2}\sum_{i,j}\frac{2\mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right)}{1+\mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right)}
\end{aligned}$$

where in the second line we denote $\Delta_{i,j} := \mathbb{P}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^\infty(\mathbb{P}_1,\dots,\mathbb{P}_N)}\left(\mathbf{X}_i\neq\mathbf{X}_j\right)$ and in the last line we use that $\Delta_{i,j}\le\frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}$ with the chosen coupling. Similarly, in the case of product distributions, with the same matching but the product coupling $\pi^\otimes(\mathbb{p}_1^{\otimes n},\dots,\mathbb{p}_N^{\otimes n})$ we obtain as a consequence of Lemma 3.4.6

$$\begin{aligned}
&\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right)\\
&\qquad\ge \frac{1}{2N^2}\sum_{i,j}\Bigg(\Bigg(1-(1-e^{-\epsilon})\frac{2\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}{1+\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}\Bigg)^n\\
&\qquad\qquad -2ne^{-\epsilon}\delta\frac{2\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}{1+\mathrm{TV}\left(\mathbb{p}_i,\mathbb{p}_j\right)}\Bigg)\ .
\end{aligned}$$

When $\delta = 0$, the generic bound of Theorem 3.3.2 applied with the Fano matching technique described in Theorem 3.3.3 and the coupling $\pi^\infty(\mathbb{P}_1,\dots,\mathbb{P}_N)$ leads to

$$\begin{aligned}
&\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right)\\
&\qquad\ge 1-\frac{1+\frac{\epsilon}{N^2}\sum_{i=1}^N\sum_{j=1}^N\mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^\infty(\mathbb{P}_1,\dots,\mathbb{P}_N)}\left(d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)\right)}{\ln N}\\
&\qquad\overset{\text{Lemma 3.4.5}}{\ge} 1-\frac{1+\frac{\epsilon}{N^2}\sum_{i=1}^N\sum_{j=1}^N n\Delta_{i,j}}{\ln N}\\
&\qquad\ge 1-\frac{1+\frac{n\epsilon}{N^2}\sum_{i=1}^N\sum_{j=1}^N\frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}}{\ln N}
\end{aligned}$$

where in the second line we denote $\Delta_{i,j} := \mathbb{P}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^\infty(\mathbb{P}_1,\dots,\mathbb{P}_N)}\left(\mathbf{X}_i\neq\mathbf{X}_j\right)$ and in the last line we use that $\Delta_{i,j}\le\frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}$ with the chosen coupling. Similarly, in the case of product distributions, with the same matching but the coupling $\pi^\otimes(\mathbb{p}_1^{\otimes n},\dots,\mathbb{p}_N^{\otimes n})$ we obtain as a consequence of Lemma 3.4.6

$$\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right)\neq i\right)\ge 1-\frac{1+\frac{n\epsilon}{N^2}\sum_{i=1}^N\sum_{j=1}^N\frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}}{\ln N}\ .$$

□

Proof of Theorem 3.2.4. When $\mathfrak{M}$ is ρ-zCDP, the generic bound of Theorem 3.3.2 applied with the Fano matching technique described in Theorem 3.3.4 and the coupling $\pi^{\infty}(\mathbb{P}_1, \dots, \mathbb{P}_N)$ leads to

$$\begin{aligned}
&\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \\
&\qquad \geq 1 - \frac{1 + \frac{\rho}{N^2}\sum_{i=1}^N \sum_{j=1}^N \mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^{\infty}(\mathbb{P}_1,\dots,\mathbb{P}_N)}\left(d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)^2\right)}{\ln N} \\
&\qquad \overset{\text{Lemma 3.4.5}}{\geq} 1 - \frac{1 + \frac{\rho}{N^2}\sum_{i=1}^N \sum_{j=1}^N n^2\Delta_{i,j}}{\ln N} \\
&\qquad \geq 1 - \frac{1 + \frac{n^2\rho}{N^2}\sum_{i=1}^N \sum_{j=1}^N \frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}}{\ln N}
\end{aligned}$$

where in the second line we denote $\Delta_{i,j} := \mathbb{P}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^{\infty}(\mathbb{P}_1,\dots,\mathbb{P}_N)}\left(\mathbf{X}_i \neq \mathbf{X}_j\right)$ and in the last line we use that $\Delta_{i,j} \leq \frac{2\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}{1+\mathrm{TV}(\mathbb{P}_i,\mathbb{P}_j)}$ with the chosen coupling. Similarly, in the case of product distributions, with the same matching but with the product coupling $\pi^{\otimes}(\mathbb{p}_1^{\otimes n}, \dots, \mathbb{p}_N^{\otimes n})$ we obtain,

$$\begin{aligned}
&\max_{i\in\{1,\dots,N\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_i,\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq i\right) \\
&\qquad \geq 1 - \frac{1 + \frac{\rho}{N^2}\sum_{i=1}^N \sum_{j=1}^N \mathbb{E}_{(\mathbf{X}_1,\dots,\mathbf{X}_N)\sim\pi^{\otimes}(\mathbb{p}_1^{\otimes n},\dots,\mathbb{p}_N^{\otimes n})}\left(d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)^2\right)}{\ln N} \\
&\qquad \overset{\text{Lemma 3.4.6}}{\geq} 1 - \frac{1 + \frac{\rho}{N^2}\sum_{i=1}^N \sum_{j=1}^N \left(n^2\delta_{i,j}^2 + n\delta_{i,j}\right)}{\ln N} \\
&\qquad \geq 1 - \frac{1 + \frac{n^2\rho}{N^2}\sum_{i=1}^N \sum_{j=1}^N \left(\left(\frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}\right)^2 + \frac{1}{n}\frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}\right)}{\ln N}
\end{aligned}$$

where in the second line we denote $\delta_{i,j} := \mathbb{P}_{(X_1,\dots,X_N)\sim\pi^{\infty}(\mathbb{p}_1,\dots,\mathbb{p}_N)}\left(X_i \neq X_j\right)$ and in the last line we use that $\delta_{i,j} \leq \frac{2\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}{1+\mathrm{TV}(\mathbb{p}_i,\mathbb{p}_j)}$ with the chosen coupling.

□

3.5 A note on Assouad's method

As the reduction to a testing problem between multiple hypotheses, Assouad's lemma relies on similar ideas, where the packing has to be parametrized by a hypercube. Its advantage over tools like Fano's lemma is that it only makes tests between pairs of hypotheses (instead of all of them at the same time). The cost of this is that the control of the packing is slightly more difficult.

Suppose that the set of distributions of interest $\mathcal{P}$ contains a family of distributions $(\mathbb{P}_\omega)_{\omega\in\{0,1\}^m}$ for a certain positive integer m. If the loss function (taken quadratic for

simplicity) can be decomposed as (where in $\|\mathbb{P}_\omega - \mathbb{P}_{\omega'}\|$, the difference between distributions should be interpreted as the difference of the features that we are trying to estimate corresponding to those distributions)

$$\forall \omega, \omega' \in \{0,1\}^m, \quad \|\mathbb{P}_\omega - \mathbb{P}_{\omega'}\|^2 \geq 2\tau \sum_{i=1}^{m} \mathbb{1}_{\omega \neq \omega} \ , \tag{3.18}$$

then the minimax risk can be lower-bounded as

$$\begin{aligned} &\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}, \hat{\pi}}(\|\hat{\pi}(\mathbf{X}) - \pi\|^2) \\ &\qquad \geq \frac{\tau}{16} \sum_{i=1}^{m} \inf_{\substack{\mathfrak{M} \text{ s.t. } \mathcal{C} \\ \Psi : \text{codom}(\mathfrak{M}) \to \{0,1\}}} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\omega^{i,0}}, \mathfrak{M}} \left(\Psi \left(\mathfrak{M}(\mathbf{X}) \right) \neq 0 \right) + \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\omega^{i,1}}, \mathfrak{M}} \left(\Psi \left(\mathfrak{M}(\mathbf{X}) \right) \neq 1 \right) \ . \end{aligned} \tag{3.19}$$

where $\mathbb{P}_{\omega^{i,0}}$ and $\mathbb{P}_{\omega^{i,1}}$ are the *mixture* distributions

$$\mathbb{P}_{\omega^{i,0}} := \frac{1}{2^{m-1}} \sum_{\omega \in \{0,1\}^m | \omega_i = 0} \mathbb{P}_\omega \quad \text{and} \quad \mathbb{P}_{\omega^{i,0}} := \frac{1}{2^{m-1}} \sum_{\omega \in \{0,1\}^m | \omega_i = 1} \mathbb{P}_\omega \ .$$

The proof is classical and can be found in [Acharya et al., 2021e]. The term

$$\mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\omega^{i,0}}, \mathfrak{M}} \left(\Psi \left(\mathfrak{M}(\mathbf{X}) \right) \neq 0 \right) + \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{\omega^{i,1}}, \mathfrak{M}} \left(\Psi \left(\mathfrak{M}(\mathbf{X}) \right) \neq 1 \right)$$

characterizes the *testing* difficulty between $\mathbb{P}_{\omega^{i,0}}$ and $\mathbb{P}_{\omega^{i,0}}$. It can be controlled by Le Cam's lemma, and by its variants when working under privacy (see [Acharya et al., 2021e, Lalanne et al., 2023b] for differential privacy and [Lalanne et al., 2023b] for concentrated differential privacy).

Using this technique usually leads to better lower-bounds in the case of (ϵ, δ)-differential privacy when $\delta \neq 0$ or with concentrated differential privacy.

Chapter 4

Examples of lower-bounds on parametric models

The origin of this chapter, and the use of the first person. This chapter is based on the article [Lalanne et al., 2023b], written by Aurélien Garivier[1], Rémi Gribonval[2], and by myself. In this chapter, I will try to respect the following rule : the use of the first person of the plural (we, our, ...) represents all the above-mentioned people, while the use of the first person of the singular (I, my, ...) represents myself.

This chapter illustrates the results of the last chapter on three simple, fully worked out parametric examples. In particular, it shows that the problem class has a huge importance on the provable degradation of utility due to privacy. In certain scenarios, it shows that maintaining privacy results in a noticeable reduction in performance only when the level of privacy protection is very high. Conversely, for other problems, even a modest level of privacy protection can lead to a significant decrease in performance.

It also demonstrates that the DP-SGLD algorithm, a private convex solver, can be employed for maximum likelihood estimation with a high degree of confidence, as it provides near-optimal results with respect to both the size of the sample and the level of privacy

[1] https://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/index.html/

[2] https://people.irisa.fr/Remi.Gribonval/

protection. This algorithm is applicable to a broad range of parametric estimation procedures, including exponential families.

Finally, it gives bibliographical pointers to many recent research articles studying similar problems of private parametric estimation problems.

4.1 Parametric unidimensional examples

First, let us start with unidimensional examples.

4.1.1 Bernoulli model

The first application is the estimation of the proportion of a population that satisfies a certain property. It is a prime example of the application of Le Cam's lemma Fact 3.1.1 and its private counterparts Theorem 3.2.1 and Theorem 3.2.2. When we consider the parametric Bernoulli model

$$(\mathcal{B}(\theta))_{\theta\in\Theta}\,, \qquad \Theta = (0,1)\,,$$

a classical and simple estimator for estimating the true parameter θ^* from i.i.d. samples $X_1,\dots,X_n$ drawn according to $\mathcal{B}(\theta^*)$ is via the empirical average

$$\hat{\theta} := \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

The quadratic risk of this estimator is

$$\mathbb{E}\left((\theta^*-\hat{\theta})^2\right) = \frac{\theta^*(1-\theta^*)}{n} \le \frac{1/4}{n}\,.$$

In order to find lower bounds on the minimax risk (with or without privacy constraints), let us investigate an $\Omega = \frac{\alpha}{4}$-packing[3] with $\theta_1 := \frac{1+\alpha}{2}$ and $\theta_2 := \frac{1}{2}$.

Regular Minimax Risk. By the master bound (3.6), Le Cam's lemma (Fact 3.1.1) and Pinsker's inequality (see [Tsybakov, 2009, Lemma 2.5]),

$$\begin{aligned}\mathfrak{M}_n &\ge (\alpha/4)^2\cdot\frac{1}{2}\left(1-\mathrm{TV}\left(\mathcal{B}(\theta_1)^{\otimes n},\mathcal{B}(\theta_2)^{\otimes n}\right)\right)\\ &\ge \frac{\alpha^2}{32}\left(1-\sqrt{\mathrm{KL}\left(\mathcal{B}(\theta_1)^{\otimes n}\|\,\mathcal{B}(\theta_2)^{\otimes n}\right)/2}\right)\\ &= \frac{\alpha^2}{32}\left(1-\sqrt{n\mathrm{KL}\left(\mathcal{B}(\theta_1)\|\,\mathcal{B}(\theta_2)\right)/2}\right).\end{aligned}$$

where we used the tensorization property of the KL divergence (see [van Erven & Harremoës, 2014, Theorem 28]). We can observe that when $\alpha\in[0,1/2]$,

$$\mathrm{KL}\left(\mathcal{B}(\theta_1)\|\,\mathcal{B}(\theta_2)\right) \le \alpha^2\,.$$

[3] With $d(\cdot,\cdot) = |\cdot - \cdot|$, see Section 3.1.3: an Ω-packing must satisfy $d(\theta_i,\theta_j)\ge 2\Omega$, $i\neq j$.

Indeed, let us note $g(x) = \frac{1+x}{2}\ln(1+x) + \frac{1-x}{2}\ln(1-x) - x^2$. We have that $\frac{dg(x)}{dx}(x) = \frac{\ln(1+x)+\ln(1-x)}{2} - 2x$ and since $g(0) = 0$ and $x \mapsto \ln(1+x)$ is 2-Lipschitz on $[-1/2, 1/2]$, we have that $g(x) \leq 0, \quad \forall x \in [0, 1/2]$. Hence, when $\alpha \in [0, 1/2]$,

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{B}(\theta_1) \| \mathcal{B}(\theta_2)\right) &= \left(\theta_1 \ln\left(\frac{\theta_1}{\theta_2}\right) + (1-\theta_1)\ln\left(\frac{1-\theta_1}{1-\theta_2}\right)\right) \\
&= \left(\frac{1+\alpha}{2}\ln(1+\alpha) + \frac{1-\alpha}{2}\ln(1-\alpha)\right) \\
&\leq \alpha^2 \,.
\end{aligned}$$

So, with $\alpha = \frac{1}{\sqrt{n}}$, as soon as $n \geq 4$, we obtain that

$$\mathfrak{M}_n \geq \frac{\alpha^2}{32}\left(1 - \sqrt{n\alpha^2/2}\right) = \frac{1/160}{n} = \Omega\left(\frac{1}{n}\right) \,,$$

which concludes that the non-private minimax rate is lower bounded by a quantity of the order of $\frac{1}{n}$ and in particular, that the empirical mean estimator $\hat{\theta}$ is minimax optimal in term of rates of convergence. Furthermore, any private minimax rate also has to be of the order of at least $\frac{1}{n}$.

Minimax Risk with ϵ-Differential Privacy. By the private master lower bound (3.10) and the product form of Le Cam's lemma for $(\epsilon, 0)$-DP (see Theorem 3.2.1) combined with the last inequality in Lemma 3.4.6 we obtain

$$\begin{aligned}
\mathfrak{M}_n^{(\epsilon\text{-DP})} &\geq (\alpha/4)^2 \cdot \frac{1}{2} e^{-n\epsilon \mathrm{TV}(\mathcal{B}(\theta_1), \mathcal{B}(\theta_2))} \\
&\geq \frac{\alpha^2}{32} e^{-n\epsilon\sqrt{\mathrm{KL}(\mathcal{B}(\theta_1)\|\mathcal{B}(\theta_2))/2}} \\
&= \frac{\alpha^2}{32} e^{-\sqrt{(n\epsilon)^2\alpha^2/2}}
\end{aligned}$$

where we used again Pinsker's inequality.

So, with $\alpha = \frac{1}{n\epsilon}$, when $n\epsilon \geq 2$, we obtain that

$$\mathfrak{M}_n^{(\epsilon\text{-DP})} \geq \frac{1/32}{(n\epsilon)^2} e^{-\sqrt{1/2}} \geq \frac{1/80}{(n\epsilon)^2} = \Omega\left(\frac{1}{(n\epsilon)^2}\right) \,.$$

ρ-zero Concentrated Differential Privacy. Similarly, by the product form of Le Cam's lemma for ρ-zCDP (see Theorem 3.2.2), we get with $\alpha = \frac{1}{n\sqrt{\rho}}$ when $n\sqrt{\rho} \geq 2$,

$$\begin{aligned}
\mathfrak{M}_n^{(\rho\text{-zCDP})} &\geq \frac{\alpha^2}{32}\left(1 - n\sqrt{\rho/2}\mathrm{TV}\left(\mathcal{B}(\theta_1), \mathcal{B}(\theta_2)\right)\right) \\
&\geq \frac{\alpha^2}{32}\left(1 - n\sqrt{\rho \mathrm{KL}\left(\mathcal{B}(\theta_1) \| \mathcal{B}(\theta_2)\right)/4}\right) \\
&= \frac{\alpha^2}{32}\left(1 - \sqrt{n^2\rho\alpha^2/4}\right) = \frac{1/64}{n^2\rho} \\
&= \Omega\left(\frac{1}{n^2\rho}\right) .
\end{aligned}$$

Matching Upper Bounds. Consider the Laplace mechanism $\mathfrak{M}(\mathbf{X}) := \frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n\epsilon}\mathrm{Lap}(1)$. It is an $(\epsilon, 0)$-DP estimator $\mathbf{X}$ [Dwork & Roth, 2014] and its quadratic risk is $O\left(\frac{1}{n} + \frac{1}{(n\epsilon)^2}\right)$. Likewise, the Gaussian mechanism $\mathfrak{M}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i + \frac{2}{n\sqrt{\rho}}\mathcal{N}(0,1)$ is ρ-zCDP [Bun & Steinke, 2016] and its one is $O\left(\frac{1}{n} + \frac{1}{n^2\rho}\right)$. Combined with the lower bounds established so far and with Lemma 3.4.7, this allows to conclude that in fact

$$\mathfrak{M}_n^{(\epsilon\text{-DP})} = \Theta\left(\max\left\{\frac{1}{n}, \frac{1}{(n\epsilon)^2}\right\}\right) ,$$

and that this optimal rate is achieved with the Laplace mechanism, while

$$\mathfrak{M}_n^{(\rho\text{-zCDP})} = \Theta\left(\max\left\{\frac{1}{n}, \frac{1}{n^2\rho}\right\}\right) ,$$

which is an optimal rate achieved by the Gaussian mechanism.

The Cost of Privacy. An interesting observation for both definitions of privacy is that there exist regimes ($\epsilon \ll 1/\sqrt{n}$ or $\rho \ll 1/n$) for which the minimax rate of convergence is degraded compared to the non private one. In other words, privacy has an unavoidable cost on utility, no matter the mechanism used. Conversely, the order of magnitude of the minimax risk is not degraded otherwise.

4.1.2 Uniform support model

We consider the parametric model

$$(\mathbb{p}_\theta := \mathcal{U}([0,\theta]))_{\theta\in\Theta} , \qquad \Theta = (0,1] .$$

To exploit Le Cam's lemma we will need to control the total variation between two distributions. In this model, it can be done explicitly. The total variation between $\mathbb{p}_{\theta_1}^{\otimes n}$ and $\mathbb{p}_{\theta_2}^{\otimes n}$ can be computed as

$$\mathrm{TV}\left(\mathbb{p}_{\theta_1}^{\otimes n}, \mathbb{p}_{\theta_2}^{\otimes n}\right) = 1 - \int_{[0,1]^n} \min\left(\pi_{\mathbb{p}_{\theta_1}^{\otimes n}}, \pi_{\mathbb{p}_{\theta_2}^{\otimes n}}\right) = 1 - \left(\frac{\min\left(\theta_1,\theta_2\right)}{\max\left(\theta_1,\theta_2\right)}\right)^n .$$

Non-Private Minimax Risk. By the (non-private) master lower bound (3.6) and Le Cam's lemma (Fact 3.1.1), applied to the $\frac{1}{2n}$-packing $\theta_1 = 1 - \frac{1}{n}$ and $\theta_2 = 1$, we have

$$\mathfrak{M}_n \geq \frac{e^{-1}}{8n^2} = \Omega\left(\frac{1}{n^2}\right) .$$

where we used that $1 - \mathrm{TV}\left(\mathbb{p}_{\theta_1}^{\otimes n}, \mathbb{p}_{\theta_2}^{\otimes n}\right) = \left(1 - \frac{1}{n}\right)^n \geq e^{-1}$. Furthermore, as we now show, the estimator $\max \mathbf{X}$ achieves this rate of convergence when $X_1, \ldots, X_n \sim \mathcal{U}([0, \theta^*])$ are independent. Indeed, for any $t \in [0, \theta^*]$,

$$\mathbb{P}\left(\max \mathbf{X} < t\right) = \Pi_{i=1}^n \mathbb{P}\left(X_i < t\right) = \left(\frac{t}{\theta^*}\right)^n .$$

Hence, $\max \mathbf{X}$ has a density $\pi_{\max \mathbf{X}}$ with respect to the Lebesgue measure where

$$\forall t \in \mathbb{R}, \quad \pi_{\max \mathbf{X}}(t) = \mathbb{1}_{[0,\theta^*]}(t) \frac{n t^{n-1}}{\theta^{*n}} ,$$

so that

$$\begin{aligned}
\mathbb{E}(\max \mathbf{X}) &= \int_0^{\theta^*} t \left(\frac{n t^{n-1}}{\theta^{*n}}\right) dt = \frac{n}{n+1}\theta^* , \\
\mathbb{V}(\max \mathbf{X}) &= \int_0^{\theta^*} t^2 \left(\frac{n t^{n-1}}{\theta^{*n}}\right) dt - \left[\mathbb{E}(\max \mathbf{X})\right]^2 = \theta^{*2} \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) .
\end{aligned}$$

By the bias-variance tradeoff, the quadratic risk of $\max \mathbf{X}$ is thus $O\left(\frac{\theta^{*2}}{n^2}\right)$. In particular, this proves that the non-private minimax rate of convergence is $\Theta\left(\frac{1}{n^2}\right)$ and that $\max \mathbf{X}$ achieves this minimax rate of convergence.

Minimax Risk with ϵ-Differential Privacy. By the private master lower bound (3.10) and the product form of Le Cam's private lemma for ϵ-DP on product distributions (see Theorem 3.2.1 with $\delta = 0$) with the $\frac{1}{2n\epsilon}$-packing $\theta_1 = 1 - \frac{1}{n\epsilon}$ and $\theta_2 = 1$ we have when $n\epsilon > 1$

$$\mathfrak{M}_n^{(\epsilon\text{-DP})} \geq \frac{e^{-1}}{8(n\epsilon)^2} = \Omega\left(\frac{1}{(n\epsilon)^2}\right) ,$$

In particular, the rate is degraded compared to the non-private one as soon as ϵ is decreasing.

Minimax Risk with ρ-zero Concentrated Differential Privacy. Similarly, using the product form of Le Cam's private lemma for ρ-zCDP on product distributions (see Theorem 3.2.2) and the $\frac{1}{2n\sqrt{\rho}}$-packing $\theta_1 = 1 - \frac{1}{n\sqrt{\rho}}$ and $\theta_2 = 1$ gives that when $n\sqrt{\rho} > 1$,

$$\mathfrak{M}_n^{(\rho\text{-zCDP})} \geq \frac{1 - \frac{1}{\sqrt{2}}}{8n^2\rho} = \Omega\left(\frac{1}{n^2\rho}\right) .$$

In particular, the rate is degraded compared to the non-private one as soon as ρ is decreasing.

This example shows that when the stochastic noise due to sampling shrinks too fast (here $\max \mathbf{X}$ has quadratic risk $O(1/n^2)$), then the noise due to privacy becomes predominant. In particular, we do not observe a distinction on the rate at which ϵ or ρ tends to 0 in order the conclude to a degradation of the minimax risk. It is systematically degraded.

4.2 Parametric multidimensional examples

One dimensional models already allow exhibiting degradation that is due to privacy. However, things are more interesting when looking at *multidimensional* examples. Indeed, in this setup, dimensionality amplifies the degradation that is due to privacy.

For instance, with Gaussians where the usual estimation quadratic risk is of the order $\Theta\left(\frac{d}{n}\right)$, we will see in the following that this rate becomes $\Omega\left(\frac{d}{n} + \frac{d^2}{(n\epsilon)^2}\right)$ under ϵ-differential privacy. The privacy overhead degrades quadratically in the dimension, whereas the regular estimation rate only degrades linearly.

4.2.1 Gaussian model

The second application is the estimation of the unknown mean $\theta^* \in \mathbb{R}^d$ of multivariate normally distributed data with fixed covariance matrix $\sigma^2 I_d$. When we consider the parametric model $\left(\mathcal{N}(\theta, \sigma^2 I_d)\right)_{\theta \in \Theta}, \Theta = \mathbb{R}^d$, a classical and simple estimator for estimating the mean θ^* from i.i.d. samples $X_1, \ldots, X_n$ is the empirical average $\hat{\theta} := \frac{1}{n}\sum_{i=1}^n X_i$. The quadratic risk of this estimator is

$$\mathbb{E}\left(\|\theta^* - \hat{\theta}\|^2\right) = \frac{\sigma^2 d}{n} \,. \tag{4.1}$$

If we were to apply Le Cam's lemma Fact 3.1.1 or its private counterparts Theorem 3.2.1 and Theorem 3.2.2, the parameter that tunes the dimensionality d would not be captured by the resulting minimax lower bounds which would thus be overly optimistic. This example forces us to use Fano's lemma Fact 3.1.2 or its private counterparts Theorem 3.2.3 or Theorem 3.2.4 in order to have a chance to capture this phenomenon.

The total variation that appears in Fano's inequality is controlled via Pinsker's inequality in terms of a Kullback-Leibler divergence, which in the case of isotropic Gaussians is known to be proportional to the squared Euclidean distance.

$$\forall \theta_1, \theta_2 \in \Theta, \quad \mathrm{KL}\left(\mathcal{N}(\theta_1, \sigma^2 I_d) \middle\| \mathcal{N}(\theta_2, \sigma^2 I_d)\right) = \frac{\|\theta_2 - \theta_1\|^2}{2\sigma^2} \,. \tag{4.2}$$

This enables the use of packing results for the Euclidean norm, and minimax bounds valid in the more general case where the KL divergence is controlled by the Euclidean norm between parameters.

Packing Choice. In high dimension, the packing is chosen with an exponential number of hypotheses. A good way to obtain well-spread points is to use Varshamov–Gilbert's theorem

Fact 4.2.1 (Varshamov–Gilbert's theorem [Rigollet & Hütter, 2015, Lemma 5.12])**.** *For any* $\zeta \in \left(0, \frac{1}{2}\right)$ *and for every dimension* $d \geq 1$ *there exist* $N \geq e^{\frac{\zeta^2 d}{2}}$ *and* $w_1, \dots, w_N \in \{0,1\}^d$ *such that,*

$$i \neq j \implies d_{\text{ham}}(w_i, w_j) \geq \left(\frac{1}{2} - \zeta\right) d \,.$$

Minimax Lower Bounds. We obtain the following minimax lower bounds that we factorized in a single result:

Proposition 4.2.2. *Let* $(\mathbb{p}_\theta)_{\theta \in \Theta}$ *be a family of probability distributions on the same measurable space and* Θ *be a subset of* $\mathbb{R}^d$ *with* $d \geq 66$ *that contains a ball of radius* r_0 *for the euclidean distance. Assume that* $\gamma > 0$ *is such that*

$$\forall \theta_1, \theta_2 \in \Theta, \quad \text{KL}(\mathbb{p}_{\theta_1} \| \mathbb{p}_{\theta_2}) \leq \gamma \|\theta_2 - \theta_1\|^2. \tag{4.3}$$

Then we have the following results on the minimax rates:

$$\mathfrak{M}_n \geq \frac{\min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64\sqrt{n\gamma}}\right)^2 d}{32} = \Omega\left(\frac{d}{n\gamma}\right) \,,$$

$$\begin{aligned}
\mathfrak{M}_n^{(\epsilon\text{-}DP)} &\geq \frac{\max\left(\min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64\sqrt{n\gamma}}\right), \min\left(\frac{r_0}{\sqrt{d}}, \frac{\sqrt{d}}{64^2\sqrt{2}n\epsilon\sqrt{\gamma}}\right)\right)^2 d}{32} \\
&= \Omega\left(\max\left\{\frac{d}{n\gamma}, \frac{d^2}{(n\epsilon)^2\gamma}\right\}\right) \,, \\
\mathfrak{M}_n^{(\rho\text{-}zCDP)} &\geq \frac{\max\left(\min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64\sqrt{n\gamma}}\right), \min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64^2 2\sqrt{2}n\sqrt{\rho\gamma}}\right)\right)^2 d}{32} \\
&= \Omega\left(\max\left\{\frac{d}{n\gamma}, \frac{d}{n^2\rho\gamma}\right\}\right) \,,
\end{aligned}$$

when $\rho < 1$*. Note that all the asymptotic expressions are taken when* $r_0 > C\sqrt{d}$ *for a positive constant* C *i.e. when the parameter space is not "too small".*

Proof. Without loss of generality, let us suppose that 0 is the center of the ball of radius r_0 (without loss of generality because we are going to work on a neighborhood of 0 but it can be translated to any point). Varshamov–Gilbert's theorem (Fact 4.2.1) with $\zeta = \frac{1}{4}$ allows

us to consider N and $w_1, \dots, w_N$ and to define a packing of the form $\theta_1 := \alpha w_1, \dots, \theta_N := \alpha w_N$ such that

$$i \neq j \implies \frac{\alpha^2 d}{4} \leq \|\theta_i - \theta_j\|^2 \leq \alpha^2 d\,.$$

This yields an $\Omega = \alpha\sqrt{d}/4$-packing with respect to the Euclidean metric. Since 0 is in the interior of Θ, all the θ_i's are in Θ provided that α is small enough. By the (non-private) master lower bound (3.6) and Fano's lemma (Fact 3.1.2),

$$\begin{aligned}
\mathfrak{M}_n &\geq (\alpha\sqrt{d}/4)^2 \cdot \left(1 - \frac{1 + \frac{1}{N}\sum_i \mathrm{KL}\left(\mathbb{p}_{\theta_i}^{\otimes n} \middle\| \frac{1}{N}\sum_j \mathbb{p}_{\theta_j}^{\otimes n}\right)}{\ln N}\right) \\
&\overset{\text{Jensen}}{\geq} (\alpha\sqrt{d}/4)^2 \cdot \left(1 - \frac{1 + \frac{1}{N^2}\sum_{i,j} \mathrm{KL}\left(\mathbb{p}_{\theta_i}^{\otimes n} \middle\| \mathbb{p}_{\theta_j}^{\otimes n}\right)}{\ln N}\right) \\
&= (\alpha\sqrt{d}/4)^2 \cdot \left(1 - \frac{1 + \frac{1}{N^2}\sum_{i,j} n\mathrm{KL}\left(\mathbb{p}_{\theta_i} \| \mathbb{p}_{\theta_j}\right)}{\ln N}\right) \\
&\overset{(4.3)}{\geq} \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{1}{N^2}\sum_{i,j} n\gamma\|\theta_i - \theta_j\|^2}{\ln N}\right) \\
&\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + n\gamma\alpha^2 d}{d/32}\right)\,,
\end{aligned}$$

where in the last line we used that $N \geq e^{d/32}$ and $\|\theta_i - \theta_j\|^2 \leq \alpha^2 d$. With $\alpha := \min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64\sqrt{n\gamma}}\right)$ when $d \geq 66$ leads to

$$\mathfrak{M}_n \geq \frac{\min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64\sqrt{n\gamma}}\right)^2 d}{32} = \Omega\left(\frac{d}{n\gamma}\right)\,.$$

For ϵ-DP and ρ-zCDP, the first term in the max expressed in Proposition 4.2.2 is a direct consequence of the above bound and of Lemma 3.4.7 so we now concentrate on the other term. By the private master lower bound (3.10) and Fano's lemma for product distributions and $(\epsilon, 0)$-DP (see Theorem 3.2.3), arguments as above show that

$$\begin{aligned}
\mathfrak{M}_n^{(\epsilon\text{-DP})} &\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{2n\epsilon}{N^2}\sum_{i,j} \mathrm{TV}\left(\mathbb{p}_{\theta_i}, \mathbb{p}_{\theta_j}\right)}{\ln N}\right) \\
&\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{2n\epsilon}{N^2}\sum_{i,j} \sqrt{\mathrm{KL}\left(\mathbb{p}_{\theta_i} \| \mathbb{p}_{\theta_j}\right)/2}}{\ln N}\right) \\
&\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{2n\epsilon}{N^2}\sum_{i,j} \sqrt{\gamma/2}\|\theta_i - \theta_j\|}{\ln N}\right) \\
&\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + 2n\epsilon\alpha\sqrt{\gamma/2}\sqrt{d}}{d/32}\right)\,.
\end{aligned}$$

Again, setting $\alpha := \min\left(\frac{r_0}{\sqrt{d}}, \frac{\sqrt{d}}{64^2\sqrt{2}n\epsilon\sqrt{\gamma}}\right)$ when $d \geq 66$ allows to conclude that

$$\begin{aligned}\mathfrak{M}_n^{(\epsilon\text{-DP})} &\geq \frac{\min\left(\frac{r_0}{\sqrt{d}}, \frac{\sqrt{d}}{64^2\sqrt{2}n\epsilon\sqrt{\gamma}}\right)^2 d}{32} \\ &= \Omega\left(\frac{d^2}{(n\epsilon)^2\gamma}\right) .\end{aligned}$$

Similarly, by Fano's lemma for product distributions and ρ-zCDP (see Theorem 3.2.4),

$$\begin{aligned}\mathfrak{M}_n^{(\rho\text{-zCDP})} &\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{4n^2\rho}{N^2}\sum_{i,j}\frac{1}{2n}\mathrm{TV}\left(\mathbb{p}_{\theta_i}, \mathbb{p}_{\theta_j}\right) + \mathrm{TV}\left(\mathbb{p}_{\theta_i}, \mathbb{p}_{\theta_j}\right)^2}{\ln N}\right) \\ &\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{4n^2\rho}{N^2}\sum_{i,j}\frac{1}{2n}\sqrt{\mathrm{KL}\left(\mathbb{p}_{\theta_i} \| \mathbb{p}_{\theta_j}\right)/2} + \mathrm{KL}\left(\mathbb{p}_{\theta_i} \| \mathbb{p}_{\theta_j}\right)/2}{\ln N}\right) \\ &\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \frac{4n^2\rho}{N^2}\sum_{i,j}\frac{1}{2n}\sqrt{\gamma/2}\|\theta_i - \theta_j\| + \gamma\|\theta_i - \theta_j\|^2/2}{\ln N}\right) \\ &\geq \frac{\alpha^2 d}{16} \quad \left(1 - \frac{1 + \left(2\sqrt{2}n\rho\alpha\sqrt{\gamma d} + 2n^2\rho\gamma\alpha^2 d\right)}{d/32}\right) ,\end{aligned}$$

and setting $\alpha := \min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64^2 2\sqrt{2}n\sqrt{\rho\gamma}}\right)$ when $d \geq 66$ concludes that (because $\rho \leq 1$)

$$\begin{aligned}\mathfrak{M}_n^{(\rho\text{-zCDP})} &\geq \frac{\min\left(\frac{r_0}{\sqrt{d}}, \frac{1}{64^2 2\sqrt{2}n\sqrt{\rho\gamma}}\right)^2 d}{32} \\ &= \Omega\left(\frac{d}{n^2\rho\gamma}\right) .\end{aligned}$$

□

Note that the constraint $d \geq 66$ can be relaxed to smaller constants by changing the ζ in the application of Varshamov–Gilbert's theorem at the cost of changing the constants in the minimax lower bounds. Likewise, the constraint $\rho < 1$ can be replaced by $\rho < M$ for any positive constant M at the cost again of worse constants. Since we aim to use this result in high dimension and with high privacy, those hypotheses are natural in order to simplify the expressions.

About the choice of the norm. For the estimation, we chose to use the squared l_2 norm as the measure of performance, since it is the one people are the most used to. However, in the literature of differential privacy, a more common practice is to use the total variation distance (see Section 4.3 for an overview).

4.2.2 Continuous exponential families and maximum likelihood

For many other parametric models, the statistician typically would like to consider the maximum likelihood estimator. Given $X_1, \dots, X_n$ i.i.d. random variables of distribution $\mathbb{p}_{\theta^*}$, the maximum likelihood estimator has value

$$\hat{\theta}_{\mathrm{ML}} \in \arg\max_{\theta \in \Theta} \left\{ l(\theta) := \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \right\} , \tag{4.4}$$

where f is the log-likelihood. The parametric model with respect to a reference measure μ is thus

$$\forall X, \quad \frac{d\mathbb{p}_\theta}{d\mu}(X) := e^{f(X,\theta)}, \qquad \theta \in \Theta ,$$

where $\frac{d\mathbb{p}_\theta}{d\mu}$ is the Radon-Nikodym density of $\mathbb{p}_\theta$ with respect to μ and Θ is often a closed, convex subset of $\mathbb{R}^d$ with nonempty interior. This setup covers, in particular, exponential families [Van der Vaart, 1998] with $f(X, \theta) = \theta^T T(X) - \ln(Z(\theta))$ associated with some statistic T and normalization factor $Z(\theta)$. This section first presents a lower bound on the minimax risk for the private estimation in such parametric models and then studies the optimality properties of the Differentially Private Stochastic Gradient Langevin Dynamics (DP-SGLD) of [Ryffel et al., 2022] for this specific task based on the existing upper bounds for this private convex optimizer.

On the regularity of f and the estimation complexity

First, we may assume that the parametric model is not degenerate in the sense that f satisfies

$$\forall \theta \in \Theta, \quad \int \nabla_\theta f(X, \theta) d\mathbb{p}_\theta(X) = 0 . \tag{4.5}$$

This hypothesis is for instance satisfied in the Gaussian model presented previously. Indeed, in this case $\forall X, \nabla_\theta f(\theta + X, \theta) + \nabla_\theta f(\theta - X, \theta) = 0$ and $\forall X, \frac{d\mathbb{p}_\theta}{d\mu}(\theta + X) = \frac{d\mathbb{p}_\theta}{d\mu}(\theta - X)$. This hypothesis is more generally satisfied in the broader model of the exponential families (see [Boucheron et al., 2019, Théorème 4.10]). Under such hypothesis, we have the following lemma which will allow to leverage Proposition 4.2.2:

Lemma 4.2.3. *If $(\mathbb{P}_\theta)_{\theta \in \Theta}$ satisfies the property (4.5) and if f is concave and β-smooth in its second argument, then*

$$\forall \theta_1, \theta_2 \in \Theta, \mathrm{KL}\left(\mathbb{p}_{\theta_1} \| \mathbb{p}_{\theta_2}\right) \leq \frac{\beta}{2} \|\theta_2 - \theta_1\|^2 .$$

Note that the family $(\mathbb{P}_\theta)_{\theta \in \Theta}$ directly depends on f. In particular, for the Gaussian Model, $\beta = \frac{1}{\sigma^2}$, we recover the classical upper bound on the KL divergence between multivariate normal distributions, which is in fact in this case, an equality.

Proof. Because of concavity in the second argument of f and the fact that it is β-smooth, we have the following result:

$$\forall \theta_1, \theta_2 \in \Theta, \forall x, \quad f(x,\theta_1) + \nabla_\theta f(x,\theta_1)^T(\theta_2 - \theta_1) \leq f(x,\theta_2) + \frac{\beta}{2}\|\theta_1 - \theta_2\|^2 .$$

As a consequence,

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{p}_{\theta_1} \| \mathbb{p}_{\theta_2}\right) &= \int \ln\left(\frac{d\mathbb{p}_{\theta_1}}{d\mathbb{P}_{\theta_2}}\right) d\mathbb{p}_{\theta_1} = \int \left(f(X,\theta_1) - f(X,\theta_2)\right) d\mathbb{p}_{\theta_1}(X) \\
&\leq \int \left(-\nabla_\theta f(X,\theta_1)^T(\theta_2 - \theta_1) + \frac{\beta}{2}\|\theta_1 - \theta_2\|^2\right) d\mathbb{p}_{\theta_1}(X) \\
&\overset{(4.5)}{=} \int \frac{\beta}{2}\|\theta_1 - \theta_2\|^2 d\mathbb{p}_{\theta_1}(X) = \frac{\beta}{2}\|\theta_1 - \theta_2\|^2 .
\end{aligned}$$

□

We may apply Proposition 4.2.2 with $\gamma = \beta/2$ and we obtain that

$$\mathfrak{M}_n^{(\rho\text{-zCDP})} = \Omega\left(\max\left\{\frac{d}{n^2\beta\rho}, \frac{d}{n\beta}\right\}\right) \tag{4.6}$$

Under the hypotheses of Proposition 4.2.2: d is big enough, ρ is small enough and the interior of the parameter space is big enough. In particular, this gives us a lower bound to compare any private estimator to.

Private maximum likelihood

In general, $\hat{\theta}_{\mathrm{ML}}$ has no closed form formula. Even when it has some, the closed form formula usually does not respect differential privacy.

The problem (4.4) is typically addressed via numerical optimization: instead of considering its explicit maximum, a provably converging sequence is constructed. This requires some assumptions on the log-likelihood f. A convenient combination of hypotheses is that f is λ-strongly concave, β-smooth and L-Lipschitz in its second argument: then, the stochastic gradient ascend algorithm converges rapidly to $\hat{\theta}_{\mathrm{ML}}$ [Beck, 2017]. Exponential families typically obey those requirements with $\beta := \sup_{\theta\in\Theta} \lambda_{\max}(C_\theta)$ and, $\lambda := \inf_{\theta\in\Theta} \lambda_{\min}(C_\theta)$ where $C_\theta := \mathrm{Cov}_{X\sim\mathbb{P}_\theta}(T(X))$ and $\lambda_{\min}(C)$ (resp. $\lambda_{\max}(C)$) denotes the smallest (resp. largest) eigenvalue of a matrix C (see [Boucheron et al., 2019, Théorème 4.10]).

The issue of privacy can be addressed directly in the optimization procedure. DP-SGD [Abadi et al., 2016] is an adaptation of the Stochastic Gradient Descent method where the gradient is first clipped and then noised. The privacy guarantees are based on the moment accountant method or on the composition of Renyi differential privacy [Mironov, 2017]. The results are obtained under very general hypotheses on the objective function, but

are based on a pessimistic scenario where an adversary may observe every gradient in the optimizer. Recent work based on Langevin diffusion [Chourasia et al., 2021, Ryffel et al., 2022] has adapted the Gradient Descent algorithm and the Stochastic Gradient Descent algorithm in order to have privacy guarantees with tighter utility bounds at the price of stronger hypotheses on the objective function which is required to have a compact domain and to be strongly convex.

Building on DP-SGLD by [Ryffel et al., 2022], we consider its adaptation for maximum likelihood DP-SGML (Algorithm 1). For a batch $\mathcal{B} \subseteq \{1, \dots, n\}$, the batch log-likelihood is defined as

$$l_{\mathcal{B}}(\theta) := \frac{1}{\#(\mathcal{B})} \sum_{i \in \mathcal{B}} f(X_i, \theta) \ .$$

For a closed convex set Θ, Π_Θ refers to the projection onto Θ.

Data: $X_1, \dots, X_n$, f, step sizes $(\eta_k)_{k \geq 0}$, batch size m, noise variance σ^2, initial parameter θ_0, stopping time K.
for $k = 0, \dots, K-1$ **do**
  Sample batch $\mathcal{B}_k$ from $X_1, \dots, X_n$ with replacement of size m ;
  Compute $\nabla l_{\mathcal{B}_k}(\theta_k) = \frac{1}{\#(\mathcal{B}_k)} \sum_{i \in \mathcal{B}_k} \nabla_\theta f(X_i, \theta_k)$;
  Update parameter $\theta_{k+1} = \Pi_\Theta \left(\theta_k + \eta_k \nabla l_{\mathcal{B}_k}(\theta_k) + \sqrt{2\eta_k} \mathcal{N}(0, \sigma^2 I_d) \right)$.
end
return θ_K

Algorithm 1: DP-SGML: Differentially Private Stochastic Gradient Maximum Likelihood

A choice of the parameters $(\eta_k)_{k \geq 0}$, σ^2, θ_0 and K is suggested by the privacy-utility theorem Fact 4.2.4 which is a direct corollary of [Ryffel et al., 2022].

Fact 4.2.4 (Utility and Privacy of Algorithm 1, Fixed Step Size)**.** *Assume that f is λ-strongly concave, β-smooth and L-Lipschitz in its second argument on Θ. Consider any $\rho > 0$, an integer $n \geq 1$, a batch size m and set*

$$\sigma^2 := \frac{4L^2}{\rho \lambda n^2}, \quad K := \frac{2\beta}{\lambda} \ln \left(\frac{\rho n^2}{d} \right), \quad \xi^2 := \mathbb{E}_{\mathcal{B}} \left(\| \nabla l_{\mathcal{B}}(\theta_{\mathrm{ML}}) \|^2 \right)$$

Given a collection $\mathbf{X}$ of n arbitrary samples, consider $\mathfrak{M}(\mathbf{X}) = \theta_K$ obtained using DP-SGML with $\theta_0 \sim \Pi_\Theta \left(\mathcal{N}(0, \frac{2\sigma^2}{\lambda} I_d) \right)$ and constant step size $\eta = \frac{1}{2\beta}$. This mechanism satisfies ρ-zCDP. Moreover, if $\mathbf{X}$ is such that the solution θ_{ML} of (4.4) is in the interior of Θ, then

$$\mathbb{E} \left(\| \theta_{\mathrm{ML}} - \theta_K \|^2 \right) = O \left(\frac{\beta d L^2}{\rho \lambda^3 n^2} \right) + \frac{\xi^2}{2\lambda^2}$$

where the expectation is with respect to initialization, random batch sampling, and noise addition in the parameter update step.

Indeed, the direct application of [Ryffel et al., 2022, Theorem 4.1] gives the privacy guarantee, and that

$$\mathbb{E}\left(l(\theta_{\mathrm{ML}}) - l(\theta_K)\right) = O\left(\frac{\beta d L^2}{\rho \lambda^2 n^2}\right) + \frac{\xi^2}{4\lambda} .$$

Furthermore, by λ-strong concavity of l,

$$l(\theta_{\mathrm{ML}}) - l(\theta_K) \geq \nabla l(\theta_{\mathrm{ML}})(\theta_K - \theta_{\mathrm{ML}}) + \frac{\lambda}{2}\|\theta_L - \theta_{\mathrm{ML}}\|^2$$

and since θ_{ML} is in the interior of Θ, $\nabla l(\theta_{\mathrm{ML}}) = 0$ which concludes the proof. The term $\xi^2 := \mathbb{E}_{\mathcal{B}}\left(\|\nabla l_{\mathcal{B}}(\theta_{\mathrm{ML}})\|^2\right)$ is due to the stochastic noise of the batch sampling. Indeed, even though $\nabla l(\theta_{\mathrm{ML}}) = 0$, this is not necessarily the case when working on batches. This term depends on the batch size m and can be made arbitrarily small by choosing m large enough.

About minimax optimality

The quadratic risk of any (private or not) solver $\mathfrak{M}$ can be decomposed (by the triangle inequality and since $(a+b)^2 \leq 2a^2 + 2b^2, \forall a, b \geq 0$) as:

$$\mathbb{E}\left(\|\theta^* - \mathfrak{M}(\mathbf{X})\|^2\right) \leq 2\left(\mathbb{E}\Big(\|\theta^* - \theta_{\mathrm{ML}}\|^2\Big) + \mathbb{E}\Big(\|\theta_{\mathrm{ML}} - \mathfrak{M}(\mathbf{X})\|^2\Big)\right) \tag{4.7}$$

where the expectation is over the draw of $\mathbf{X}$ and, in the case of a private solver, on the intrinsic randomness of $\mathfrak{M}$.

The first term in the right hand side of (4.7) only depends on the properties of the "ideal" maximum likelihood estimator in this parametric model. Under mild assumptions, it is asymptotically normal – for example, in exponential families (see [Van der Vaart, 1998, Theorem 4.6]): we have

$$\sqrt{n}\left(\theta^* - \theta_{\mathrm{ML}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, C_{\theta^*}^{-1}\right) ,$$

and

$$\mathbb{E}\Big(\|\theta^* - \theta_{\mathrm{ML}}\|^2\Big) = O\left(\frac{d}{n\lambda}\right) . \tag{4.8}$$

The second term in (4.7) depends on the solver, which here can be controlled with Fact 4.2.4. As a consequence, the ratio between the error of estimation and the minimax risk which is lower-bounded in (4.6) can be bounded as follows:

$$\begin{aligned}\frac{\mathbb{E}\Big(\|\theta^* - \mathfrak{M}(\mathbf{X})\|^2\Big)}{\mathfrak{M}_n^{(\rho\text{-zCDP})}} &\overset{(4.7)\&(4.6)}{=} O\left(\frac{\mathbb{E}\Big(\|\theta^* - \theta_{\mathrm{ML}}\|^2\Big) + \mathbb{E}\Big(\|\theta_{\mathrm{ML}} - \mathfrak{M}(\mathbf{X})\|^2\Big)}{\max\left\{\frac{d}{n^2\beta\rho}, \frac{d}{n\beta}\right\}}\right) \\ &= O\left(\frac{n\beta}{d}\mathbb{E}\Big(\|\theta^* - \theta_{\mathrm{ML}}\|^2\Big) + \frac{n^2\beta\rho}{d}\mathbb{E}\Big(\|\theta_{\mathrm{ML}} - \mathfrak{M}(\mathbf{X})\|^2\Big)\right) .\end{aligned}$$

In particular, for the fixed step-size (see Fact 4.2.4), when θ_{ML} is in the interior of Θ and when the variance term due to the clipped gradient is negligible (i.e., when be batch size is big enough to have $\frac{\xi^2}{4\lambda} = O\left(\frac{\beta d L^2}{\rho \lambda^2 n^2}\right)$), the second term is $O\left(\frac{\beta^2 L^2}{\lambda^3}\right)$.

All in all, the ratio between the risk of DP-SGML for maximum likelihood in exponential families when the maximum likelihood estimator is in the interior of the search set is

$$\frac{\mathbb{E}\Big(\|\theta^* - \mathfrak{M}(\mathbf{X})\|^2\Big)}{\mathfrak{M}_n^{(\rho\text{-zCDP})}} = O\left(\frac{\beta}{\lambda} + \frac{\beta^2 L^2}{\lambda^3}\right) .$$

DP-SGML optimally captures the variation in the sample size n, in the privacy parameter ρ, and to some extent, in the dimensionality d (to some extent because even if d vanishes in the expressions, L, β and λ may vary with d). This proves what we call the near-minimax optimality of DP-SG(L)D for performing inference via maximum likelihood in a broad class of parametric models.

4.3 Other parametric models in the literature

Many interesting parametric estimation procedures have been studied in the literature. Table 4.1 presents some of the interesting contributions, without necessarily being exhaustive.

Article	*Model(s)*
[Smith, 2011]	Broad class of models with asymptotically normal, low privacy regime.
[Barber & Duchi, 2014]	Mean estimation.
[Diakonikolas et al., 2015]	Discrete structured distributions, tv distance.
[Karwa & Vadhan, 2018]	Gaussian mean, unidimensional.
[Bun et al., 2019] [Bun et al., 2021]	Hypothesis selection, tv distance. Finite product distributions, tv distance. Gaussian means in high dimensions, tv distance. Sum of independent random variables, tv distance. Piecewise polynomial finite density, tv distance. Mixtures, tv distance. Supervised learning, tv distance.
[Kamath et al., 2019]	Gaussian covariances in high dimensions, scaled Frobenius distance. Gaussian means in high dimensions, tv distance. Product distributions, tv distance.
[Biswas et al., 2020]	Gaussian means in high dimensions, scaled l_2 distance. Gaussian covariances in high dimensions, scaled Frobenius distance.
[Kamath et al., 2020]	Mean of heavy tailed distributions, l_2 distance.
[Acharya et al., 2021e]	Finite distributions, tv distance. Finite distributions, l_2 distance. Finite product distributions, tv distance. Finite mixtures of Gaussian means in high dimension, tv distance
[Aden-Ali et al., 2021]	Gaussian means in high dimensions, tv distance.
[Cai et al., 2021]	Subgaussian mean. Linear regression.
[Brown et al., 2021]	Gaussian means in high dimensions, scaled l_2 distance.
[Cai et al., 2021]	Subgaussian mean. Linear regression.
[Kamath et al., 2022]	Stochastic convex optimization with heavy tailed data.
[Singhal, 2023]	Bernoulli product distributions.
[Kamath et al., 2023a]	Exponential families.
[Kamath et al., 2023b]	Exponential families.

Table 4.1: Bibliography on private parametric estimation

Chapter 5

Nonparametric density estimation

The origin of this chapter, and the use of the first person. This chapter is based on the article [Lalanne et al., 2023a], written by Aurélien Garivier[1], Rémi Gribonval[2], and by myself. In this chapter, I will try to respect the following rule : the use of the first person of the plural (we, our, ...) represents all the above-mentioned people, while the use of the first person of the singular (I, my, ...) represents myself.

We address here the problem of privately estimating a probability density, which fits in this line of work. Given $\mathbf{X} := (X_1, \dots, X_n) \sim \mathbb{P}_\pi^{\otimes n}$, where $\mathbb{P}_\pi$ refers to a distribution of probability that has a density π with respect to the Lebesgue measure on $[0, 1]$, how to estimate π privately? Technically, what metrics or hypothesis should be set on π? What is the cost of privacy? Are the methods known so far optimal? Such are the questions that are investigated in the rest of this chapter.

Related work. Non-parametric density estimation has been an important topic of research in statistics for many decades now. Among the vast literature on the topic, let us just mention the important references [Györfi et al., 2002, Tsybakov, 2009].

Recently, the interest for private statistics has shone a new light on this problem. Remark-

[1] https://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/index.html/

[2] https://people.irisa.fr/Remi.Gribonval/

able early contributions [Wasserman & Zhou, 2010, Hall et al., 2013] adapted histogram estimators, so-called projection estimators and kernel estimators to satisfy the privacy constraint. They conclude that the minimax rate of convergence, $n^{-2\beta/(2\beta+1)}$, where n is the sample size and β is the (Sobolev) smoothness of the density, is not affected by *global* privacy. However, an important implicit hypothesis in this line of work is that ϵ, the parameter that decides how private the estimation needs to be, is supposed not to depend on the sample size. This hypothesis may seem disputable, and more importantly, it fails to precisely characterize the tradeoff between utility and privacy. To the best of our knowledge, the only piece of work that studies this problem under *global* privacy when ϵ is not supposed constant is [Barber & Duchi, 2014]. They study histogram estimators on Lipschitz distributions for the integrated risk. They conclude that the minimax risk of estimation is $\max\left(n^{-2/3} + (n\epsilon)^{-1}\right)$, showing how small ϵ can be before the minimax risk of estimation is degraded. Our chapter extends such results to high degrees of smoothness, to other definitions of *global* differential privacy, and to other risks.

The literature under the much stricter notion of *local* privacy is a lot richer. Contrary to *global* privacy, *local* privacy requires that each data holder anonymizes its data before them being communicated to an aggregator. It is a stronger definition of privacy than global differential privacy. A remarkable early piece of work [Duchi et al., 2016] has brought a nice toolbox for deriving minimax lower bounds under local privacy that has proven to give sharp results for many problems. As a result, the problem of non-parametric density estimation (or its analogous problem of non-parametric regression) has been extensively studied under local privacy. For instance, [Butucea et al., 2019] investigates the elbow effect and questions of adaptivity over Besov ellipsoids. [Kroll, 2021] and [Schluttenhofer & Johannes, 2022] study the density estimation problem at a given point with an emphasis on adaptivity. Universal consistency properties have recently been derived in [Györfi & Kroll, 2023]. Analogous regression problems have been studied in [Berrett et al., 2021] and in [Györfi & Kroll, 2022]. Finally, the problem of optimal non-parametric testing has been studied in [Lam-Weil et al., 2022].

Contributions. In this chapter, we investigate the impact of *global* privacy when the privacy budget is not constant. We treat multiple definitions of *global* privacy and different levels of smoothness for the densities of interest.

In terms of upper-bounds, we analyze histogram and projection estimators at a resolution that captures the impact of the privacy and smoothness parameters. We also prove new lower bounds using the classical packing method combined with new tools that characterize the testing difficulty under global privacy from [Acharya et al., 2021e, Kamath et al., 2022, Lalanne et al., 2023b].

In particular, for Lipschitz densities and under pure differential privacy, we recover the results of [Barber & Duchi, 2014] with a few complements. We then extend the estimation on this class of distributions to the context of concentrated differential privacy [Bun &

Steinke, 2016], a more modern definition of privacy that is compatible with stochastic processes relying on Gaussian noise. We finally investigate higher degrees of smoothness by looking at periodic Sobolev distributions. The main results are summarized in Table 5.

	ϵ-DP	ρ-zCDP
Lipschitz Equation (5.2)	Upper-bound: $O\left(\max\left\{n^{-2/3}, (n\epsilon)^{-1}\right\}\right)$ [Barber & Duchi, 2014] & Theorem 5.1.2 Lower-bounds: -Pointwise: $\Omega\left(\max\left\{n^{-2/3}, (n\epsilon)^{-1}\right\}\right)$ Theorem 5.1.3 & Corollary 5.1.4 -Integrated: $\Omega\left(\max\left\{n^{-2/3}, (n\epsilon)^{-1}\right\}\right)$ [Barber & Duchi, 2014] & Theorem 5.1.6	Upper-bound: $O\left(\max\left\{n^{-2/3}, (n\sqrt{\rho})^{-1}\right\}\right)$ Theorem 5.1.2 Lower-bounds: -Pointwise: $\Omega\left(\max\left\{n^{-2/3}, (n\sqrt{\rho})^{-1}\right\}\right)$ Theorem 5.1.3 & Corollary 5.1.4 -Integrated: $\Omega\left(\max\left\{n^{-2/3}, (n\sqrt{\rho})^{-1}\right\}\right)$ Theorem 5.1.6
Periodic Sobolev Smoothness β Equation (5.21)	Upper-bounds: -Pure DP: $O\left(\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+3/2}}\right\}\right)$ Theorem 5.2.3 -Relaxed: $\tilde{O}\left(\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+1}}\right\}\right)$ $\tilde{O}$ hides polylog factors. Corollary 5.2.5 Lower-bound: $\Omega\left(\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+1}}\right\}\right)$ Theorem 5.2.4	Upper-bound: $O\left(\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\sqrt{\rho})^{-\frac{2\beta}{\beta+1}}\right\}\right)$ Theorem 5.2.3 Lower-bound: $\Omega\left(\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\sqrt{\rho})^{-\frac{2\beta}{\beta+1}}\right\}\right)$ Theorem 5.2.4

Table 5.1: Summary of the results

5.1 Histogram Estimators and Lipschitz Densities

Histogram estimators approximate densities with a piecewise continuous function by counting the number of points that fall into each bin of a partition of the support. Since those numbers follow binomial distributions, the study of histogram estimators is rather simple. Besides, they are particularly interesting when privacy is required, since the sensitivity of a histogram query is bounded independently of the number of bins. They were first studied in this setup in [Wasserman & Zhou, 2010], while [Barber & Duchi, 2014] provided new lower-bounds that did not require a constant privacy budget.

As a warm-up, this section proposes a new derivation of known results in more modern lower-bounding frameworks [Acharya et al., 2021e, Kamath et al., 2022, Lalanne et al., 2023b], and then extends these upper-bounds and lower-bounds to the case of zCDP. Furthermore, it also covers the pointwise risk as well as the infinite-norm risk.

Let $h > 0$ be a given bandwidth or binsize. In order to simplify the notation, we suppose without loss of generality that $1/h \in \mathbb{N} \setminus \{0\}$ (if the converse is true, simply take $h' = 1/\lceil 1/h \rceil$ where $\lceil x \rceil$ refers to the smallest integer bigger than x). $[0,1]$ is partitioned in $\frac{1}{h}$

sub-intervals of length h, which are called the bins of the histogram. Let $Z_1, \dots, Z_{1/h}$ be independent and identically distributed random variables with the same distribution as a random variable Z that is supposed to be centered and to have a finite variance. Given a dataset $\mathbf{X} = (X_1, \dots, X_n)$, the (randomized) histogram estimator is defined for $x \in [0,1]$ as

$$\hat{\pi}^{\text{hist}}(\mathbf{X})(x) := \sum_{b \in \text{bins}} \mathbb{1}_b(x) \frac{1}{nh} \left(\sum_{i=1}^{n} \mathbb{1}_b(X_i) + Z_b \right) . \tag{5.1}$$

We indexed the Z's by a bin instead of an integer without ambiguity. Note that by taking Z almost-surely constant to 0, one recovers the usual (non-private) histogram estimator of a density.

5.1.1 General utility of histogram estimators

Characterizing the utility of (5.1) typically requires assumptions on the distribution π to estimate. The class of L-Lipschitz densities is defined as

$$\Theta_L^{\text{Lip}} := \left\{ \pi \in \mathcal{C}^0([0,1], \mathbb{R}_+) \,\middle|\, \begin{cases} \forall x, y \in [0,1], |\pi(y) - \pi(x)| \leq L|y - x| , \\ \int_{[0,1]} \pi = 1 . \end{cases} \right\} . \tag{5.2}$$

The following general-purpose lemma gives an upper-bound on the error that the histogram estimator makes on Lipschitz distributions:

Lemma 5.1.1 (General utility of (5.1))**.** *There exists $C_L > 0$, a positive constant that only depends on L, such that*

$$\sup_{x_0 \in [0,1]} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}^{\text{hist}}} \left(\left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0) - \pi(x_0) \right)^2 \right) \leq C_L \left(h^2 + \frac{1}{nh} + \frac{\mathbb{V}(Z)}{n^2 h^2} \right) .$$

Proof. Let $\pi \in \Theta_L^{\text{Lip}}$, $x_0 \in [0,1]$. The classical bias-variance decomposition gives that

$$\mathbb{E} \left(\left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0) - \pi(x_0) \right)^2 \right) = \left(\mathbb{E} \left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0) \right) - \pi(x_0) \right)^2 + \mathbb{V} \left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0) \right) .$$

For any $x \in [0,1]$, we note $\text{bin}(x)$ the bin of the histogram in which x falls into. Notice that, for any $x_0 \in [0,1]$ and any integer i, the random variable $\mathbb{1}_{\text{bin}(x_0)}(X_i)$ follows a Bernoulli distribution of probability of success $\int_{\text{bin}(x_0)} \pi$. Let us first study the bias, using

the definition (5.1) of $\hat{\pi}^{\text{hist}}$

$$\begin{aligned}
\left|\mathbb{E}\left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0)\right) - \pi(x_0)\right| &= \left|\frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left(\mathbb{1}_{\text{bin}(x_0)}(X_i)\right) - \pi(x_0)\right| \\
&= \left|\frac{n\int_{\text{bin}(x_0)}\pi(x)dx}{nh} - \pi(x_0)\right| \\
&= \frac{1}{h}\left|\int_{\text{bin}(x_0)}(\pi(x)-\pi(x_0))dx\right| \\
&\leq \frac{1}{h}\int_{\text{bin}(x_0)}|\pi(x)-\pi(x_0)|\,dx \\
&\leq \frac{L}{h}\int_{\text{bin}(x_0)}|x-x_0|\,dx \leq \frac{Lh}{2}\,.
\end{aligned}$$

Let us now look at the variance. By independence of X_i's and Z_j's,

$$\begin{aligned}
\mathbb{V}\left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0)\right) &= \frac{1}{n^2h^2}\left(\sum_{i=1}^{n}\mathbb{V}\left(\mathbb{1}_{\text{bin}(x_0)}(X_i)\right) + \mathbb{V}\left(Z_{\text{bin}(x_0)}\right)\right) \\
&= \frac{1}{n^2h^2}\left(n\left(\int_{\text{bin}(x_0)}\pi\right)\left(1-\int_{\text{bin}(x_0)}\pi\right) + \mathbb{V}\left(Z\right)\right) \\
&\leq \frac{1}{nh^2}\left(\int_{\text{bin}(x_0)}\pi\right) + \frac{\mathbb{V}\left(Z\right)}{n^2h^2}.
\end{aligned}$$

Since π is L-Lipschitz on $[0,1]$ and has to integrate to 1 (because it is a density), π is uniformly bounded from above by $L+1$ on $[0,1]$. Hence, $\int_{\text{bin}(x_0)}\pi \leq (L+1)h$ and the result follows. □

The term h^2 corresponds to the bias of the estimator. The variance term $\frac{1}{nh} + \frac{\mathbb{V}(Z)}{n^2h^2}$ exhibits two distinct contributions : the sampling noise $\frac{1}{nh}$ and the privacy noise $\frac{\mathbb{V}(Z)}{n^2h^2}$. In particular, the utility of $\hat{\pi}^{\text{hist}}$ changes depending whether the variance is dominated by the sampling noise or by the privacy noise.

5.1.2 Privacy and bin size tuning

$\hat{\pi}^{\text{hist}}(\mathbf{X})$ is a simple function of the bin count vector

$$f(\mathbf{X}) := \left(\sum_{i=1}^{n}\mathbb{1}_{b_1}(X_i), \ldots, \sum_{i=1}^{n}\mathbb{1}_{b_{1/h}}(X_i)\right)\,.$$

In particular, since the bins form a partition of $[0,1]$, changing the value of one of the X's can change the values of at most two components of $f(\mathbf{X})$ by at most 1. Hence, the l_1 and l_2 sensitivities of f are respectively 2 and $\sqrt{2}$. By a direct application of the Laplace or Gaussian mechanisms, and by choosing the binsize that minimizes the variance, we obtain the following privacy-utility result :

Theorem 5.1.2 (Privacy and utility of (5.1) - DP case)**.** *Given $\epsilon > 0$, using $\hat{\pi}^{\text{hist}}$ with $h = \max(n^{-1/3}, (n\epsilon)^{-1/2})$ and $Z = \frac{2}{\epsilon}\mathcal{L}(1)$, where $\mathcal{L}(1)$ refers to a random variable following a Laplace distribution of parameter 1, leads to an ϵ-DP procedure. Furthermore, in this case, there exists $C_L > 0$, a positive constant that only depends on L, such that*

$$\sup_{x_0\in[0,1]} \sup_{\pi\in\Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}^{\text{hist}}} \left(\left(\hat{\pi}^{\text{hist}}(\mathbf{X})(x_0) - \pi(x_0) \right)^2 \right) \leq C_L \max\left\{ n^{-2/3}, (n\epsilon)^{-1} \right\} .$$

Furthermore, given $\rho > 0$, using $\hat{\pi}^{hist}$ with $h = \max(n^{-1/3}, (n\sqrt{\rho})^{-1/2})$ and $Z = \sqrt{\frac{1}{\rho}}\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ refers to a random variable following a centered Gaussian distribution of variance 1, leads to a ρ-zCDP procedure. Furthermore, in this case, there exists $C_L > 0$, a positive constant that only depends on L, such that

$$\sup_{x_0\in[0,1]} \sup_{\pi\in\Theta_L^{Lip}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}^{hist}} \left(\left(\hat{\pi}^{hist}(\mathbf{X})(x_0) - \pi(x_0) \right)^2 \right) \leq C_L \max\left\{ n^{-2/3}, (n\sqrt{\rho})^{-1} \right\} .$$

Note that this bound is uniform in x. In particular, by integration on $[0,1]$, the same bound also holds for the integrated risk (in L^2 norm). As expected, the optimal bin size h depends on the sample size n and on the parameter (ϵ or ρ) tuning the privacy.

5.1.3 Lower-bounds and minimax optimality

All lower-bounds will be investigated in a minimax sense. Given a class Π of admissible densities, a semi-norm $\|\cdot\|$ on a space containing the class Π, and a non-decreasing positive function Φ such that $\Phi(0) = 0$, the minimax risk is defined as

$$\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\pi\in\Pi} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}} \Phi(\|\hat{\pi}(\mathbf{X}) - \pi\|) ,$$

where $\mathcal{C}$ is a condition that must satisfy the estimator (privacy in our case).

General framework. A usual technique for the derivation of minimax lower bounds on the risk uses a reduction to a testing problem (see [Tsybakov, 2009]). Indeed, if a family $\Pi' := \{\pi_1, \dots, \pi_m\} \subset \Pi$ of cardinal m is an Ω-packing of Π (that is if $i \neq j \implies \|\pi_i - \pi_j\| \geq 2\Omega$), then a lower bound is given by

$$\begin{aligned} \inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\pi\in\Pi} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}} \Phi(\|\hat{\pi}(\mathbf{X}) - \pi\|) & \\ \geq \Phi(\Omega) \inf_{\substack{\hat{\pi} \text{ s.t. } \mathcal{C} \\ \Psi:\text{codom}(\hat{\pi})\to\{1,\dots,m\}}} \max_{i\in\{1,\dots,m\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\pi_i}^{\otimes n},\hat{\pi}} \left(\Psi\left(\hat{\pi}(\mathbf{X})\right) \neq i \right) . & \end{aligned} \tag{5.3}$$

For more details, see [Duchi et al., 2016, Acharya et al., 2021e, Lalanne et al., 2023b]. The right-hand side characterizes the difficulty of discriminating the distributions of the packing by a statistical test. Independently on the condition $\mathcal{C}$, it can be lower-bounded using information-theoretic results such a Le Cam's lemma [Rigollet & Hütter, 2015, Lemma 5.3] or Fano's lemma [Giraud, 2021, Theorem 3.1]. When $\mathcal{C}$ is a local privacy condition,

[Duchi et al., 2016] provides analogous results that take privacy into account. Recent work [Acharya et al., 2021e, Kamath et al., 2022, Lalanne et al., 2023b] provides analogous forms for multiple notions of *global* privacy. When using this technique, finding good lower-bounds on the minimax risk boils down to finding a packing of densities that are far enough from one another without being too easy to discriminate with a statistical test.

It is interesting to note that for the considered problem, this technique does not yield satisfying lower-bounds with ρ-zCDP every time Fano's lemma is involved. Systematically, a small order is lost. To circumvent that difficulty, we had to adapt Assouad's technique to the context of ρ-zCDP. Similar ideas have been used in [Duchi et al., 2016] for lower-bounds under *local* differential privacy and in [Acharya et al., 2021e] for regular *global* differential privacy. To the best of our knowledge, such a technique has never been used in the context of global *concentrated* differential privacy, and is presented in Remark 5.1.5. In all the proofs of the lower-bounds, we systematically presented both approaches whenever there is a quantitative difference. This difference could be due to small suboptimalities in Fano's lemma for concentrated differential privacy, or simply to the use of a suboptimal packing.

Pointwise lower-bound

The first lower-bound that will be investigated is with respect to the pointwise risk. Pointwise, that is to say given $x_0 \in [0,1]$, the performance of the estimator $\hat{\pi}$ is measured by how well it approximates π at x_0 with the quadratic risk $\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\left((\hat{\pi}(\mathbf{X})(x_0)-\pi(x_0))^2\right)$. Technically, it is the easiest since it requires a "packing" of only two elements, which gives the following lower-bound:

Theorem 5.1.3 (Pointwise lower-bound)**.** *There exists $C_L > 0$, a positive constant depending only on L such that, for any $x_0 \in [0,1]$, there exist $n_0(x_0, L) \in \mathbb{N}$ and $c_0(x_0, L) > 0$ such that for any $n \geq n_0$, and any $\alpha \geq c_0/n$*

$$\inf_{\hat{\pi}\text{s.t.}\mathcal{C}} \sup_{\pi\in\Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\left((\hat{\pi}(\mathbf{X})(x_0)-\pi(x_0))^2\right) \geq C_L^{-1}\max\left\{n^{-2/3},(n\alpha)^{-1}\right\}, \tag{5.4}$$

where $\alpha = \epsilon$ when the condition $\mathcal{C}$ is the ϵ-DP condition and $\alpha = \sqrt{\rho}$ when $\mathcal{C}$ is ρ-zCDP.

Proof. Let $x_0 \in [0,1]$. As explained above, finding a "good" lower-bound can be done by finding and analyzing a "good" packing of the parameter space. Namely, in this case, we have to find distributions on $[0,1]$ that have a L-Lipschitz density (w.r.t. Lebesgue's measure) such that the densities are far from one another at x_0, but such that it is not extremely easy to discriminate them with a statistical test. We propose to use a packing $\{\mathbb{P}_f, \mathbb{P}_g\}$ of two elements where g is the constant function on $[0,1]$ (hence $\mathbb{P}_g$ is the uniform distribution) and f deviates from g by a small triangle centered at x_0. The two densities are represented in Figure 5.1. After analyzing various quantities about these densities,

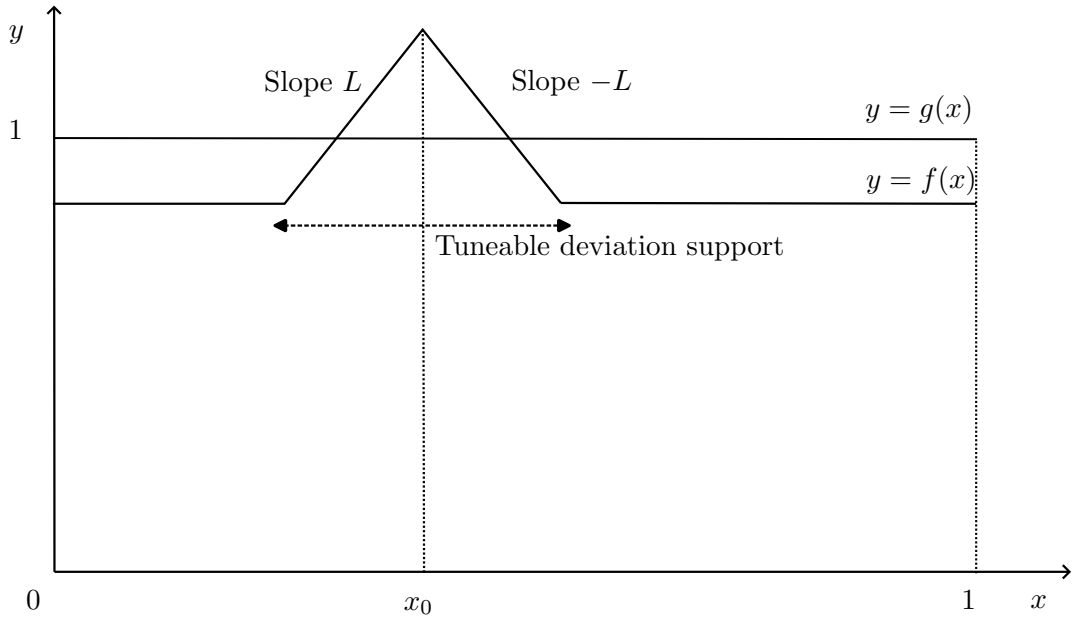


Figure 5.1: Packing for Theorem 5.1.3

such as their distance at x_0, their KL divergences or their TV distance, we leverage Le Cam-type results to conclude.

Packing construction. We define the functions $f_{L,x_0,h}, \forall h > 0$ as

$$\forall x \in [0,1], \quad f_{L,x_0,h}(x) := \begin{cases} 1 - Lh^2 \text{ if } x \in [0, x_0 - h) \cup [x_0 + h, 1], \\ 1 - Lh^2 + Lh + L(x - x_0) \text{ if } x \in [x_0 - h, x_0) \\ 1 - Lh^2 + Lh - L(x - x_0) \text{ if } x \in [x_0, x_0 + h) \end{cases} . \tag{5.5}$$

Note that as soon as $h \leq \min\{x_0, 1 - x_0\}$, $f_{L,x_0,h} \in \Theta_L^{\text{Lip}}$. The case $x_0 \in \{0, 1\}$ is treated in the exact same fashion, but by considering functions that only contain "half of a spike" centered on x_0. Furthermore, let us note g the function that is constant to 1 on $[0, 1]$ (we have $g \in \Theta_L^{\text{Lip}}$).

We start by redefining the total variation distance between two probability distributions, and we recall some useful alternative expressions that are used in the proofs of this chapter. Given $(\mathcal{U}, \mathcal{T})$ a set $\mathcal{U}$ equipped with a σ-algebra $\mathcal{T}$, and two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ two probability distributions on $\mathcal{U}$, and compatible with $\mathcal{T}$, the total variation distance $\text{TV}(\cdot, \cdot)$ between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as

$$\text{TV}(\mathbb{P}_1, \mathbb{P}_2) := \sup_{\mathcal{S} \in \mathcal{T}} |\mathbb{P}_1(\mathcal{S}) - \mathbb{P}_2(\mathcal{S})| .$$

Furthermore, when $\mathbb{P}_1, \mathbb{P}_2$ are dominated by a common σ-finite measure μ on $(\mathcal{U}, \mathcal{T})$, by noting $p_1 := \frac{d\mathbb{P}_1}{d\mu}$ and $p_2 := \frac{d\mathbb{P}_2}{d\mu}$, the Radon-Nikodym derivatives of $\mathbb{P}_1$ and $\mathbb{P}_2$ with respect

to μ, the following alternative expressions to the total variation can be useful :

$$\begin{aligned}
\mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right) &:= \sup_{\mathcal{S}\in\mathcal{T}} |\mathbb{P}_1(\mathcal{S}) - \mathbb{P}_2(\mathcal{S})| \\
&= \mathbb{P}_1(\{p_1 > p_2\}) - \mathbb{P}_2(\{p_1 > p_2\}) \\
&= \int_{\{p_1>p_2\}} p_1 - p_2 d\mu \\
&= \int_{\{p_2\geq p_1\}} p_2 - p_1 d\mu \\
&= \frac{1}{2}\int_{\mathcal{U}} |p_1 - p_2| d\mu \\
&= 1 - \int_{\mathcal{U}} \min(p_1, p_2) d\mu \ .
\end{aligned}$$

These expressions simply come from considering the events $\{p_1 > p_2\}$ and $\{p_2 \geq p_1\}$ that form a partition of $\mathcal{U}$, and from the relation $|a - b| = a + b - 2\min(a, b)$ for any real numbers a and b.

Jumping back to our original proof, when $f_{L,x_0,h} \in \Theta_L^{\mathrm{Lip}}$, we can compute the total variation between $\mathbb{P}_{f_{L,x_0,h}}$ and $\mathbb{P}_g$ the distributions of probability with densities $f_{L,x_0,h}$ and g with respect to Lebesgue's measure on $[0, 1]$,

$$\begin{aligned}
\mathrm{TV}\left(\mathbb{P}_{f_{L,x_0,h}}, \mathbb{P}_g\right) &= 1 - \int_{[0,1]} \min\left(f_{L,x_0,h}, g\right) \\
&\overset{\text{Constant part}}{\leq} 1 - \int_{[0,1]} 1 - Lh^2 dx \\
&= Lh^2 \ .
\end{aligned} \tag{5.6}$$

Another important measure of discrepancy between probability distributions is the so-called KL divergence. For two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ such that $\mathbb{P} \ll \mathbb{Q}$ (absolute continuity), it is defined as

$$\mathrm{KL}\left(\mathbb{P} \| \mathbb{Q}\right) = \int \ln\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P} \ .$$

Back to our problem, for h in a neighborhood of 0, we also have the following Taylor expansion on their KL divergence:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_g \| \mathbb{P}_{f_{L,x_0,h}}\right) &= \int_{[0,1]} \ln\left(\frac{g}{f_{L,x_0,h}}\right) g \\
&= \ln\left(\frac{1}{1-Lh^2}\right)(1-2h) + 2\int_0^h \ln\left(\frac{1}{1-Lh^2+Lt}\right) dt \\
&\leq C\left(h^3 + O(h^4)\right) \ ,
\end{aligned} \tag{5.7}$$

where C is a positive constant depending only on L the O only hides constant factors. Furthermore, $|g(x_0) - f_{L,x_0,h}(x_0)| = L|h^2 - h|$ and $\{g, f_{L,x_0,h}\}$ is thus a $\frac{L}{2}|h^2 - h|$ packing of Θ_L^{Lip} w.r.t the seminorm $f, g \mapsto \|f - g\| \coloneqq |f(x_0) - g(x_0)|$.

Recovering the usual lower-bound By the classical minimax reduction as hypothesis testing Equation (5.3),

$$
\begin{aligned}
&\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left((\hat{\pi}(\mathbf{X})(x_0) - \pi(x_0))^2 \right) \\
&\quad \geq \frac{L^2}{4} \left(h^2 - h\right)^2 \inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \inf_{\Psi : \Theta_L^{\text{Lip}} \to \{0,1\}} \max \Big\{ \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_g^{\otimes n}, \hat{\pi}} \left(\Psi(\hat{\pi}(\mathbf{X})) \neq 0 \right), \\
&\qquad\qquad \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{f_{L,x_0,h}}^{\otimes n}, \hat{\pi}} \left(\Psi(\hat{\pi}(\mathbf{X})) \neq 1 \right) \Big\} \\
&\quad \overset{\textit{Fact } 3.1.1}{\geq} \frac{L^2}{8}(h^2 - h)^2 \left(1 - \text{TV}\left(\mathbb{P}_g^{\otimes n}, \mathbb{P}_{f_{L,x_0,h}}^{\otimes n}\right)\right) \\
&\quad \overset{\text{Pinsker}}{\geq} \frac{L^2}{8}(h^2 - h)^2 \left(1 - \sqrt{\text{KL}\left(\mathbb{P}_g^{\otimes n} \| \, \mathbb{P}_{f_{L,x_0,h}}^{\otimes n}\right)/2}\right) \\
&\quad \overset{\text{Tensorization}}{=} \frac{L^2}{8}(h^2 - h)^2 \left(1 - \sqrt{n\text{KL}\left(\mathbb{P}_g \| \, \mathbb{P}_{f_{L,x_0,h}}\right)/2}\right) \\
&\quad \overset{(5.7)}{\geq} \frac{L^2}{8}(h^2 - h)^2 \left(1 - \sqrt{\frac{n}{2}\left(\frac{h^3 L^2}{3} + O(h^4)\right)}\right) .
\end{aligned}
\tag{5.8}
$$

The second inequality comes from the so-called Le Cam's lemma [Rigollet & Hütter, 2015] that lower-bounds the testing difficulty (without further constraints) between two distributions. The next inequality comes from the so-called Pinsker's inequality [Tsybakov, 2009], that states that for two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, $\text{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\text{KL}(\mathbb{P} \| \, \mathbb{Q})/2}$. The last inequality is the result of the so-called tensorization property of the KL divergence that states that for two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, and for an integer $n \geq 1$, $\text{KL}(\mathbb{P}^{\otimes n} \| \, \mathbb{Q}^{\otimes n}) \leq n\text{KL}(\mathbb{P} \| \, \mathbb{Q})$.

When possible (i.e. when n is big enough), setting $h = \left(\frac{1}{4nL^2}\right)^{1/3}$ leads to, for n big enough (so that $h - h^2 \geq h/2$ and $|O(h^4)| \leq \frac{h^3 L^2}{3}$),

$$
\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left((\hat{\pi}(\mathbf{X})(x_0) - \pi(x_0))^2 \right) \geq \frac{L^2}{64} \left(\frac{1}{4L^2}\right)^{2/3} n^{-2/3} .
$$

This implies the first lower bound.

ϵ-DP overhead. By Equation (5.8) and by Le Cam's lemma for differential privacy on product distributions Theorem 3.2.1,

$$\inf_{\hat{\pi}\ \epsilon\text{-DP}} \sup_{\pi\in\Theta_L^{\mathrm{Lip}}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\left((\hat{\pi}(\mathbf{X})(x_0)-\pi(x_0))^2\right) \geq \frac{L^2}{8}(h^2-h)^2 e^{-n\epsilon \mathrm{TV}\left(\mathbb{P}_{f_{L,x_0,h}},\mathbb{P}_g\right)} \overset{(5.6)}{\geq} \frac{L^2}{8}(h^2-h)^2 e^{-Ln\epsilon h^2} .$$

When possible (i.e. when $n\epsilon$ is big enough), setting $h = 1/\sqrt{n\epsilon}$ leads to, for $n\epsilon$ big enough (so that $h-h^2 \geq h/2$),

$$\inf_{\hat{\pi}\ \epsilon\text{-DP}} \sup_{\pi\in\Theta_L^{\mathrm{Lip}}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\left((\hat{\pi}(\mathbf{X})(x_0)-\pi(x_0))^2\right) \geq \frac{L^2 e^{-L}}{32}(n\epsilon)^{-1} .$$

ρ-zCDP overhead. By Le Cam's lemma for zero-concentrated differential privacy on product distributions Theorem 3.2.2 in (5.8),

$$\begin{aligned}\inf_{\hat{\pi}\ \epsilon\text{-DP}} \sup_{\pi\in\Theta_L^{\mathrm{Lip}}} &\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\left((\hat{\pi}(\mathbf{X})(x_0)-\pi(x_0))^2\right) \\ &\geq \frac{L^2}{8}(h^2-h)^2\left(1-n\sqrt{\rho/2}\mathrm{TV}\left(\mathbb{P}_{f_{L,x_0,h}},\mathbb{P}_g\right)\right) \\ &\overset{(5.6)}{\geq} \frac{L^2}{8}(h^2-h)^2\left(1-n\sqrt{\rho/2}Lh^2\right) .\end{aligned}$$

When possible (i.e. when $n\sqrt{\rho}$ is big enough), setting $h = \left(\frac{1}{\sqrt{2}Ln\sqrt{\rho}}\right)^{1/2}$ leads to, for $n\sqrt{\rho}$ big enough (so that $h-h^2 \geq h/2$),

$$\inf_{\hat{\pi}\ \rho\text{-zCDP}} \sup_{\pi\in\Theta_L^{\mathrm{Lip}}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\left((\hat{\pi}(\mathbf{X})(x_0)-\pi(x_0))^2\right) \geq \frac{L}{64}(n\sqrt{\rho})^{-1} .$$

□

Additionally, we can notice that, when applied to any fixed $x_0 \in [0,1]$, Theorem 5.1.3 immediately gives the following corollary for the control in infinite norm :

Corollary 5.1.4 (Infinite norm lower-bound)**.** *There exists $C_L > 0$, a positive constant depending only on L such that there exist $n_0(L) \in \mathbb{N}$ and $c_0(L) > 0$ such that for any $n \geq n_0$, and any $\alpha \geq c_0/n$*

$$\inf_{\hat{\pi}\mathrm{s.t.}\mathcal{C}} \sup_{\pi\in\Theta_L^{\mathrm{Lip}}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\|\hat{\pi}(\mathbf{X})-\pi\|_\infty^2 \geq C_L^{-1}\max\left\{n^{-2/3},(n\alpha)^{-1}\right\} , \tag{5.9}$$

where $\alpha = \epsilon$ when the condition $\mathcal{C}$ is the ϵ-DP condition and $\alpha = \sqrt{\rho}$ when $\mathcal{C}$ is ρ-zCDP.

On the optimality and on the cost of privacy. Theorem 5.1.2, Theorem 5.1.3 and Corollary 5.1.4 give the following general result : Under ϵ-DP or under ρ-zCDP, histogram estimators have minimax-optimal rates of convergence against distributions with Lipschitz densities, for the pointwise risk or the risk in infinite norm. In particular, in the *low privacy* regime ("large" α), the usual minimax rate of estimation of $n^{-\frac{2}{3}}$ is not degraded. This includes the early observations of [Wasserman & Zhou, 2010] in the case of constant α (ϵ or $\sqrt{\rho}$). However, in the *high privacy* regimes ($\alpha \ll n^{-\frac{1}{3}}$), these results prove a systematic degradation of the estimation. Those regimes are the same as in [Barber & Duchi, 2014], the metrics on the other hand are different.

Remark 5.1.5 (Assouad's lemma/method)**.** As the reduction to a testing problem between multiple hypotheses, Assouad's lemma relies on similar ideas, where the packing has to be parametrized by a hypercube. Its advantage over tools like Fano's lemma is that it only makes tests between pairs of hypotheses (instead of all of them at the same time). The cost of this is that the control of the packing is slightly more difficult.

Suppose that the set of distributions of interest $\mathcal{P}$ contains a family of distributions $(\mathbb{P}_\omega)_{\omega\in\{0,1\}^m}$ for a certain positive integer m. If the loss function (taken quadratic for simplicity) can be decomposed as

$$\forall \omega, \omega' \in \{0,1\}^m, \quad \|\mathbb{P}_\omega - \mathbb{P}_{\omega'}\|_{L^2}^2 \geq 2\tau \sum_{i=1}^m \mathbb{1}_{\omega_i \neq \omega_i'} = 2\tau d_{\text{ham}}\left(\omega, \omega'\right) \ , \tag{5.10}$$

then the minimax risk can be lower-bounded as (the proof is classical and can be found in [Acharya et al., 2021e, Section 5.4])

$$\begin{aligned} &\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P},\hat{\pi}}(\|\hat{\pi}(\mathbf{X}) - \pi\|_{L^2}^2) \\ &\qquad \geq \frac{\tau}{16} \sum_{i=1}^m \inf_{\mathfrak{M} \text{ s.t. } \mathcal{C} \atop \Psi:\text{codom}(\mathfrak{M})\to\{0,1\}} \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\omega^{i,0}},\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq 0\right) + \\ &\qquad\qquad\qquad\qquad \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\omega^{i,1}},\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq 1\right) \ . \end{aligned} \tag{5.11}$$

where $\mathbb{P}_{\omega^{i,0}}$ and $\mathbb{P}_{\omega^{i,1}}$ are the *mixture* distributions

$$\mathbb{P}_{\omega^{i,0}} := \frac{1}{2^{m-1}} \sum_{\omega\in\{0,1\}^m | \omega_i = 0} \mathbb{P}_\omega \quad \text{and} \quad \mathbb{P}_{\omega^{i,0}} := \frac{1}{2^{m-1}} \sum_{\omega\in\{0,1\}^m | \omega_i = 1} \mathbb{P}_\omega \ . \tag{5.12}$$

The term

$$\mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\omega^{i,0}},\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq 0\right) + \mathbb{P}_{\mathbf{X}\sim\mathbb{P}_{\omega^{i,1}},\mathfrak{M}}\left(\Psi\left(\mathfrak{M}(\mathbf{X})\right) \neq 1\right)$$

characterizes the *testing* difficulty between $\mathbb{P}_{\omega^{i,0}}$ and $\mathbb{P}_{\omega^{i,1}}$. It can be controlled by Le Cam's lemma, and by its variants when working under privacy (see [Acharya et al., 2021e, Lalanne et al., 2023b] for differential privacy and [Lalanne et al., 2023b] for concentrated differential privacy).

Integrated lower-bound

The lower-bound of Theorem 5.1.3 is interesting, but its pointwise (or in infinite norm in the case of Corollary 5.1.4) nature means that much global information is possibly lost. Instead, one can look at the integrated risk $\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\|\hat{\pi}(\mathbf{X})-\pi\|_{L^2}^2$. Given Lemma 5.1.1 and the fact that we work on probability distributions with a compact support, upper-bounding this quantity is straightforward.

The lower-bound for the integrated risk is given by :

Theorem 5.1.6 (Integrated lower-bound)**.** *There exists $C_L > 0$, a positive constant depending only on L such that, there exist $n_0(L) \in \mathbb{N}$ and $c_0(L) > 0$ such that for any $n \geq n_0$, and any $\alpha \geq c_0/n$*

$$\inf_{\hat{\pi}\ s.t.\ \mathcal{C}} \sup_{\pi\in\Theta_L^{Lip}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}}\|\hat{\pi}(\mathbf{X})-\pi\|_{L^2}^2 \geq C_L^{-1}\max\left\{n^{-2/3},(n\alpha)^{-1}\right\}$$

where $\alpha = \epsilon$ when $\mathcal{C}$ is the ϵ-DP condition, and $\alpha = \sqrt{\rho}$ when $\mathcal{C}$ is the ρ-zCDP condition.

Proof. If we were to use the same packing (see Figure 5.1) as in the proof of Theorem 5.1.3, the lower-bounds would not be good. Indeed, moving from the pointwise difference to the L^2 norm significantly diminishes the distances in the packing. Instead, we will use the same idea of deviating from a constant function by triangles, except that we authorize more than one deviation. More specifically, we consider a packing consisting of densities f_ω's where the ω's are a well-chosen family of $\{0,1\}^m$ (m is fixed in the proof) [Van der Vaart, 1998]. Then, for a given $\omega \in \{0,1\}^m$, f_ω has a triangle centered on $\frac{i}{m+1}$ iff $w_i \neq 0$. We then leverage Fano-type inequalities, and we use Assouad's method in order to find the announced lower-bounds.

For any $\omega \in \{0,1\}^m$ different from 0 and any $h > 0$, we define the function $g_{L,\omega,h}$ as

$$g_{L,\omega,h} := \frac{1}{\|\omega\|_1}\sum_{i=1}^m \omega_i f_{\|\omega\|_1 L,\frac{i}{m+1},h}\ , \tag{5.13}$$

where the functions f are defined in (5.5). Note that $g_{L,\omega,h}$ is L-Lipschitz and that as soon as $h \leq h_m := \frac{1}{2(m+1)}$ it is also a valid density so that $g_{L,\omega,h} \in \Theta_L^{\mathrm{Lip}}$. Notice that the function $g_{L,\omega,h}$ is constant to $1-\|\omega\|_1 Lh^2$ everywhere except on each interval $\left[\frac{i}{m+1}-h,\frac{i}{m+1}+h\right]$ with i such that $\omega_i \neq 0$, on which it deviates by a triangle of slopes $+L$ and $-L$.

By denoting by K the triangle kernel such that $K(t) = \int_{-\infty}^t L\mathbb{1}_{[-h,0]}(t') - L\mathbb{1}_{[(0,h]}(t')dt'$,

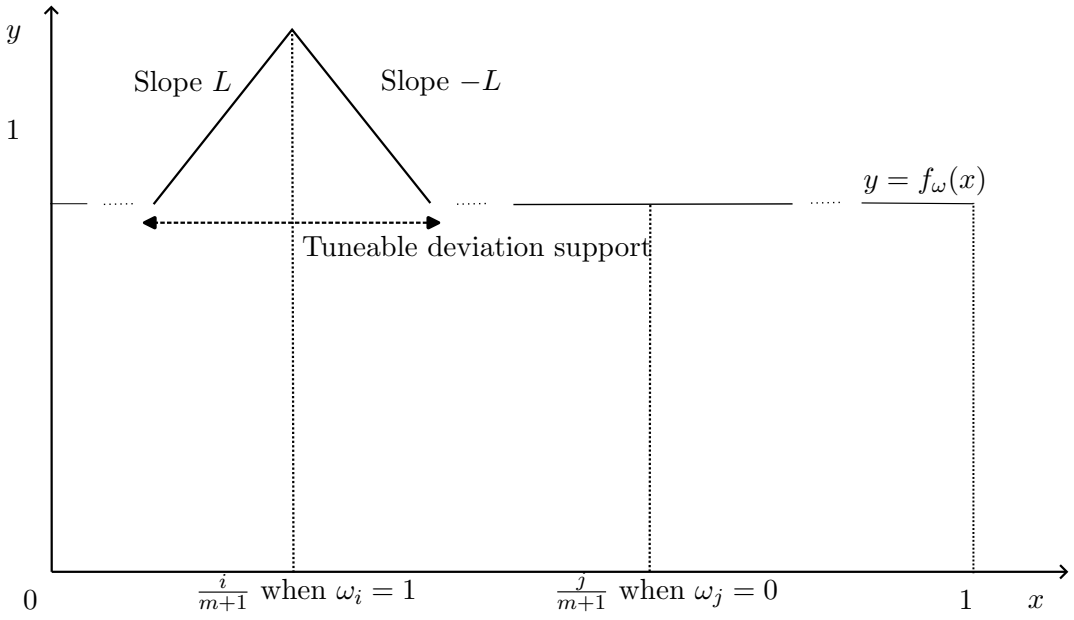


Figure 5.2: Packing for Theorem 5.1.6

it might be easier to visualize $g_{L,\omega,h}$ as

$$\forall t \in [0,1], \quad g_{L,\omega,h}(t) = 1 - \|\omega\|_1 \int K + \sum_{i=1}^{m} \omega_i K\left(t - \frac{i}{m+1}\right) ,$$

where $\int K = Lh^2$.

For $\omega, \omega' \in \{0,1\}^m$ and for h small enough (i.e. $h \le h_m$), we can bound the total variation between $\mathbb{P}_{g_{L,\omega,h}}$ and $\mathbb{P}_{g_{L,\omega',h}}$ as

$$\begin{aligned} \mathrm{TV}\left(\mathbb{P}_{g_{L,\omega,h}}, \mathbb{P}_{g_{L,\omega',h}}\right) &= 1 - \int_{[0,1]} \min\left(g_{L,\omega,h}, g_{L,\omega',h}\right) \\ &\overset{\text{Constant part}}{\le} 1 - \min\left(1 - \|\omega\|_1 Lh^2, 1 - \|\omega'\|_1 Lh^2\right) \\ &= \max\left(\|\omega\|_1, \|\omega'\|_1\right) Lh^2 \le mLh^2 . \end{aligned} \tag{5.14}$$

$$
\begin{aligned}
&\mathrm{TV}\left(\mathbb{P}_{g_{L,\omega,h}}, \mathbb{P}_{g_{L,\omega',h}}\right) \\
&\qquad = \frac{1}{2}\int \left|g_{L,\omega,h} - g_{L,\omega',h}\right| \\
&\qquad = \frac{1}{2}\int \left| \|(\omega'\|_1 - \|\omega\|_1) \int K + \sum_{i=1}^{m} (\omega_i' - \omega_i) K\left(\cdot - \frac{i}{m+1}\right)\right| \\
&\qquad \leq \frac{1}{2}\int \left|\|\omega'\|_1 - \|\omega\|_1\right| \int K + \sum_{i=1}^{m} \left|\omega_i' - \omega_i\right| K\left(\cdot - \frac{i}{m+1}\right) \\
&\qquad = \frac{1}{2}\left(\left|\|\omega'\|_1 - \|\omega\|_1\right| + d_{\mathrm{ham}}\left(\omega, \omega'\right)\right) \int K \\
&\qquad \leq mLh^2 \ .
\end{aligned}
\tag{5.15}
$$

The KL divergence between $\mathbb{P}_{g_{L,\omega,h}}$ and $\mathbb{P}_g$, with g the density constant equal to 1 on $[0,1]$, satisfies

$$
\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{g_{L,\omega,h}} \middle\| \mathbb{P}_g\right) \\
&\qquad = \int_{[0,1]} \ln\left(g_{L,\omega,h}\right) g_{L,\omega,h} \\
&\qquad = \ln\left(1 - \|\omega\|_1 Lh^2\right)\left(1 - \|\omega\|_1 Lh^2\right)\left(1 - \|\omega\|_1 2h\right) \\
&\qquad\qquad + 2\|\omega\|_1 \int_0^h \ln\left(1 - \|\omega\|_1 Lh^2 + Lt\right)\left(1 - \|\omega\|_1 Lh^2 + Lt\right) dt \\
&\qquad \overset{\ln(1+\cdot)\leq\cdot}{\leq} \left(-\|\omega\|_1 Lh^2\right)\left(1 - \|\omega\|_1 Lh^2\right)\left(1 - \|\omega\|_1 2h\right) \\
&\qquad\qquad + 2\|\omega\|_1 \int_0^h \left(-\|\omega\|_1 Lh^2 + Lt\right)\left(1 - \|\omega\|_1 Lh^2 + Lt\right) dt \\
&\qquad \overset{\text{Calculus}}{\leq} \frac{L^2}{3}\|\omega\|_1 h^3 (2 - 3\|\omega\|_1 h) \ .
\end{aligned}
\tag{5.16}
$$

Finally, we lower bound the squared L^2 distance between $g_{L,\omega,h}$ and $g_{L,\omega',h}$:

$$
\begin{aligned}
&\int_{[0,1]} \left(g_{L,\omega,h} - g_{L,\omega',h}\right)^2 \\
&\quad = \sum_{i=1}^{m} \mathbb{1}_{\omega_i \neq \omega_i'} \int_{\frac{i}{m+1}-h}^{\frac{i}{m+1}+h} \left(\left(\|\omega'\|_1 - \|\omega\|_1\right) \int K + (\omega_i - \omega_i') K\left(t - \frac{i}{m+1}\right)\right)^2 dt \\
&\quad \geq \sum_{i=1}^{m} \mathbb{1}_{\omega_i \neq \omega_i'} \int_{\frac{i}{m+1}-h}^{\frac{i}{m+1}+h} \left(K\left(t - \frac{i}{m+1}\right) - \left|\|\omega\|_1 - \|\omega'\|_1\right| \int K\right)^2 dt \\
&\quad \geq \sum_{i=1}^{m} \mathbb{1}_{\omega_i \neq \omega_i'} \int_{\frac{i}{m+1}-h}^{\frac{i}{m+1}+h} \left\{ \left(K\left(t - \frac{i}{m+1}\right)\right)^2 \right. \\
&\qquad\qquad \left. -2K\left(t - \frac{i}{m+1}\right) \left|\|\omega\|_1 - \|\omega'\|_1\right| \int K \right\} dt \\
&\quad \geq d_{\text{ham}}\left(\omega, \omega'\right) \left(\int K^2 - 2m \left(\int K \right)^2 \right) \\
&\quad \geq 2 d_{\text{ham}}\left(\omega, \omega'\right) L^2 \left(\frac{h^3}{3} - m h^4 \right) \\
&\quad = \frac{2 d_{\text{ham}}\left(\omega, \omega'\right) L^2 \left(h^3 - 3 m h^4\right)}{3} .
\end{aligned}
\tag{5.17}
$$

By the Varshamov-Gilbert theorem [Tsybakov, 2009, Lemma 2.7], as long as $m \geq 8$, there exist $M \in \mathbb{N}$ and $\omega^{(0)}, \ldots, \omega^{(M)} \in \{0,1\}^m$ such that $M \geq 2^{m/8}$, $\omega^{(0)} = \{0\}^m$ and $i \neq j \implies d_{\text{ham}}\left(\omega^{(i)}, \omega^{(j)}\right) \geq m/8$. According to (5.17), the family $\left(g_{L,\omega^{(i)},h}\right)_{i=1,\ldots,M}$ is then an $\Omega := \frac{1}{2}\sqrt{\frac{mL^2(h^3 - 3mh^4)}{12}}$ packing of Θ_L^{Lip} for the L^2 distance.

Recovering the usual lower-bound. By Equation (5.3) with $\Phi(\cdot) \coloneqq (\cdot)^2$ and $\|\cdot\|$ the L^2 norm,

$$\begin{aligned}
&\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) \\
&\quad \geq \frac{mL^2 \left(h^3 - 3mh^4\right)}{48} \inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \inf_{\Psi : \Theta_L^{\text{Lip}} \to \{0,1\}} \max_{i=1,\dots,M} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}_{g_{L,\omega^{(i)},h}}^{\otimes n}, \hat{\pi}} \left(\Psi(\hat{\pi}(\mathbf{X})) \neq i \right) \\
&\quad \overset{\textit{Fact } 3.1.2}{\geq} \frac{mL^2 \left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + \frac{1}{M} \sum_{1 \leq i \leq M} \mathrm{KL}\left(\mathbb{P}_{g_{L,\omega^{(i)},h}}^{\otimes n} \middle\| \mathbb{P}_g^{\otimes n} \right)}{\ln(M)} \right) \\
&\quad \overset{\text{Tensorization}}{=} \frac{mL^2 \left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + \frac{n}{M} \sum_{1 \leq i \leq M} \mathrm{KL}\left(\mathbb{P}_{g_{L,\omega^{(i)},h}} \middle\| \mathbb{P}_g \right)}{\ln(M)} \right) \\
&\quad \overset{(5.16)\&\|\omega\|_1 \leq m, M \geq 2^{m/8}}{\geq} \frac{mL^2 \left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + \frac{L^2}{3} nmh^3(2 - 3mh)}{\ln(2) m/8} \right) .
\end{aligned} \tag{5.18}$$

So, by choosing $m = \lceil n^{1/3} \rceil$ and $h = \frac{c}{m}$ where c is a positive constant small enough we get, for n big enough,

$$\inf_{\hat{\pi} \ \epsilon\text{-DP}} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) \geq C^{-1} (n)^{-2/3} ,$$

where C is a positive constant depending only on L.

ϵ-DP overhead. By the same reduction and Fano's lemma for differential privacy on product distributions (Theorem 3.2.3), we get for any $h \leq h_m$,

$$\begin{aligned}
&\inf_{\hat{\pi} \ \epsilon\text{-DP}} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) \\
&\qquad \geq \frac{mL^2 \left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + \frac{n\epsilon}{M^2} 2 \sum_{1 \leq i,j \leq M} \mathrm{TV}\left(\mathbb{P}_{g_{L,\omega^{(i)},h}}, \mathbb{P}_{g_{L,\omega^{(j)},h}} \right)}{\ln(M)} \right) \\
&\qquad \overset{(5.15)\&M \geq 2^{m/8}}{\geq} \frac{mL^2 \left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + 2n\epsilon mLh^2}{\ln(2) m/8} \right) .
\end{aligned}$$

So, by choosing $m = \lceil \sqrt{n\epsilon} \rceil$ and $h = \frac{c}{m}$ where c is small enough a positive constant (depending only on L), we get, as soon as $\min(n, n\epsilon)$ is big enough,

$$\inf_{\hat{\pi} \ \epsilon\text{-DP}} \sup_{\pi \in \Theta_L^{\text{Lip}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) \geq C'^{-1} (n\epsilon)^{-1} ,$$

where C' is a positive constant depending only on L.

ρ-zCDP overhead. For ρ-zCDP, we present the proof using both Fano's lemma and Assouad's method. We will see that Assouad gives better results

Fano version. By again the same reduction and Fano's lemma for zero-concentrated differential privacy (Theorem 3.2.4), denoting $t_{i,j} := \mathrm{TV}\left(\mathbb{P}_{g_{L,\omega^{(i)},h}}, \mathbb{P}_{g_{L,\omega^{(j)},h}}\right)$, we get for any $h \leq h_m$,

$$\begin{aligned}
\inf_{\hat{\pi}\ \rho\text{-zCDP}} \sup_{\pi \in \Theta_L^{\mathrm{Lip}}} \mathbb{E}_{\mathbf{X}\sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) \\
&\geq \frac{mL^2\left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + \frac{n^2\rho}{M^2} 4 \sum_{1\leq i,j\leq M} \left(\frac{1}{n} t_{i,j} + t_{i,j}^2\right)}{\ln(M)}\right) \\
&\overset{(5.15)}{\geq} \frac{mL^2\left(h^3 - 3mh^4\right)}{48} \left(1 - \frac{1 + n^2\rho 4 \left(\frac{mLh^2}{n} + m^2L^2h^4\right)}{\ln(2)m/8}\right) .
\end{aligned}$$

So, by choosing $m = \left\lceil \left(n\sqrt{\rho}\right)^{\frac{2}{3}} \right\rceil$ and $h = \frac{c}{m}$ for c small enough (depending only on L), if $\frac{n}{\rho}$ is big enough, we get that

$$\inf_{\hat{\pi}\ \text{s.t.}\ \rho\text{-zCDP}} \sup_{\pi \in \Theta_L^{\mathrm{Lip}}} \mathbb{E}_{\mathbf{X}\sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) \geq C''^{-1} (n\sqrt{\rho})^{-4/3}$$

where C'' is a positive constant depending only on L.

Assouad version. From Equation (5.17), we can see that when $h := \frac{c}{m}$ for a positive c that is small enough, the condition expressed in Equation (5.10) is satisfied for $\tau = \Omega(h^3)$. To apply (5.12), the only missing ingredient is to bound the testing difficulties between the mixtures on the hypercube.

In the sequel, $\mathbb{P}_\omega$ is used as a short for $\mathbb{P}_{g_{L,\omega,h}}$. We need to bound the total variation

between the mixtures on the hypercube (see (5.12)) as

$$\begin{aligned}
\mathrm{TV}\left(\mathbb{P}_{\omega^{i,0}}, \mathbb{P}_{\omega^{i,1}}\right) &= \mathrm{TV}\left(\frac{1}{2^{m-1}} \sum_{\omega \in \{0,1\}^m | \omega_i = 0} \mathbb{P}_{g_{L,\omega,h}}, \frac{1}{2^{m-1}} \sum_{\omega \in \{0,1\}^m | \omega_i = 1} \mathbb{P}_{g_{L,\omega,h}}\right) \\
&= \frac{1}{2} \frac{1}{2^{m-1}} \int \left| \sum_{\omega \in \{0,1\}^m | \omega_i = 0} g_{L,\omega,h} - \sum_{\omega \in \{0,1\}^m | \omega_i = 1} g_{L,\omega,h} \right| \\
&= \frac{1}{2^m} \int \left| \sum_{\omega_1,\ldots,\omega_{i-1},\omega_{i+1}\ldots,\omega_m \in \{0,1\}} \left(g_{L,(\omega_1,\ldots,\omega_{i-1},0,\omega_{i+1}\ldots,\omega_m),h} - \right.\right. \\
&\qquad \left.\left. g_{L,(\omega_1,\ldots,\omega_{i-1},1,\omega_{i+1}\ldots,\omega_m),h}\right) \right| \\
&\leq \frac{1}{2^m} \sum_{\omega_1,\ldots,\omega_{i-1},\omega_{i+1}\ldots,\omega_m \in \{0,1\}} \int \left| g_{L,(\omega_1,\ldots,\omega_{i-1},0,\omega_{i+1}\ldots,\omega_m),h} - \right. \\
&\qquad \left. g_{L,(\omega_1,\ldots,\omega_{i-1},1,\omega_{i+1}\ldots,\omega_m),h} \right| \\
&\overset{\textit{Equation (5.15)}}{\leq} \frac{1}{2^m} \sum_{\omega_1,\ldots,\omega_{i-1},\omega_{i+1}\ldots,\omega_m \in \{0,1\}} 2Lh^2 \\
&= O\left(h^2\right) \ .
\end{aligned}$$

All in all, by using Le Cam's lemma for product distribution and ρ-zCDP Theorem 3.2.2, and by Equation (5.11),

$$\inf_{\hat{\pi}\ \rho\text{-zCDP}} \sup_{\pi \in \Theta^{\mathrm{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \left(\int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \right) = \Omega\left(mh^3\right)\left(1 - n\sqrt{\rho} O\left(h^2\right)\right). \tag{5.19}$$

Setting $h \approx \left(n\sqrt{\rho}\right)^{\frac{-1}{2}}$ concludes the proof. □

Since the lower-bounds of Theorem 5.1.6 match the upper-bounds of Theorem 5.1.2, we conclude that the corresponding estimators are optimal in terms of minimax rate of convergence.

5.2 Projection Estimators and Periodic Sobolev Densities

The Lipschitz densities considered in Section 5.1 are general enough to be applicable in many problems. However, this level of generality becomes a curse in terms of rate of estimation. Indeed, as we have seen, the optimal rate of estimation is $\max\left(n^{-2/3}, (n\epsilon)^{-1}\right)$. To put it into perspective, for many parametric estimation procedures, the optimal rate of convergence usually scales as $\max\left(n^{-1}, (n\epsilon)^{-2}\right)$ [Acharya et al., 2021e]. This section studies the estimation of smoother distributions, for different smoothness levels, at the

cost of generality. In particular, it establishes that the smoother the distribution class is, the closer the private rate of estimation is to $\max\left(n^{-1}, (n\epsilon)^{-2}\right)$. In other words, it means that the more regular the density is supposed to be, the closer we get to the difficulty of parametric estimation.

When the density of interest π is in $L^2([0,1])$, it is possible to approximate it by projections. Indeed, $L^2([0,1])$ being a separable Hilbert space, there exists a countable orthonormal family $(\phi_i)_{i\in\mathbb{N}\setminus\{0\}}$ that is a Hilbert basis. In particular, if $\theta_i := \int_{[0,1]} \pi\, \phi_i$ then

$$\sum_{i=1}^{N} \theta_i \phi_i \underset{N\to+\infty}{\overset{L^2}{\longrightarrow}} \pi \ .$$

Let N be a positive integer, $Z_1, \dots, Z_N$ be independent and identically distributed random variables with the same distribution as a centered random variable Z having a finite variance. Given a dataset $\mathbf{X} = (X_1, \dots, X_n)$, that is also independent of $Z_1, \dots, Z_N$, the (randomized) projection estimator is defined as

$$\hat{\pi}^{\text{proj}}(\mathbf{X}) = \sum_{i=1}^{N} \left(\hat{\theta}_i + \frac{1}{n} Z_i\right) \phi_i \quad \text{where} \quad \hat{\theta}_i := \frac{1}{n}\sum_{j=1}^{n} \phi_i(X_j) \ . \tag{5.20}$$

The truncation order N and the random variable Z are tuned later to obtain the desired levels of privacy and utility. $L^2([0,1])$ has many well known Hilbert bases, hence multiple choices for the family $(\phi_i)_{i\in\mathbb{N}\setminus\{0\}}$. For instance, orthogonal polynomials, wavelets, or the Fourier basis, are often great choices for projection estimators. Because of the privacy constraint however, it is better to consider a *uniformly bounded* Hilbert basis [Wasserman & Zhou, 2010], which is typically not the case with a polynomial or wavelet basis. From now on, this work will focus on the following Fourier basis :

$$\begin{aligned} \phi_1(x) &= 1 \\ \phi_{2k}(x) &= \sqrt{2}\sin\left(2\pi k x\right) \quad k \geq 1 \\ \phi_{2k+1}(x) &= \sqrt{2}\cos\left(2\pi k x\right) \quad k \geq 1 \ . \end{aligned}$$

5.2.1 General utility of projection estimators

By the Parseval formula, the truncation resulting of approximating the density π on a finite family of N orthonormal functions induces a bias term that accounts for $\sum_{i\geq N+1}\theta_i^2$ in the mean square error. Characterizing the utility of $\hat{\pi}^{\text{proj}}$ requires controlling this term, and this is usually done by imposing that π is in a Sobolev space. We recall the definition given in [Tsybakov, 2009]: given $\beta \in \mathbb{N}\setminus\{0\}$ and $L > 0$, the class $\Theta^{\text{Sob}}_{L,\beta}$ of Sobolev densities of parameters β and L is defined as

$$\Theta^{\text{Sob}}_{L,\beta} := \left\{ \pi \in \mathcal{C}^{\beta}([0,1], \mathbb{R}_+) \;\middle|\; \begin{cases} \pi^{(\beta-1)} \text{ is absolutely continuous} \ , \\ \int_{[0,1]} \left(\pi^{(\beta)}\right)^2 \leq L^2 \ , \\ \int_{[0,1]} \pi = 1 \ . \end{cases} \right\} \ .$$

For a function f, we used the notation $f^{(\beta)}$ to refer to its derivative of order β. In addition, the class $\Theta^{\mathrm{PSob}}_{L,\beta}$ of periodic Sobolev densities of parameters β and L is defined as

$$\Theta^{\mathrm{PSob}}_{L,\beta} := \left\{ \pi \in \Theta^{\mathrm{Sob}}_{L,\beta} \middle| \forall j \in \{0, \dots, \beta - 1\}, \pi^{(j)}(0) = \pi^{(j)}(1) \right\} . \tag{5.21}$$

Finally, we recall the following general-purpose lemma [Tsybakov, 2009] that allows controlling the truncation bias :

Fact 5.2.1 (Ellipsoid reformulation [Tsybakov, 2009])**.** *A non-negative function π with integral 1 belongs to $\Theta^{PSob}_{L,\beta}$ if and only if $\sum_{i=1}^{\infty} a_i^{2\beta}\theta_i^2 \leq \frac{L^2}{\pi^{2\beta}}$, where $a_j := j$ if j is even and $a_j := j-1$ if j is odd.*

In this class, one can characterize the utility of projection estimators with the following lemma:

Lemma 5.2.2 (General utility of (5.20))**.** *There is a constant $C_{L,\beta} > 0$, depending only on L, β, such that*

$$\sup_{\pi \in \Theta^{PSob}_{L,\beta}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}^{proj}} \|\hat{\pi}^{proj}(\mathbf{X}) - \pi\|_{L^2}^2 \leq C_{L,\beta} \left(\frac{1}{N^{2\beta}} + \frac{N}{n} + \frac{N\mathbb{V}(Z)}{n^2} \right) .$$

Proof. Let $\pi \in \Theta^{\mathrm{PSob}}_{L,\beta}$. We have,

$$\begin{aligned}
\mathbb{E}\left(\int_{[0,1]} \left(\hat{\pi}^{\mathrm{proj}}(\mathbf{X}) - \pi \right)^2 \right) &\overset{\text{Parseval}}{=} \mathbb{E}\left(\sum_{i=1}^{N} \left(\hat{\theta}_i - \theta_i + \frac{1}{n} Z_i \right)^2 + \sum_{i=N+1}^{+\infty} \theta_i^2 \right) \\
&= \sum_{i=1}^{N} \mathbb{E}\left(\left(\hat{\theta}_i - \theta_i + \frac{1}{n} Z_i \right)^2 \right) + \sum_{i=N+1}^{+\infty} \theta_i^2 .
\end{aligned}$$

Furthermore, for any i, since Z is centered

$$\begin{aligned}
\mathbb{E}\left(\hat{\theta}_i \right) &= \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^{n} \phi_i(X_j) \right) \\
&= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left(\phi_i(X_j) \right) \overset{X_j \text{ i.i.d.}}{=} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}} \phi_i(X_1) \\
&= \int \pi \phi_i = \theta_i
\end{aligned}$$

Hence, for any i, since Z_i is independent from the dataset

$$\mathbb{E}\left(\left(\hat{\theta}_i - \theta_i + \frac{1}{n}Z_i\right)^2\right) = \mathbb{V}\left(\hat{\theta}_i\right) + \frac{1}{n^2}\mathbb{V}\left(Z_i\right)$$
$$\overset{\text{Independence of } X_j}{=} \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{V}\left(\phi_i\left(X_j\right)\right) + \frac{1}{n^2}\mathbb{V}\left(Z_i\right)$$
$$\overset{|\phi_i|\leq\sqrt{2}}{\leq} \frac{2}{n} + \frac{1}{n^2}\mathbb{V}\left(Z\right) \ .$$

Finally, with $a_j \coloneqq j-1$, Fact 5.2.1 allows bounding $\sum_{i=m+1}^{+\infty}\theta_i^2$ as

$$\sum_{i=N+1}^{+\infty}\theta_i^2 \leq \frac{1}{N^{2\beta}}\sum_{i=N+1}^{+\infty}a_i^{2\beta}\theta_i^2 \leq \frac{1}{N^{2\beta}}\sum_{i=1}^{+\infty}a_i^{2\beta}\theta_i^2 \overset{Fact\ 5.2.1}{\leq} \frac{1}{N^{2\beta}}\frac{L^2}{\pi^{2\beta}} \ .$$

This yields the conclusion with $C_{L,\beta} \coloneqq \max(2, L^2/\pi^{2\beta})$. □

5.2.2 Privacy and bias tuning

The estimator $\hat{\pi}^{\text{proj}}(\mathbf{X})$ is a function of the sums $\left(\sum_{j=1}^{n}\phi_1(X_j), \ldots, \sum_{j=1}^{n}\phi_N(X_j)\right)$. In particular, it is possible to use Laplace and Gaussian mechanisms on this function in order to obtain privacy. Since the functions $|\phi_i|$ are bounded by $\sqrt{2}$ for any i, the l_1 sensitivity of this function is $2\sqrt{2}N$ and its l_2 sensitivity is $2\sqrt{2}\sqrt{N}$. Applying the Laplace and the Gaussian mechanism and tuning N to optimize the utility of Lemma 5.2.2 gives the following result:

Theorem 5.2.3 (Privacy and utility of (5.20))**.** *Given any $\epsilon > 0$ and truncation order N, using $\hat{\pi}^{proj}$ with $Z = \frac{2N\sqrt{2}}{\epsilon}\mathcal{L}(1)$, where $\mathcal{L}(1)$ refers to a random variable following a Laplace distribution of parameter 1, leads to an ϵ-DP procedure. Moreover, there exists $C_{L,\beta} > 0$, a positive constant that only depends on L and β, such that if N is on the order of* $\min\left(n^{\frac{1}{2\beta+1}}, (n\epsilon)^{\frac{1}{\beta+3/2}}\right)$,

$$\sup_{\pi\in\Theta_{L,\beta}^{PSob}}\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}^{proj}}\|\hat{\pi}^{proj}(\mathbf{X}) - \pi\|_{L^2}^2 \leq C_{L,\beta}\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+3/2}}\right\} \ .$$

Furthermore, given any $\rho > 0$, and truncation order N, using $\hat{\pi}^{proj}$ with $Z = \frac{2\sqrt{N}}{\sqrt{\rho}}\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ refers to a random variable following a centered Gaussian distribution of variance 1, leads to a ρ-zCDP procedure. Moreover, there exists $C_{L,\beta} > 0$, a positive constant that only depends on L and β, such that, if N is of the order of $\min\left(n^{\frac{1}{2\beta+1}}, \left(n\sqrt{\rho}\right)^{\frac{1}{\beta+1}}\right)$

$$\sup_{\pi\in\Theta_{L,\beta}^{PSob}}\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}^{proj}}\|\hat{\pi}^{proj}(\mathbf{X}) - \pi\|_{L^2}^2 \leq C_{L,\beta}\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\sqrt{\rho})^{-\frac{2\beta}{\beta+1}}\right\} \ .$$

We now discuss these guarantees depending on the considered privacy regime.

Low privacy regimes. According to Theorem 5.2.3, when the privacy-tuning parameters are not too small (i.e. when the estimation is not too private), the usual rate of convergence $n^{-\frac{2\beta}{2\beta+1}}$ is not degraded. In particular, for constant ϵ or ρ, this recovers the results of [Wasserman & Zhou, 2010].

High privacy regimes. Furthermore, Theorem 5.2.3 tells that in high privacy regimes ($\epsilon \ll n^{-\frac{\beta-1/2}{2\beta+1}}$ or $\rho \ll n^{-\frac{2\beta+2}{2\beta+1}}$), the provable guarantees of the projection estimator are degraded compared to the usual rate of convergence. Is this degradation constitutive of the estimation problem, or is it due to a suboptimal upper-bound? Section 5.2.3 shows that this excess of risk is in fact almost optimal.

5.2.3 Lower-bounds

As with the integrated risk on Lipschitz distributions, obtaining lower-bounds for the class of periodic Sobolev densities is done by considering a packing with many elements. The idea of the packing is globally the same as for histograms, except that the uniform density is perturbed with a general C^∞ kernel with compact support instead of simple triangles. In the end, we obtain the following result:

Theorem 5.2.4 (Integrated lower-bound)**.** *Given $L, \beta > 0$ there exists constants $C_{L,\beta} > 0$, $n_0(L,\beta) \in \mathbb{N}$, and $c_0(L,\beta) > 0$, such that for any $n \geq n_0$, and any $\alpha \geq c_0/n$*

$$\inf_{\hat{\pi}\ s.t.\ \mathcal{C}} \sup_{\pi \in \Theta^{PSob}_{L,\beta}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}} \|\hat{\pi}(\mathbf{X}) - \pi\|_{L^2}^2 \geq C_{L,\beta}^{-1} \max\left\{ n^{-\frac{2\beta}{2\beta+1}}, (n\alpha)^{-\frac{2\beta}{\beta+1}} \right\}$$

where $\alpha = \epsilon$ when $\mathcal{C}$ is the ϵ-DP condition, and $\alpha = \sqrt{\rho}$ when $\mathcal{C}$ is the ρ-zCDP condition.

Proof. As with the proof of Theorem 5.1.6, this lower-bound is based on the construction of a packing of densities f_ω's where the ω's are a well-chosen family of $\{0,1\}^m$ (m is fixed in the proof). Then, for a given $\omega \in \{0,1\}^m$, f_ω deviates from a constant function around $\frac{i}{m+1}$ if, and only if, $w_i \neq 0$. Contrary to the proof of Theorem 5.1.6 however, the deviation cannot be by a triangle : Indeed, such a function wouldn't even be differentiable. Instead, we use a deviation by a C^∞ kernel with compact support. Even if the complete details are given in the full proof, Figure 5.3 gives a general illustration of the packing. Again, Fano-type inequalities (for the ϵ-DP case), and Assouad's lemma (for the ρ-zCDP case) are used to conclude.

Let us consider the following well-known function :

$$\forall x \in \mathbb{R}, \quad K_0(x) := e^{-\frac{1}{1-x^2}} \mathbb{1}_{(-1,1)}(x)\ .$$

We can notice that for any $\beta > 0$ there exists $\nu > 0$ such that the kernel $K(x) := \nu K_0(2x)$ satisfies $K \in \mathcal{C}^\infty(\mathbb{R}, [0,+\infty))$, $\int \left(K^{(\beta)}\right)^2 \leq 1$ and $K(x) > 0$ iff $x \in (-1/2, 1/2)$. Furthermore, for any $i \in \mathbb{N}$, $K^{(i)}(x) = 0$ for every $x \in (-\infty, -1/2] \cup [1/2, +\infty)$.

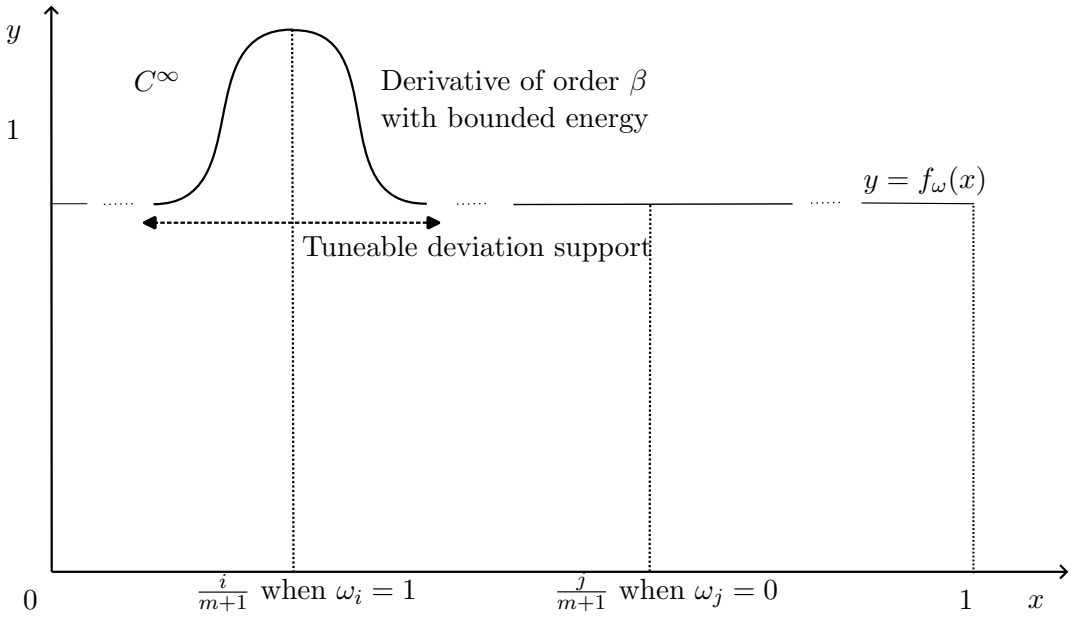


Figure 5.3: Packing for Theorem 5.2.4

Packing construction. Let $m \in \mathbb{N} \setminus \{0\}$ that will be fixed later. For any $h > 0$, and $\omega \in \{0,1\}^m$, we define the function $g_{L,\beta,\omega,h}$ as,

$$\forall x \in [0,1], \quad g_{L,\beta,\omega,h}(x) \coloneqq 1 - \|\omega\|_1 L h^{\beta+1} \int K + L h^\beta \sum_{i=1}^m \omega_i K\left(\frac{x - \frac{i}{m+1}}{h}\right) . \tag{5.22}$$

Note that when $h < \frac{1}{m+1}$ we have $\int_0^1 g_{L,\beta,\omega,h} = 1$; when $mh \int \left(K^{(\beta)}\right)^2 \leq 1$, we have $g_{L,\beta,\omega,h} \geq 0$; and when both hold we have $g_{L,\beta,\omega,h} \in \Theta^{\mathrm{PSob}}_{L,\beta}$. Indeed, under these hypotheses

$$\begin{aligned}
\int \left(g^{(\beta)}_{L,\beta,\omega,h}\right)^2 &= \int \left(L h^\beta \sum_{i=1}^m \omega_i \left(x \mapsto K\left(\frac{x - \frac{i}{m+1}}{h}\right)\right)^{(\beta)}\right)^2 \\
&= \int \left(L \sum_{i=1}^m \omega_i K^{(\beta)}\left(\frac{\cdot - \frac{i}{m+1}}{h}\right)\right)^2 \\
&\overset{\text{disjoint support}}{=} L^2 \sum_{i=1}^m \omega_i \int K^{(\beta)}\left(\frac{\cdot - \frac{i}{m+1}}{h}\right)^2 \\
&\leq L^2 mh \int \left(K^{(\beta)}\right)^2 \leq L^2 .
\end{aligned}$$

In the sequel of this proof, this hypothesis will always be satisfied asymptotically for all the values of m and h that will be considered. From now on, we may consider it valid.

Given $h > 0$ and $\omega, \omega' \in \{0,1\}^m$, when $g_{L,\beta,\omega,h}, g_{L,\beta,\omega',h} \in \Theta^{\mathrm{PSob}}_{L,\beta}$, we can bound the total variation between $\mathbb{P}_{g_{L,\beta,\omega,h}}$ and $\mathbb{P}_{g_{L,\beta,\omega',h}}$ as,

$$
\begin{aligned}
&\mathrm{TV}\left(\mathbb{P}_{g_{L,\beta,\omega,h}}, \mathbb{P}_{g_{L,\beta,\omega',h}}\right) \\
&\qquad = \frac{1}{2}\int \left| g_{L,\beta,\omega,h} - g_{L,\beta,\omega',h} \right| \\
&\qquad = \frac{1}{2}\int \left| \|(\omega'\|_1 - \|\omega\|_1) L h^{\beta+1} \int K + \sum_{i=1}^{m} (\omega'_i - \omega_i) L h^{\beta} K\left(\frac{\cdot - \frac{i}{m+1}}{h}\right) \right| \\
&\qquad \le \frac{1}{2}\int \left| \|\omega'\|_1 - \|\omega\|_1 \right| L h^{\beta+1} \int K + \sum_{i=1}^{m} |\omega'_i - \omega_i| L h^{\beta} K\left(\frac{\cdot - \frac{i}{m+1}}{h}\right) \\
&\qquad = \frac{1}{2}\Big(\left| \|\omega'\|_1 - \|\omega\|_1 \right| + d_{\mathrm{ham}}\left(\omega, \omega'\right) \Big) L h^{\beta+1} \int K \\
&\qquad \le m L h^{\beta+1} \, .
\end{aligned}
\tag{5.23}
$$

The KL divergence between $\mathbb{P}_{g_{L,\beta,\omega,h}}$ and $\mathbb{P}_g$, the uniform distribution on $[0,1]$, is bounded as

$$
\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{g_{L,\beta,\omega,h}} \middle\| \mathbb{P}_g\right) = \int_{[0,1]} \ln\left(g_{L,\beta,\omega,h}\right) g_{L,\beta,\omega,h} \\
&= \int_{[0,1]\setminus \cup_{i:\omega_i \neq 0}\left[\frac{i}{m+1} - \frac{h}{2}, \frac{i}{m+1} + \frac{h}{2}\right]} \ln\left(1 - \|\omega\|_1 L h^{\beta+1} \int K\right)\left(1 - \|\omega\|_1 L h^{\beta+1} \int K\right) dt \\
&\qquad + \|\omega\|_1 \int_{-\frac{h}{2}}^{\frac{h}{2}} \ln\left(1 - \|\omega\|_1 L h^{\beta+1} \int K + L h^{\beta} K\left(\frac{t}{h}\right)\right) \\
&\qquad\qquad \left(1 - \|\omega\|_1 L h^{\beta+1} \int K + L h^{\beta} K\left(\frac{t}{h}\right)\right) dt \\
&\overset{\ln(1+\cdot)\le\cdot}{\le} (1 - \|\omega\|_1 h)\left(-\|\omega\|_1 L h^{\beta+1} \int K\right)\left(1 - \|\omega\|_1 L h^{\beta+1} \int K\right) \\
&\qquad + \|\omega\|_1 \int_{-\frac{h}{2}}^{\frac{h}{2}} \left(-\|\omega\|_1 L h^{\beta+1} \int K + L h^{\beta} K\left(\frac{t}{h}\right)\right) \\
&\qquad\qquad \left(1 - \|\omega\|_1 L h^{\beta+1} \int K + L h^{\beta} K\left(\frac{t}{h}\right)\right) dt \\
&\overset{\text{Calculus}}{\le} \|\omega\|_1 L^2 h^{2\beta+1} \int K^2 \le m L^2 h^{2\beta+1} \int K^2 \, .
\end{aligned}
\tag{5.24}
$$

Finally, the squared L^2 distance between $g_{L,\beta,\omega,h}$ and $g_{L,\beta,\omega',h}$ can be lower bounded as,

$$\begin{aligned}
&\int_{[0,1]} \left(g_{L,\beta,\omega,h} - g_{L,\beta,\omega',h}\right)^2 \\
&\quad = \sum_{i=1}^{m} \mathbb{1}_{\omega_i \neq \omega_i'} \\
&\qquad \int_{\frac{i}{m+1}-\frac{h}{2}}^{\frac{i}{m+1}+\frac{h}{2}} \left(Lh^{\beta+1}\left(\|\omega'\|_1 - \|\omega\|_1\right) \int K + (\omega_i - \omega_i') Lh^{\beta} K\left(\frac{t-\frac{i}{m+1}}{h}\right)\right)^2 dt \\
&\quad \geq \sum_{i=1}^{m} \mathbb{1}_{\omega_i \neq \omega_i'} \int_{\frac{i}{m+1}-\frac{h}{2}}^{\frac{i}{m+1}+\frac{h}{2}} \left(Lh^{\beta} K\left(\frac{t-\frac{i}{m+1}}{h}\right) - Lh^{\beta+1}\left|\|\omega\|_1 - \|\omega'\|_1\right| \int K\right)^2 dt \\
&\quad \geq \sum_{i=1}^{m} \mathbb{1}_{\omega_i \neq \omega_i'} \int_{\frac{i}{m+1}-\frac{h}{2}}^{\frac{i}{m+1}+\frac{h}{2}} \left\{ \left(Lh^{\beta} K\left(\frac{t-\frac{i}{m+1}}{h}\right)\right)^2 \right. \\
&\qquad\qquad \left. -2Lh^{\beta} K\left(\frac{t-\frac{i}{m+1}}{h}\right) Lh^{\beta+1}\left|\|\omega\|_1 - \|\omega'\|_1\right| \int K \right\} dt \\
&\quad \geq d_{\text{ham}}\left(\omega,\omega'\right) L^2 h^{2\beta+1} \left(\int K^2 - 2mh\left(\int K\right)^2\right) .
\end{aligned} \tag{5.25}$$

By the Varshamov-Gilbert theorem [Tsybakov, 2009, Lemma 2.7], as long as $m \geq 8$, there exist $M \in \mathbb{N}$ and $\omega^{(0)}, \ldots, \omega^{(M)} \in \{0,1\}^m$ such that $M \geq 2^{m/8}$, $\omega^{(0)} = \{0\}^m$ and $i \neq j \implies d_{\text{ham}}\left(\omega^{(i)}, \omega^{(j)}\right) \geq m/8$. According to (5.25), the family $\left(g_{L,\beta,\omega^{(i)},h}\right)_{i=1,\ldots,M}$ is then a $\Omega = \frac{1}{2}\sqrt{\frac{m}{8} L^2 h^{2\beta+1}\left(\int K^2 - 2mh\left(\int K\right)^2\right)}$ packing of $\Theta^{\text{PSob}}_{L,\beta}$ for the L^2 distance.

Recovering the usual lower-bound. By Equation (5.3) with $\Phi(\cdot) \coloneqq (\cdot)^2$ and $\|\cdot\|$ the L^2 norm,

$$\begin{aligned}
&\inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \sup_{\pi \in \Theta^{\mathrm{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \\
&\quad \geq \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \\
&\qquad \inf_{\hat{\pi} \text{ s.t. } \mathcal{C}} \inf_{\Psi : \Theta^{\mathrm{PSob}}_{L,\beta} \to \{0,1\}} \max_{i=1,\dots,M} \mathbb{P}_{\mathbf{X} \sim \mathbb{P}^{\otimes n}_{g_{L,\beta,\omega^{(i)},h}}, \hat{\pi}} \left(\Psi(\hat{\pi}(\mathbf{X})) \neq i \right) \\
&\quad \overset{\textit{Fact } 3.1.2}{\geq} \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \\
&\qquad \left(1 - \frac{1 + \frac{1}{M} \sum_{1 \leq i \leq M} \mathrm{KL}\left(\mathbb{P}^{\otimes n}_{g_{L,\beta,\omega^{(i)},h}} \middle\| \mathbb{P}^{\otimes n}_{g} \right)}{\ln(M)} \right) \\
&\quad \overset{\text{Tensorization}}{=} \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \\
&\qquad \left(1 - \frac{1 + \frac{n}{M} \sum_{1 \leq i \leq M} \mathrm{KL}\left(\mathbb{P}_{g_{L,\beta,\omega^{(i)},h}} \middle\| \mathbb{P}_{g} \right)}{\ln(M)} \right) \\
&\quad \overset{(5.24)\&M \geq 2^{m/8}}{\geq} \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \left(1 - \frac{1 + nmL^2 h^{2\beta+1} \int K^2}{\ln(2) m/8} \right) .
\end{aligned} \tag{5.26}$$

Finally, setting $m = \left\lceil n^{\frac{1}{2\beta+1}} \right\rceil$ and $h = \frac{c}{m}$ for c small enough gives that, for n big enough,

$$\inf_{\hat{\pi} \ \epsilon\text{-DP}} \sup_{\pi \in \Theta^{\mathrm{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \geq C^{-1} n^{-\frac{2\beta}{2\beta+1}} ,$$

where C is a positive constant depending only on L and β.

ϵ-DP overhead. By the same reduction and Fano's lemma for differential privacy on product distributions (Theorem 3.2.3), we get

$$\begin{aligned}
&\inf_{\hat{\pi} \ \epsilon\text{-DP}} \sup_{\pi \in \Theta^{\mathrm{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \\
&\quad \geq \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \\
&\qquad \left(1 - \frac{1 + \frac{n\epsilon}{M^2} 2 \sum_{1 \leq i,j \leq M} \mathrm{TV}\left(\mathbb{P}_{g_{L,\beta,\omega^{(i)},h}}, \mathbb{P}_{g_{L,\beta,\omega^{(j)},h}} \right)}{\ln(M)} \right) \\
&\quad \overset{(5.23)}{\geq} \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \left(1 - \frac{1 + 2n\epsilon m L h^{\beta+1} \int K}{\ln(2) m/8} \right) .
\end{aligned}$$

Setting $m = \left\lceil (n\epsilon)^{\frac{1}{\beta+1}} \right\rceil$ and $h = \frac{c}{m}$ for c small enough leads to, for $n\epsilon$ big enough,

$$\inf_{\hat{\pi}\ \epsilon\text{-DP}} \sup_{\pi \in \Theta^{\text{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}} \int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \geq C'^{-1} (n\epsilon)^{-\frac{2\beta}{\beta+1}} \ ,$$

where C' is a constant depending only on L and β.

ρ-zCDP overhead. For ρ-zCDP, we present the proof using both Fano's lemma and Assouad's method. We will see that Assouad gives better results.

Fano version. By again the same reduction and Fano's lemma for zero-concentrated differential privacy (Theorem 3.2.4), denoting $t_{i,j} := \text{TV}\left(\mathbb{P}_{g_{L,\beta,\omega^{(i)},h}}, \mathbb{P}_{g_{L,\beta,\omega^{(j)},h}}\right)$, we get

$$\begin{aligned}
&\inf_{\hat{\pi}\ \rho\text{-zCDP}} \sup_{\pi \in \Theta^{\text{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}} \int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \\
&\qquad \geq \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \\
&\qquad\qquad \left(1 - \frac{1 + \frac{n^2\rho}{M^2} 4 \sum_{1\leq i,j\leq M} \frac{1}{n} t_{i,j} + t_{i,j}^2}{\ln(M)} \right) \\
&\qquad \overset{(5.23)}{\geq} \frac{L^2}{32} m h^{2\beta+1} \left(\int K^2 - 2mh \left(\int K \right)^2 \right) \\
&\qquad\qquad \left(1 - \frac{1 + 4n^2\rho \left(\frac{mLh^{\beta+1} \int K}{n} + \left(mLh^{\beta+1} \int K \right)^2 \right)}{\ln(2)m/8} \right) .
\end{aligned}$$

So, by choosing $m = \left\lceil \left(n\sqrt{\rho}\right)^{\frac{2}{2\beta+1}} \right\rceil$ and $h = \frac{c}{m}$ for c small enough, if $n\sqrt{\rho}$ and $\frac{n}{\left(n\sqrt{\rho}\right)^{\frac{2\beta}{2\beta+1}}} = \left(n\sqrt{\rho}\right)^{\frac{1}{2\beta+1}} / \sqrt{\rho}$ are big enough,

$$\inf_{\hat{\pi}\ \epsilon\text{-DP}} \sup_{\pi \in \Theta^{\text{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n},\hat{\pi}} \int_{[0,1]} (\hat{\pi}(\mathbf{X}) - \pi)^2 \geq C''^{-1} \left(n\sqrt{\rho}\right)^{-\frac{2\beta}{\beta+1/2}} \ ,$$

where C'' is a constant depending only on L and β.

Assouad version. From Equation (5.25), we can see that when $h := \frac{c}{m}$ for a positive c that is small enough, the condition expressed in Equation (5.10) is satisfied for $\tau = \Omega(h^{2\beta+1})$. To apply (5.12), the only missing ingredient is to bound the testing difficulties between the mixtures on the hypercube.

In the sequel, $\mathbb{P}_\omega$ is used as a short for $\mathbb{P}_{g_{L,\beta,\omega,h}}$. We need to bound the total variation between the mixtures on the hypercube (denoted $\mathbb{P}_{\omega^{i,0}}$ and $\mathbb{P}_{\omega^{i,1}}$, cf (5.12)) as

$$
\begin{aligned}
&\mathrm{TV}\left(\mathbb{P}_{\omega^{i,0}}, \mathbb{P}_{\omega^{i,1}}\right) \\
&\quad = \frac{1}{2}\frac{1}{2^{m-1}} \int \left| \sum_{\omega \in \{0,1\}^m | \omega_i = 0} g_{L,\beta,\omega,h} - \sum_{\omega \in \{0,1\}^m | \omega_i = 1} g_{L,\beta,\omega,h} \right| \\
&\quad = \frac{1}{2^m} \int \Bigg| \sum_{\omega_1,\dots,\omega_{i-1},\omega_{i+1}\dots,\omega_m \in \{0,1\}} \big(g_{L,\beta,(\omega_1,\dots,\omega_{i-1},0,\omega_{i+1}\dots,\omega_m),h} - \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad g_{L,\beta,(\omega_1,\dots,\omega_{i-1},1,\omega_{i+1}\dots,\omega_m),h} \big) \Bigg| \\
&\quad \leq \frac{1}{2^m} \sum_{\omega_1,\dots,\omega_{i-1},\omega_{i+1}\dots,\omega_m \in \{0,1\}} \int \Big| g_{L,\beta,(\omega_1,\dots,\omega_{i-1},0,\omega_{i+1}\dots,\omega_m),h} - \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad g_{L,\beta,(\omega_1,\dots,\omega_{i-1},1,\omega_{i+1}\dots,\omega_m),h} \Big| \\
&\quad \overset{\textit{Equation (5.23)}}{\leq} \frac{1}{2^m} \sum_{\omega_1,\dots,\omega_{i-1},\omega_{i+1}\dots,\omega_m \in \{0,1\}} 2Lh^{\beta+1} \int K \\
&\quad = O\left(h^{\beta+1}\right) .
\end{aligned}
$$

All in all, by using Le Cam's lemma for product distribution and ρ-zCDP (Theorem 3.2.2), and by leveraging Equation (5.11), with $\tau = \Omega(h^{2\beta+1})$,

$$
\inf_{\hat{\pi}\ \rho\text{-zCDP}} \sup_{\pi \in \Theta^{\mathrm{PSob}}_{L,\beta}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_\pi^{\otimes n}, \hat{\pi}} \int_{[0,1]} \left(\hat{\pi}(\mathbf{X}) - \pi\right)^2 = \Omega\left(mh^{2\beta+1}\right)\left(1 - n\sqrt{\rho} O\left(h^{\beta+1}\right)\right). \quad (5.27)
$$

Setting $h \approx \left(n\sqrt{\rho}\right)^{\frac{-1}{\beta+1}}$ and $m = c/h$ for c small enough concludes the proof by yielding a lower bound $\Omega\left(\left(n\sqrt{\rho}\right)^{-\frac{2\beta}{2\beta+1}}\right)$. □

In comparison with the upper-bounds of Theorem 5.2.3, for ϵ-DP the lower-bound *almost* matches the guarantees of the projection estimator. In particular, the excess of risk in the high privacy regime is close to being optimal. Section 5.2.4 explains how to bridge the gap even more, at the cost of relaxation.

Under ρ-zCDP, the lower-bounds and upper-bounds actually match. We conclude that projection estimators with ρ-zCDP obtain minimax-optimal rates of convergence.

5.2.4 Near minimax optimality via relaxation

An hypothesis that we can make on the sub-optimality of the projection estimator against ϵ-DP mechanisms is that the $l_1$sensitivity of the estimation of N Fourier coefficients scales as N whereas its l_2 sensitivity scales as $\sqrt{N}$. Traditionally, the Gaussian mechanism [Dwork et al., 2006a, Dwork et al., 2006b] has allowed to use the l_2 sensitivity instead of the l_1 one at the cost of introducing a relaxation term δ in the privacy guarantees, leading to (ϵ, δ)-DP.

[Bun & Steinke, 2016] states that if a mechanism $\mathfrak{M}$ is ρ-zCDP, then it is $\left(\rho + 2\sqrt{\rho \ln(1/\delta)}, \delta\right)$-DP for any $\delta > 0$. Applying this result and Theorem 5.2.3 with δ of the form $\frac{1}{n^\gamma}$ for a positive exponent γ gives the following result :

Corollary 5.2.5 (Privacy and utility of (5.20) with relaxation)**.** *For any $\epsilon > 0$ and $\gamma > 0$, defining $\tilde{\rho} := \frac{1}{16}\frac{\epsilon^2}{\ln(n^\gamma)}$ and using $\hat{\pi}^{proj}$ with $Z = \frac{2\sqrt{N}}{\sqrt{\tilde{\rho}}}\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ refers to a random variable following a centered Gaussian distribution of variance 1, leads to an $\left(\epsilon, \frac{1}{n^\gamma}\right)$-DP procedure if $\epsilon \leq 8 \ln(n^\gamma)$. Furthermore, there exists $C_{L,\beta} > 0$, a positive constant that only depends on L and β, such that if N is of the order of $\min\left(n^{\frac{1}{2\beta+1}}, \left(n\sqrt{\tilde{\rho}}\right)^{\frac{1}{\beta+1}}\right)$ then*

$$\sup_{\pi \in \Theta^{PSob}_{L,\beta}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\pi^{\otimes n}, \hat{\pi}^{proj}} \|\hat{\pi}^{proj}(\mathbf{X}) - \pi\|^2_{L^2} \leq C_{L,\beta} \max\left\{n^{-\frac{2\beta}{2\beta+1}}, P_{\beta,\gamma}(\ln(n))(n\epsilon)^{-\frac{2\beta}{\beta+1}}\right\},$$

where $P_{\beta,\gamma}$ is a polynomial expression depending on β and γ.

Proof. Indeed, when $\epsilon \leq 8\ln(n^\gamma)$, $\tilde{\rho} \leq 2\sqrt{\tilde{\rho}\ln(n^\gamma)}$. The resulting mechanism is thus $\left(4\sqrt{\tilde{\rho}\ln(n^\gamma)}, \frac{1}{n^\gamma}\right)$-DP □

In order to understand the implications of this result, one must understand the role of δ in (ϵ, δ)-differential privacy. It is usually interpreted as the probability of the procedure not respecting the ϵ-DP condition [Dwork & Roth, 2014]. Hence, with probability δ, the result is not guaranteed to be private. A general rule of thumb for choosing δ is to take it much smaller than $1/n$ so that each individual of the database only has a small chance of seeing its data leak [Dwork & Roth, 2014]. Choosing $\delta = 1/n^\gamma$ for $\gamma > 1$ is hence considered a good choice for δ.

With this relaxation, the upper-bound of Corollary 5.2.5 matches the lower-bound of Theorem 5.2.4 for ϵ-DP up to polylog factors.

Chapter 6

Quantile function estimation

The origin of this chapter, and the use of the first person. This chapter is based on two articles. The first one [Lalanne et al., 2023d] was written by Aurélien Garivier[1], Clément Gastaud (research engineer at Sarus Technologies [2]), Nicolas Grislain (CEO of Sarus Technologies), Rémi Gribonval[3], and by myself. The second one [Lalanne et al., 2023c] was written by Aurélien Garivier, Rémi Gribonval, and by myself. In this chapter, I will try to respect the following rule : the use of the first person of the plural (we, our, ...) represents all the above-mentioned people. In particular, I will not specify which set of authors contributed to each result for brevity. I encourage the reader to refer to the articles for more clarifications. The use of the first person of the singular (I, my, ...) represents myself.

Any probability distribution $\mathbb{P}$ on $[0,1]$ is fully characterized by its cumulative distribution function (CDF) defined by

$$F_{\mathbb{P}}(t) := \mathbb{P}\big((-\infty, t]\big), \quad \forall t \in \mathbb{R}\,.$$

The central topic of this chapter is the quantile function (QF), $F_{\mathbb{P}}^{-1}$, defined as the generalized inverse of $F_{\mathbb{P}}$:

$$F_{\mathbb{P}}^{-1}(p) = \inf\Big\{t \in \mathbb{R} \mid p \leq F_{\mathbb{P}}(t)\Big\}, \quad \forall p \in [0,1]\,,$$

[1] https://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/index.html/

[2] https://www.sarus.tech/

[3] https://people.irisa.fr/Remi.Gribonval/

with the convention $\inf \emptyset = +\infty$. When $\mathbb{P}$ is absolutely continuous w.r.t. Lebesgue's measure with a density that is bounded away from 0, $F_{\mathbb{P}}$ and $F_{\mathbb{P}}^{-1}$ are bijective and are inverse from one another.

A well-known result is that, under mild hypotheses on $\mathbb{P}$, if $U \sim \mathcal{U}([0,1])$ (U follows a uniform distribution on $[0,1]$), then $F_{\mathbb{P}}^{-1}(U) \sim \mathbb{P}$ [Devroye, 1986]. In other words, knowing $F_{\mathbb{P}}^{-1}$ allows to generate data with distribution $\mathbb{P}$. It makes the estimation of $F_{\mathbb{P}}^{-1}$ a key component in data generation. Indeed, privately learning the quantile function would then allow generating infinitely many new coherent samples at no extra cost on privacy.

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, this section studies the *private* estimation of $F_{\mathbb{P}}^{-1}(p_j)$ from these samples at prescribed values $\{p_1, \ldots, p_m\} \subset (0,1)$.

6.1 Empirical quantiles proxy

Without privacy and under mild hypotheses on the distribution, it is well-known [Van der Vaart, 1998] that for each $p \in (0,1)$, the quantity $X_{(E(np))}$ is a good estimator of $F_{\mathbb{P}}^{-1}(p)$, where $X_{(1)}, \ldots, X_{(n)}$ are the order statistic of $X_1, \ldots, X_n$ (i.e. a permutation of the observations such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$) and $E(x)$ denotes the largest integer smaller or equal to x. The quantity $X_{(E(np))}$ is called the empirical (as opposed to statistical) quantile of the dataset $(X_1, \ldots, X_n)$ (as opposed to the distribution $\mathbb{P}$) of order p. In this chapter, we will use private estimators of the empirical quantiles as proxies for private estimators of the statistical ones.

While the computation of private *empirical* quantiles has led to a rich literature, much less is known on the statistical properties of the resulting algorithms seen as estimators of the *statistical* quantiles of an underlying distribution, compared to more traditional ways of estimating a distribution.

Early approaches for solving the private empirical quantile computation used the Laplace mechanism [Dwork et al., 2006a, Dwork et al., 2006b] but the high sensitivity of the quantile query made it of poor utility. Smoothed sensitivity-based approaches followed [Nissim et al., 2007] and managed to achieve greatly improved utility.

The current state of the art for the computation of *a single empirical private quantile* [Smith, 2011] is an instantiation of the so-called exponential mechanism [McSherry & Talwar, 2007] (see Chapter 2) with a specific utility function that we will denote QExp (for exponential quantile) in the rest of this section. It is implemented in many DP software libraries [Allen, , IBM,].

For the computation of *multiple empirical private quantiles*, the problem gets more complicated. Indeed, with differential privacy, every access to the dataset has to be accounted for in the overall privacy budget. Luckily, and part of the reasons why differential privacy became so popular in the first place, composition theorems [Dwork et al., 2006b, Kairouz et al., 2015, Dong et al., 2019, Dong et al., 2020, Abadi et al., 2016] (see Chapter 2) give general rules for characterizing the privacy budget of an algorithm depending on the privacy budgets of its subroutines. It is hence possible to estimate multiple empirical quantiles privately by separately estimating each empirical quantile privately (using the techniques presented above) and by updating the overall privacy budget with composition theorems. The algorithm IndExp (see Section 6.1.3) builds on this framework. However, recent research has shown that such approaches are suboptimal. For instance, [Gillenwater et al., 2021] presented an algorithm (JointExp) based on the exponential mechanism again, with a utility function tailored for the joint computation of multiple private empirical quantiles directly. JointExp became the state of the art for about a year. It can be seen as a generalization of QExp, and the associated clever sampling algorithm is interesting in itself. Yet, more recently, [Kaplan et al., 2022] demonstrated that an ingenious use of a composition theorem (as opposed to a more straightforward direct independent application) yields a simple recursive computation using QExp that achieves the best empirical performance to date. We will refer to their algorithm as RecExp (for recursive exponential). Furthermore, contrary to JointExp, RecExp is endowed with strong utility guarantees [Kaplan et al., 2022] in terms of the quality of estimation of the *empirical* quantiles. There is still ongoing work studying this problem, in particular for getting rid of the bounded assumption [Durfee, 2023].

6.1.1 Motivations for empirical quantiles

In fact, $E(np_j) \approx np_j$ has no link with $F_{\mathbb{P}}$ a priori. In contrast, from a statistical point of view, the quantity of interest is the deviation w.r.t. the statistical quantiles $(F_{\mathbb{P}}^{-1}(p_1), \dots, F_{\mathbb{P}}^{-1}(p_m))$. We circumvent that difficulty with the following general purpose lemma :

Lemma 6.1.1 (Concentration of empirical quantiles)**.** *If* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* $\pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely, then for any* $p \in (0,1)$ *and* $\gamma > 0$ *such that* $\gamma < \min\left(F_X^{-1}(p), 1 - F_X^{-1}(p)\right)$*, we have*

$$\mathbb{P}\Big(\sup_{k \in J} |X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\Big) \leq 2e^{-\frac{\gamma^2 \underline{\pi}^2}{8p} n} + 2e^{-\frac{\gamma^2 \underline{\pi}^2}{8(1-p)} n},$$

where

$$J := \left\{ \max\left(-E(np) + 1, -E\left(\frac{1}{2} n\gamma\underline{\pi}\right) + 1\right), \dots, \min\left(n - E(np), E\left(\frac{1}{2} n\gamma\underline{\pi}\right) - 1\right) \right\}.$$

The integer set J may be viewed as an error buffer : As long as an algorithm returns a point with an *order* error falling into J (compared to $E(np)$), the error on the *statistical* estimation will be small.

Proof. We define

$$\bar{N} := \sum_{i=1}^{n} \mathbb{1}_{(F_X^{-1}(p)+\gamma,+\infty)}(X_i) \ .$$

Let $k \in \{-E(np)+1, \dots, n-E(np)\}$. We have the following event inclusion:

$$\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) \subset \left(\bar{N} \geq n-(E(np)+k)\right) \subset \left(\bar{N} \geq n(1-p)-k-1\right) \ .$$

$\bar{N}$ being a sum of independent Bernoulli random variables, we introduce $\eta := 1-p-\gamma\underline{\pi}$, a natural upper bound on the probability of success of each of these Bernoulli random variables. Hence, by multiplicative Chernoff bounds, whenever $\frac{\gamma\underline{\pi}}{\eta} - \frac{k+1}{n\eta} \geq 0$, which is equivalent to $k \leq n\gamma\underline{\pi} - 1$,

$$\begin{aligned}\mathbb{P}\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) &\leq \mathbb{P}\left(\bar{N} \geq n\eta\left(1+\frac{\gamma\underline{\pi}}{\eta}-\frac{k+1}{n\eta}\right)\right) \\ &\leq e^{-n\eta\left(\frac{\gamma\underline{\pi}}{\eta}-\frac{k+1}{n\eta}\right)^2 / \left(2+\frac{\gamma\underline{\pi}}{\eta}-\frac{k+1}{n\eta}\right)} \ .\end{aligned}$$

By going further and imposing that $k+1 \leq \frac{1}{2}n\gamma\underline{\pi}$, we get

$$\mathbb{P}\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) \leq e^{-\frac{n\eta}{4}\left(\frac{\gamma\underline{\pi}}{\eta}\right)^2 / \left(2+\frac{\gamma\underline{\pi}}{2\eta}\right)} \ .$$

Finally, by noticing that $\eta\left(\frac{\gamma\underline{\pi}}{\eta}\right)^2 / \left(2+\frac{\gamma\underline{\pi}}{2\eta}\right) = \frac{\gamma^2\underline{\pi}^2}{2(1-p)-\frac{3}{2}\gamma\underline{\pi}} \geq \frac{\gamma^2\underline{\pi}^2}{2(1-p)}$,

$$\mathbb{P}\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) \leq e^{-\frac{\gamma^2\underline{\pi}^2}{8(1-p)}n} \ .$$

Now, looking at the other inequality, we define

$$\underline{N} := \sum_{i=1}^{n} \mathbb{1}_{(-\infty,F_X^{-1}(p)-\gamma)}(X_i) \ .$$

Like previously,

$$\left(X_{(E(np)+k)} < F_X^{-1}(p)-\gamma\right) \subset \left(\underline{N} \geq E(np)+k\right) \subset \left(\underline{N} \geq np+k-1\right) \ .$$

With the exact same techniques as previously, imposing the condition $k-1 \geq -\frac{1}{2}n\gamma\underline{\pi}$ gives

$$\mathbb{P}\left(X_{(E(np)+k)} < F_X^{-1}(p)-\gamma\right) \leq e^{-\frac{\gamma^2\underline{\pi}^2}{8p}n} \ .$$

Thus, under the various conditions specified for k, by union bound,

$$\mathbb{P}\left(|X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\right) \leq e^{-\frac{\gamma^2 \underline{\pi}^2}{8p}n} + e^{-\frac{\gamma^2 \underline{\pi}^2}{8(1-p)}n} .$$

Now define $I := \{k \in \{-E(np), \dots, n - E(np)\} | |X_{(E(np)+k)} - F_X^{-1}(p)| \leq \gamma\}$. Notice that since $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, I is an integer interval. Which means that if $a \in I \leq b \in I$, then $[a, b] \cap \mathbb{Z} \subset I$. As a consequence, if $|X_{(E(np)+k)} - F_X^{-1}(p)| \leq \gamma$ for two integers k_1 and k_2, it is also the case for all the integers between them. By union bound, we get

$$\mathbb{P}\left(\sup_{k \in J} |X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\right) \leq 2e^{-\frac{\gamma^2 \underline{\pi}^2}{8p}n} + 2e^{-\frac{\gamma^2 \underline{\pi}^2}{8(1-p)}n} ,$$

where

$$J := \left\{\max\left(-E(np) + 1, -E\left(\frac{1}{2}n\gamma\underline{\pi}\right) + 1\right), \dots, \min\left(n - E(np), E\left(\frac{1}{2}n\gamma\underline{\pi}\right) - 1\right)\right\} .$$

□

This result strengthens the approach of using private empirical quantiles as proxies for the statistical ones. Furthermore, it will be at the core of many of our results.

6.1.2 Exponential quantile

Given n points $X_1, \dots, X_n \in [0, 1]$ and $p \in (0, 1)$, the QExp mechanism, introduced by [Smith, 2011], is an instantiation of the exponential mechanism w.r.t. μ the Lebesgue's measure on $[0, 1]$, with utility function u_{QExp} such that, for any $q \in [0, 1]$,

$$u_{\text{QExp}}\big((X_1, \dots, X_n), q\big) := -\big|\#\left(\{i | X_i < q\}\right) - E(np)\big| .$$

The sensitivity of u_{QExp} is 1 for both replacement and addition/deletion neighboring relations. As the density of QExp is constant on all the intervals of the form $(X_{(i)}, X_{(i+1)})$, a sampling algorithm for QExp is to first sample an interval (which can be done by sampling a point in a finite space) and then to uniformly sample a point in this interval. This algorithm has complexity $O(n)$ if the points are sorted and $O(n \log n)$ otherwise. Its utility (as measured by a so-called "empirical error") is controlled, cf [Kaplan et al., 2022] Lemma A.1. This is summarized as follows

Fact 6.1.2 (Empirical Error of QExp)**.** *Consider fixed real numbers $X_1, \dots, X_n \in [0, 1]$ that satisfy $\min_i X_{(i+1)} - X_{(i)} \geq \Delta > 0$ with the convention $X_{(0)} = 0$ and $X_{(n+1)} = 1$. Denote q the (random) output of QExp on this dataset, for the estimation of a single empirical quantile of order p, and*

$$\mathfrak{E} := \left|\#\left(\{i | X_i < q\}\right) - E(np)\right| ,$$

the empirical error *of QExp. For any* $\beta \in (0,1)$*, we have*

$$\mathbb{P}\left(\mathfrak{E} \geq 2\frac{\ln\left(\frac{1}{\Delta}\right) + \ln\left(\frac{1}{\beta}\right)}{\epsilon}\right) \leq \beta \ .$$

This private mechanism for empirical quantiles is used in many of the DP libraries [Wilson et al., 2019, Allen, , IBM, , Labs, 2022, Berghel et al., 2022, OpenMinded, 2022, Johnson et al., 2020].

6.1.3 Independent exponential quantiles

Given $p_1, \ldots, p_m \in (0,1)$, IndExp privately estimates the empirical quantiles of order $p_1, \ldots, p_m$ by evaluating each quantile independently using QExp and the simple composition property (see Chapter 2). Each quantile is estimated with a privacy budget of $\frac{\epsilon}{m}$. The complexity is $O(mn)$ if the points are sorted, $O(mn + n\log n)$ otherwise.

6.1.4 Joint exponential quantiles

For JointExp, we suppose that the support of the distribution is $[a,b]$ and not simply $[0,1]$. The readers may read those results with $a=0$ and $b=1$ in mind. We suppose that the vector containing the orders of the quantiles to estimate is sorted ($\mathbf{p} = (p_1, \ldots, p_m) \in (0,1)^{m\nearrow}$), and that $\mathbf{X} = (X_1, \ldots, X_m)$ is sorted as well. For any $\mathbf{X} \in \mathcal{X}^n$, $\mathbf{q} = (q_1, \ldots, q_m)$ sorted output candidate, and $\mathbf{p} \in (0,1)^m$, we use the convention $X_{i\leq 0} = q_{i\leq 0} = a$, $X_{i\geq n+1} = q_{i\geq m+1} = b$, $p_{i\leq 0} = 0$ and $p_{i\geq m+1} = 1$. Finally, for the brevity of notation, vectors are interpreted when needed as the set containing their components.

The JointExp mechanism (introduced by [Gillenwater et al., 2021]) is an instanciation of an exponential mechanism with utility function

$$u_{\mathrm{JE}}(\mathbf{X}, \mathbf{q}) := -\frac{1}{2}\sum_{i=1}^{m+1} \left|\delta^{\mathrm{JE}}(i, \mathbf{X}, \mathbf{q})\right| \ ,$$

(which is of sensitivity 1 for the replacement neighboring relation and of 1/2 for the addition/deletion one) where

$$\delta^{\mathrm{JE}}(i, \mathbf{X}, \mathbf{q}) := n(p_i - p_{i-1}) - \#\left(\mathbf{X} \cap (q_{i-1}, q_i]\right) \ .$$

This mechanism works by penalizing the result whenever the number of data points in each quantile interval ($\#\left(\mathbf{X} \cap (q_{i-1}, q_i]\right)$) deviates from what should be expected ($n(p_i - p_{i-1})$). We can notice that this mechanism is the same as QExp when $m = 1$.

In the original article, [Gillenwater et al., 2021] also provide a sampling algorithm of complexity $O(mn\log(n) + nm^2)$.

6.1.5 Recursive exponential quantiles

Introduced by [Kaplan et al., 2022], RecExp is based on the following idea : Suppose that we already have a private estimate, q_i, of the empirical quantile of order p_i for a given i. Estimating the empirical quantiles of orders $p_j > p_i$ should be possible by only looking at the data points that are bigger than q_i, and similarly for the empirical quantiles of orders $p_j < p_i$. Representing this process as a tree, the addition or removal of an element in the dataset only affects at most one child of each node and at most one node per level of depth in the tree. The "per-level" composition of mechanisms comes for free in terms of privacy budget, hence only the tree depth matters for composition. By choosing a certain order on the quantiles to estimate, this depth can be bounded by $\log_2 m + 1$. More details can be found in the original article [Kaplan et al., 2022].

When using QExp with privacy budget $\frac{\epsilon}{\log_2 m+1}$ for estimating the individual empirical quantiles, RecExp is ϵ-DP with the addition/removal neighborhing relation. This remains valid with the replacement relation if we replace ϵ by $\epsilon/2$, as the replacement relation can be seen as a two-steps addition/removal relation. RecExp has a complexity of $O(n \log m)$ if the points are sorted and $O(n \log(nm))$ otherwise. The following control of its empirical error is adapted from [Kaplan et al., 2022] Theorem 3.3.

Fact 6.1.3 (Empirical Error of RecExp)**.** *Consider fixed real numbers $X_1, \dots, X_n \in [0,1]$ that satisfy $\min_i X_{(i+1)} - X_{(i)} \geq \Delta > 0$ with the convention $X_{(0)} = 0$ and $X_{(n+1)} = 1$. Denote $(q_1, \dots, q_m)$ the (random) return of RecExp on this dataset, for the estimation of m empirical quantiles of orders $(p_1, \dots, p_m)$, and*

$$\mathfrak{E} := \max_j \left| \#\left(\{i | X_i < q_j\}\right) - E(np_j) \right| ,$$

the empirical error *of RecExp. For any $\beta \in (0,1)$, we have*

$$\mathbb{P}\left(\mathfrak{E} \geq 2(\log_2 m + 1)^2 \frac{\ln\left(\frac{1}{\Delta}\right) + \ln(m) + \ln\left(\frac{1}{\beta}\right)}{\epsilon} \right) \leq \beta .$$

6.1.6 Quantiles with inverse sensitivity

At first glance, there is no connection between the theory of the Inverse Sensitivity (see Chapter 2 for a reminder) and JointExp. The first one is born from the need to build a general mechanism that is endowed with optimality properties [Asi & Duchi, 2020b, Asi & Duchi, 2020a] for a broad class of problems, while the second comes from the idea that good empirical quantiles should separate the data points proportionally. In the case of the estimation of a single quantile (i.e. $m = 1$), it was observed [Asi & Duchi, 2020b] that the two algorithms are similar. Here we prove that, up to minor differences, this remains true with an arbitrary number of quantiles. For this, we provide the precise expression of the inverse sensitivity function for the multiquantile problem.

Recall that the theory of inverse sensitivity aims at mimicing the behavior of a deterministic function f that is defined on the set of all datasets. For the problem of estimating multiple empirical quantiles, we thus need to define such function. We propose to apply the theory of inverse sensitivity to the function

$$f : \mathbf{X} \mapsto (X_{(\lceil np_1 \rceil)}, \dots, X_{(\lceil np_m \rceil)}) \, .$$

For the inverse sensitivity, we will always work with the replacement neighboring relation. Deriving the expression of the inverse sensitivity for a dataset $\mathbf{X}$ and an output candidate $\mathbf{q}$ boils down to answering the question: What is the minimal number of points from $\mathbf{X}$ that need to be changed in order to obtain a vector that has $\mathbf{q}$ as its empirical quantiles? Theorem 6.1.4 solves this question for Lebesgue-almost-any $\mathbf{q}$.

Theorem 6.1.4. *For any $\mathbf{X} \in \mathfrak{X}^n$ and $\mathbf{q} \in ([a,b] \setminus \mathbf{X})^{m\nearrow}$ without collision,*

$$\begin{aligned} -u_{IS}(\mathbf{X}, \mathbf{q}) = \frac{1}{2} \sum_{i=2}^{m+1} |\delta(i, \mathbf{X}, \mathbf{q})| + \sum_{i=2}^{m} \mathbb{1}_{\mathbb{R}_+} (\delta(i, \mathbf{X}, \mathbf{q})) \\ + \frac{1}{2} |\delta_{\texttt{closed}}(1, \mathbf{X}, \mathbf{q})| + \mathbb{1}_{\mathbb{R}_+} (\delta_{\texttt{closed}}(1, \mathbf{X}, \mathbf{q})) \end{aligned}$$

with

$$\begin{aligned} \delta(i, \mathbf{X}, \mathbf{q}) &= \#(\mathbf{X} \cap (q_{i-1}, q_i]) - (\lceil np_i \rceil - \lceil np_{i-1} \rceil) \\ \delta_{\texttt{closed}}(i, \mathbf{X}, \mathbf{q}) &= \#(\mathbf{X} \cap [q_{i-1}, q_i]) - (\lceil np_i \rceil - \lceil np_{i-1} \rceil) \, . \end{aligned}$$

Proof. If $\mathbf{Y} \in f^{-1}(\mathbf{q})$ then:

- Each "bin" has the right number of points: $\delta(i, \mathbf{Y}, \mathbf{q}) = 0$, $i \in \{2 \dots m+1\}$, and $\delta_{\texttt{closed}}(1, \mathbf{Y}, \mathbf{q}) = 0$.

- Every point of $\mathbf{q}$ appears in $\mathbf{Y}$: $\mathbf{q} \subseteq \mathbf{Y}$.

Then we can understand the modifications that have to be made to $\mathbf{X}$ in order to obtain a $\mathbf{Y} \in Q^{-1}(\mathbf{q})$. For the first condition, some points have to be moved from bins in excess to bins in deficit. This procedure accounts for $\sum_{i=2}^{m+1} \delta(i, \mathbf{X}, \mathbf{q})_+ + \delta_{\text{closed}}(1, \mathbf{X}, \mathbf{q})_+$ operations which can be reformulated as $\frac{1}{2} \sum_{i=2}^{m+1} |\delta(i, \mathbf{X}, \mathbf{q})| + \frac{1}{2} |\delta_{\text{closed}}(1, \mathbf{X}, \mathbf{q})|$. For the second condition, we have to make sure that for all i, q_i belongs to the dataset. For a bin in strict deficit, at least a point has to be added to it due to the first condition. Hence, we can make sure to add the associated quantile at no extra cost. For a bin in excess on the

other hand, since by hypothesis $\mathbf{q} \cap \mathbf{X} = \emptyset$, a point in the bin will have to be replaced by the associated quantile at an extra cost of 1. In the end, we find the desired result. □

The case when $\mathbf{q}$ has collisions or shares some common points with the dataset is more difficult. Luckily, those cases can be neglected when considering the sampling mechanism. Indeed, the inverse sensitivity mechanism has a density that is absolutely continuous w.r.t. Lebesgue's measure, the expression of the resulting mechanism can be further simplified (see Corollary 6.1.5) by modifying the density on outcomes of null Lebesgue measure.

Corollary 6.1.5. *For any* $\mathbf{X} \in \mathfrak{X}^n$, $\mathcal{E}_{u_{IS}}^{(2/\epsilon)}(\mathbf{X})$ *has the same output distribution as* $\mathcal{E}_{\tilde{u}_{IS}}^{(2/\epsilon)}(\mathbf{X})$ *where* $\forall \mathbf{X} \in \mathfrak{X}^n$, $\forall \mathbf{q} \in O$,

$$-\tilde{u}_{IS}(\mathbf{X}, \mathbf{q}) = \frac{1}{2} \sum_{i=1}^{m+1} |\delta(i, \mathbf{X}, \mathbf{q})| + \sum_{i=1}^{m} \mathbb{1}_{\mathbb{R}_+}(\delta(i, \mathbf{X}, \mathbf{q})) \ .$$

Remark 6.1.6. One can check that $|\tilde{u}_{\mathrm{IS}}(\mathbf{X}, \mathbf{q}) - u_{\mathrm{JE}}(\mathbf{X}, \mathbf{q})| \leq 2(m+1)$ and thus the distributions differ significantly only on outcomes of high utility (when the number of misclassified points is of the order $O(m)$). The bad outcomes are almost equally penalized and for this reason, we can expect the two algorithms to perform almost identically when n is large enough. This is indeed confirmed by numerical examples, as illustrated in Section 6.3.5. As a consequence, we will mainly focus on JointExp for the rest of this chapter, all the results being applicable to IS as well (with some minor tweaks).

Sampling from that distribution. For simplicity, $\mathbf{X}$ is assumed to be sorted. The sampling density of the inverse sensitivity mechanism is constant on sets $\big([X_{i_1}, X_{i_1+1}) \times \ldots \times \cdots \times [X_{i_m}, X_{i_m+1})\big) \cap [a, b]^{m\nearrow}$ for $\mathbf{i} = (i_1, \ldots, i_m) \in O'$ where

$$O' = \{\mathbf{i} \in \{0, \ldots, n\}^m, 0 \leq i_1 \leq \cdots \leq i_m \leq m\} \ .$$

Hence, a finite sampling algorithm for the inverse sensitivity mechanism is to:

- sample $\mathbf{i} = (i_1, \ldots, i_m) \in O'$ under $\mathbb{P}_{O'}$;
- sample q'_j uniformly in $[X_{i_j}, X_{i_j+1})$, independently for all j in $\{1 \ldots m\}$;
- output $(q'_j)_{j \in \{1 \ldots m\}}$ sorted by increasing order;

with the probability $\mathbb{P}_{O'}$ defined on O' as

$$\mathbb{P}_{O'}(\mathbf{i}) \propto \frac{1}{\gamma(\mathbf{i})} \Pi_{j=1}^{m+1} \phi(i_{j-1}, i_j, j) \Pi_{j=1}^{m} \tau(i_j) \tag{6.1}$$

where, if we denote by $\text{count}_{\mathbf{i}}(i)$ the number of occurrences of integer i in the ordered tuple $\mathbf{i}$,

$$\forall \mathbf{i} \in O', \gamma(\mathbf{i}) = \Pi_{i=0}^{m} \text{count}_{\mathbf{i}}(i)! \ ,$$
$$\forall i \in \{0, \dots, m\}, \tau(i) = X_{i+1} - X_i \ ,$$

and for $0 \leq i, i' \leq m$ and $1 \leq j \leq m+1$,

$$\phi(i, i', j) = \begin{cases} 0, & \text{if } i' < i \\ e^{-\frac{\epsilon}{2}\left(\frac{1}{2}|\hat{\delta}(i,i',m+1)|\right)}, & \text{if } j = m+1 \\ e^{-\frac{\epsilon}{2}\left(\frac{1}{2}|\hat{\delta}(i,i',j)|\right) + \mathbb{1}_{\mathbb{R}_+}(\hat{\delta}(i,i',j))}, & \text{otherwise} \end{cases}$$

with $\hat{\delta}(i, i', j) = i' - i - (\lceil np_j \rceil - \lceil np_{j-1} \rceil)$.

Since O' has a finite cardinality bounded by $(n+1)^m$, it is possible to compute the probability of all the elements in that space and to sample this way. However, the fact that this complexity is exponential in m makes it unusable in practice. [Gillenwater et al., 2021] present an algorithm that allows to sample from any distribution that factorizes in an analog form of (6.1) that has a complexity (both in time and space) of $O(n^2 m + m^2 n)$. Furthermore, if the function $\phi(i, i', j)$ can be rewritten as $\phi'(i' - i, j)$ (which is the case in our problem), the complexity becomes $O(mn \log n + m^2 n)$. Overall, in order to sample efficiently from the inverse sensitivity mechanism, one can use Algorithm 1 proposed by [Gillenwater et al., 2021] by taking great care of using a sensitivity of 1 (instead of 2) and by replacing the function ϕ by the one used in this chapter.

6.2 Statistical utility

As promised at the beginning of the chapter, we now derive bounds on the statistical utility of the above-mentioned mechanisms when used as statistical estimators of the underlying quantiles.

6.2.1 Controlling the gaps

A difficulty is that the guaranteed utility of the empirical quantiles depends on the minimum gap in the order statistics Δ. For many distributions, this quantity can be as small as we want, and the guarantees on the empirical error of QExp, IndExp and RecExp can be made as poor as we want [Lalanne et al., 2023d]. However, by imposing a simple condition on the density, the following lemma tells that the minimum gap in the order statistic is "not too small".

Lemma 6.2.1 (Concentration of the gaps)**.** *Consider $n \geq 1$ and $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely. Denote $\Delta_i = X_{(i)} - X_{(i-1)}$, $1 \leq i \leq n+1$, with the convention $X_{(0)} = 0$ and $X_{(n+1)} = 1$. For any $\gamma > 0$ such that $\gamma < \frac{1}{4\bar{\pi}}$, we have*

$$\mathbb{P}\left(\min_{i=1}^{n+1} \Delta_i > \frac{\gamma}{n^2}\right) \geq e^{-4\bar{\pi}\gamma} .$$

Proof. The following fact is a direct consequence of Lemma 2.1 in Chapter 5 of [Devroye, 1986].

Fact 6.2.2 (Concentration of the gaps for uniform samples)**.** *Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} U([0,1])$, the uniform distribution on $[0,1]$. Denoting $\Delta_1 := X_{(1)}, \Delta_2 := X_{(2)} - X_{(1)}, \dots, \Delta_n := X_{(n)} - X_{(n-1)}$, and $\Delta_{n+1} := 1 - X_{(n)}$, for any $\gamma > 0$ such that $\gamma < \frac{1}{n+1}$,*

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) = (1 - (n+1)\gamma)^n .$$

We give a proof here for completeness. The first step consists in characterizing the distribution of $(\Delta_1, \dots, \Delta_n)$. Let $h : \mathbb{R}^n \to \mathbb{R}$ be a positive Borelian function. By the transfer theorem,

$$\int h(\Delta_1, \dots, \Delta_n) d\mathbb{P}(\Delta_1, \dots, \Delta_n) = \int h(X_{(1)}, X_{(2)} - X_{(1)}, \dots, X_{(n)} - X_{(n-1)}) d\mathbb{P}(X_{(1)}, \dots, X_{(n)}) .$$

Furthermore, $(X_{(1)}, \dots, X_{(n)})$ follows a uniform distribution on the set of n ordered points in $[0,1]$. Hence,

$$\int h(\Delta_1, \dots, \Delta_n) d\mathbb{P}(\Delta_1, \dots, \Delta_n) = n! \int h(X_1, X_2 - X_1, \dots, X_n - X_{n-1}) \mathbb{1}_{0 \leq X_1 \leq \dots \leq X_n \leq 1} dX_1 \dots dX_n .$$

Finally, the variable swap $\delta_1 = X_1, \delta_2 = X_2 - X_1, \dots, \delta_n = X_n - X_{n_1}$ that has a jacobian of 1, same as its inverse (both transformations are triangular matrices with only 1's on the diagonal), gives that

$$\int h(\Delta_1, \dots, \Delta_n) d\mathbb{P}(\Delta_1, \dots, \Delta_n) = n! \int h(\delta_1, \dots, \delta_n) \mathbb{1}_{0 \leq \delta_1, \dots, 0 \leq \delta_n, \sum_{i=1}^n \delta_i \leq 1} d\delta_1 \dots d\delta_n .$$

The last equation means that $(\Delta_1, \dots, \Delta_n)$ follows a uniform distribution on the simplex $\left\{0 \leq \Delta_1, \dots, \Delta_n \leq 1, \sum_{i=1}^n \Delta_i \leq 1\right\}$. The probability $\mathbb{P}\left(\min_i \Delta_i > \gamma\right)$ may now be computed as

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) = n! \int \mathbb{1}_{\gamma < \delta_1, \dots, \gamma < \delta_n, \sum_{i=1}^n \delta_i < 1-\gamma} \mathbb{1}_{0 \leq \delta_1, \dots, 0 \leq \delta_n, \sum_{i=1}^n \delta_i \leq 1} d\delta_1 \dots d\delta_n,$$

and by considering the variable swap $\delta_i' := \frac{\delta_i - \gamma}{1-(n+1)\gamma}$ (which is separable) of which the jacobian of the inverse is $(1-(n+1)\gamma)^n$,

$$\begin{aligned}\mathbb{P}\left(\min_i \Delta_i > \gamma\right) &= n!(1-(n+1)\gamma)^n \int \mathbb{1}_{0 < \delta_1', \dots, 0 < \delta_n', \sum_{i=1}^n \delta_i' < 1} d\delta_1' \dots d\delta_n' \\ &= (1-(n+1)\gamma)^n \ .\end{aligned}$$

This concludes the proof of Fact 6.2.2. Now, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely. In particular, the data is not necessary uniform. By a coupling argument, if $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} U([0,1])$, $\left(F_\pi^{-1}(U_1), \dots, F_\pi^{-1}(U_n)\right)$ has the same distribution as $(X_1, \dots, X_n)$. We can furthermore notice that

$$\forall p, q \in (0,1), \epsilon > 0, \quad , |p-q| > \epsilon \implies \left|F_\pi^{-1}(p) - F_\pi^{-1}(q)\right| > \frac{\epsilon}{\bar{\pi}} \ .$$

Indeed, the lower bound $\pi \geq \underline{\pi}$ ensures that F_π is a bijection and that so does its inverse. The upper bound $\bar{\pi} \geq \pi$ ensures that F_π cannot grow too fast, and thus that its inverse is not too flat. Formally,

$$\forall a, b, \quad |F_\pi(b) - F_\pi(a)| = \left|\int_a^b \pi\right| \leq \bar{\pi}|b-a|.$$

In particular, it holds for $b = F_\pi^{-1}(p)$ and $a = F_\pi^{-1}(q)$.

Consequently, if $\Delta_1' := U_{(1)}, \Delta_2' := U_{(2)} - U_{(1)}, \dots, \Delta_n' := U_{(n)} - U_{(n-1)}$, and $\Delta_{n+1}' := 1 - U_{(n)}$,

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) \geq \mathbb{P}\left(\min_i \Delta_i' > \bar{\pi}\gamma\right) = (1-(n+1)\bar{\pi}\gamma)^n \ .$$

Finally, let us simplify this expression to a easy-to-handle one. If $\gamma < \frac{n}{2\bar{\pi}}$,

$$\mathbb{P}\left(\min_i \Delta_i > \frac{\gamma}{n^2}\right) = \left(1 - \frac{n+1}{n}\frac{\bar{\pi}\gamma}{n}\right)^n \geq \left(1 - \frac{2n}{n}\frac{\bar{\pi}\gamma}{n}\right)^n = \left(1 - \frac{2\bar{\pi}\gamma}{n}\right)^n \ .$$

Furthermore, for any $x \in (0, 1/2)$ and $n \geq 1$, by the Taylor-Lagrange formula, there exist $c \in \left(-\frac{x}{n}, 0\right)$

$$\left(1 - \frac{x}{n}\right)^n = e^{n\ln\left(1-\frac{x}{n}\right)} = e^{n\left(-\frac{x}{n} - \frac{1}{2}\frac{1}{(1+c)^2}\frac{x^2}{n^2}\right)}$$

And so, when $n \geq 1$,

$$\left(1 - \frac{x}{n}\right)^n \geq e^{-2x}$$

In definitive, when $n \geq 1$ and $\gamma < \frac{1}{4\bar{\pi}}$

$$\mathbb{P}\left(\min_i \Delta_i > \frac{\gamma}{n^2}\right) \geq e^{-4\bar{\pi}\gamma} .$$

□

6.2.2 High probability bounds

This subsection leverages the previous results in order to provide high-probability bounds for QExp, IndExp and RecExp.

Theorem 6.2.3 (Statistical utility of QExp)**.** *Consider $n \geq 1$ and $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely. Denote q the (random) result of QExp on $(X_1, \ldots, X_n)$ for the estimation of the quantile of order p, where $\min(p, 1-p) > 2/n$. For any $\gamma \in (0, \frac{2\min(p,1-p)}{\underline{\pi}})$*

$$\mathbb{P}\big(|q - F_\pi^{-1}(p)| > \gamma\big) \leq 4n\sqrt{2e\bar{\pi}}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32}} + 4e^{-\frac{\gamma^2\underline{\pi}^2}{8}n} .$$

Proof. Let us start with the general idea. We fix a buffer size K and define QC (for quantile concentration) the event "Any error of at most K points in the order statistic compared to $X_{(E(np))}$ induces an error of at most γ on the statistical estimation of $F_\pi^{-1}(p)$". The probability $\mathbb{P}(QC^c)$ is controlled by Lemma 6.1.1.
We fix a gap size $\Delta > 0$ and define the event G (for gaps) $\min_i \Delta_i \geq \Delta$, so that $\mathbb{P}(G^c)$ is controlled by Lemma 6.2.1.
Then, we notice that

$$\begin{aligned}
&\mathbb{P}\big(|q - F_\pi^{-1}(p)| > \gamma\big) \\
&\quad \leq \mathbb{P}\big(|q - F_\pi^{-1}(p)| > \gamma \big| QC, G\big) + \mathbb{P}(QC^c) + \mathbb{P}(G^c) \\
&\quad \leq \mathbb{P}\big(\mathfrak{E} \geq K+1 \big| QC, G\big) + \mathbb{P}(QC^c) + \mathbb{P}(G^c) \ ,
\end{aligned}$$

where $\mathfrak{E}$ refers to the empirical error of QExp. Using Fact 6.1.2 for a suited β controls $\mathbb{P}\big(\mathfrak{E} \geq K+1 \Big| QC, G\big)$. Tuning the values of K, Δ and β concludes the proof.

For simplicity, let us assume that $E\left(\frac{1}{2}n\gamma\underline{\pi}\right) - 1 \leq \min\left(E(np) - 1, n - E(np)\right)$, which is for instance the case when $\gamma < \frac{2\min(p,1-p)}{\underline{\pi}}$, which we suppose. Furthermore, suppose that

$\frac{1}{2}n\gamma\underline{\pi} \geq 2$, which is for instance the case when $n > 2/\min(p, 1-p)$ thank to the hypothesis on γ. By noting $K := E\left(\frac{1}{4}n\gamma\underline{\pi}\right)$, Lemma 6.1.1 says that,

$$\mathbb{P}\left(\sup_{k\in\{-K,\dots,K\}} |X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\right) \leq 4e^{-\frac{\gamma^2\underline{\pi}^2}{8\max(p,(1-p))}n},$$

We call QC (for *quantile concentration*) the complementary of this last event. Let $\delta > 0$ that satisfies $\delta < \frac{1}{4\bar{\pi}}$. We define the event $G := \left(\min_i \Delta_i > \frac{\delta}{n^2}\right)$ (for *gaps*). Lemma 6.2.1 ensures that

$$\mathbb{P}(G^c) \leq 1 - e^{-4\bar{\pi}\delta}.$$

Conditionally to QC, denoting by q the output of QExp, $|q - F_\pi^{-1}(p)| > \gamma \implies \mathfrak{E} \geq K - 1 \geq K/2$ whenever $n \geq 4/(\gamma\underline{\pi})$. By also working conditionally to G, and in order to apply Fact 6.1.2, we look for a $\beta > 0$ such that

$$K/2 = 2\frac{\ln(n^2) + \ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{1}{\beta}\right)}{\epsilon},$$

which gives

$$\beta = \frac{n^2}{\delta}e^{-\frac{\epsilon E\left(\frac{1}{4}n\gamma\underline{\pi}\right)}{4}}.$$

Note that even if Fact 6.1.2 is stated for $\beta \in (0,1)$, its conclusion remains obviously true for $\beta \geq 1$.

Finally,

$$\begin{aligned}\mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma\right) &\leq \mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma, QC, G\right) + \mathbb{P}(QC^c) + \mathbb{P}(G^c)\\ &\leq \frac{en^2}{\delta}e^{-\frac{\epsilon n\gamma\underline{\pi}}{16}} + 1 - e^{-4\bar{\pi}\delta} + 4e^{-\frac{\gamma^2\underline{\pi}^2}{8\max(p,1-p)}n},\end{aligned}$$

and by fixing $\delta := \frac{n\sqrt{e}}{2\sqrt{2\bar{\pi}}}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32}}$, because $1 - e^{-4\bar{\pi}\delta} \leq 8\bar{\pi}\delta$ for any $\delta > 0$,

$$\mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma\right) \leq 4n\sqrt{2e\bar{\pi}}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32}} + 4e^{-\frac{\gamma^2\underline{\pi}^2}{8\max(p,(1-p))}n}.$$

□

Applying this result to IndExp (ϵ becomes $\frac{\epsilon}{m}$) together with a union bound gives the following result :

Corollary 6.2.4 (Statistical utility of IndExp)**.** *Consider $n \geq 1$ and $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$*

almost surely. Denote $\mathbf{q} := (q_1, \dots, q_m)$ *the (random) result of IndExp on* $(X_1, \dots, X_n)$ *for the estimation of the quantiles of orders* $\mathbf{p} := (p_1, \dots, p_m)$*, where* $\min_i[\min(p_i, 1-p_i)] > 2/n$*. For each* $\gamma \in \left(0, \frac{2\min_i[\min(p_i,1-p_i)]}{\underline{\pi}}\right)$ *we have*

$$\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq 4nm\sqrt{2e\overline{\pi}}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32m}} + 4me^{-\frac{\gamma^2\underline{\pi}^2}{8}n},$$

where $F_\pi^{-1}(\mathbf{p}) = (F_\pi^{-1}(p_1), \dots, F_\pi^{-1}(p_m))$*.*

Proof. IndExp is the application of m independent QExp procedures but with privacy parameter $\frac{\epsilon}{m}$ in each. A union bound on the events that check if each quantile is off by at least γ gives the result by Theorem 6.2.3. □

So, there exist a polynomial expression P and two positive constants C_1 and C_2 depending only on the distribution such that, under mild hypotheses,

$$\mathbb{P}\left(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq P(n,m)\max\left(e^{-C_1\frac{\epsilon n\gamma}{m}}, e^{-C_2\gamma^2 n}\right).$$

We factorized the polynomial expression since it plays a minor role compared to the values in the exponential.

Statistical complexity of IndExp. The term $P(n,m)e^{-C_2\gamma^2 n}$ simply comes from the concentration of the empirical quantiles around the statistical ones. It is independent of the private nature of the estimation. It is the price that one usually expects to pay without the privacy constraint.

Privacy overhead of IndExp. The term $P(n,m)e^{-C_1\frac{\epsilon n\gamma}{m}}$ can be called the privacy overhead. It is the price paid for using this specific private algorithm for the estimation. For IndExp, if we want it to be constant, ϵn has to roughly scale as m times a polynomial expression in $\log_2 m$. As we will see later in Theorem 6.2.5, RecExp behaves much better, with $n\epsilon$ having to scale only as a polynomial expression in $\log_2 m$.

A privacy overhead of this type is not only an artifact due to a given algorithm (even if a suboptimal algorithm can make it worse), but in fact a constituent part of the private estimation problem, associated to a necessary price to pay, as captured by several works on generic lower bounds valid for *all* private estimators [Duchi et al., 2013, Duchi et al., 2014, Acharya et al., 2021e, Acharya et al., 2018, Acharya et al., 2021a, Acharya et al., 2021c, Acharya et al., 2021d, Acharya et al., 2021b, Barnes et al., 2020a, Barnes et al.,

2020b, Barnes et al., 2019, Kamath et al., 2022, Butucea et al., 2019, Lalanne et al., 2023b, Berrett & Butucea, 2019, Steinberger, 2023, Kroll, 2021].

With a similar proof technique as in the one of Theorem 6.2.3, the following result gives the statistical utility of RecExp :

Theorem 6.2.5 (Statistical utility of RecExp)**.** *Consider $n \geq 1$ and $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely. Denote $\mathbf{q} := (q_1, \dots, q_m)$ the result of RecExp on $(X_1, \dots, X_n)$ for the quantiles of orders $\mathbf{p} := (p_1, \dots, p_m)$, where $\min_i[\min(p_i, 1-p_i)] > 2/n$. For any $\gamma \in (0, \frac{2\min_i[\min(p_i,1-p_i)]}{\underline{\pi}})$ we have*

$$\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq 4n\sqrt{2e\bar{\pi}m}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32\log_2(2m)^2}} + 4me^{-\frac{\gamma^2\underline{\pi}^2}{8}n} .$$

Proof. For simplicity, let us assume that $E\left(\frac{1}{2}n\gamma\underline{\pi}\right) - 1 \leq \min\left(E(np_1) - 1, n - E(np_m)\right)$, which is for instance the case when $\gamma < \frac{2\min_i \min(p_i,1-p_i)}{\underline{\pi}}$, which we suppose. Furthermore, suppose that $\frac{1}{2}n\gamma\underline{\pi} \geq 2$, which is for instance the case when $n > 2/\min_i \min(p_i, 1-p_i)$ thank to the hypothesis on γ. By noting $K := E\left(\frac{1}{4}n\gamma\underline{\pi}\right)$, Lemma 6.1.1 says that for any $i \in \{1, \dots, m\}$,

$$\mathbb{P}\left(\sup_{k\in\{-K,\dots,K\}} |X_{(E(np_i)+k)} - F_X^{-1}(p_i)| > \gamma\right) \leq 4e^{-\frac{\gamma^2\underline{\pi}^2}{8C_{p_1,\dots,p_m}}n} ,$$

where $C_{p_1,\dots,p_m} := \max_i\left(\max\left(p_i, (1-p_i)\right)\right)$. We define the event QC (for *quantile concentration*),

$$QC := \bigcap_{i=1}^{m}\left(\sup_{k\in\{-K,\dots,K\}} |X_{(E(np_i)+k)} - F_X^{-1}(p_i)| \leq \gamma\right) .$$

By union bounds,

$$\mathbb{P}\left(QC^c\right) \leq 4me^{-\frac{\gamma^2\underline{\pi}^2}{8C_{p_1,\dots,p_m}}n} .$$

Let $\delta > 0$ that satisfies $\delta < \frac{1}{4\bar{\pi}}$. We define the event $G := \left(\min_i \Delta_i > \frac{\delta}{n^2}\right)$ (for *gaps*). Lemma 6.2.1 ensures that

$$\mathbb{P}\left(G^c\right) \leq 1 - e^{-4\bar{\pi}\delta} .$$

Conditionally to QC, denoting by $\mathbf{q}$ the output of RecExp, $\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma \implies \mathfrak{E} \geq K - 1 \geq K/2$ whenever $n \geq 4/(\gamma\underline{\pi})$. By also working conditionally to G, and in order to

apply Fact 6.1.3, we look for a $\beta > 0$ such that

$$K/2 = 2(\log_2 m + 1)^2 \frac{\ln(n^2) + \ln\left(\frac{1}{\delta}\right) + \ln m + \ln\left(\frac{1}{\beta}\right)}{\epsilon} ,$$

which gives

$$\beta = \frac{n^2 m}{\delta} e^{-\frac{\epsilon E\left(\frac{1}{4} n \gamma_{\underline{\pi}}\right)}{4(\log_2 m+1)^2}} .$$

Note that even if Fact 6.1.3 is stated for $\beta \in (0,1)$, its conclusion remains obviously true for $\beta \geq 1$.

Finally,

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) &\leq \mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma, QC, G\right) + \mathbb{P}\left(QC^c\right) + \mathbb{P}\left(G^c\right) \\ &\leq \frac{en^2 m}{\delta} e^{-\frac{\epsilon n \gamma_{\underline{\pi}}}{32(\log_2 m+1)^2}} + 1 - e^{-4\bar{\pi}\delta} + 4me^{-\frac{\gamma^2 {\underline{\pi}}^2}{8C_{p_1,\ldots,p_m}} n},\end{aligned}$$

and by fixing $\delta := \frac{n\sqrt{em}}{2\sqrt{2\bar{\pi}}} e^{-\frac{\epsilon n \gamma_{\underline{\pi}}}{32(\log_2 m+1)^2}}$, we get that,

$$\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq 4n\sqrt{2e\bar{\pi}m} e^{-\frac{\epsilon n \gamma_{\underline{\pi}}}{32(\log_2 m+1)^2}} + 4me^{-\frac{\gamma^2 {\underline{\pi}}^2}{8C_{p_1,\ldots,p_m}} n} .$$

□

As with Corollary 6.2.4, we can simplify this expression as

$$\begin{aligned}&\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \\ &\qquad \leq P(n,m) \max\left(e^{-C_1 \frac{\epsilon n \gamma}{(\log_2 m)^2}}, e^{-C_2 \gamma^2 n}\right) ,\end{aligned}$$

where P is a polynomial expression and C_1 and C_2 are constants, all depending only on the distribution.

Statistical complexity of RecExp. On the one hand the statistical term of this expression, which is independent of ϵ, is the same as with IndExp. This is natural since the considered statistical estimation problem is unchanged, only the privacy mechanism employed to solve it under a DP constraint was changed.

Privacy overhead of RecExp. On the other hand the privacy overhead $P(n,m)e^{-C_1 \frac{\epsilon n \gamma}{(\log_2 m)^2}}$ is much smaller than the one of IndExp. The scaling of ϵn to reach a prescribed probability went from approximately linear in m to roughly a polynomial expression in $\log_2 m$.

In particular and to the best of our knowledge, this scaling in m places RecExp much ahead of its competitors (the algorithms that compute multiple private empirical quantiles) for the task of statistical estimation.

Remark 6.2.6. All the results presented in this subsection require a uniform lower-bound on the density of the distribution from which the data is being sampled. Note that via some minor adaptations in the proofs, all the results can be adapted to the less restrictive hypothesis that the density is lower-bounded on a neighborhood of the statistical quantiles only.

6.2.3 Provable suboptimalities

Private quantile estimators often focus on estimating the quantile function at specific points $p_1, \dots, p_m$, which is probably motivated by a combination of practical considerations (algorithms to estimate and representing finitely many numbers are easier to design and manipulate than algorithms to estimate a function) and of intuitions about privacy (estimating the whole quantile function could increase privacy risks compared to estimating it on specific points). It is however well-documented in the (non-private) statistical literature that, under regularity assumptions on the quantile function, it can also be approximated accurately from functional estimators, see e.g. [Györfi et al., 2002, Tsybakov, 2009].

Building on this, this section considers a simple private histogram estimator of the density [Wasserman & Zhou, 2010] in order to estimate the quantile function in functional infinite norm. This allows of course to estimate the quantile function at $(p_1, \dots, p_m)$ for arbitrary m. As a natural consequence, we show that when m is very high, for a given privacy level RecExp has suboptimal utility guarantees and is beaten by the guarantees of the histogram estimator. Theorem 6.2.10 and Theorem 6.2.5 give a decision criterion (by comparing the upper bounds) to decide whether to use RecExp or a histogram estimator for the estimation problem.

Motivation: lower bounds for IndExp and RecExp

Lower-bounding the density of the exponential mechanism for u_{QExp} gives a general lower-bound on its utility:

Lemma 6.2.7 (Utility of QExp; Lower Bound)**.** *Let $X_1, \dots, X_n \in [0,1]$. Denoting by q the result of QExp on $(X_1, \dots, X_n)$ for the quantile of order p, we have for any $t \in [0,1]$ and any positive $\gamma \in (0, \frac{1}{4}]$,*

$$\mathbb{P}\Big(|q - t| > \gamma\Big) \geq \frac{1}{2} e^{-\frac{n\epsilon}{2}} .$$

Note that this holds without any constraint relating p,n, or γ.

Proof. By definition of u_{QExp} we have $-n \leq u_{\mathrm{QExp}}\big((X_1,\dots,X_n),q\big) \leq 0$ for any input, hence using that $0 \leq \gamma \leq 1/4$ we get

$$\begin{aligned}
\mathbb{P}\bigg(|q-t|>\gamma\bigg) &= \frac{\int_{[0,1]\setminus[t-\gamma,t+\gamma]} e^{\frac{\epsilon}{2}u_{\mathrm{QExp}}\left((X_1,\dots,X_n),q\right)}dq}{\int_{[0,1]} e^{\frac{\epsilon}{2}u_{\mathrm{QExp}}\left((X_1,\dots,X_n),q\right)}dq} \\
&\geq \frac{\int_{[0,1]\setminus[t-\gamma,t+\gamma]} e^{-\frac{\epsilon}{2}n}dq}{\int_{[0,1]} e^{0}dq} \\
&\geq (1-2\gamma)e^{-\frac{\epsilon}{2}n} \\
&\geq \frac{1}{2}e^{-\frac{\epsilon}{2}n}\ .
\end{aligned}$$

□

As a consequence, if the points $X_1,\dots,X_n$ are randomized, the probability that QExp makes an error bigger than γ on the estimation of a quantile of the distribution is at least $\frac{1}{2}e^{-\frac{n\epsilon}{2}}$. A direct consequence is that for any $\gamma \in (0,\frac{1}{4}]$, the statistical utility of IndExp has a is lower-bounded:

$$\mathbb{P}\bigg(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty>\gamma\bigg) \geq \frac{1}{2}e^{-\frac{n\epsilon}{2m}}\ ,$$

and the statistical utility of RecExp is also lower-bounded:

$$\mathbb{P}\bigg(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty>\gamma\bigg) \geq \frac{1}{2}e^{-\frac{n\epsilon}{2(\log_2 m+1)}}\ .$$

These are consequences of lower-bounds on the estimation error of the first statistical quantile estimated by each algorithm in its respective computation graph (with privacy level ϵ/m for IndExp; $\epsilon/(\log_2 m+1)$ for RecExp).

In particular, for both algorithms, utility becomes arbitrarily bad when m increases. This is not a behavior that would be expected from any optimal algorithm. The rest of this section studies a better estimator for high values of m.

Histogram density estimator

The histogram density estimator is a well-known estimator of the density of a distribution of probability. Despite its simplicity, a correct choice of the bin size can even make it minimax optimal for the class of Lipschitz densities.

Under differential privacy, this estimator was first adapted and studied by [Wasserman & Zhou, 2010]. It is studied both in terms of integrated squared error and in Kolmogorov-Smirnov distance. In the sequel, we need a control in infinite norm. We hence determine the histogram concentration properties for this metric.

Given a a bin size $h > 0$ that satisfies $\frac{1}{h} \in \mathbb{N}$, we partition $[0,1]$ in $\frac{1}{h}$ intervals of length h. The intervals of this partition are called the bins of the histogram. Given $\frac{1}{h}$ i.i.d. centered Laplace distributions of parameter 1, $(\mathcal{L}_b)_{b\in\text{bins}}$, we define $\hat{\pi}^{\text{hist}}$, an estimator of the supposed density π of the distribution as: for each $t \in [0,1]$,

$$\hat{\pi}^{\text{hist}}(t) := \sum_{b\in\text{bins}} \mathbb{1}_b(t) \frac{1}{nh} \left(\sum_{i=1}^{n} \mathbb{1}_b(X_i) + \frac{2}{\epsilon}\mathcal{L}_b \right) .$$

The function that, given the bins of a histogram, counts the number of points that fall in each bin of the histogram has a sensitivity of 2 for the replacement neighboring relation. Indeed, replacing a point by another changes the counts of at most two (consecutive) bins by one. Hence, the construction of the Laplace mechanism ensures that $\hat{\pi}^{\text{hist}}$ is ϵ-DP.

Note that, as a common practice, we divided by n freely in terms of privacy budget in the construction of $\hat{\pi}^{\text{hist}}$. This is possible because we work with the replacement neighboring relation. The size n of the datasets is fixed and is a constant of the problem.

The deviation between π and $\hat{\pi}^{\text{hist}}$ can be controlled.

Lemma 6.2.8 (Utility of $\hat{\pi}^{\text{hist}}$; Density estimation)**.** *Consider $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that π is L-Lipschitz for some positive constant L, and the private histogram density estimator $\hat{\pi}^{hist}$ with bin size h. For any $\gamma > Lh$, we have*

$$\mathbb{P}\left(\|\hat{\pi}^{hist} - \pi\|_\infty > \gamma \right) \leq \frac{1}{h} e^{-\frac{\gamma h n \epsilon}{4}} + \frac{2}{h} e^{-\frac{h^2(\gamma - Lh)^2}{4} n} .$$

Proof. Let us consider a specific bin of the histogram b. Let $\gamma > 0$. Denoting by $\|\cdot\|_{\infty,b}$ the infinite norm restrained to the support of b, which is a semi-norm, we have

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty,b} > \gamma \right) = \mathbb{P}\left(\left\| \frac{1}{nh} \left(\sum_{i=1}^{n} \mathbb{1}_b(X_i) + \frac{2}{\epsilon}\mathcal{L} \right) - \pi \right\|_{\infty,b} > \gamma \right)$$

where $\mathcal{L} \sim \text{Lap}(1)$, a centered Laplace distribution of parameter 1. So,

$$\begin{aligned}\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty,b} > \gamma\right) &= \mathbb{P}\left(\left\|\left(\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}_b(X_i) - \pi\right) + \frac{2}{nh\epsilon}\mathcal{L}\right\|_{\infty,b} > \gamma\right) \\ &\overset{\text{triangular inequality}}{\leq} \mathbb{P}\left(\left\|\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}_b(X_i) - \pi\right\|_{\infty,b} > \gamma/2\right) \\ &\quad + \mathbb{P}\left(\left|\frac{2}{nh\epsilon}\mathcal{L}\right| > \gamma/2\right)\end{aligned}$$

Let us first control the first term. Since π is L Lipschitz, $\forall x \in b, \left|\pi(x) - \frac{1}{h}\int_b \pi\right| \leq \frac{Lh}{2}$. So, when $\gamma > Lh$,

$$\left(\left\|\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}_b(X_i) - \pi\right\|_{\infty,b} > \gamma/2\right) \subset \left(\left|\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}_b(X_i) - \frac{1}{h}\int_b \pi\right| > \gamma/2 - Lh/2\right) .$$

Finally, notice that the family $(\mathbb{1}_b(X_i))_i$ is a family of i.i.d. Bernoulli random variables of probability of success $\int_b \pi$. By Hoeffding's inequality,

$$\mathbb{P}\left(\left\|\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}_b(X_i) - \pi\right\|_{\infty,b} > \gamma/2\right) \leq 2e^{-\frac{h^2(\gamma-Lh)^2}{4}n} .$$

The second term is controlled via a tail bound on the Laplace distribution as

$$\begin{aligned}\mathbb{P}\left(\left|\frac{2}{nh\epsilon}\mathcal{L}\right| > \gamma/2\right) &= \mathbb{P}\left(|\mathcal{L}| > \frac{\gamma nh\epsilon}{4}\right) \\ &= \int_{\frac{\gamma nh\epsilon}{4}}^{\infty} e^{-t}dt \\ &= e^{-\frac{\gamma hn\epsilon}{4}} .\end{aligned}$$

So, if $\gamma > Lh$,

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty,b} > \gamma\right) \leq 2e^{-\frac{h^2(\gamma-Lh)^2}{4}n} + e^{-\frac{\gamma hn\epsilon}{4}} .$$

Finally, a union bound on all the bins gives that if $\gamma > Lh$,

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty} > \gamma\right) \leq \frac{2}{h}e^{-\frac{h^2(\gamma-Lh)^2}{4}n} + \frac{1}{h}e^{-\frac{\gamma hn\epsilon}{4}} .$$

□

Application to quantile function estimation

In order to use $\hat{\pi}^{\text{hist}}$ as an estimator of the quantile function, we need to properly define a quantile function estimator associated with it. Indeed, even if $\hat{\pi}^{\text{hist}}$ estimates a density of

probability, it does not necessary integrate to 1 and can even be negative at some locations. Given any integrable function $\hat{\pi}$ on $[0,1]$, we define its generalized quantile function

$$F_{\hat{\pi}}^{-1}(p) = \inf\left\{q \in [0,1] | \int_0^q \hat{\pi} \geq p\right\}, \forall p \in [0,1] ,$$

with the convention $\inf \emptyset = 1$. Even if this quantity has no reason to behave as a quantile function, the following lemma tells that $F_{\hat{\pi}}^{-1}$ is close to an existing quantile function when $\hat{\pi}$ is close to its corresponding density.

Lemma 6.2.9 (Inversion of density estimators)**.** *Consider a density π on $[0,1]$ w.r.t. Lebesgue's measure such that $\pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely. If $\hat{\pi}$ is an integrable function that satisfies $\|\hat{\pi} - \pi\|_\infty \leq \alpha$, and if $p \in [0,1]$ is such that $\left[F_\pi^{-1}(p) - \frac{2\alpha}{\underline{\pi}}, F_\pi^{-1}(p) + \frac{\alpha}{\underline{\pi}}\right] \subset (0,1)$, then*

$$\left|F_\pi^{-1}(p) - F_{\hat{\pi}}^{-1}(p)\right| \leq \frac{2\alpha}{\underline{\pi}} .$$

Proof. We have,

$$\begin{aligned} F_{\hat{\pi}}\left(F_\pi^{-1}(p) + \frac{\alpha}{\underline{\pi}}\right) &\overset{\|\hat{\pi}-\pi\|_\infty \leq \alpha}{\geq} F_\pi\left(F_\pi^{-1}(p) + \frac{\alpha}{\underline{\pi}}\right) - \alpha \\ &\overset{\pi \geq \underline{\pi}}{\geq} F_\pi\left(F_\pi^{-1}(p)\right) + \frac{\alpha}{\underline{\pi}}\underline{\pi} - \alpha \\ &= F_\pi\left(F_\pi^{-1}(p)\right) = p . \end{aligned}$$

So,

$$F_{\hat{\pi}}^{-1}(p) \leq F_\pi^{-1}(p) + \frac{\alpha}{\underline{\pi}} .$$

Furthermore, for any $t \in \left[\frac{2\alpha}{\underline{\pi}}, F_\pi^{-1}(p)\right]$,

$$\begin{aligned} F_{\hat{\pi}}\left(F_\pi^{-1}(p) - t\right) &\overset{\|\hat{\pi}-\pi\|_\infty \leq \alpha}{\leq} F_\pi\left(F_\pi^{-1}(p) - t\right) + \alpha \\ &\overset{\pi \geq \underline{\pi}}{\leq} F_\pi\left(F_\pi^{-1}(p)\right) - t\underline{\pi} + \alpha \\ &< F_\pi\left(F_\pi^{-1}(p)\right) - \frac{2\alpha}{\underline{\pi}}\underline{\pi} + \alpha \\ &= F_\pi\left(F_\pi^{-1}(p)\right) - \alpha < p . \end{aligned}$$

So, for any $t \in \left(\frac{2\alpha}{\underline{\pi}}, F_\pi^{-1}(p)\right)$;

$$F_{\hat{\pi}}^{-1}(p) \geq F_\pi^{-1}(p) - t ,$$

and finally,

$$F_{\hat{\pi}}^{-1}(p) \geq F_{\pi}^{-1}(p) - \frac{2\alpha}{\underline{\pi}} .$$

□

A direct consequence of Lemma 6.2.8 and Lemma 6.2.9 is Theorem 6.2.10. It controls the deviation of the generalized quantile function based on $\hat{\pi}^{\text{hist}}$ to the true quantile function.

Theorem 6.2.10 (Utility of $F_{\hat{\pi}^{\text{hist}}}^{-1}$; Quantile function estimation)**.** *Consider* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* π *is* L*-Lipschitz for some positive constant* L *and that* $\pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely, and* $h < \underline{\pi}/(4L)$ *such that* $\frac{1}{h} \in \mathbb{N}$*. Let* $F_{\hat{\pi}^{hist}}^{-1}$ *be the quantile function estimator associated with the private histogram density estimator* $\hat{\pi}^{hist}$ *with bin size* h*. Consider* $\gamma_0 \in (2Lh/\underline{\pi}, 1/2)$*,* $I := F_\pi\big((\gamma_0, 1-\gamma_0)\big)$*, and* $\|\cdot\|_{\infty,I}$ *the sup-norm of functions on the interval* I*. We have*

$$\mathbb{P}\left(\|F_{\hat{\pi}^{hist}}^{-1} - F_{\pi}^{-1}\|_{\infty,I} > \gamma\right) \leq \frac{1}{h} e^{-\frac{\gamma\underline{\pi} hn\epsilon}{8}} + \frac{2}{h} e^{-\frac{h^2}{4}\left(\frac{\gamma\underline{\pi}}{2} - Lh\right)^2 n} ;,$$

whenever $\gamma \in (2Lh/\underline{\pi}, \gamma_0)$*.*

Proof. Given $\gamma \in \left(\frac{2Lh}{\underline{\pi}}, \gamma_0\right)$, $\frac{\gamma\underline{\pi}}{2} \geq \frac{2\underline{\pi}Lh}{2\underline{\pi}} = Lh$. So, Lemma 6.2.9 applies and gives that

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_\infty > \frac{\gamma\underline{\pi}}{2}\right) \leq \frac{1}{h} e^{-\frac{\gamma\underline{\pi} hn\epsilon}{8}} + \frac{2}{h} e^{-\frac{h^2}{4}\left(\frac{\gamma\underline{\pi}}{2} - Lh\right)^2 n} .$$

Furthermore, $I = F_\pi\left((\gamma_0, 1-\gamma_0\right)$. So,

$$\forall p \in I, \quad \gamma_0 < F_\pi^{-1}(p) < 1 - \gamma_0 .$$

In particular, when $\hat{\pi}^{\text{hist}}$ satisfies $\|\hat{\pi}^{\text{hist}} - \pi\| \leq \frac{\gamma\underline{\pi}}{2}$, Lemma 6.2.8 applies and gives

$$\forall p \in I, \quad |F_{\hat{\pi}^{\text{hist}}}^{-1}(p) - F_\pi^{-1}(p)| \leq \gamma .$$

This is equivalent to

$$\forall p \in I, \quad \|F_{\hat{\pi}^{\text{hist}}}^{-1}(p) - F_\pi^{-1}(p)\|_{\infty,I} \leq \gamma .$$

Finally,

$$\mathbb{P}\left(\|F_{\hat{\pi}^{\text{hist}}}^{-1} - F_{\pi}^{-1}\|_{\infty,I} > \gamma\right) \leq \frac{1}{h} e^{-\frac{\gamma\underline{\pi} hn\epsilon}{8}} + \frac{2}{h} e^{-\frac{h^2}{4}\left(\frac{\gamma\underline{\pi}}{2} - Lh\right)^2 n} .$$

□

Analysis of Theorem 6.2.10. As with Theorem 6.2.3 and Theorem 6.2.5, the upper-bound provided by Theorem 6.2.10 can be split in two terms : The error that one usually expects without privacy constraint, $\frac{2}{h}\exp(-\frac{h^2}{4}(\frac{\gamma\pi}{2}-Lh)^2n)$, and the one that come from the private algorithm, $\frac{1}{h}\exp(-\frac{\gamma\pi hn\epsilon}{8})$. The assumption $\frac{\gamma\pi}{2} > Lh$ ensures that the bin size h and the desired level of precision γ are compatible.

Computational aspects. $\hat{\pi}^{\text{hist}}$ is constant on each bin. Hence, it can be stored in a single array of size $\frac{1}{h}$. If the data points are sorted, this array can be filled with a single pass over all data points and over the array. Then, given $p_1, \dots, p_m \in (0,1)$ sorted, estimating $F^{-1}_{\hat{\pi}^{\text{hist}}}(p_1), \dots, F^{-1}_{\hat{\pi}^{\text{hist}}}(p_m)$ can be done with a single pass over $p_1, \dots, p_m$ and over the array that stores $\hat{\pi}^{\text{hist}}$. Indeed, it is done by "integration" of the array until the thresholds of the p_i's are reached. The overall complexity of this procedure is $O\left(n+m+\frac{1}{h}\right)$ to which must be added $O(n\log n)$ if the data is not sorted and $O(m\log m)$ if the targeted quantiles p_i are not sorted.

Comparison with RecExp. Comparing this histogram-based algorithm to RecExp is more difficult than comparing RecExp to IndExp. First of all, the results are qualitatively different. Indeed, RecExp estimates the quantile function on a finite number of points and the histogram estimator estimates it on an interval. The second result is stronger in the sense that when the estimation is done on an interval, it is done for any finite number of points in that interval. However, the error of RecExp for that finite number of points may be smaller than the one given by the histogram on the interval. Then, the histogram depends on a meta parameter h. With a priori information on the distribution, it can be tuned using Theorem 6.2.10. Aditionally, the hypothesis required are different : Theorem 6.2.5 does not require the density to be Lipschitz contrary to Theorem 6.2.10. Finaly, we can observe that the histogram estimator is not affected by the lower bounds described in Section 6.2.3. Hence, when all the hypotheses are met, there will obviously always be a number m of targeted quantiles above which it is better to use histograms.

Remark 6.2.11. Notice that the hypothesis of Lipschitzness of the density is only useful for the histogram estimators. In particular the guarantees of QExp, IndExp, and RecExp do not require such hypothesis. This section thus presented a *strict subclass* of the problem on which RecExp may be suboptimal.

Remark 6.2.12. We would like to highlight the fact that histograms are used as an illustration of the suboptimality of RecExp on some instances of the problem. In particular, it does not imply that they are the state of the art on such instances. It is very possible that other mechanisms perform well in such cases [Blocki et al., 2012, Alabi et al., 2022]. In fact, provided that the inversion from the cumulative distribution function of the distribution to its quantile function is easy (which is typically the case when the density is uniformly lower-bounded), we expect that many private CDF estimators will behave similarly or better on these specific instances [Bun et al., 2015, Kaplan et al., 2020, Drechsler et al., 2022, Denisov et al., 2022, Henzinger & Upadhyay, 2022].

6.2.4 Experimental results

For the experiments, we benchmarked the different estimators on beta distributions, as they allow to easily tune the Lipschitz constants of the densities, which is important for characterizing the utility of the histogram estimator.

Figure 6.1 represents the performance of the estimator as a function of m. We estimate the quantiles of orders $\mathbf{p} = \left(\frac{1}{4} + \frac{1}{2(m+1)}, \dots, \frac{1}{4} + \frac{m}{2(m+1)}\right)$ since it allows us to stay in the regions where the density is not too small.

IndExp *vs* RecExp *vs* Histograms. Figure 6.1, confirms our claims about the scaling in m of IndExp and RecExp. Indeed, even if IndExp quickly becomes unusable, RecExp stays at a low error until really high values of m. The conclusions of Section 6.2.3 also seem to be verified : Even if RecExp performs well for small to intermediate values of m, there is always a certain value of m for which it becomes worse than the histogram estimator. This shift of regime occurs between $m \approx 10$ for the distribution Beta(0.5, 0.5) and $m \approx 40$ for the distribution Beta(2, 5).

Error of the histogram-based approach. The shape of the error for the histogram estimator is almost flat. Again, it is compatible with Theorem 6.2.10 : The control in infinite norm is well suited for the histograms.

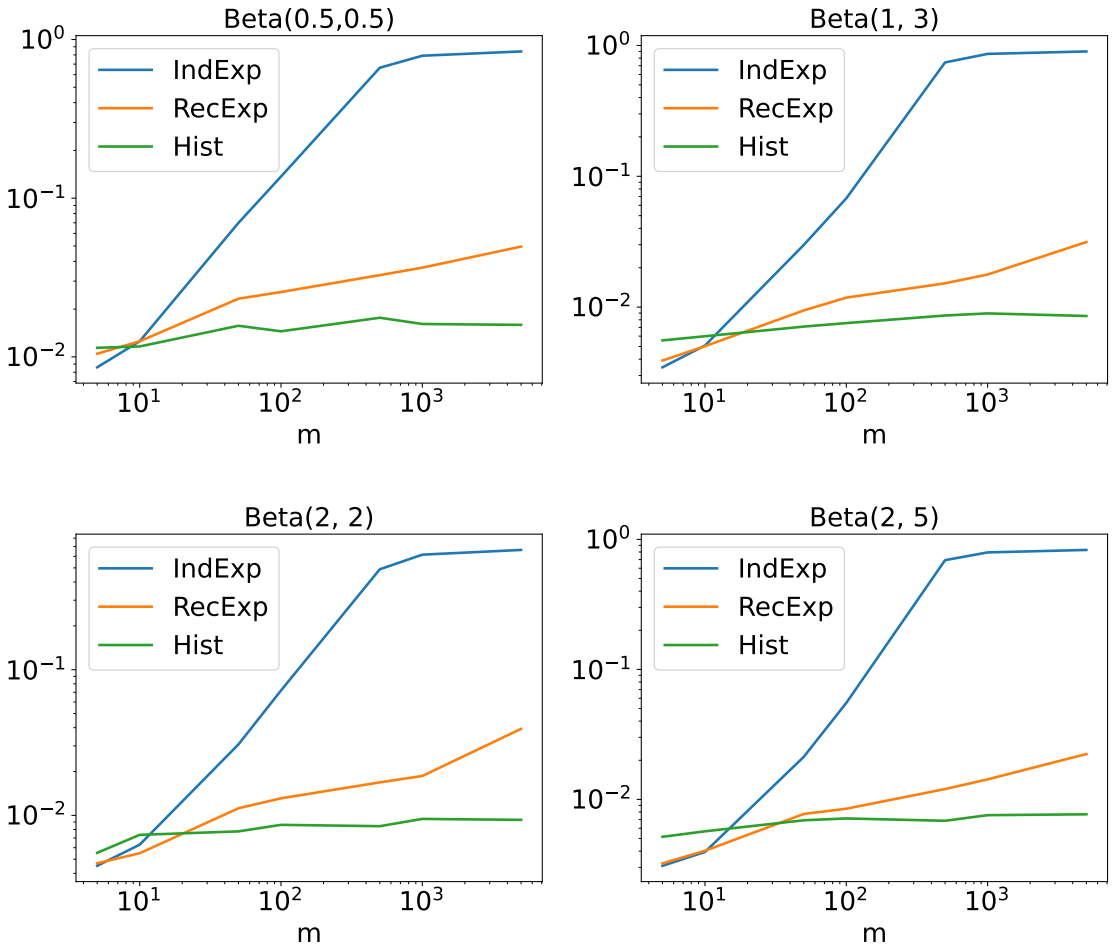
Role of the Lipschitz constant. By crossing the shape of the beta distributions and Figure 6.1, a pattern becomes clear : The distributions on which the histogram estimator performs best (i.e. the distributions on which it becomes the best estimator for the lowest possible value of m) are the distributions with the smallest Lipschitz constant. This was expected since the guarantees of utility of Theorem 6.2.10 get poorer the higher this quantity is.

6.2.5 The case of JointExp and the inverse sensitivity

For the specific case of JointExp and the inverse sensitivity mechanism (due to their similarities we only focus on JointExp), we do not have high probability bounds as satisfying as the ones of QExp, IndExp or RecExp, in the sense that the scaling in m is much poorer. However, we are still able to provide consistency results at fixed m and ϵ which is the first consistency result for this method.

For this, we first need two technical lemmas.

Lemma 6.2.13. *Let $\tilde{X}$ be a real random variable with density $\pi_{\tilde{X}}$ and $p \in (0,1)$. We suppose that $\pi_{\tilde{X}} \geq \pi_{min} > 0$ on an open neighborhood $\mathcal{N}$ of $F_{\tilde{X}}^{-1}(p)$. If we have access to*



The vertical axis reads the error $\mathbb{E}\left(\|\hat{\mathbf{q}} - F^{-1}(\mathbf{p})\|_\infty\right)$ where $\mathbf{p} = \left(\frac{1}{4} + \frac{1}{2(m+1)}, \ldots, \frac{1}{4} + \frac{m}{2(m+1)}\right)$ for different values of m, $n = 10000$, $\epsilon = 0.1$, $\hat{\mathbf{q}}$ is the private estimator, and $\mathbb{E}$ is estimated by Monte-Carlo averaging over 50 runs. The histogram is computed on 200 bins.

Figure 6.1: Numerical performance of the different private estimators

$\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ *i.i.d. realisations of* $\tilde{X}$ *then for every* $\gamma > 0$, *if* $n \geq \frac{2}{\gamma \pi_{min}}$,

$$[F_{\tilde{X}}^{-1}(p) - \gamma, F_{\tilde{X}}^{-1}(p) + \gamma] \subset \mathcal{N} \implies \mathbb{P}\left(\left|F_{\tilde{X}}^{-1}(p) - \tilde{X}_{(\lceil np \rceil)}\right| > \gamma\right) \leq e^{-n\left(\frac{\gamma^2 \pi_{min}^2}{8(1-p)}\right)} + e^{-n\left(\frac{\gamma^2 \pi_{min}^2}{8p}\right)}$$

Proof. Let $\gamma > 0$ such that $[F_{\tilde{X}}^{-1}(p) - \gamma, F_{\tilde{X}}^{-1}(p) + \gamma] \subset \mathcal{N}$. Let us define

$$N = \sum_{i=1}^{n} \mathbb{1}_{(F_{\tilde{X}}^{-1}(p) + \gamma, +\infty)}(\tilde{X}_i) \, .$$

N is a sum of n independent Bernoulli random variable with probabilities of success lower than $\eta = 1 - p - \gamma\pi_{\min}$. If $\tilde{X}_{(\lceil np \rceil)} > F_{\tilde{X}}^{-1}(p) + \gamma$, then $N \geq n(1-p)$. So,

$$\begin{aligned}
&\mathbb{P}\left(\tilde{X}_{(\lceil np \rceil)} > F_{\tilde{X}}^{-1}(p) + \gamma\right) \leq \mathbb{P}\left(N \geq n(1-p) - 1\right) \\
&= \mathbb{P}\left(N \geq n\eta\left(1 + \frac{\gamma\pi_{\min}}{\eta} - \frac{1}{n\eta}\right)\right) \\
&\leq e^{-n\eta\left(\frac{\gamma\pi_{\min}}{\eta} - \frac{1}{n\eta}\right)^2 / \left(2 + \frac{\gamma\pi_{\min}}{\eta} - \frac{1}{n\eta}\right)}
\end{aligned}$$

where line 3 is deduced from line 2 by a multiplicative Chernoff bounds. If we further impose that $n \geq \frac{2}{\gamma\pi_{\min}}$,

$$\begin{aligned}
&\mathbb{P}\left(\tilde{X}_{(\lceil np \rceil)} > F_{\tilde{X}}^{-1}(p) + \gamma\right) \leq e^{-\frac{n\eta}{4}\left(\frac{\gamma\pi_{\min}}{\eta}\right)^2 / \left(2 + \frac{\gamma\pi_{\min}}{\eta}\right)} \\
&\leq e^{-\frac{n}{4}\left(\frac{\gamma^2\pi_{\min}^2}{2(1-p) - \gamma\pi_{\min}}\right)} \leq e^{-n\left(\frac{\gamma^2\pi_{\min}^2}{8(1-p)}\right)}
\end{aligned}$$

Looking at the event $\left(\tilde{X}_{(np)} < F_{\tilde{X}}^{-1}(p) - \gamma\right)$ and a union bound give the expected result. □

Lemma 6.2.14. *Let $\tilde{X}$ be a real random variable with density $\pi_{\tilde{X}}$ and $p \in (0,1)$. We suppose that $\pi_{max} \geq \pi_{\tilde{X}} \geq \pi_{min} > 0$ on an interval I of $\mathbb{R}$. If we note $N = \sum_{i=1}^{n} \mathbb{1}_I(\tilde{X}_i)$ the number of points that fall in I, we have*

$$\mathbb{P}\left(N \geq 2n\lambda(I)\pi_{max}\right) \leq e^{-\frac{n\lambda(I)\pi_{max}}{3}},$$

$$\mathbb{P}\left(N \leq \frac{1}{2}n\lambda(I)\pi_{min}\right) \leq e^{-\frac{n\lambda(I)\pi_{min}}{8}}.$$

.

Proof. This is a simple application of multiplicative Chernoff bounds to the sum N of independent Bernoulli random variables. □

Finally, the consistency result states that

Theorem 6.2.15. *If X is a random variable with density π_X w.r.t. Lebesgue measure that is piecewise continuous and if there exists $\beta > 0$ such that $\pi_X > 0$ and is continuous on $\cup_{i=1}^{n}[F_{\tilde{X}}^{-1}(p_i) - \beta, F_{\tilde{X}}^{-1}(p_i) + \beta]$, then, denoting by $\mathbf{q}$ the output of JointExp applied to $\mathbf{X}$ with constant ϵ,*

$$\mathbb{P}\left(\|\mathbf{q} - F_X^{-1}(\mathbf{p})\|_\infty > \beta\right) = o_n(1).$$

Proof. Let $0 < \gamma < \beta$ such that $\pi_{\tilde{X}} > 0$ on $\mathcal{O} := \cup_{i=1}^{n}[F_{\tilde{X}}^{-1}(p_i) - \beta, F_{\tilde{X}}^{-1}(p_i) + \beta]$. We note $\pi_{\min} = \inf_{\mathcal{O}} \pi_{\tilde{X}}$ and $\pi_{\max} = \sup_{\mathcal{O}} \pi_{\tilde{X}}$. We also define the following events:

$$A : \forall i, \left|\tilde{X}_{(\lceil np_i \rceil)} - F_{\tilde{X}}^{-1}(p_i)\right| \leq \gamma \ ,$$

$$B : \forall i, \#(\tilde{\mathbf{X}} \cap [F_{\tilde{X}}^{-1}(p_i) + \gamma, F_{\tilde{X}}^{-1}(p_i) + \beta]) \geq \frac{1}{2} n(\beta - \gamma)\pi_{\min} \text{ and}$$
$$\#(\tilde{\mathbf{X}} \cap [F_{\tilde{X}}^{-1}(p_i) - \beta, F_{\tilde{X}}^{-1}(p_i) - \gamma]) \geq \frac{1}{2} n(\beta - \gamma)\pi_{\min} \ ,$$

$$C : \forall i, \#(\tilde{\mathbf{X}} \cap [F_{\tilde{X}}^{-1}(p_i) - \gamma, F_{\tilde{X}}^{-1}(p_i) + \gamma]) \leq 2n2\gamma\pi_{\max} \ .$$

Then we can compute,

$$\frac{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty > \beta | A, B, C\right)}{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \beta | A, B, C\right)} \leq \frac{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty > \beta | A, B, C\right)}{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \gamma | A, B, C\right)}$$

Conditionally to A and B, $-u_{\text{JE}}(\tilde{\mathbf{X}}, \mathbf{q}) \leq \frac{1}{2}\left(\frac{1}{2}n(\beta - \gamma)\pi_{\min}\right) \implies \|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \beta$. Furthermore, conditionally to A and C, $\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \gamma \implies -u_{\text{JE}}(\tilde{\mathbf{X}}, \mathbf{q}) \leq \frac{1}{2}\left(4(m+1)n\gamma\pi_{\max}\right)$. So,

$$\frac{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty > \beta | A, B, C\right)}{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \gamma | A, B, C\right)} \leq \frac{(b-a)^m}{(2\gamma)^m/m!} e^{-\frac{\epsilon}{4}\left(\frac{(\beta-\alpha)\pi_{\min}}{2} - 4(m+1)\gamma\pi_{\max}\right)n}$$

and by fixing $\gamma = \frac{\beta\pi_{\min}}{16(m+1)\pi_{\max} + 2\pi_{\min}}$ we end up with

$$\frac{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty > \beta | A, B, C\right)}{\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \beta | A, B, C\right)}$$
$$\leq \frac{2^m(b-a)^m m!}{\beta^m}\left(\frac{4(m+1)\pi_{\max} + \pi_{\min}/2}{\pi_{\min}}\right)^m e^{-\frac{\epsilon\beta\pi_{\min}}{16}n} \ .$$

We can use Lemma 6.2.13, Lemma 6.2.14 and union bounds to obtain the following for n big enough:

$$\mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty > \beta\right)$$
$$\leq \mathbb{P}\left(\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty > \beta | A, B, C\right) + \mathbb{P}(A^c) + \mathbb{P}(B^c) + \mathbb{P}(C^c)$$
$$\leq \frac{2^m(b-a)^m m!}{\beta^m}\left(\frac{4(m+1)\pi_{\max} + \pi_{\min}/2}{\pi_{\min}}\right)^m e^{-\frac{\epsilon\beta\pi_{\min}}{16}n}$$
$$+ \sum_{i=1}^{m} e^{-n\left(\frac{\beta^2\pi_{\min}^4}{8(1-p_i)(16(m+1)\pi_{\max}+2\pi_{\min})^2}\right)} + \sum_{i=1}^{m} e^{-n\left(\frac{\beta^2\pi_{\min}^4}{8p_i(16(m+1)\pi_{\max}+2\pi_{\min})^2}\right)}$$
$$+ me^{-n\frac{\beta\pi_{\min}\pi_{\max}}{24(m+1)\pi_{\max}+3\pi_{\min}}} + 2me^{-n\frac{\pi_{\min}}{8}\left(\beta - \frac{\beta\pi_{\min}}{16(m+1)\pi_{\max}+2\pi_{\min}}\right)} \ .$$

□

6.3 JointExp and atomic distributions

The theoretical guarantees of JointExp (and also all the algorithms that use it such as IndExp or RecExp) hold true for *continuous* distributions. This section observes numerically and proves theoretically that in fact, when the distribution involves isolated atoms, this mechanism fails completely. It also provides a simple solution to this problem, and demonstrates the improvement both theoretically and experimentally.

6.3.1 JointExp fails on atomic distributions

In order to understand the origin of this weakness of JointExp, we analyse the density of the distribution of its output. This density is constant on the *"blocks"*

$$\big([X_{i_1}, X_{i_1+1}) \times \ldots \times \cdots \times [X_{i_m}, X_{i_m+1})\big) \cap [a,b]^{m\nearrow}$$

for each $\mathbf{i} = (i_1, \ldots, i_m) \in O'$ where

$$O' = \{\mathbf{i} \in \{0, \ldots, n\}^m, i_1 \leq \cdots \leq i_m\} \ .$$

The probability of the output of JointExp being in a given block is proportional to the volume of this block.

What can happen in practice is that even though a block is interesting in terms of utility level, its volume can in fact be close to zero if the data points are close. The volume can even be zero in case of equality, hence this block is never selected by the exponential mechanism. This phenomenon occurs particularly often for data drawn from distributions with isolated atoms: asymptotically, the dataset will almost surely contain collisions among the data points as n grows and JointExp will fail on the corresponding quantiles.

To formally capture this phenomenon, from now on, $\mathbf{X}$ is supposed to be a collection of n i.i.d. samples of a random variable X with distribution $\mathbb{P}_X$ and with cumulative distribution function (CDF) F_X.

Proposition 6.3.1. *Suppose that there exist $q \in (a,b)$ and $\eta > 0$ such that $I := (q-\eta, q+\eta) \subset [a,b]$ satisfies $\mathbb{P}_X(\{q\}) > 0$ and $\mathbb{P}_X(I \setminus \{q\}) = 0$. Then there exist some probability vectors $\mathbf{p}$ such that, denoting by $\mathbf{q}$ the output of JointExp applied to $\mathbf{X}$ with constant ϵ,*

$$\mathbb{E}_{\mathbf{X},\ \mathit{JointExp}}\left(\|\mathbf{q} - F_X^{-1}(\mathbf{p})\|_\infty\right) = \Omega_n(1) \ , \tag{6.2}$$

where we use the vector notation $F_X^{-1}(\mathbf{p}) = (F_X^{-1}(p_1), \ldots, F_X^{-1}(p_m))$. Furthermore, the Lebesgue measure of the set of problematic probability vectors is lower bounded by $\mathbb{P}_X(\{q\})^m/(m!)$.

Proof. Since $\mathbb{P}_X(\{q\}) > 0$ and $\mathbb{P}_X(I \setminus \{q\}) = 0$, there exists a nonempty interval A of $[0,1]$ such that $\{q\} = F_X^{-1}(A)$ with $\lambda(A) \geq \mathbb{P}_X(\{q\})$, λ referring to Lebesgue measure. Let us prove that any $\mathbf{p}$ with at least one component in A satisfies (6.2).

For this, assume that $\mathbf{p}$ has its i^{th} entry p_i in A. Due to the structure of $\mathbb{P}_X$, $\mathbb{P}_X(\mathbf{X} \cap (I \setminus \{q\}) \neq \emptyset) = 0$, hence almost surely it holds that for every j we have either $|X_j - q| \geq \eta > 0$ or $|X_j - q| = 0$. Remember that the output density is a mixture of uniforms on the sets $\big([X_{i_1}, X_{i_1+1}) \times \ldots \times \cdots \times [X_{i_m}, X_{i_m+1})\big) \cap [a,b]^m \nearrow$ for $\mathbf{i} = (i_1, \ldots, i_m) \in O'$. If the i^{th} component of the output q_i was to be sampled from a data interval that doesn't admit q in its closure, then $\|\mathbf{q} - F_X^{-1}(\mathbf{p})\|_\infty \geq \eta$. If on the other hand q_i was to be sampled from a data interval that does admit q in its closure, then it belongs to an interval $[X_k, X_{k+1})$ for some k such that $q \in [X_k, X_{k+1}]$ and $X_{k+1} - X_k \geq \eta$. Conditionally to the fact that there are $m' \leq m$ other quantiles that are sampled from $[X_k, X_{k+1}]$, the conditional expectation of $\|\mathbf{q} - F_X^{-1}(\mathbf{p})\|_\infty$ can be lower by a (strictly) positive functional ($f(\eta, m')$) of η and m' (because the corresponding slice of the output is uniform on $[X_k, X_{k+1}]^{m'} \nearrow$.

This shows that the risk can be lower bounded by a quantity in $\text{Conv}\{\eta, f(\eta,1), \ldots, f(\eta,m)\}$ which is then bigger than $\min\{\eta, f(\eta,1), \ldots, f(\eta,m)\}$ which is positive. □

This result shows that for certain data distributions with isolated atoms, JointExp is not consistent, even asymptotically, on many instances of the estimation problem (i.e. not on unrealistic corner cases). This behavior is all the more counterintuitive as one would think that on datasets with a lot of collisions, very little noise would be needed to ensure privacy since the points are already indistinguishable.

Example 6.3.2. Consider the private estimation of the median (i.e. $m = 1$ quantile, and $\mathbf{p} = (1/2)$) on $[a,b] = [-1,1]$. Since $m = 1$, JointExp coincides with ExponentialQuantile, and when all data points are equal to 0 (i.e. $\mathbb{P}_X = \delta_0$) its output is uniformly distributed in $[-1,1]$ *whatever the sample size* n as long as it is even.

When considering estimation on real-world distributions, many real-life datasets show *accumulation points* and can be modeled as continuous distributions with some Diracs at specific points. A famous example is the revenue statistics of the US Census Bureau: many participants in surveys are not qualified to have some category of revenue (too young or not investing in some assets) hence the presence of accumulations at the zero value for these categories. In fact, any continuous variable that is censored, conditional on some other variable or generated by mimetic agents tending to repeat exactly some values, will show accumulation points where JointExp has great chances to fail.

This type of failure may seem surprising given a) the strong connection between JointExp and the Inverse Sensitivity; and b) existing performance guarantees for *smoothed* Inverse Sensitivity mechanisms [Asi & Duchi, 2020b, Asi & Duchi, 2020a]. Indeed, while JointExp is not smoothed, smoothing convolves the output distribution with a max kernel, increasing the volume of the maximum of the distribution to circumvent the difficulties raised by isolated atoms. This approach is perfectly viable when JointExp is applied with $m = 1$ (so in the case of RecExp), however, when m is strictly bigger, the problem becomes intractable.

As a tractable alternative, we propose a heuristic algorithm based on noise addition prior to the application of JointExp, and we show that this mechanism is endowed with privacy and consistency guarantees. Note that the exposed problems with atomic distribution also occur for highly concentrated continuous distributions.

Another possible solution would be to discretize the output space. However, the resulting algorithm would have a complexity of $O(f(m, n, \delta) + 1/\delta^m)$ where δ is the precision of the discretization and f is some function. Since this is exponential in the number of quantiles, it suffers from the curse of dimensionality, and we argue that jittering is a better alternative.

6.3.2 Introducing the HSJointExp algorithm

Since JointExp has a density that is constant on the blocks

$$\big([X_{i_1}, X_{i_1+1}) \times \ldots \times \cdots \times [X_{i_m}, X_{i_m+1})\big) \cap [a, b]^{m\nearrow}$$

for $\mathbf{i} = (i_1, \ldots, i_m) \in O'$, it fails when the blocks that have a great utility (i.e. the ones leading to interesting quantile candidates) have a volume that is too small. By adding noise to the data points, we ensure a minimal volume for the blocks, and in particular for the interesting regions, while only shifting the empirical quantiles of the dataset by a small amount.

Let $w_1, \ldots, w_n$ be i.i.d variables, and let

$$\tilde{\mathbf{X}} = (X_1 + w_1, \ldots, X_n + w_n)\ . \tag{6.3}$$

The Heuristically Smoothed JointExp (HSJointExp) is defined as the algorithm that applies JointExp on the noisy data $\tilde{\mathbf{X}}$.

Other possible solutions could be to discretize the output space [Xiao et al., 2010], or to use the smoothing trick of the inverse sensitivity mechanism [Asi & Duchi, 2020b, Asi & Duchi, 2020a, Asi et al., 2023]. However, those two approaches are computationally hard and suffer from the curse of dimensionality.

Let us now discuss the choice of the distribution $\mathbb{P}_w$ of the $(w_i)'s$. Discrete noise distributions (for instance Bernoulli noise scaled by some $\alpha > 0$: $\frac{w}{\alpha} \sim \mathcal{B}(\frac{1}{2})$) may seem interesting because they lead to easily tuneable data gaps. However, this often just creates new instances where JointExp fails. Indeed, adding discrete noise to data distributions with accumulation points creates new accumulation points.

For this reason, we focus in the sequel on continuous noise distributions with a density denoted by π_w. The density $\pi_{\tilde{X}}$ of the noisy data $\tilde{X}$ is hence given by the convolution formula,

$$\forall t \in \mathbb{R}, \quad \pi_{\tilde{X}}(t) = \int \pi_w(t-x)\mathbb{P}_X(dx) \ . \tag{6.4}$$

A typical choice of noise discussed in the sequel is the uniform distribution on the interval $[-\alpha, \alpha]$.

Before discussing the choice of the scale parameter α, we remark that HSJointExp consists of the addition of i.i.d. noise prior to running JointExp. Its privacy guarantees are thus a direct consequence of the following generic composition lemma.

Proposition 6.3.3. *Let* $\mathbf{w}$ *be a random variable on* $\mathbb{R}^n$ *with probability distribution* $\mathbb{P}_{\mathbf{w}}$ *that is invariant by permutations of the components of the vector. If* $\mathfrak{M}$ *is* ϵ*-DP on* $\mathcal{X}^n$*, then* $\mathbf{X} \mapsto \mathfrak{M}(\texttt{proj}_{\mathcal{X}^n}(\mathbf{X}+\mathbf{w}))$ *is also* ϵ*-DP.*

Proof. Let $\mathfrak{M}$ be a ϵ-DP algorithm on $\mathcal{X}^n$, $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^n$ such that $\mathbf{X} \sim \mathbf{X}'$. Then, for every $\mathbf{w} \in \mathbb{R}^n, \text{proj}_{\mathcal{X}^n}(\mathbf{X}+\mathbf{w}) \sim \text{proj}_{\mathcal{X}^n}(\mathbf{X}'+\sigma(\mathbf{w}))$ for a specific permutation of the components σ. For each measurable set $\mathcal{S} \subset O$ we get

$$\begin{aligned}
&\mathbb{P}(\mathfrak{M}(\text{proj}_{\mathcal{X}^n}(\mathbf{X}+\mathbf{w})) \in \mathcal{S}) \\
&= \int_{\mathbb{R}^n} \mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\text{proj}_{\mathcal{X}^n}(\mathbf{X}+\mathbf{w})) \in \mathcal{S})\mathbb{P}_{\mathbf{w}}(d\mathbf{w}) \\
&\leq e^{\epsilon} \int_{\mathbb{R}^n} \mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\text{proj}_{\mathcal{X}^n}(\mathbf{X}'+\sigma(\mathbf{w}))) \in \mathcal{S})\mathbb{P}_{\mathbf{w}}(d\sigma(\mathbf{w})) \\
&= e^{\epsilon} \int_{\mathbb{R}^n} \mathbb{P}_{\mathfrak{M}}(\mathfrak{M}(\text{proj}_{\mathcal{X}^n}(\mathbf{X}'+\mathbf{w})) \in \mathcal{S})\mathbb{P}_{\mathbf{w}}(d\mathbf{w}) \\
&= e^{\epsilon}\mathbb{P}(\mathfrak{M}(\text{proj}_{\mathcal{X}^n}(\mathbf{X}'+\mathbf{w})) \in \mathcal{S})
\end{aligned}$$

which completes the proof. □

The projection step `proj` onto the data space $\mathcal{X}^n$ is necessary because JointExp needs to know the range of the data. Note that $\mathcal{X}^n$ could be replaced by any set of the form $[a-\delta_{\alpha,n}, b+\delta_{\alpha,n}]^n$ where $\delta_{\alpha,n}$ is a quantity that depends on α and n. So for instance, if the

noise follows a uniform distribution on the interval $[-\alpha, \alpha]$, projecting on $[a-\alpha, b+\alpha]^n$ (does nothing) and then running JointExp on $[a-\alpha, b+\alpha]$ ensures that no point will overflow.

6.3.3 Consistency of HSJointExp on constant data

In order to give some insight on the general analysis of HSJointExp, and to explain the choice that we suggest for the amplitude α of the noise, we start by discussing the simple setting of Example 6.3.2 where $X_i \equiv 0$ and JointExp is known to fail. We consider uniform noise with distribution $d\mathbb{P}_w(w) = \frac{\mathbb{1}_{[-\alpha,\alpha]}(w)}{2\alpha} dw$, and HSJointExp returns the output of ExponentialQuantile/JointExp with $m = 1$ on the noisy data $\tilde{X}$:

$$M := \mathcal{E}^{(2/\epsilon)}_{u_{\mathrm{JE}}}(\tilde{\mathbf{X}}).$$

The true median of the dataset is 0, and we study the quadratic risk $\mathbb{E}(M^2)$ of our mechanism. Note that the classical way of analyzing exponential mechanisms is to use the utility bounds found in [McSherry & Talwar, 2007]. However, here we do not have the required level of control on the normalization factor. We hence go for a more direct way of controlling the output distribution. Denoting by $N(x,y) = \sum_{i=1}^n \mathbb{1}_{[x,y)}(0+w_i)$ the number of noisy points falling in the interval $[x,y)$, we define the event

$$G := \big\{N(-\alpha, -\alpha/4) \geq n/4\big\} \cap \big\{N(\alpha/4, \alpha) \geq n/4\big\} \,.$$

Since $N(-\alpha,-\alpha/4) \overset{\mathcal{L}}{=} N(\alpha/4,\alpha) \sim \mathcal{B}(n, 3/8)$, by Hoeffding's inequality, the probability of G is a least $1 - 2\exp(-n/32)$. Moreover, on the event G, for every $x \in [-\alpha/4, \alpha/4]$ one has $N(-\alpha, x) \geq n/4$ and $N(x,\alpha) \geq n/4$; hence, the minimal number of sample points that need to be changed so as to reach a median equal to x is at most $\delta^{\mathrm{JE}}(1, \tilde{X}, x) = \big|n/2 - N(-1,x)\big| \leq n/4$ (see Figure 6.2), and $-u_{\mathrm{JE}}(\tilde{X}, x) \leq n/8$. On the other hand, for every $x \notin [-\alpha,\alpha]$, $\delta^{\mathrm{JE}}(1,\tilde{X},x) = n/2$ and $u_{\mathrm{JE}}(\tilde{X},x) = n/4$. Since the density of M at $x \in [-1,1]$ is equal to $\exp\big(-u_{\mathrm{JE}}(\tilde{X},x)\epsilon/2\big) / \int_{-1}^1 \exp\big(-u_{\mathrm{JE}}(\tilde{X},t)\epsilon/2\big)dt$,

$$\begin{aligned}
\mathbb{P}\big(|M| > \alpha \big| G\big) &\leq \frac{\mathbb{P}\big(|M| > \alpha \big| G\big)}{\mathbb{P}\big(|M| \leq \alpha/4 \big| G\big)} \\
&\leq \frac{2 \times e^{-n\epsilon/8}}{\alpha/2 \times e^{-n\epsilon/16}} = \frac{4e^{-n\epsilon/16}}{\alpha} \,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left(M^2\right) &\leq 1^2 \left(\mathbb{P}(\bar{G}) + \mathbb{P}\big(|M| > \alpha \big| G\big)\right) + \alpha^2 \, \mathbb{P}\big(|M| \leq \alpha \big| G\big) \\
&\leq e^{-n/32} + \frac{4e^{-n\epsilon/16}}{\alpha} + \alpha^2 \,.
\end{aligned}$$

Choosing $\alpha = e^{-n\epsilon/48}$ yields

$$\mathbb{E}\left(M^2\right) \leq 5e^{-n\epsilon/24} + e^{-n/32} \,.$$

We conclude that, contrary to JointExp, HSJointExp is here consistent as soon as $n\epsilon \to \infty$, which is anyway a necessary condition. Besides, the analysis provides a simple and generic way to tune the noise amplitude α as a function of n and ϵ.

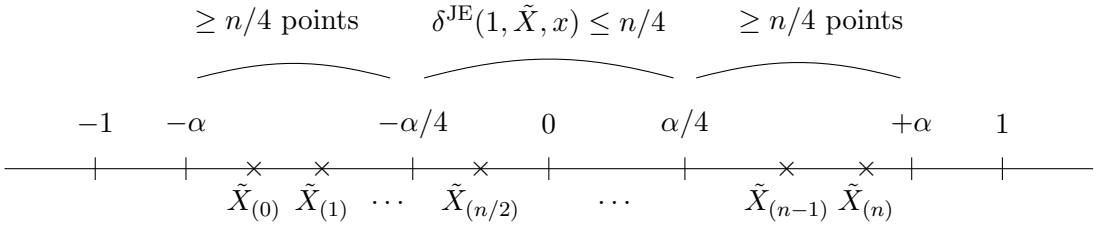


Figure 6.2: $\delta^{\mathrm{JE}}(1, \tilde{X}, x)$ is bounded by $n/4$ for $-\alpha/4 \leq x \leq \alpha/4$ on the event G.

6.3.4 General Consistency of HSJointExp

The consistency of HSJointExp is based on the idea of leveraging the consistency of JointExp on continuous distributions. We decompose the error on the estimation in two terms: The error measuring the gap between the quantiles of $\mathbb{P}_{\mathbf{X}}$ and the ones of $\mathbb{P}_{\tilde{\mathbf{X}}}$ and the error made by JointExp on the estimation of the quantiles of $\mathbb{P}_{\tilde{\mathbf{X}}}$.

The first term can be controlled by the following general purpose proposition.

Proposition 6.3.4. *For any non-increasing $f : \mathbb{R} \to [0,1]$ such that $\forall t \geq 0, \mathbb{P}(|w| > t) \leq f(t)$, then for every $p \in (0,1)$, for every $t \geq 0$ such that $1 - f(t) > 0$,*

$$F_{\tilde{X}}^{-1}(p) \leq F_X^{-1}\left(\frac{p}{1-f(t)}\right) + t\,,$$

$$\sup_{\delta \in (0,p)} -F_{-X}^{-1}\left(\frac{1-p+\delta}{1-f(t)}\right) - t \leq F_{\tilde{X}}^{-1}(p)\ .$$

Proof. Let $t \geq 0$ such that $1 - f(t) > 0$,

$$\begin{aligned}
&\mathbb{P}\left(X + w \leq F_X^{-1}\left(\frac{p}{1-f(t)}\right) + t\right) \\
&\geq \mathbb{P}\left(X + w \leq F_X^{-1}\left(\frac{p}{1-f(t)}\right) + t, |w| \leq t\right) \\
&\geq \mathbb{P}\left(X \leq F_X^{-1}\left(\frac{p}{1-f(t)}\right), |w| \leq t\right) \\
&\geq \mathbb{P}\left(X \leq F_X^{-1}\left(\frac{p}{1-f(t)}\right)\right) \mathbb{P}\left(|w| \leq t\right) \\
&\geq \frac{p}{1-f(t)}\left(1 - f(t)\right) \geq p\,.
\end{aligned}$$

So, $F_{\tilde{X}}^{-1}(p) \leq F_X^{-1}\left(\frac{p}{1-f(t)}\right) + t$. Let $\delta \in (0,p)$, the same arguments give

$$\mathbb{P}\left(X + w \leq -F_{-X}^{-1}\left(\frac{1-p+\delta}{1-f(t)}\right) - t\right) \leq p - \delta < p$$

which allows concluding with the desired result. □

For instance, when applied to some noise with distribution $d\mathbb{P}_w(w) = \frac{\mathbb{1}_{[-\alpha,\alpha]}(w)}{2\alpha}dw$ with $t = \alpha$ and $f(t) = 0$, if F_X is continuous and strictly increasing on a neighborhood of $F_X^{-1}(p)$, we can say that $|F_X^{-1}(p) - F_{\tilde{X}}^{-1}(p)| \leq \alpha$.

The second error term can be controlled with Theorem 6.2.15 assuming that we fall into its hypothesis. By adding some uniform noise in $[-\alpha, \alpha]$, we then obtain the following result:

Theorem 6.3.5. *If the distribution of X is a mixture of a finite number of Diracs in (a, b) and of a random variable Y with a continuous density π_Y on $[a, b]$ w.r.t. Lebesgue's measure such that $\pi_Y > 0$ on $[a, b] \setminus \mathcal{O}$ where $\mathcal{O}$ is a finite union of intervals and $\pi_Y = 0$ on $\mathcal{O}$, then for any precision δ and Lebesgue-almost-any probability vector $\mathbf{p}$, there exists a noise level $\alpha > 0$ such that the ϵ-DP estimator $\mathbf{q}$ based on HSJointExp satisfies*

$$\|\mathbf{q} - F_X^{-1}(\mathbf{p})\|_\infty \leq \delta$$

with high probability (as n grows).

Proof. We tune the noise w to have density $d\mathbb{P}_w(w) = \frac{\mathbb{1}_{[-\delta/2,\delta/2]}(w)}{\delta}dw$. Under the hypothesis, F_X^{-1} has a finite number of discontinuity points. We can apply Proposition 6.3.4 with $t = \delta/2$ and $f(t) = 0$ to get that for Lebesgue-almost-any $\mathbf{p}$,

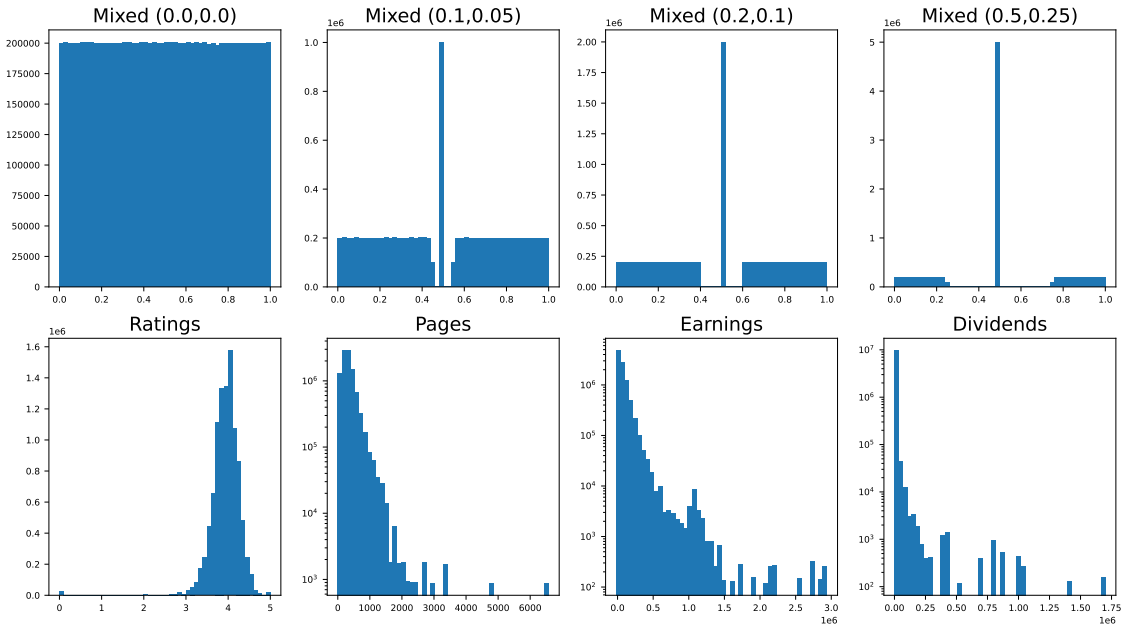
$$\|F_{\tilde{X}}^{-1}(\mathbf{p}) - F_X^{-1}(\mathbf{p})\|_\infty \leq \delta/2 \,.$$

In order to conclude, we can describe the density $\pi_{\tilde{X}}$ of the noisy random variable. It is piecewise continuous on $[a, b]$, $\pi_{\tilde{X}} > 0$ on $[a, b] \setminus \mathcal{O}'$ where $\mathcal{O}'$ is a finite union of intervals and $\pi_{\tilde{X}} = 0$ on $\mathcal{O}'$. Consequently, there only are a finite number of p's in $(0, 1)$ such that it is not possible to find a $\beta > 0$ such that $\pi_{\tilde{X}} > 0$ on $[F_{\tilde{X}}^{-1}(p) - \beta, F_{\tilde{X}}^{-1}(p) + \beta]$ and where $\pi_{\tilde{X}}$ is continuous on that interval. Any $\mathbf{p}$ that has no such p as any of its components qualifies and we can apply Theorem 6.2.15 to get that

$$\|\mathbf{q} - F_{\tilde{X}}^{-1}(\mathbf{p})\|_\infty \leq \delta/2$$

with high probability. We get the result by the triangle inequality. □

Theorem 6.3.5 states in particular that many distributions that satisfy the hypothesis of Proposition 6.3.1 and on which JointExp is not consistent also satisfy the hypothesis of Theorem 6.3.5 and HSJointExp can thus achieve arbitrary levels of precision on them (provided n is large enough).



Histograms representing $n = 10^7$ data points sampled from the original distributions and binned in 50 bins. Note that for Pages, Earnings and Dividends, the vertical axis is in $\log_{10}$-scale.

Figure 6.3: Distributions used for experiments

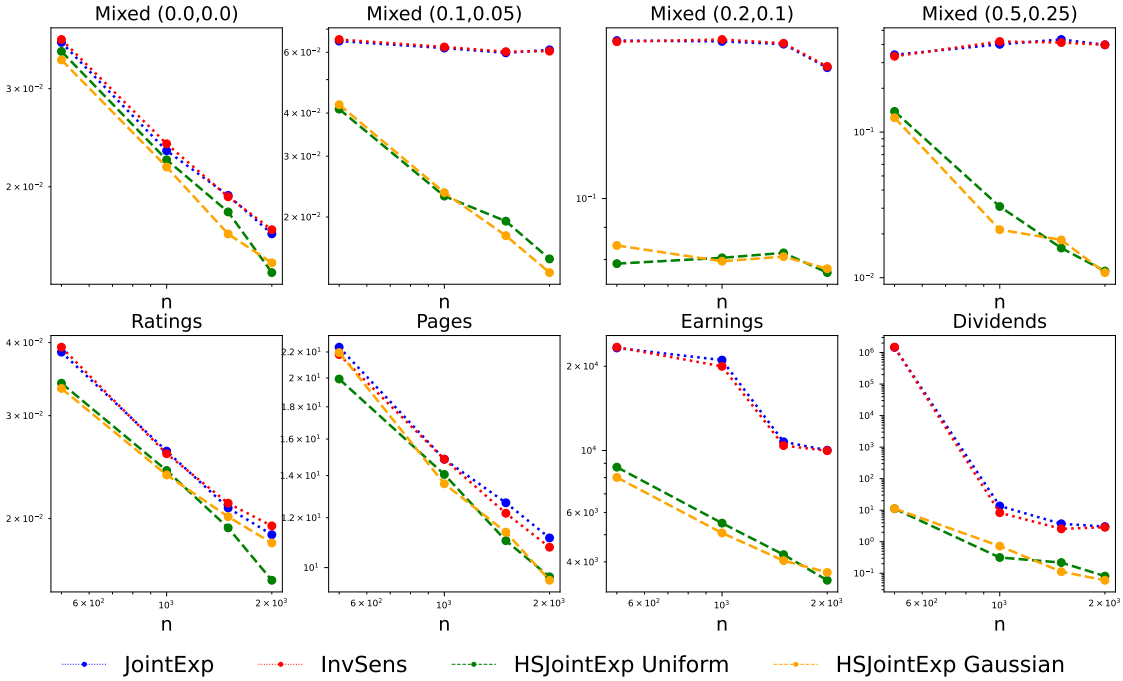
As highlighted by Section 6.3.3, working on much stricter distribution classes can lead to numerically tractable optimal levels of noise.

6.3.5 Numerical Results

This section presents the behaviors of JointExp, the Inverse Sensitivity mechanism and HSJointExp on synthetic and on real-world distributions. In particular, 6.3.5 is devoted to the presentation of the distributions of interest. Section 6.3.5 numerically studies the performance of the algorithms on the above-mentioned distributions. And finally, Section 6.3.5 looks at the possible numerical gain of privacy resulting from the noise addition.

Distributions

We claimed that HSJointExp has a huge advantage over regular JointExp in the case of distributions with isolated atoms. In order to test it numerically, we propose to do so with synthetic data in the first place. Indeed, in allows us to tune various interesting quantities. For real world distributions, it is harder to identify which ones satisfy the condition of having isolated atoms. We propose to evaluate the performance of the algorithms by identifying a real-world distribution with the empirical distribution of a real-world dataset. The concentration of this dataset (i.e. how peaked its histogram is) is then the decisive criterion: The more concentrated it is, the more suboptimal JointExp/IS is expected to be compared to the smoothed variants.



The vertical axis reads the error $\mathbb{E}\left(\|\hat{\mathbf{q}} - F^{-1}(\mathbf{p})\|_\infty\right)$ where $\mathbf{p} = \left(\frac{1}{m+1}, \ldots, \frac{m}{m+1}\right)$ for $m = 8$, $\epsilon = 1$, $\hat{\mathbf{q}}$ is the private estimator, and $\mathbb{E}$ is estimated by Monte-Carlo averaging over 50 runs. For HSJointExp Uniform and Gaussian, the optimal noise level on a discretization of $\log_{10}$-resolution 2 of $[10^{-10}, 10^4]$ is selected. Note that both axis are in $\log_{10}$-scale.

Figure 6.4: Error of the estimators as a function of n

Mixed distributions (synthetic). For $p \in [0,1]$ and $\delta \in [0, 1/2]$, we define the *Mixed* distribution of parameters (p, δ) as the distribution with support in $[0, 1/2 - \delta] \cup \{1/2\} \cup [1/2 + \delta, 1]$ such that if a random variable X follows this distribution, we have $\mathbb{P}(X = 1/2) = p$, $\mathbb{P}(X \in [0, 1/2 - \delta]) = \mathbb{P}(X \in [1/2 + \delta, 1])$, and conditionally to the event $(X \in [0, 1/2 - \delta])$ or to the event $(X \in [1/2 + \delta, 1])$, X is uniform. In particular, the mixed distribution of parameters $(0, 0)$ is the uniform distribution on $[0, 1]$. In order to better visualize such distributions, sampled histograms are represented in Figure 6.3. The parameters ϵ and δ allow tuning, respectively, the probability of the atom and its isolation. The bigger they are, the more HSJointExp is expected to outperform the non-smoothed variants.

Pages and Ratings (real-world). The distributions that we call *Pages* and *Ratings* correspond to the empirical distributions of a collection of ratings and of number of pages of books from the Goodreads-Books dataset [Soumik,]. Gillenwater et al. [Gillenwater et al., 2021] used the same datasets as numerical evidences of the performance of JointExp for estimating empirical quantiles. Again, sampled histograms are represented in Figure 6.3. The distributions look relatively smooth (i.e. not too peaked and with a relatively small support), and as a result, we can expect the gap between JointExp/IS and HSJointExp

to be negligible.

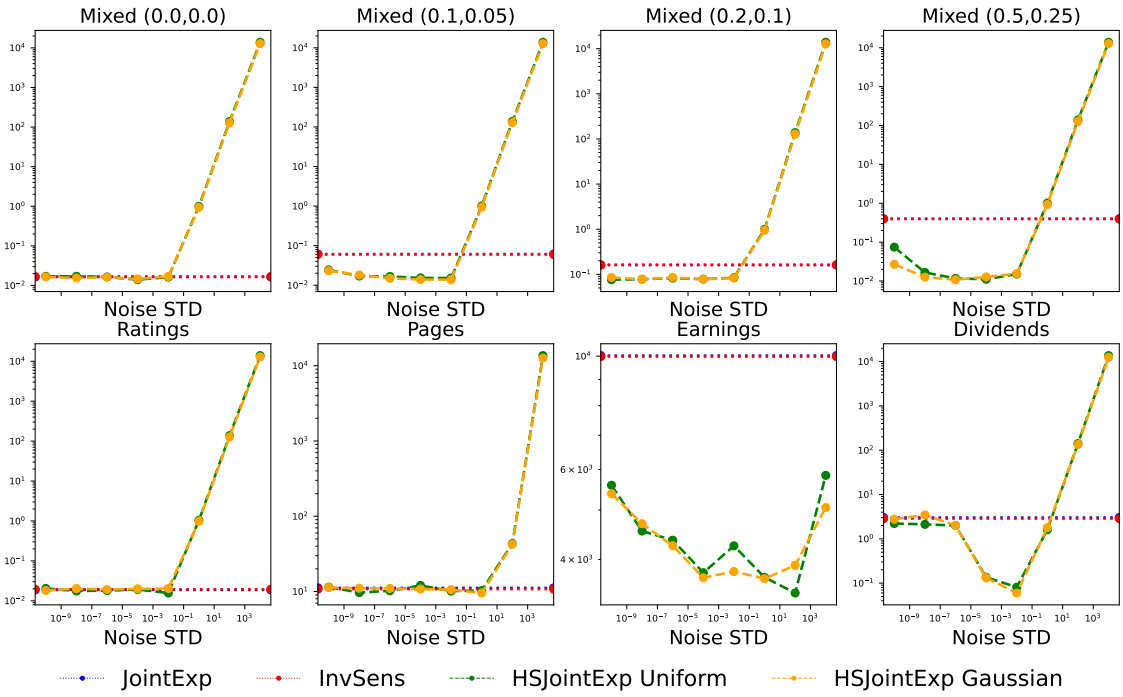
Earnings and Dividends (real-world). The distributions that we call *Earnings* and *Dividends* correspond respectively to the personal incomes and personal incomes from dividends categories of the US 2021 Census [Bureau, 2021]. Again, sampled histograms are represented in Figure 6.3. We can notice that contrary to the previous two real-world distributions, these two are much more concentrated. For Earnings, the concentration is due to the existence of categories of extremely high revenues. As a consequence, the support of the distribution is necessary big, and the algorithms that seek for privately estimating the quantiles have little information about the localization of the data points. On the other hand, the vast majority of people declare revenues below \$500 500, resulting in the high concentration of the distribution close to 0. For Dividends, the support is smaller, but since a big part of the population simply does not have any revenues from dividends, the distribution shows an accumulation point at 0. With both distributions, we expect the smoothing operation to vastly improve the performance of JointExp/IS.

Numerical Performance

Figure 6.4 and Figure 6.5 Compare the performance of JointExp, the Inverse Sensitivity mechanism and two variants of HSJointExp with uniform and Gaussian noise structure respectively on the distributions presented in Figure 6.3.

Complements on HSJointExp Uniform and Gaussian. The mechanism that we call *HSJointExp Uniform* is the application of JointExp post addition of centered uniform noise. If $[a, b]$ was our estimate of the support of the distribution, we apply JointExp on $[a - \sigma\sqrt{3}, b + \sigma\sqrt{3}]$ where σ is the standard deviation of the noise. In *HSJointExp Gaussian*, the centered uniform noise is replaced by centered Gaussian noise. The support of the resulting distribution is now infinite, and the projection step is therefore mandatory. We chose to project the data points in $[a - 5\sigma, b + 5\sigma]$ where σ is the standard deviation of the noise in order to make sure that most of the points will remain untouched by the projection step.

Analyzing the results of Figure 6.4. The first important fact to notice is the similar performance of JointExp and the Inverse Sensitivity mechanism, confirming the theoretical results. The second is the similar performance of HSJointExp Uniform and Gaussian, showing that the structure of the noise, given that it is regular enough, is not of critical importance. Finally, and probably the most important, we can compare the performance of JointExp/IS and of HSJointExp. On Mixed$(0, 0)$ (i.e. the uniform distribution on $[0, 1]$), Ratings and Pages, the two algorithms perform identically. This is what we expected given the smoothness of the distributions. On more concentrated distributions like Earnings and Dividends on the other hand, we see that HSJointExp vastly improves the performance of JointExp, sometimes by multiple orders of magnitude. Finally, Mixed$(0.1, 0.05)$, Mixed$(0.2, 0.1)$ and Mixed$(0.5, 0.25)$ demonstrate that the more isolated and probable the atoms of the distribution are, the more suboptimal JointExp is compared to the smoothed variants.



The vertical axis reads the error $\mathbb{E}\left(\|\hat{\mathbf{q}} - F^{-1}(\mathbf{p})\|_\infty\right)$ where $n = 2000$, $\mathbf{p} = \left(\frac{1}{m+1}, \ldots, \frac{m}{m+1}\right)$ for $m = 8$, $\epsilon = 1$, $\hat{\mathbf{q}}$ is the private estimator, and $\mathbb{E}$ is estimated by Monte-Carlo averaging over 50 runs. For HSJointExp Uniform and Gaussian, the optimal noise level on a discretization of $\log_{10}$-resolution 2 of $[10^{-10}, 10^4]$ is selected. Note that both axis are in $\log_{10}$-scale. The horizontal axis reads the standard deviation of the smoothing noise. JointExp and the Inverse Sensitivity mechanism are represented by horizontal bars, since they do not depend on the noise level.

Figure 6.5: Dependence on the smoothing level

Analyzing the results of Figure 6.5. Figure 6.5 shows the same results as Figure 6.4 but with an emphasis on the dependence on the noise level. For instance, we can see that when the smoothing operation allows for better performance, it is often the case for a large range of smoothing levels. Finally, we can numerically observe two limit behaviors that are quite intuitive : When the noise level tends to 0, HSJointExp performs as JointExp. Indeed, in this case, the smoothing trick has almost no effect on the distribution. When the noise level tends to $+\infty$ on the other hand, the performance of HSJointExp is terrible. This is also quite intuitive, since the smoothed distribution has lost almost all correlation with the original distribution. For all these reasons, we recommend tuning the noise as in the extreme case of the Dirac (see Section 6.3.3) since this value is small enough to not fall in the regime where the performance are degraded by the smoothing, but it still greatly improves the performance on degenerated distributions.

Privacy Amplification

A final property that we would like to explore is the possible amplification of privacy of HSJointExp. Indeed, adding Laplace or Gaussian noise to bounded quantities is a common way to make them private [Dwork et al., 2006b]. Furthermore, it is well known that some preprocessing steps (prior to the application of an already private mechanism) increase the provable privacy of the overall mechanism. This is for instance the case with subsampling [Balle et al., 2018b]. Consequently, one would think that adding noise to the data does not only preserve the privacy guarantees of the original mechanism (as stated by Proposition 6.3.3), but has reasonable chances to make it more private. In order to evaluate the actual privacy of our mechanism, we investigate its privacy loss:

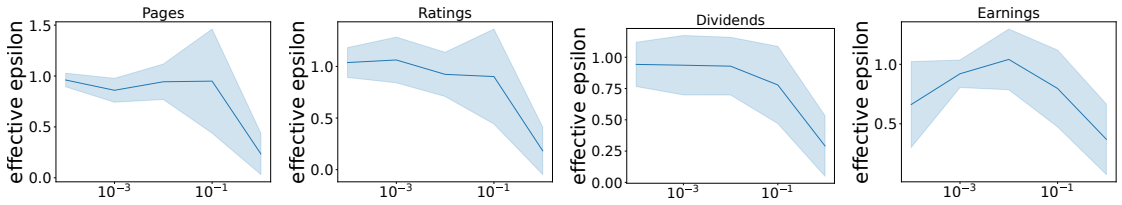
$$\mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{q}) := \frac{d\mathbb{P}/d\mathbf{q}\left(\text{JointExp}(\tilde{\mathbf{X}}) = \mathbf{q}\right)}{d\mathbb{P}/d\mathbf{q}\left(\text{JointExp}(\tilde{\mathbf{Y}}) = \mathbf{q}\right)}$$

for $\mathbf{Y} \sim \mathbf{X}$ and $\mathbf{q} \in O$ where $d\mathbb{P}/d\mathbf{q}\left(\text{JointExp}(\tilde{\mathbf{X}}) = \mathbf{q}\right)$ refers to the value of the density of HSJointExp applied to $\mathbf{X}$ at $\mathbf{q}$. For a given dataset $\mathbf{X}$, we define

$$\epsilon_{\text{eff}} := \sup_{\mathbf{X} \sim \mathbf{Y}} \sup_{\mathbf{q}} \log\left(\mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{q})\right)$$

the effective difficulty of distinguishing $\mathbf{X}$ from any of its neighbors. We always have that $\epsilon_{\text{eff}} \leq \epsilon$ but we would like to measure the difference between the two and its dependence on the noise level.

In Figure 6.6 we numerically estimate ϵ_{eff} in the following setup: For each of the datasets (noted $\mathbf{X}$), we estimate the median using HSJointExp with Laplace noise tuned with $\epsilon = 1$. We estimate $\mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{q})$ for any $\mathbf{Y} \sim \mathbf{X}$ by discretizing the search space of $\mathbf{Y}$ and by Monte Carlo averaging to integrate with respect to the noise.



The horizontal axis represents the standard deviation of the noise divided by the length of the support of the distribution. $\epsilon = 1$.

Figure 6.6: Evolution of ϵ_{eff} for the median estimation

The variance of the resulting ϵ_{eff} is high, but we can see two regimes: For low values of the noise, the privacy of the mechanism is unchanged. For high values of noise, on the other hand, $\epsilon_{\text{eff}} < \epsilon$ and differentiating the datasets from their neighbors is harder.

By crossing the results with Figure 6.5 however, it seems that the privacy amplification only occurs for values of the noise for which the utility of HSJointExp is already degraded compared to regular JointExp.

Conclusion

This thesis studied the impact of privacy on the statistical difficulty of estimation and learning problems. It started in Chapter 1 by experimentally demonstrating that sparsity in neural networks can be leveraged as a practical defense against membership inference attacks.

In order to provide strong privacy guarantees, Chapter 2 introduced the gold standard definition of differential privacy, and presented important privacy-preserving mechanisms. This definition of privacy allows for strong privacy guarantees, and is usually achieved by adding randomized noise to deterministic mechanisms.

Chapter 3 then presented a unified framework for deriving lower-bounds on the statistical testing difficulty between distributions under various notions of privacy. this characterization of the testing difficulty can in turn be used to characterize the estimation difficulty : how hard is it to learn distributions under privacy ? The rest of the thesis was devoted to the study of multiple examples.

As presented in Chapter 4, on unidimensional learning (or estimation) of regular-enough parametric problems (e.g. Bernoulli), the typical rate of estimation can usually be expressed as $\Theta\left(\frac{1}{n}\right)$. When constrained to be ϵ-differentially private, this rate of estimation typically becomes $\Theta\left(\max\left(\frac{1}{n}, \frac{1}{n^2\epsilon^2}\right)\right)$. Two regimes must be distinguished : in the regime $\epsilon = \Omega\left(\frac{1}{\sqrt{n}}\right)$, privacy has no real impact on the statistical complexity of the problem. It can basically be obtained "for free". In the regime $\epsilon \ll \frac{1}{\sqrt{n}}$ on the other hand, privacy has a necessary cost on estimation. In particular, those upper-bounds were recovered for the multiquantiles problem Chapter 6 in the regime where m, the number of quantiles, is fixed.

When dropping the parametric assumption, and when replacing it with an assumption

of β periodic-Sobolev densities, Chapter 5 demonstrated that the rate of estimation of the problem is $\tilde{\Theta}\left(\max\left(n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+1}}\right)\right)$ under (ϵ, δ)-DP for reasonably small δ (i.e. $\delta \ll 1/n$). This indicates that with only little assumptions on the regularity of the density (e.g. $\beta = 1$), the rate of estimation is degraded compared to the parametric case, and that the degradation due to privacy occurs much sooner than in the parametric case. On the contrary, when the regularity is assumed high (e.g. $\beta \approx \infty$), one recovers the parametric rate of estimation.

For learning high-dimensional parametric distributions (e.g. Gaussian distributions), Chapter 4 showed that the typical rate of estimation is $\Omega\left(\max\left(\frac{d}{n}, \frac{d^2}{n^2\epsilon^2}\right)\right)$. When compared to the non-private rate of estimation of $\Theta\left(\frac{d}{n}\right)$, one can see that *dimensionality disproportionately affects the privacy overhead*. In other words, it means that the privacy overhead has a worse scaling in the dimensionality than the usual statistical rate of estimation (d^2 instead of d).

These last observations suggest that in order to move forward, dimensionality and smoothness assumptions will probably play central roles in the next steps of private machine learning. For the smoothness, it has already been demonstrated that certain specific architectures are more prone to high accuracy with privacy than others [Papernot et al., 2021, Tramèr & Boneh, 2021]. For dimensionality, even if the typical degradation in d^2 sounds like a fatality, one must remember that it is a worst case scenario. Many empirical evidences suggest that the way data is represented in memory and the models that are used to process it are grossly overparametrized. In other words, the effective dimensionality of the problem is much lower than the one of its representation (the ambient one). Since in practice, algorithms like DP-SGD [Abadi et al., 2016] add a level of noise that is proportional to the dimension of the gradient [Yu et al., 2021, Tramèr & Boneh, 2021, Shen et al., 2021, Kurakin et al., 2022], working with the effective dimension instead of the ambient one can be a solution.

Smart techniques can be used in order to reduce the dimensionality of the problem. The input dimension can be reduced using ad-hoc techniques such as Fourier or Wavelet thresholding, thus getting rid of a part of the noise. Similar techniques can be applied with learned basis (e.g. PCA or dictionary learning). When one is interested in the overall dimension (i.e. data + model), other techniques have recently been proposed in order to reduce it during training. [Yu et al., 2021] and [Zhou et al., 2021] project or approximate the gradients on low-dimensional subspaces during the training. [Zhang et al., 2021c] focuses on subproblems where the gradients are sparse. As demonstrated in Chapter 1 and in [Adamczewski & Park, 2023], pruning neural networks is also a good way to reduce the dimensionality. On specific tasks, *compressive learning* can be used in order to only keep what's needed for the learning task [Schellekens et al., 2019a, Schellekens et al., 2019b, Chatalic et al., 2021]. Finally, transfer learning (i.e. the fact of fine-tuning a model that has been pre-trained on another task) can be used to efficiently reduce the dimension-

ality, and unlocked high accuracy on complex problems like ImageNet [De et al., 2022]. As for mixed privacy [Golatkar et al., 2022], this last solution requires being private only with respect to a fraction of the training data, and might be seen as a form of cheating. However, this assumption is perfectly realistic for certain types of data. For instance, with images, it wouldn't be unrealistic to consider images taken on the street to not require privacy protection.

A problem with transfer learning as presented in [De et al., 2022] is that it requires a lot of annotated data, which is expansive. I personally believe that *autosupervision* (or self-supervised learning) can be the key of high accuracy with privacy for a broad class of problems, and on a tight budget. Autosupervision aims at learning meaningful low-dimensional embeddings of the data from *unlabeled* data only. This embedding can then be used alternatively to the one used in [De et al., 2022]. In particular, recent methods allow for similar accuracies between self-supervised embeddings and supervised embeddings for many downstream tasks [He et al., 2020, Grill et al., 2020, Caron et al., 2020, Gidaris et al., 2021, Zheng et al., 2023]. For images, we can see that there is a huge intersection between the images that do not require privacy, and the ones that are easily collected. For instance, by strapping a camera to a car for a hew hours, it would be possible to collect a huge dataset. Applying autosupervision to this dataset (without annotations) may lead to good embeddings, suitable for private fine-tuning on the task at hand.

As the last words of this thesis, I would like to say that I really enjoyed working on these subjects, and in the working environment that was provided to me. Despite a somewhat stale first half of the thesis because of COVID-19 that severely impacted human interactions (including scientific), the much freer second half allowed me to discover the thriving and exciting world of research. Finally, I would like to thank once again the people without whom this thesis wouldn't have been possible : Aurélien Garivier and Rémi Gribonval for their kind supervision, and Nicolas Grislain for introducing me to the ins and outs of differential privacy.

Bibliography

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, 308–318. https://doi.org/10.1145/2976749.2978318 41, 56, 95, 132, 172

[Abowd, 2018] Abowd, J. M. (2018). The us census bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867–2867. 22

[Acar et al., 2018] Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.*, 51(4), 79:1–79:35. https://doi.org/10.1145/3214303 19

[Acharya et al., 2021a] Acharya, J., Canonne, C., Singh, A. V., & Tyagi, H. (2021a). Optimal rates for nonparametric density estimation under communication constraints. *Advances in Neural Information Processing Systems*, volume 34, 26754–26766. https://proceedings.neurips.cc/paper_files/paper/2021/file/e1021d43911ca2c1845910d84f40aeae-Paper.pdf 24, 145

[Acharya et al., 2021b] Acharya, J., Canonne, C. L., Freitag, C., Sun, Z., & Tyagi, H. (2021b). Inference under information constraints iii: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory*, 2(1), 253–267. https://doi.org/10.1109/JSAIT.2021.3053569 24, 145

[Acharya et al., 2021c] Acharya, J., Canonne, C. L., Mayekar, P., & Tyagi, H. (2021c). Information-constrained optimization: can adaptive processing of gradients help?

CoRR, abs/2104.00979. https://arxiv.org/abs/2104.00979 24, 145

[Acharya et al., 2021d] Acharya, J., Canonne, C. L., Sun, Z., & Tyagi, H. (2021d). Unified lower bounds for interactive high-dimensional estimation under information constraints. *CoRR*, abs/2010.06562. https://arxiv.org/abs/2010.06562 24, 145

[Acharya et al., 2018] Acharya, J., Sun, Z., & Zhang, H. (2018). Differentially private testing of identity and closeness of discrete distributions. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 6879–6891. https://proceedings.neurips.cc/paper/2018/hash/7de32147a4f1055bed9e4faf3485a84d-Abstract.html 145

[Acharya et al., 2021e] Acharya, J., Sun, Z., & Zhang, H. (2021e). Differentially private assouad, fano, and le cam. *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, 48–78. http://proceedings.mlr.press/v132/acharya21a.html 30, 53, 63, 64, 65, 66, 67, 68, 84, 99, 101, 102, 105, 106, 111, 118, 145

[Adamczewski & Park, 2023] Adamczewski, K. & Park, M. (2023). Differential privacy meets neural network pruning. *CoRR*, abs/2303.04612. https://doi.org/10.48550/arXiv.2303.04612 41, 172

[Aden-Ali et al., 2021] Aden-Ali, I., Ashtiani, H., & Kamath, G. (2021). On the sample complexity of privately learning unbounded high-dimensional gaussians. *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, 185–216. http://proceedings.mlr.press/v132/aden-ali21a.html 99

[Alabi et al., 2022] Alabi, D., Ben-Eliezer, O., & Chaturvedi, A. (2022). Bounded space differentially private quantiles. *CoRR*, abs/2201.03380. https://arxiv.org/abs/2201.03380 153

[Allen,] Allen, J. e. a. (*). *Smartnoise core differential privacy library*. https://github.com/opendp/smartnoise-core. 131, 135

[Aloupis et al., 2013] Aloupis, G., Pérez-Rosés, H., Pineda-Villavicencio, G., Taslakian, P., & Trinchet-Almaguer, D. (2013). Fitting voronoi diagrams to planar tessela-

tions. *Combinatorial Algorithms - 24th International Workshop, IWOCA 2013, Rouen, France, July 10-12, 2013, Revised Selected Papers*, volume 8288 of *Lecture Notes in Computer Science*, 349–361. https://doi.org/10.1007/978-3-642-45278-9_30 18

[Angel & Spinka, 2021] Angel, O. & Spinka, Y. (2021). *Pairwise optimal coupling of multiple random variables*. 78

[Ash & Bolker, 1985] Ash, P. F. & Bolker, E. D. (1985). Recognizing dirichlet tessellations. *Geometriae Dedicata*, 19, 175–206. https://doi.org/10.1007/BF00181470 18

[Asi & Duchi, 2020a] Asi, H. & Duchi, J. C. (2020a). Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/a267f936e54d7c10a2bb70dbe6ad7a89-Abstract.html 31, 55, 56, 136, 160

[Asi & Duchi, 2020b] Asi, H. & Duchi, J. C. (2020b). Near instance-optimality in differential privacy. *CoRR*, abs/2005.10630. https://arxiv.org/abs/2005.10630 31, 55, 56, 136, 160

[Asi et al., 2023] Asi, H., Ullman, J. R., & Zakynthinou, L. (2023). From robustness to privacy and back. *CoRR*, abs/2302.01855. https://doi.org/10.48550/arXiv.2302.01855 56, 160

[Aurenhammer, 1987] Aurenhammer, F. (1987). Recognising polytopical cell complexes and constructing projection polyhedra. *J. Symb. Comput.*, 3(3), 249–255. https://doi.org/10.1016/S0747-7171(87)80003-2 18

[Avella-Medina et al., 2021] Avella-Medina, M., Bradshaw, C., & Loh, P. (2021). Differentially private inference via noisy optimization. *CoRR*, abs/2103.11003. https://arxiv.org/abs/2103.11003 56, 57

[Bach, 2021] Bach, F. (2021). Learning theory from first principles. *Online version*. 16

[Backstrom et al., 2007] Backstrom, L., Dwork, C., & Kleinberg, J. M. (2007). Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural

steganography. *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, 181–190. https://doi.org/10.1145/1242572.1242598 16

[Balle et al., 2018a] Balle, B., Barthe, G., & Gaboardi, M. (2018a). Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 6280–6290. https://proceedings.neurips.cc/paper/2018/hash/3b5020bb891119b9f5130f1fea9bd773-Abstract.html 50

[Balle et al., 2018b] Balle, B., Barthe, G., & Gaboardi, M. (2018b). Privacy amplification by subsampling: Tight analyses via couplings and divergences. *CoRR*, abs/1807.01647. http://arxiv.org/abs/1807.01647 169

[Barber & Duchi, 2014] Barber, R. F. & Duchi, J. C. (2014). *Privacy and statistical risk: Formalisms and minimax bounds*. 30, 99, 101, 102, 111

[Bargiotas et al., 2022] Bargiotas, I., Wang, D., Mantilla, J., Quijoux, F., Moreau, A., Vidal, C., Barrois, R., Nicolaï, A., Audiffren, J., Labourdette, C., Bertin-Hugaul, F., Oudre, L., Buffat, S., Yelnik, A., Ricard, D., Vayatis, N., & Vidal, P.-P. (2022). Preventing falls: the use of machine learning for the prediction of future falls in individuals without history of fall. *Journal of Neurology*, 270. https://doi.org/10.1007/s00415-022-11251-3 13

[Barnes et al., 2020a] Barnes, L. P., Chen, W.-N., & Ozgur, A. (2020a). Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3), 645–659. https://doi.org/10.1109/JSAIT.2020.3039461 145

[Barnes et al., 2019] Barnes, L. P., Han, Y., & Ozgur, A. (2019). Fisher information for distributed estimation under a blackboard communication protocol. *2019 IEEE International Symposium on Information Theory (ISIT)*, 2704–2708. https://doi.org/10.1109/ISIT.2019.8849821 24, 145

[Barnes et al., 2020b] Barnes, L. P., Han, Y., & Özgür, A. (2020b). Lower bounds for learning distributions under communication constraints via fisher information. *Journal of Machine Learning Research*, 21, Paper No. 236, 30. https://jmlr.csail.mit.edu/papers/volume21/19-737/19-737.pdf 24, 145

[Bartlett & Mendelson, 2002] Bartlett, P. L. & Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3, 463–482. http://jmlr.org/papers/v3/bartlett02a.html 22

[Bassily et al., 2019] Bassily, R., Feldman, V., Talwar, K., & Thakurta, A. G. (2019). Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 11279–11288. https://proceedings.neurips.cc/paper/2019/hash/3bd8fdb090f1f5eb66a00c84dbc5ad51-Abstract.html 56

[Bassily et al., 2014] Bassily, R., Smith, A. D., & Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, 464–473. https://doi.org/10.1109/FOCS.2014.56 56

[Bauschke & Combettes, 2011] Bauschke, H. H. & Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer. https://doi.org/10.1007/978-1-4419-9467-7 57

[Beck, 2017] Beck, A. (2017). *First-Order Methods in Optimization* . Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974997 95

[Belkin et al., 2019] Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Science*, 116(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116 29, 33

[Bellet et al., 2018] Bellet, A., Guerraoui, R., Taziki, M., & Tommasi, M. (2018). Personalized and private peer-to-peer machine learning. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, 473–481. https://proceedings.mlr.press/v84/bellet18a.html 19

[Bennett et al., 2007] Bennett, J., Lanning, S., et al. (2007). The netflix prize. *Proceedings of KDD cup and workshop*, volume 2007, 35. 14

[Berghel et al., 2022] Berghel, S., Bohannon, P., Desfontaines, D., Estes, C., Haney, S.,

Hartman, L., Hay, M., Machanavajjhala, A., Magerlein, T., Miklau, G., Pai, A., Sexton, W., & Shrestha, R. (2022). Tumult Analytics: a robust, easy-to-use, scalable, and expressive framework for differential privacy. *arXiv preprint arXiv:2212.04133*. 135

[Berrett & Butucea, 2019] Berrett, T. & Butucea, C. (2019). *Classification under local differential privacy*. https://arxiv.org/abs/1912.04629 145

[Berrett et al., 2021] Berrett, T. B., Györfi, L., & Walk, H. (2021). Strongly universally consistent nonparametric regression and classification with privatised data. *Electronic Journal of Statistics*, 15(1), 2430 – 2453. https://doi.org/10.1214/21-EJS1845 101

[Biswas et al., 2020] Biswas, S., Dong, Y., Kamath, G., & Ullman, J. R. (2020). Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/a684eceee76fc522773286a895bc8436-Abstract.html 99

[Blocki et al., 2012] Blocki, J., Blum, A., Datta, A., & Sheffet, O. (2012). The johnson-lindenstrauss transform itself preserves differential privacy. *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, 410–419. https://doi.org/10.1109/FOCS.2012.67 153

[Blumer et al., 1989] Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4), 929–965. https://doi.org/10.1145/76359.76371 21

[Bonawitz et al., 2019] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al. (2019). Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1, 374–388. 19

[Boucheron et al., 2019] Boucheron, Boyer, & Ryder (2019). Cours de statistiques, ens ulm, fimfa ens. *Lectures at the École Normale Supérieure*. https://stephane-v-boucheron.fr/files/stats/notes-17-18.pdf 94, 95

[Bousquet & Elisseeff, 2002] Bousquet, O. & Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.*, 2, 499–526. http://jmlr.org/papers/v2/

bousquet02a.html 22

[Boyd et al., 2011] Boyd, S. P., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1), 1–122. https://doi.org/10.1561/2200000016 58

[Brat et al., 2020] Brat, G. A., Weber, G. M., Gehlenborg, N., Avillach, P., Palmer, N. P., Chiovato, L., Cimino, J. J., Waitman, L. R., Omenn, G. S., Malovini, A., Moore, J. H., Beaulieu-Jones, B. K., Tibollo, V., Murphy, S. N., L'Yi, S., Keller, M. S., Bellazzi, R., Hanauer, D. A., Serret-Larmande, A., Gutiérrez-Sacristán, A., Holmes, J. J., Bell, D. S., Mandl, K. D., Follett, R. W., Klann, J. G., Murad, D. A., Scudeller, L., Bucalo, M., Kirchoff, K. G., Craig, J. B., Obeid, J. S., Jouhet, V., Griffier, R., Cossin, S., Moal, B., Patel, L. P., Bellasi, A., Prokosch, H., Kraska, D., Sliz, P., Tan, A. L. M., Ngiam, K. Y., Zambelli, A., Mowery, D. L., Schiver, E., Devkota, B., Bradford, R. L., Daniar, M., Daniel, C., Benoit, V., Bey, R., Paris, N., Serre, P., Orlova, N., Dubiel, J., Hilka, M., Jannot, A., Bréant, S., Leblanc, J., Griffon, N., Burgun, A., Bernaux, M., Sandrin, A., Salamanca, E., Cormont, S., Ganslandt, T., Gradinger, T., Champ, J., Boeker, M., Martel, P., Esteve, L., Gramfort, A., Grisel, O., Leprovost, D., Moreau, T., Varoquaux, G., Vie, J., Wassermann, D., Mensch, A., Caucheteux, C., Haverkamp, C., Lemaitre, G., Bosari, S., Krantz, I. D., South, A. M., Cai, T., & Kohane, I. S. (2020). International electronic health record-derived COVID-19 clinical course profiles: the 4ce consortium. *npj Digit. Medicine*, 3. https://doi.org/10.1038/s41746-020-00308-0 13

[Brown et al., 2021] Brown, G., Gaboardi, M., Smith, A. D., Ullman, J. R., & Zakynthinou, L. (2021). Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 7950–7964. https://proceedings.neurips.cc/paper/2021/hash/42778ef0b5805a96f9511e20b5611fce-Abstract.html 99

[Bun et al., 2019] Bun, M., Kamath, G., Steinke, T., & Wu, Z. S. (2019). Private hypothesis selection. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 156–167. https://proceedings.neurips.cc/paper/2019/hash/9778d5d219c5080b9a6a17bef029331c-Abstract.html 99

[Bun et al., 2021] Bun, M., Kamath, G., Steinke, T., & Wu, Z. S. (2021). Private hypothesis selection. *IEEE Trans. Inf. Theory*, 67(3), 1981–2000. https://doi.org/10.1109/TIT.2021.3049802 99

[Bun et al., 2015] Bun, M., Nissim, K., Stemmer, U., & Vadhan, S. P. (2015). Differentially private release and learning of threshold functions. *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, 634–649. https://doi.org/10.1109/FOCS.2015.45 153

[Bun & Steinke, 2016] Bun, M. & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, volume 9985 of *Lecture Notes in Computer Science*, 635–658. https://doi.org/10.1007/978-3-662-53641-4_24 30, 46, 47, 48, 70, 74, 75, 88, 102, 129

[Bureau, 2021] Bureau, U. S. C. (2021). *2021 annual social and economic supplements*. https://www.census.gov/data/datasets/2021/demo/cps/cps-asec-2021.html. 167

[Butucea et al., 2019] Butucea, C., Dubois, A., Kroll, M., & Saumard, A. (2019). Local differential privacy: Elbow effect in optimal density estimation and adaptation over besov ellipsoids. *CoRR*, abs/1903.01927. http://arxiv.org/abs/1903.01927 101, 145

[Cai et al., 2021] Cai, T. T., Wang, Y., & Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5), 2825–2850. 99

[Caron et al., 2020] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html 173

[Chatalic et al., 2021] Chatalic, A., Schellekens, V., Houssiau, F., de Montjoye, Y.-A., Jacques, L., & Gribonval, R. (2021). Compressive Learning with Privacy Guarantees. *Information and Inference*. https://doi.org/10.1093/imaiai/iaab005 172

[Chourasia et al., 2021] Chourasia, R., Ye, J., & Shokri, R. (2021). Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34, 14771–14781. 96

[Cyffers et al., 2023] Cyffers, E., Bellet, A., & Basu, D. (2023). From noisy fixed-point iterations to private ADMM for centralized and federated learning. https://doi.org/10.48550/arXiv.2302.12559 58

[Czernichow et al., 2020] Czernichow, S., Beeker, N., Rives-Lange, C., Guerot, E., Diehl, J.-L., Katsahian, S., Hulot, J.-S., Poghosyan, T., Carette, C., Jannot, A.-S., & AP-HP / Universities / INSERM COVID-19 research collaboration and AP-HP COVID CDR Initiative (2020). Obesity doubles mortality in patients hospitalized for severe acute respiratory syndrome coronavirus 2 in paris hospitals, france: A cohort study on 5,795 patients. *Obesity (Silver Spring, Md.)*, 28(12), 2282—2289. https://doi.org/10.1002/oby.23014 13

[Dao et al., 2022] Dao, T., Chen, B., Sohoni, N. S., Desai, A. D., Poli, M., Grogan, J., Liu, A., Rao, A., Rudra, A., & Ré, C. (2022). Monarch: Expressive structured matrices for efficient and accurate training. *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 4690–4721. https://proceedings.mlr.press/v162/dao22a.html 34, 37, 38

[Dao et al., 2020] Dao, T., Gu, A., Eichhorn, M., Rudra, A., & Ré, C. (2020). *Learning fast algorithms for linear transforms using butterfly factorizations*. 37

[Dao et al., 2021] Dao, T., Sohoni, N. S., Gu, A., Eichhorn, M., Blonder, A., Leszczynski, M., Rudra, A., & Ré, C. (2021). *Kaleidoscope: An efficient, learnable representation for all structured linear maps*. 37

[De et al., 2022] De, S., Berrada, L., Hayes, J., Smith, S. L., & Balle, B. (2022). Unlocking high-accuracy differentially private image classification through scale. *CoRR*, abs/2204.13650. https://doi.org/10.48550/arXiv.2204.13650 173

[Denisov et al., 2022] Denisov, S., McMahan, H. B., Rush, J., Smith, A. D., & Thakurta, A. G. (2022). Improved differential privacy for SGD via optimal private linear operators on adaptive streams. *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/271ec4d1a9ff5e6b81a6e21d38b1ba96-Abstract-Conference.html 153

[Devroye, 1986] Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer. https://doi.org/10.1007/978-1-4613-8643-8 31, 131, 140

[Diakonikolas et al., 2015] Diakonikolas, I., Hardt, M., & Schmidt, L. (2015). Differentially private learning of structured discrete distributions. *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2566–2574. https://proceedings.neurips.cc/paper/2015/hash/2b3bf3eee2475e03885a110e9acaab61-Abstract.html 99

[Ding et al., 2017] Ding, B., Kulkarni, J., & Yekhanin, S. (2017). Collecting telemetry data privately. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 3571–3580. https://proceedings.neurips.cc/paper/2017/hash/253614bbac999b38b5b60cae531c4969-Abstract.html 22

[Dinur & Nissim, 2003] Dinur, I. & Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, 202–210. https://doi.org/10.1145/773153.773173 16

[Dong et al., 2020] Dong, J., Durfee, D., & Rogers, R. (2020). Optimal differential privacy composition for exponential mechanisms. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 2597–2606. http://proceedings.mlr.press/v119/dong20a.html 132

[Dong et al., 2019] Dong, J., Roth, A., & Su, W. J. (2019). Gaussian differential privacy. *CoRR*, abs/1905.02383. http://arxiv.org/abs/1905.02383 45, 47, 132

[Drechsler et al., 2022] Drechsler, J., Globus-Harris, I., Mcmillan, A., Sarathy, J., & Smith, A. (2022). Nonparametric Differentially Private Confidence Intervals for the Median. *Journal of Survey Statistics and Methodology*, 10(3), 804–829. https://doi.org/10.1093/jssam/smac021 153

[Dubost et al., 2020] Dubost, C., Oudre, L., Labourdette, C., Vayatis, N., & Vidal, P.-P. (2020). Quantitative assessment of consciousness during anesthesia without EEG data. *Journal of Clinical Monitoring and Computing*. https://doi.org/10.1007/s10877-020-00553-4 13

[Duchi et al., 2013] Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013). Local privacy and statistical minimax rates. *51st Annual Allerton Conference on Communication,*

Control, and Computing, Allerton 2013, Allerton Park & Retreat Center, Monticello, IL, USA, October 2-4, 2013, 1592. https://doi.org/10.1109/Allerton.2013.6736718 20, 61, 62, 63, 145

[Duchi et al., 2014] Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2014). *Local privacy, data processing inequalities, and statistical minimax rates.* https://arxiv.org/abs/1302.3203 145

[Duchi et al., 2016] Duchi, J. C., Wainwright, M. J., & Jordan, M. I. (2016). Minimax optimal procedures for locally private estimation. *CoRR*, abs/1604.02390. http://arxiv.org/abs/1604.02390 101, 105, 106

[Dudley, 1967] Dudley, R. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3), 290–330. https://doi.org/https://doi.org/10.1016/0022-1236(67)90017-1 22

[Durfee, 2023] Durfee, D. (2023). *Unbounded differentially private quantile and maximum estimation.* 132

[Dwork et al., 2015] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. L. (2015). Preserving statistical validity in adaptive data analysis. *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, 117–126. https://doi.org/10.1145/2746539.2746580 22

[Dwork et al., 2006a] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, volume 4004 of *Lecture Notes in Computer Science*, 486–503. https://doi.org/10.1007/11761679_29 16, 17, 25, 44, 49, 50, 129, 131

[Dwork et al., 2006b] Dwork, C., McSherry, F., Nissim, K., & Smith, A. D. (2006b). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, 265–284. https://doi.org/10.1007/11681878_14 16, 17, 23, 25, 44, 49, 50, 129, 131, 132, 169

[Dwork & Roth, 2014] Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4), 211–407. https://doi.org/10.1561/0400000042 45, 88, 129

[Dwork & Rothblum, 2016] Dwork, C. & Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*. 46, 74

[Dwork et al., 2010] Dwork, C., Rothblum, G. N., & Vadhan, S. P. (2010). Boosting and differential privacy. *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, 51–60. https://doi.org/10.1109/FOCS.2010.12 49, 74

[Erlingsson et al., 2014] Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). RAPPOR: randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, 1054–1067. https://doi.org/10.1145/2660267.2660348 22

[Frankle & Carbin, 2019] Frankle, J. & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=rJl-b3RcF7 34, 37

[Frankle et al., 2021] Frankle, J., Dziugaite, G. K., Roy, D., & Carbin, M. (2021). Pruning neural networks at initialization: Why are we missing the mark? *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. https://openreview.net/forum?id=Ig-VyQc-MLK 37

[Fredrikson et al., 2015] Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*, 1322–1333. https://doi.org/10.1145/2810103.2813677 16

[Ganesh et al., 2023] Ganesh, A., Haghifam, M., Steinke, T., & Thakurta, A. (2023). *Faster differentially private convex optimization via second-order methods*. https://arxiv.org/abs/2305.13209 56, 57

[Ganesh et al., 2022] Ganesh, A., Thakurta, A., & Upadhyay, J. (2022). Langevin diffu-

sion: An almost universal algorithm for private euclidean (convex) optimization. *CoRR*, abs/2204.01585. https://doi.org/10.48550/arXiv.2204.01585 56, 57

[Gidaris et al., 2021] Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., & Pérez, P. (2021). Obow: Online bag-of-visual-words generation for self-supervised learning. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 6830–6840. https://doi.org/10.1109/CVPR46437.2021.00676 173

[Gillenwater et al., 2021] Gillenwater, J., Joseph, M., & Kulesza, A. (2021). Differentially private quantiles. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 3713–3722. http://proceedings.mlr.press/v139/gillenwater21a.html 31, 55, 132, 135, 139, 166

[Giraud, 2021] Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003158745 62, 105

[Golatkar et al., 2022] Golatkar, A., Achille, A., Wang, Y., Roth, A., Kearns, M., & Soatto, S. (2022). Mixed differential privacy in computer vision. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 8366–8376. https://doi.org/10.1109/CVPR52688.2022.00819 173

[Gonon et al., 2023a] Gonon, A., Brisebarre, N., Riccietti, E., & Gribonval, R. (2023a). A path-norm toolkit for modern networks: consequences, promises and challenges. *CoRR*, abs/2310.01225. https://doi.org/10.48550/ARXIV.2310.01225 21

[Gonon et al., 2023b] Gonon, A., Zheng, L., Lalanne, C., Le, Q.-T., Lauga, G., & Pouliquen, C. (2023b). Can sparsity improve the privacy of neural networks? *GRETSI 2023 - XXIXème Colloque Francophone de Traitement du Signal et des Images*. https://hal.science/hal-04062317 16, 28, 33

[Gopi et al., 2022] Gopi, S., Lee, Y. T., & Liu, D. (2022). Private convex optimization via exponential mechanism. *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, 1948–1989. https://proceedings.mlr.press/v178/gopi22a.html 56

[Grill et al., 2020] Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H.,

Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent - A new approach to self-supervised learning. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html 173

[Györfi et al., 2002] Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer. https://doi.org/10.1007/b97848 100, 147

[Györfi & Kroll, 2022] Györfi, L. & Kroll, M. (2022). On rate optimal private regression under local differential privacy. *arXiv preprint arXiv:2206.00114*. 101

[Györfi & Kroll, 2023] Györfi, L. & Kroll, M. (2023). Multivariate density estimation from privatised data: universal consistency and minimax rates. *Journal of Nonparametric Statistics*, 0(0), 1–23. https://doi.org/10.1080/10485252.2022.2163634 101

[Haddouche & Guedj, 2022] Haddouche, M. & Guedj, B. (2022). Online pac-bayes learning. *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/a4d991d581accd2955a1e1928f4e6965-Abstract-Conference.html 22

[Haddouche & Guedj, 2023] Haddouche, M. & Guedj, B. (2023). Wasserstein pac-bayes learning: A bridge between generalisation and optimisation. *CoRR*, abs/2304.07048. https://doi.org/10.48550/arXiv.2304.07048 22

[Haddouche et al., 2020] Haddouche, M., Guedj, B., Rivasplata, O., & Shawe-Taylor, J. (2020). Upper and lower bounds on the performance of kernel PCA. *CoRR*, abs/2012.10369. https://arxiv.org/abs/2012.10369 22

[Haddouche et al., 2021] Haddouche, M., Guedj, B., Rivasplata, O., & Shawe-Taylor, J. (2021). Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), 1330. https://doi.org/10.3390/e23101330 22

[Haddouche et al., 2023] Haddouche, M., Guedj, B., & Wintenberger, O. (2023). Optimistic dynamic regret bounds. *CoRR*, abs/2301.07530. https://doi.org/10.48550/arXiv.2301.07530 22

[Hall et al., 2013] Hall, R., Rinaldo, A., & Wasserman, L. A. (2013). Differential privacy for functions and functional data. *J. Mach. Learn. Res.*, 14(1), 703–727. https://doi.org/10.5555/2567709.2502603 101

[Hartvigsen, 1992] Hartvigsen, D. (1992). Recognizing voronoi diagrams with linear programming. *INFORMS J. Comput.*, 4(4), 369–374. https://doi.org/10.1287/ijoc.4.4.369 18

[Haussler, 1995] Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *J. Comb. Theory, Ser. A*, 69(2), 217–232. 22

[He et al., 2021] He, F., Wang, B., & Tao, D. (2021). Tighter generalization bounds for iterative differentially private learning algorithms. *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, 802–812. https://proceedings.mlr.press/v161/he21a.html 22, 74

[He et al., 2020] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. B. (2020). Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975 173

[He et al., 2016] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. https://doi.org/10.1109/CVPR.2016.90 39

[Henzinger & Upadhyay, 2022] Henzinger, M. & Upadhyay, J. (2022). Constant matters: Fine-grained complexity of differentially private continual observation using completely bounded norms. *CoRR*, abs/2202.11205. https://arxiv.org/abs/2202.11205 153

[Homer et al., 2008] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. W. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8), e1000167. 16

[Hu et al., 2022] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022).

Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), 235:1–235:37. https://doi.org/10.1145/3523273 29, 34, 36, 37

[Huang et al., 2020] Huang, Y., Su, Y., Ravi, S., Song, Z., Arora, S., & Li, K. (2020). Privacy-preserving learning via deep net pruning. *CoRR*, abs/2003.01876. https://arxiv.org/abs/2003.01876 41

[Hui et al., 2021] Hui, B., Yang, Y., Yuan, H., Burlina, P., Gong, N. Z., & Cao, Y. (2021). Practical blind membership inference attack via differential comparisons. *Proceedings 2021 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2021.24293 29, 34

[IBM,] IBM (*). *Smartnoise core differential privacy library*. https://github.com/IBM/differential-privacy-library. 131, 135

[Iyengar et al., 2019] Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., & Wang, L. (2019). Towards practical differentially private convex optimization. *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, 299–316. https://doi.org/10.1109/SP.2019.00001 56

[Johnson et al., 2020] Johnson, N., Near, J. P., Hellerstein, J. M., & Song, D. (2020). Chorus: a programming framework for building scalable differential privacy mechanisms. *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 535–551. https://doi.org/10.1109/EuroSP48549.2020.00041 135

[Jung et al., 2021] Jung, S., Oudre, L., Truong, C., Dorveaux, E., Gorintin, L., Vayatis, N., & Ricard, D. (2021). Adaptive change-point detection for studying human locomotion. *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2021, Mexico, November 1-5, 2021*, 2020–2024. https://doi.org/10.1109/EMBC46164.2021.9629775 13

[Kairouz et al., 2021] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawit, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Theertha Suresh, A., Tramèr, F., Vepakomma, P.,

Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., & Zhao, S. (2021). *Advances and Open Problems in Federated Learning* 20

[Kairouz et al., 2015] Kairouz, P., Oh, S., & Viswanath, P. (2015). The composition theorem for differential privacy. *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 1376–1385. http://proceedings.mlr.press/v37/kairouz15.html 45, 46, 48, 132

[Kallenberg, 1993] Kallenberg, O. (1993). Lectures on the coupling method (torgny lindvall). *SIAM Review*, 35(3), 525–527. https://doi.org/10.1137/1035121 77

[Kamath et al., 2019] Kamath, G., Li, J., Singhal, V., & Ullman, J. R. (2019). Privately learning high-dimensional distributions. *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, 1853–1902. http://proceedings.mlr.press/v99/kamath19a.html 99

[Kamath et al., 2022] Kamath, G., Liu, X., & Zhang, H. (2022). Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 10633–10660. https://proceedings.mlr.press/v162/kamath22a.html 30, 67, 99, 101, 102, 106, 145

[Kamath et al., 2023a] Kamath, G., Mouzakis, A., Regehr, M., Singhal, V., Steinke, T., & Ullman, J. (2023a). *A bias-variance-privacy trilemma for statistical estimation.* 99

[Kamath et al., 2023b] Kamath, G., Mouzakis, A., & Singhal, V. (2023b). *New lower bounds for private estimation and a generalized fingerprinting lemma.* 99

[Kamath et al., 2020] Kamath, G., Singhal, V., & Ullman, J. R. (2020). Private mean estimation of heavy-tailed distributions. *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, 2204–2235. http://proceedings.mlr.press/v125/kamath20a.html 99

[Kaplan et al., 2020] Kaplan, H., Ligett, K., Mansour, Y., Naor, M., & Stemmer, U. (2020). Privately learning thresholds: Closing the exponential gap. *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Aus-*

tria], volume 125 of *Proceedings of Machine Learning Research*, 2263–2285. http://proceedings.mlr.press/v125/kaplan20a.html 153

[Kaplan et al., 2022] Kaplan, H., Schnapp, S., & Stemmer, U. (2022). Differentially private approximate quantiles. *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 10751–10761. https://proceedings.mlr.press/v162/kaplan22a.html 31, 55, 132, 134, 136

[Karwa & Vadhan, 2018] Karwa, V. & Vadhan, S. P. (2018). Finite sample differentially private confidence intervals. *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, 44:1–44:9. https://doi.org/10.4230/LIPIcs.ITCS.2018.44 99

[Kearns & Roth, 2019] Kearns, M. & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, Inc. 13

[Koltchinskii, 2001] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory*, 47(5), 1902–1914. https://doi.org/10.1109/18.930926 22

[Koltchinskii & Panchenko, 2000] Koltchinskii, V. & Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, 443–457. 22

[Konečný et al., 2016] Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). *Federated optimization: Distributed machine learning for on-device intelligence*. 19

[Kroll, 2021] Kroll, M. (2021). On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics*, 15(1), 1783 – 1813. https://doi.org/10.1214/21-EJS1830 101, 145

[Kurakin et al., 2022] Kurakin, A., Chien, S., Song, S., Geambasu, R., Terzis, A., & Thakurta, A. (2022). Toward training at imagenet scale with differential privacy. *CoRR*, abs/2201.12328. https://arxiv.org/abs/2201.12328 172

[la Tour et al., 2018] la Tour, T. D., Moreau, T., Jas, M., & Gramfort, A. (2018).

Multivariate convolutional sparse coding for electromagnetic brain signals. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 3296–3306. https://proceedings.neurips.cc/paper/2018/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html 13

[Labs, 2022] Labs, T. (2022). *Tumult Analytics*. https://tmlt.dev 135

[Lalanne et al., 2023a] Lalanne, C., Garivier, A., & Gribonval, R. (2023a). About the cost of central privacy in density estimation. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uq29MIWvIV 100

[Lalanne et al., 2023b] Lalanne, C., Garivier, A., & Gribonval, R. (2023b). On the Statistical Complexity of Estimation and Testing under Privacy Constraints. *Transactions on Machine Learning Research Journal*. https://hal.science/hal-03794374 28, 30, 41, 59, 84, 85, 101, 102, 105, 106, 111, 145

[Lalanne et al., 2023c] Lalanne, C., Garivier, A., & Gribonval, R. (2023c). Private Statistical Estimation of Many Quantiles. *ICML 2023 - 40th International Conference on Machine Learning*. https://hal.science/hal-03986170 28, 130

[Lalanne et al., 2023d] Lalanne, C., Gastaud, C., Grislain, N., Garivier, A., & Gribonval, R. (2023d). Private Quantiles Estimation in the Presence of Atoms. *Information and Inference*. https://doi.org/10.1093/imaiai/iaad030 28, 130, 139

[Lalanne et al., 2020] Lalanne, C., Rateaux, M., Oudre, L., Robert, M. P., & Moreau, T. (2020). Extraction of Nystagmus Patterns from Eye-Tracker Data with Convolutional Sparse Coding. *EMBC 2020 - 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society in conjunction with the 43rd Annual Conference of the Canadian Medical and Biological Engineering Society*, volume 42 of *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 928–931. https://hal.science/hal-03022547 13

[Lam-Weil et al., 2022] Lam-Weil, J., Laurent, B., & Loubes, J.-M. (2022). Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli*, 28(1), 579–600. 101

[Laurent & Massart, 2000] Laurent, B. & Massart, P. (2000). Adaptive estimation of a

quadratic functional by model selection. *The Annals of Statistics*, 28(5), 1302–1338. http://www.jstor.org/stable/2674095 54

[Le et al., 2022] Le, Q.-T., Zheng, L., Riccietti, E., & Gribonval, R. (2022). Fast learning of fast transforms, with guarantees. *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/ICASSP43922.2022.9747791. This paper is associated to code for reproducible research available at https://hal.inria.fr/hal-03552956 37

[Lin et al., 2021] Lin, R., Ran, J., Chiu, K. H., Chesi, G., & Wong, N. (2021). Deformable butterfly: A highly structured and sparse linear transform. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 16145–16157. https://proceedings.neurips.cc/paper/2021/hash/86b122d4358357d834a87ce618a55de0-Abstract.html 34, 38

[Long et al., 2018] Long, Y., Bindschaedler, V., Wang, L., Bu, D., Wang, X., Tang, H., Gunter, C. A., & Chen, K. (2018). *Understanding membership inferences on well-generalized learning models*. 29, 34

[Loukides et al., 2010] Loukides, G., Denny, J. C., & Malin, B. A. (2010). The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Medical Informatics Assoc.*, 17(3), 322–327. https://doi.org/10.1136/jamia.2009.002725 16

[Malach et al., 2020] Malach, E., Yehudai, G., Shalev-Schwartz, S., & Shamir, O. (2020). Proving the lottery ticket hypothesis: Pruning is all you need. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6682–6691. https://proceedings.mlr.press/v119/malach20a.html 37

[Mangold et al., 2022] Mangold, P., Bellet, A., Salmon, J., & Tommasi, M. (2022). Differentially private coordinate descent for composite empirical risk minimization. *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 14948–14978. https://proceedings.mlr.press/v162/mangold22a.html 56

[Marfoq et al., 2021] Marfoq, O., Neglia, G., Bellet, A., Kameni, L., & Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, volume 34,

15434–15447. https://proceedings.neurips.cc/paper_files/paper/2021/file/82599a4ec94aca066873c99b4c741ed8-Paper.pdf 19

[McAllester, 1999] McAllester, D. A. (1999). Some pac-bayesian theorems. *Mach. Learn.*, 37(3), 355–363. https://doi.org/10.1023/A:1007618624809 22

[McMahan et al., 2017] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html 19

[McMahan et al., 2018a] McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018a). Learning differentially private recurrent language models. *International Conference on Learning Representations*. https://openreview.net/forum?id=BJ0hF1Z0b 19

[McMahan et al., 2018b] McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018b). Learning differentially private recurrent language models. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. https://openreview.net/forum?id=BJ0hF1Z0b 56

[McSherry, 2010] McSherry, F. (2010). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Commun. ACM*, 53(9), 89–97. https://doi.org/10.1145/1810891.1810916 56

[McSherry & Talwar, 2007] McSherry, F. & Talwar, K. (2007). Mechanism design via differential privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, 94–103. https://doi.org/10.1109/FOCS.2007.41 54, 131, 162

[Mironov, 2017] Mironov, I. (2017). Rényi differential privacy. *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, 263–275. https://doi.org/10.1109/CSF.2017.11 46, 47, 49, 95

[Mohri et al., 2012] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press. http://mitpress.mit.edu/books/foundations-machine-learning-0 21

[Musavi et al., 1994] Musavi, M. T., Chan, K. H., Hummels, D. M., & Kalantri, K. (1994). On the generalization ability of neural network classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(6), 659–663. https://doi.org/10.1109/34.295911 21

[Narayanan & Shmatikov, 2006] Narayanan, A. & Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105. http://arxiv.org/abs/cs/0610105 14, 15, 16

[Narayanan & Shmatikov, 2008] Narayanan, A. & Shmatikov, V. (2008). Robust deanonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, 111–125. https://doi.org/10.1109/SP.2008.33 16

[Nissim et al., 2007] Nissim, K., Raskhodnikova, S., & Smith, A. D. (2007). Smooth sensitivity and sampling in private data analysis. *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, 75–84. https://doi.org/10.1145/1250790.1250803 54, 131

[Nissim & Stemmer, 2015] Nissim, K. & Stemmer, U. (2015). *On the generalization properties of differential privacy.* https://arxiv.org/abs/1504.05800 22

[Oneto et al., 2017] Oneto, L., Ridella, S., & Anguita, D. (2017). Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognit. Lett.*, 89, 31–38. https://doi.org/10.1016/j.patrec.2017.02.006 22

[OpenMinded, 2022] OpenMinded (2022). *Pydp.* https://github.com/OpenMined/PyDP 135

[Orseau et al., 2020] Orseau, L., Hutter, M., & Rivasplata, O. (2020). Logarithmic pruning is all you need. *Advances in Neural Information Processing Systems*, volume 33, 2925–2934. https://proceedings.neurips.cc/paper_files/paper/2020/file/1e9491470749d5b0e361ce4f0b24d037-Paper.pdf 37

[Papernot et al., 2021] Papernot, N., Thakurta, A., Song, S., Chien, S., & Erlingsson, Ú. (2021). Tempered sigmoid activations for deep learning with differential privacy. *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event,*

February 2-9, 2021, 9312–9321. https://ojs.aaai.org/index.php/AAAI/article/view/17123 172

[Paul et al., 2022] Paul, M., Chen, F., Larsen, B. W., Frankle, J., Ganguli, S., & Dziugaite, G. K. (2022). *Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask?* 37

[Peyré & Cuturi, 2019] Peyré, G. & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6), 355–607. https://doi.org/10.1561/2200000073 63, 67, 76

[Rezaei & Liu, 2021] Rezaei, S. & Liu, X. (2021). On the difficulty of membership inference attacks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7888–7896. https://doi.org/10.1109/CVPR46437.2021.00780 29, 34

[Rigaki & Garcia, 2020] Rigaki, M. & Garcia, S. (2020). A survey of privacy attacks in machine learning. *CoRR*, abs/2007.07646. https://arxiv.org/abs/2007.07646 29, 34

[Rigollet & Hütter, 2015] Rigollet, P. & Hütter, J.-C. (2015). High dimensional statistics. *MIT lecture notes for course 18S997*. https://math.mit.edu/~rigollet/PDFs/RigNotes17.pdf 26, 62, 91, 105, 109

[Rogers & Wagner, 1978] Rogers, W. H. & Wagner, T. J. (1978). A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules. *The Annals of Statistics*, 6(3), 506 – 514. https://doi.org/10.1214/aos/1176344196 22

[Ryffel et al., 2022] Ryffel, T., Bach, F. R., & Pointcheval, D. (2022). Differential privacy guarantees for stochastic gradient langevin dynamics. *CoRR*, abs/2201.11980. https://arxiv.org/abs/2201.11980 57, 94, 96, 97

[Sablayrolles et al., 2019] Sablayrolles, A., Douze, M., Ollivier, Y., Schmid, C., & Jégou, H. (2019). White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. *ICML 2019 - 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5558–5567. https://inria.hal.science/hal-02278902 29, 34

[Salem et al., 2018] Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes,

M. (2018). *Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models*. 29, 34

[Sander et al., 2022] Sander, T., Stock, P., & Sablayrolles, A. (2022). TAN without a burn: Scaling laws of DP-SGD. *CoRR*, abs/2210.03403. https://doi.org/10.48550/arXiv.2210.03403 41, 57

[Santambrogio, 2016] Santambrogio, F. (2016). *Optimal Transport for Applied Mathematicians*. Birkhäuser Cham/Springer. https://doi.org/10.1007/978-3-319-20828-2 63, 67, 76

[Sbidian et al., 2020] Sbidian, E., Josse, J., Lemaitre, G., Meyer, I., Bernaux, M., Gramfort, A., Lapidus, N., Paris, N., Neuraz, A., Lerner, I., et al. (2020). Hydroxychloroquine with or without azithromycin and in-hospital mortality or discharge in patients hospitalized for covid-19 infection: a cohort study of 4,642 in-patients in france. *MedRxiv*, 2020–06. 13

[Schellekens et al., 2019a] Schellekens, V., Chatalic, A., Houssiau, F., de Montjoye, Y.-A., Jacques, L., & Gribonval, R. (2019a). Compressive k-Means with Differential Privacy. *SPARS 2019 - Signal Processing with Adaptive Sparse Structured Representations*, 1–2. https://inria.hal.science/hal-02154820 172

[Schellekens et al., 2019b] Schellekens, V., Chatalic, A., Houssiau, F., de Montjoye, Y.-A., Jacques, L., & Gribonval, R. (2019b). Differentially Private Compressive k-Means. *ICASSP 2019 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, 7933–7937. https://doi.org/10.1109/ICASSP.2019.8682829 172

[Schluttenhofer & Johannes, 2022] Schluttenhofer, S. & Johannes, J. (2022). *Adaptive pointwise density estimation under local differential privacy*. 101

[Schoenberg et al., 2003] Schoenberg, F. P., Ferguson, T., & Li, C. (2003). Inverting Dirichlet Tessellations. *The Computer Journal*, 46(1), 76–83. https://doi.org/10.1093/comjnl/46.1.76 18

[Sebia et al., 2023] Sebia, H., Guyet, T., & Audureau, E. (2023). Une extension de la décomposition tensorielle au phénotypage temporel. *EGC 2023-23ème Conférence Francophone sur l'Extraction et Gestion des Connaissances*. 13

[Shalev-Shwartz & Ben-David, 2014] Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press. http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms 16

[Shen et al., 2021] Shen, Y., Wang, Z., Sun, R., & Shen, X. (2021). Towards understanding the impact of model size on differential private classification. *CoRR*, abs/2111.13895. https://arxiv.org/abs/2111.13895 172

[Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, 3–18. https://doi.org/10.1109/SP.2017.41 29, 34

[Simmons, 1979] Simmons, G. J. (1979). Symmetric and asymmetric encryption. *ACM Comput. Surv.*, 11(4), 305–330. https://doi.org/10.1145/356789.356793 19

[Singhal, 2023] Singhal, V. (2023). *A polynomial time, pure differentially private estimator for binary product distributions*. 99

[Smith, 2011] Smith, A. D. (2011). Privacy-preserving statistical estimation with optimal convergence rates. *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, 813–822. https://doi.org/10.1145/1993636.1993743 99, 131, 134

[Smith et al., 2017] Smith, A. D., Thakurta, A., & Upadhyay, J. (2017). Is interaction necessary for distributed private learning? *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, 58–77. https://doi.org/10.1109/SP.2017.35 56

[Song et al.,] Song, L., Shokri, R., & Mittal, P. Privacy risks of securing machine learning models against adversarial examples. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019*. 41

[Song et al., 2013] Song, S., Chaudhuri, K., & Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. *IEEE Global Conference on Signal and*

Information Processing, GlobalSIP 2013, Austin, TX, USA, December 3-5, 2013, 245–248. https://doi.org/10.1109/GlobalSIP.2013.6736861 56

[Song et al., 2021] Song, S., Steinke, T., Thakkar, O., & Thakurta, A. (2021). Evading the curse of dimensionality in unconstrained private glms. *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, 2638–2646. http://proceedings.mlr.press/v130/song21a.html 56

[Song et al., 2020] Song, S., Thakkar, O., & Thakurta, A. (2020). Characterizing private clipped gradient descent on convex generalized linear problems. *CoRR*, abs/2006.06783. https://arxiv.org/abs/2006.06783 56

[Soumik,] Soumik (*). *Goodreads-books dataset.* https://www.kaggle.com/jealousleopard/goodreadsbooks. 166

[Steinberger, 2023] Steinberger, L. (2023). *Efficiency in local differential privacy.* https://arxiv.org/abs/2301.10600 145

[Sweeney, 2000] Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000), 1–34. 16

[Sweeney, 2002] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5), 557–570. https://doi.org/10.1142/S0218488502001648 16, 45

[Tan et al., 2023] Tan, J., LeJeune, D., Mason, B., Javadi, H., & Baraniuk, R. G. (2023). A blessing of dimensionality in membership inference through regularization. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 10968–10993. https://proceedings.mlr.press/v206/tan23b.html 40

[Thakurta et al., 2017] Thakurta, A. G., Vyrros, A. H., Vaishampayan, U. S., Kapoor, G., Freudiger, J., Sridhar, V. R., & Davidson, D. (2017). Learning new words. *Granted US Patents*, 9594741. 22

[Tramèr & Boneh, 2021] Tramèr, F. & Boneh, D. (2021). Differentially private learning

needs better features (or much more data). *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. https://openreview.net/forum?id=YTWGvpFOQD- 172

[Truex et al., 2021] Truex, S., Liu, L., Gursoy, M. E., Yu, L., & Wei, W. (2021). Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 14(6), 2073–2089. https://doi.org/10.1109/TSC.2019.2897554 29, 34

[Truong & Oudre, 2022] Truong, C. & Oudre, L. (2022). Supervised change-point detection with dimension reduction, applied to physiological signals. *NeurIPS 2022 Workshop on Learning from Time Series for Health*. https://hal.science/hal-03883779 13

[Tsybakov, 2009] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer. https://doi.org/10.1007/b13794 75, 86, 100, 105, 109, 115, 119, 120, 125, 147

[Van der Vaart, 1998] Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256 31, 94, 97, 112, 131

[van Erven & Harremoës, 2014] van Erven, T. & Harremoës, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 60(7), 3797–3820. https://doi.org/10.1109/TIT.2014.2320500 46, 74, 75, 76, 86

[Vanhaesebrouck et al., 2017] Vanhaesebrouck, P., Bellet, A., & Tommasi, M. (2017). Decentralized Collaborative Learning of Personalized Models over Networks. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 509–517. https://proceedings.mlr.press/v54/vanhaesebrouck17a.html 19

[Vapni, 1995] Vapni, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer. https://doi.org/10.1007/978-1-4757-2440-0 21

[Vapnik, 2006] Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data, Second Editiontion*. Springer. https://doi.org/10.1007/0-387-34239-7 21

[Villani et al., 2009] Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer. https://doi.org/10.1007/978-3-540-71050-9 63

[Voyez et al., 2022a] Voyez, A., Allard, T., Avoine, G., Cauchois, P., Fromont, E., & Simonin, M. (2022a). Membership inference attacks on aggregated time series with linear programming. *19th International Conference on Security and Cryptography*. https://hal.archives-ouvertes.fr/hal-03726234/ 29, 34

[Voyez et al., 2022b] Voyez, A., Allard, T., Avoine, G., Cauchois, P., Fromont, E., & Simonin, M. (2022b). *Unique in the smart grid -the privacy cost of fine-grained electrical consumption data*. https://doi.org/10.48550/ARXIV.2211.07205 16

[Wagner & Eckhoff, 2018] Wagner, I. & Eckhoff, D. (2018). Technical privacy metrics: A systematic survey. *ACM Comput. Surv.*, 51(3), 57:1–57:38. https://doi.org/10.1145/3168389 16

[Wang et al., 2020] Wang, Y., Balle, B., & Kasiviswanathan, S. P. (2020). Subsampled rényi differential privacy and analytical moments accountant. *J. Priv. Confidentiality*, 10(2). https://doi.org/10.29012/jpc.723 50

[Wang et al., 2021] Wang, Y., Wang, C., Wang, Z., Zhou, S., Liu, H., Bi, J., Ding, C., & Rajasekaran, S. (2021). Against membership inference attack: Pruning is all you need. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 3141–3147. https://doi.org/10.24963/ijcai.2021/432 41

[Wasserman & Zhou, 2010] Wasserman, L. A. & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489), 375–389. https://doi.org/10.1198/jasa.2009.tm08651 101, 102, 111, 119, 122, 147, 149

[Wilson et al., 2019] Wilson, R. J., Zhang, C. Y., Lam, W., Desfontaines, D., Simmons-Marengo, D., & Gipson, B. (2019). *Differentially private sql with bounded user contribution*. 135

[Wu et al., 2017] Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., & Naughton, J. F. (2017). Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. *Proceedings of the 2017 ACM International Conference on Management of*

Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017, 1307–1322. https://doi.org/10.1145/3035918.3064047 56

[Xiao et al., 2010] Xiao, Y., Xiong, L., & Yuan, C. (2010). Differentially private data release through multidimensional partitioning. *Secure Data Management, 7th VLDB Workshop, SDM 2010, Singapore, September 17, 2010. Proceedings*, volume 6358 of *Lecture Notes in Computer Science*, 150–168. https://doi.org/10.1007/978-3-642-15546-8_11 160

[Xu et al., 2012] Xu, H., Caramanis, C., & Mannor, S. (2012). Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1), 187–193. https://doi.org/10.1109/TPAMI.2011.177 22

[Yang et al., 2020] Yang, M., Lyu, L., Zhao, J., Zhu, T., & Lam, K. (2020). Local differential privacy and its applications: A comprehensive survey. *CoRR*, abs/2008.03686. https://arxiv.org/abs/2008.03686 20

[Yeganova et al., 2001] Yeganova, L., Falk, J., & Dandurova, Y. (2001). Robust separation of multiple sets. *Nonlinear Analysis-theory Methods & Applications - NONLINEAR ANAL-THEOR METH APP*, 47, 1845–1856. https://doi.org/10.1016/S0362-546X(01)00315-7 18

[Yeom et al., 2018] Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. https://doi.org/10.1109/CSF.2018.00027 29, 34

[Yu et al., 2021] Yu, D., Zhang, H., Chen, W., & Liu, T. (2021). Do not let privacy overbill utility: Gradient embedding perturbation for private learning. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. https://openreview.net/forum?id=7aogOj_VYO0 172

[Yuan & Zhang, 2022] Yuan, X. & Zhang, L. (2022). Membership inference attacks and defenses in neural network pruning. *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, 4561–4578. https://www.usenix.org/conference/usenixsecurity22/presentation/yuan-xiaoyong 40

[Zhang et al., 2021a] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021a).

Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 107–115. https://doi.org/10.1145/3446776 29, 33

[Zhang et al., 2021b] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021b). A survey on federated learning. *Knowledge-Based Systems*, 216, 106775. https://doi.org/https://doi.org/10.1016/j.knosys.2021.106775 20

[Zhang et al., 2021c] Zhang, H., Mironov, I., & Hejazinia, M. (2021c). Wide network learning with differential privacy. *CoRR*, abs/2103.01294. https://arxiv.org/abs/2103.01294 172

[Zheng et al., 2023] Zheng, L., Puy, G., Riccietti, E., Perez, P., & Gribonval, R. (2023). Self-supervised learning with rotation-invariant kernels. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=8uu6JStuYm 173

[Zheng et al., 2022] Zheng, L., Riccietti, E., & Gribonval, R. (2022). Efficient Identification of Butterfly Sparse Matrix Factorizations. *SIAM Journal on Mathematics of Data Science*. https://inria.hal.science/hal-03362626 37

[Zhou et al., 2021] Zhou, Y., Wu, S., & Banerjee, A. (2021). Bypassing the ambient dimension: Private SGD with gradient subspace identification. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. https://openreview.net/forum?id=7dpmlkBuJFC 172