

Differentially Private Sampling from Rashomon Sets, and the Universality of Langevin Diffusion for Convex Optimization

Arun Ganesh* Abhradeep Thakurta[†] Jalaj Upadhyay[‡]

August 30, 2023

Abstract

In this paper we provide an algorithmic framework based on Langevin diffusion (LD) and its corresponding discretizations that allow us to simultaneously obtain: i) An algorithm for sampling from the exponential mechanism [MT07], whose privacy analysis does not depend on convexity and which can be stopped at anytime without compromising privacy, and ii) tight uniform stability guarantees for the exponential mechanism. As a direct consequence, we obtain optimal excess empirical and population risk guarantees for (strongly) convex losses under both pure and approximate differential privacy (DP). The framework allows us to design a DP uniform sampler from the Rashomon set. Rashomon sets are widely used in interpretable and robust machine learning, understanding variable importance, and characterizing fairness.

Note: For ease of presentation, some results appear in the previous version of this paper on arXiv (v3) that do not appear in this version, nor are subsumed by results in this version. Please see Section 1.4 for more details.

1 Introduction

Differentially private empirical risk minimization (DP-ERM) [CMS11, CYS21, INS⁺19, KST12, BST14, STU17, SCS13, STT20, WLK⁺17] and *differentially private stochastic optimization* (DP-SCO) [ALD21, BFTT19, BFGT20, FKT20, KLL21, GLL22] are two of the most widely studied problems in the differential privacy (DP) literature. The optimal algorithms for either of these settings are either based on *differentially private stochastic gradient descent* (DP-SGD) [ACG⁺16, BST14, SCS13]), or sampling from an appropriate Gibbs distribution (a.k.a. the exponential mechanism [MT07]). In this paper we revisit the sampling perspective of DP optimization and study its implications.

At a high level, the Gibbs distribution sampling problem is to generate a sample θ from the distribution with density proportional to $\exp(-\beta\mathcal{L}(\theta; D))$. Here β is known as the *inverse temperature*, and $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; d_i)$ (with $\ell : \mathbb{R}^p \times \tau \rightarrow \mathbb{R}$) is the *empirical loss function*. Our main contribution is an algorithmic framework based on Langevin Diffusion (LD) (see Figure 1) to privately (approximately) sample from a Gibbs distribution, with the following implications:

1. Our framework recovers all existing (and tight) bounds for DP-ERM [BST14], and in some cases improves on the best known bounds. The framework provides tight $O(1/n)$ -uniform stability guarantees for the Gibbs distribution on strongly convex losses and $O(1/\sqrt{n})$ -uniform stability on mildly regularized convex losses. This is much tighter than the generic $O(\epsilon)$ -uniform stability provided by

*Google Research. arunganesh@google.com. Part of this work was done at UC Berkeley while being supported in part by NSF CCF-1816861.

[†]Google DeepMind. athakurta@google.com.

[‡]Rutgers University. jalaj.upadhyay@rutgers.edu. This work was supported by the Decanal Research grant from Rutgers University.

the DP guarantee [BST14]¹. This allows us to obtain optimal DP-SCO bounds under both ϵ -DP and (ϵ, δ) -DP², and improves on [ALD21] for the ϵ -DP case. In this sense, the LD framework is universal for DP optimization.

2. The privacy guarantee of our Langevin Diffusion (LD) based algorithm does not depend on convexity of the loss function, $\mathcal{L}(\theta; D)$. Therefore, it can be used in non-convex settings without compromising privacy.
3. In the (ϵ, δ) -DP setting, we can release the complete trajectory of the LD until it reaches the stationary Gibbs distribution. As a direct consequence, our algorithm is “anytime” DP, i.e., the privacy is not compromised even if we stop before the chain converges to the stationary distribution. This is a very useful property in practice; in fact, subsequent works [SGM⁺22, RTT⁺23] have shown that the path of LD can be used to quantify predictive uncertainty.

Sampling from Rashomon sets Our framework also allows us to go beyond what can be achieved by known algorithms for private learning. In particular, it allows us to uniformly and privately sample from the *Rashomon set* [Bre01] (Definition 1.4), which has been extensively studied in interpretable and robust machine learning [RCC⁺22], understanding spectrum of variable importance [FRD19], decision making [TR13, TR14a, TR14b], measuring underspecification [MAD19], and characterizing fairness [CRC21]. At a high level, Rashomon set is the set of equally-performing models in terms of training/testing loss.

1.1 Related Work

We start by giving a brief exposition of some of the related work and a literature survey of works in Rashomon set, differentially private learning, and dynamical systems.

Rashomon set and predictive multiplicity Rashomon set has been extensively studied in applied machine learning since its conception [CRK21, FRD19, MB10, LLRB16, NR17, SRP22, SST10, TR14b] (see the survey [RCC⁺22] for more references), culminating in a recent work of [SRP22]. For example, [FRD19] leverages the Rashomon set in order to understand the spectrum of variable importance and other statistics across the set of good models, and [TR13, TR14a, TR14b] uses the Rashomon set to assist with decision making. However, it is computationally inefficient to find the simplest model in the Rashomon set, so a natural question is when should we even search for a simpler model. In a recent work, [SRP22] showed that, if there is a large Rashomon set of almost-equally-accurate models, a simple model may also be contained in it and that model is guaranteed to generalize well. For this, they defined *Rashomon ratio*, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space. They then used the insights gained from Rashomon ratios to infer whether a simpler model exists or not.

Rashomon ratios are defined in terms of volumes of the hypothesis (or model parameter) space. In a recent work, [HC22] proposed a different metric known as *Rashomon capacity*. It aims to measure the multiplicity of classifier outputs for individual samples, i.e., it measures the spread of the scores with divergence measures for probability distributions (also known as *predictive multiplicity*). In particular, this helps to distinguish Rashomon sets where different predictions are a result of highly different predicted probabilities vs sets where the predictions are different but they come from similar soft outputs.

Some applications of Rashomon sets Rashomon set has many applications, such as in interpretable and robust machine learning [FRD19, RCC⁺22], understanding spectrum of variable importance [FRD19], decision making [TR13, TR14a, TR14b], measuring underspecification [MAD19], and characterizing fairness [CRC21] to name a few. In fact, if we assume that the loss function is smoothness of a loss function, then one can obtain a tighter excess risk bound through local Rademacher complexity [BBM05]. A line of

¹The uniform stability arguments in [BST14] gives the $O(\epsilon + \frac{p}{\epsilon n})$ SCO bound for convex losses under pure-DP. This is at best $O(\sqrt{p/n})$. In contrast, we obtain the optimal $O(1/\sqrt{n} + p/\epsilon n)$.

²Through out the paper we will assume $\epsilon = O(\log(1/\delta))$.

work has also related Rashomon sets with p -hacking and robustness of estimation. The central argument there is that the Rashomon set is a set on which one might conduct a sensitivity analysis for choices made by an analyst. We refer the interested readers to the survey by [RCC⁺22].

Sampling and differential privacy Several lines of work have designed Markov chains that generate samples from distributions that are close to a given log-concave distribution. For example, there are works that give sampling algorithms with bounds on the distance to the target density in terms of Wasserstein distance [DRD20, DM17], KL-divergence [DMM19], and Renyi divergence [VW19]. For privacy, we need a bound in terms of ℓ_∞ distance. For this, the first work that performs sampling with bounded ℓ_∞ distance is by [HT10]. This was extended to bounded Lipschitz log-concave distribution by [BST14]. Since then, several works have shown efficient algorithms for sampling from log-concave distribution [MV21, GT20]. There has been some work that finds polynomial time sampling algorithms for special loss functions. For example, [AD20a, AD20b] showed efficient algorithms to sample from the exponential mechanism when the score function has a specific structure, which they call *path-length function*. The motivation of [AD20a, AD20b] was to study instance-optimality of certain wide class of statistical problems. A variant of this function (which is quasi-convex) has been recently used by [HKMN23] for robust high-dimensional parameter estimation problems, including mean and covariance estimation, when the data is picked from multivariate Gaussian distribution. Using the insights from differential privacy, a recent line of work [AT22b, AT22c] have improved (and characterized) the mixing time of the discretization of Langevin diffusion.

There has been some recent work to study the asymptotic bias introduced by the discretization of Langevin diffusion. In a recent work, [AT22a] showed that the stationary distribution of discretization of the Langevin diffusion is sub-Gaussian when the potential function is strongly convex, and is sub-exponential when the potential function is convex.

Differential privacy and dynamical systems The connection between dynamical systems and differential privacy is also not new. [CYS21] and [RBP22] study discretization of the LD algorithm as DP-(Stochastic) Gradient Langevin Dynamics (DP-SGLD). They show that under smoothness and strong convexity on the loss function $\mathcal{L}(\theta; D)$, the privacy cost of DP-SGLD converges to a stationary finite value, even when the number of time steps goes to ∞ . [WCX19] used the result by [RRT17] to prove a sub-optimal excess empirical risk of $\tilde{O}\left(\frac{p \log(1/\delta)}{\varepsilon^2 \log(n)}\right)$ for non-convex loss functions. In a concurrent, and complementary work on convex losses, [GLL22] study private optimization and show the universality of exponential mechanisms for both stochastic convex optimization and empirical risk minimization. Their analysis takes the sampling perspective when the diffusion process has completed.

It is probably important to mention that objective perturbation [CMS11, KST12] can be potentially thought of as a (near) universal algorithm for the problem classes considered in this paper, albeit the following two caveats: i) The instantiation of the algorithm for ε -DP and (ε, δ) -DP require two different noise models to be drawn from, namely, Gamma distribution, and Normal distribution, and ii) It requires the loss functions $\ell(\theta; \cdot)$ to be twice-continuously differentiable, and $\nabla_\theta^2 \ell(\theta; \cdot)$ to have a near constant rank. As mentioned in the remainder of our paper, Langevin diffusion does not require any such assumptions.³

Recently, [MV22] used continuous-time viewpoint to study the error incurred by adding a symmetric Gaussian matrix to input covariance matrix. In particular, they viewed the perturbed matrix as a continuous-time symmetric matrix diffusion, where each entry of the perturbed matrix is the value reached by a Brownian motion after the time equals to the scaling of variance required for privacy. In particular, the corresponding Brownian motion is well studied in statistical quantum physics and is known as *Dyson brownian motion*.

There is a contemporary and most closely related work of [GLL22] to ours. We defer the comparison to Section 2.5⁴.

³In particular, we can always ensure twice differentiability by convolving the loss function with the bump kernel [KST12], and then make the smoothness parameter finite but arbitrarily large which does not affect the Lipschitzness.

⁴The claim of contemporarity is also supported by the authors of [GLL22].

Langevin diffusion (LD). Let W_t be a p -dimensional Brownian motion and $\beta > 0$ be the *inverse temperature*. Then LD is the following stochastic differential equation:

$$d\theta_t = -\beta \nabla \mathcal{L}(\theta_t; D) \cdot dt + \sqrt{2} \cdot dW_t. \quad (1)$$

“Projected” Langevin diffusion. Sometimes, we only have the Lipschitz guarantee within a constrained set. We can also consider the following “projected” version of LD:

$$d\theta_t = -\beta \nabla \mathcal{L}(\theta_t; D) \cdot dt + \sqrt{2} \cdot dW_t - \nu_t \mu(dt), \forall t \geq 0 : \theta_t \in \mathcal{C}. \quad (2)$$

where μ is a measure supported on $\{t : \theta_t \in \partial \mathcal{C}\}$ and ν_t is an outer unit normal vector at θ_t for all such θ_t . See [BEL18, Section 2.1, 3.1] for a discussion of (2).

Figure 1: (Projected) Langevin diffusion

1.2 Our Contributions

Our main contribution is to design a Langevin diffusion (LD) based DP sampler for the following Gibbs distribution:

$$\exp(-\beta \max\{\mathcal{L}(\theta; D), \psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D)\}).$$

As we will see in Section 1.3, by setting $\psi = 0$ we obtain optimal DP-ERM/DP-SCO algorithms, and for $\psi > 0$ we obtain a DP Rashomon set sampler. In this section, we first state the main result followed by the uniform stability result for LD. We end with a discussion of discretization of our LD, that outputs a sample within δ total variation distance of the stationary distribution of LD at the same privacy/utility trade-off.

We start with the LD algorithm, described in Figure 1, which forms the building block for all the algorithms considered in this paper. Intuitively, one should think of (1) as the limit of noisy gradient descent and (2) as the limit of projected noisy gradient descent, both as step size $\eta \rightarrow 0$. Here and throughout this paper, $O_\delta(\cdot)$ and $\tilde{O}_\delta(\cdot)$ hides polylog factors in $1/\delta$.

Informal Theorem 1.1 (Corresponds to Theorems 2.1 and 2.3). *Assume that the loss functions are 1-Lipschitz, and the constraint set \mathcal{C} has diameter at most one. Then there exists a LD process $\{\theta_t\}_{t \geq 0}$ with stationary distribution Θ_∞ proportional to $\exp(-\beta \max\{\mathcal{L}(\theta; D), \psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D)\})$, and*

(i) Θ_∞ satisfies ϵ -DP if $\beta = O(\epsilon n)$.

(ii) If the loss function is m -strongly convex and M -smooth, then for $t = \tilde{O}_\delta\left(\frac{1}{\beta m}\right)$ and $\beta = \tilde{O}_\delta\left(m\epsilon^2 \min\{n^2, \frac{1}{M\psi}\}\right)$, releasing $\{\theta_{t'}\}_{0 \leq t' \leq t}$ is (ϵ, δ) -DP and θ_T is within total variation distance (TVD) δ of Θ_∞ .

Furthermore, the privacy guarantee only requires Lipschitzness and smoothness.

A key takeaway from the privacy guarantee of Theorem 1.1 is that, as ψ becomes smaller, one can run the LD at a higher β and thus provide stronger risk guarantees. In particular, we show an excess empirical risk bound of p/β in Theorem 2.4. In fact, if $\psi \leq 1/n^2$, then the choice of β is an approximately n times more than that in the ϵ -DP case. We believe the relation $\beta \propto \min\{1/\psi, n^2\}$ is necessary. (See Section 2.1 for a formal reasoning.)

For part (i) in Theorem 1.1, the privacy follows from the analysis of the exponential mechanism. For part (ii), we use a continuous analog of the composition theorem for Rényi-DP (see Lemma 2.2 and more discussion on our continuous time composition theorem below). To show the sampling guarantee, we show that the stationary distribution satisfies a *log Sobolev inequality* (a measure of concentration; see Definition A.19) using standard techniques known as the *Bakry-Emery criterion* and *Holley-Stroock perturbation principle* (see e.g., Appendix A of [Sch19]). The convergence guarantee then follows using the results in [VW19].

Note that part (ii) of Theorem 1.1 also shows that Θ_∞ is private via analyzing privacy of the chain rather than the stationary distribution. This, in particular, shows that the entire trajectory of the LD is private (not

just the *final iterate*). This matters in practice as works such as [SGM⁺22, RTT⁺23] have shown improved performance and uncertainty estimation from using intermediate values.

There is another advantage of analyzing the privacy of the entire chain. Unlike the sampling/utility guarantee, the privacy in part (ii) does not rely on convexity, i.e., we can use it for non-convex loss functions. Furthermore, by taking $\psi = 0$ and comparing to the Gaussian mechanism [DKM⁺06], we can see our privacy guarantee is tight up to log factors.

Uniform stability of Langevin diffusion (Section 2.3) While empirical guarantees are useful in their own regards, it is often desirable to get population risk guarantee. We derive a population risk guarantee by showing the uniform stability property of the LD on thresholded losses. This implies that any empirical accuracy guarantee for the Gibbs sampler in Theorem 1.1 also extends to population risk guarantees:

Informal Theorem 1.2 (Corresponds to Theorem 2.5). *Under the same assumptions and choice of β as in Theorem 1.1, the LD at any time (including at its stationary distribution) satisfies $O(L\sqrt{\psi/m} + L^2/(mn))$ -uniform stability.*

The proof uses the fact that the time-independent uniform stability for finite-time LD implies the same uniform stability for its Gibbs distribution, which could be of independent interest. This in particular gives optimal SCO rates under ε -DP guarantee that matches non-private bound of $O(1/\sqrt{n})$ as $\varepsilon \rightarrow \infty$, thereby, improving on the state-of-the-art results (Section 1.3).

Continuous time composition for LD (Section 2.1) We cannot use standard composition theorems of DP [DR14] because the underlying algorithm is a continuous time process. We quantify the Rényi divergence between two LD processes when run on neighboring data sets. A similar result was also provided in [CYS21, Theorem 1] only for the last iterate, θ_t . In contrast, we prove a divergence bound between the entire histories $\{\theta_{t'}\}_{0 \leq t' \leq t}$, which enables us to output weighted averages of $\theta_{t'}$'s privately. Furthermore, it is proven using only tools from the differential privacy literature and *Fatou's lemma*, providing an arguably much simpler proof.

Discretizing our LD (Section 3) In general, sampling from a continuous-time object such as LD is intractable in practice. A common technique for approximately sampling from the distribution induced by an LD is the *Stochastic Gradient Langevin Dynamics* (SGLD), which has been extensively studied in the literature (e.g. [Dal17, RRT17, CB18, CCAY⁺18, VW19, GT20, EHZ21, CEL⁺21, WT11, RBP22, CYS21]). SGLD uses T steps of noisy gradient descent with step size η , which approximates running eq. (1) for time $t = T\eta$. Using results in [CEL⁺21], we show that SGLD provides a private approximation (w.r.t. TVD) of our Gibbs sampler in a polynomial number of gradient oracle calls. One disadvantage of this result is that the oracle complexity has a worse dependence on the problem parameters than DP-SGD with standard hyperparameters. For example, DP-SGD's iteration complexity in [BFTT19] is constant w.r.t. dimensions as it goes to infinity, whereas our SGLD iteration complexity has a linear dependence on dimensions. We leave closing this gap as a question for future investigation.

1.3 Applications of Our Algorithmic Framework

Recovering DP-ERM/DP-SCO bounds (Section 4): We show that setting $\psi = 0$ for the Gibbs sampler in Theorem 1.1 retrieves the optimal DP-ERM/DP-SCO bounds, i.e., using only the LD sampler as a primitive, one can achieve all the existing bounds for DP-ERM/DP-SCO and improve some prior results (see Table 2) as corollaries. Since most of these results are known in the literature, in the next theorem, we only present the improvement exhibited in this paper.

Informal Theorem 1.3 (Corresponds to Theorems 4.1 and 4.2). *For L -Lipschitz convex losses over \mathcal{C} , there exists an ε -DP algorithm with $O\left(\frac{L\|\mathcal{C}\|_2^p}{\varepsilon n} + \frac{L\|\mathcal{C}\|_2}{\sqrt{n}}\right)$ excess population risk. Further, if the loss function is also m -strongly convex, then there is an ε -DP algorithm that has excess population risk of $O\left(\frac{L^2 p^2 \log n}{m\varepsilon^2 n^2} + \frac{L^2}{mn}\right)$.*

The best prior bounds were $O\left(\frac{p \log n}{\varepsilon n} + \frac{\log^{3/2} n}{\sqrt{n}}\right)$ for convex losses, and $O\left(\frac{p^2 \log^2 n}{m \varepsilon^2 n^2} + \frac{\log^3 n}{mn}\right)$ for m -strongly convex losses, both due to [ALD21], which lacked the ability to match the optimal non-private bounds of $O(1/\sqrt{n})$ (and $O(1/mn)$ respectively) as $\varepsilon \rightarrow \infty$. (To the best of our knowledge, this gap is inherent in the technique of [ALD21].)

In most cases, the best DP-ERM/SCO bounds can be achieved by setting $\psi = 0$ in Theorems 1.1 and 1.2 (in the convex case, after adding a quadratic regularizer to enforce strong convexity). The second result in Theorem 1.3 is the only one which does not directly apply the Rashomon sampler’s risk bounds to the (regularized) loss. Instead, we use an *iterated exponential mechanism*, which samples from a sequence of Gibbs distributions defined over an adaptively chosen sequence of sets $\mathcal{C}_k \subset \mathcal{C}_{k-1} \subseteq \dots \subseteq \mathcal{C}_0 = \mathcal{C}$. To analyze it, we compose the privacy/utility analysis obtained independently for the Gibbs distribution over each of these sets. Our analysis simplifies the analysis in [BST14] and does not require running two different algorithms (i.e., output perturbation and exponential mechanism) to obtain the optimal trade-off. We give the full description of the iterated exponential mechanism and its analysis in Section 4.

Additionally, for DP-ERM, in Appendix C we provide a lower bound for non-convex losses which demonstrate that, unlike for convex losses, it is not possible to achieve better utility (up to factors in $\log(1/\delta)$) than $\tilde{O}(p/\varepsilon n)$ even if we relax privacy guarantee to (ε, δ) -DP. We prove this result by appealing to the lower bound in [SU17].

Sampling from Rashomon Sets (Section 2) Our motivation to design uniform sampler for the Rashomon set stems from quantifying predictive multiplicity for Rashomon sets (described later) [HC22]⁵ and that it is a strict generalization of the problems of DP-ERM and DP-SCO. Given the wide applicability of Rashomon set mentioned earlier, we believe studying this problem will have more implications.

We start with the formal definition of uniform sampling from Rashomon set.

Definition 1.4 ((λ, ψ, γ) -Rashomon sampler). *Given a loss function $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; d_i)$ defined over a data set D , a constraint set \mathcal{C} , and a threshold ψ , an algorithm \mathcal{A} is (λ, ψ, γ) -sampler for the Rashomon set*

$$\mathcal{G} = \left\{ \theta \in \mathcal{C} : \mathcal{L}(\theta; D) \leq \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi \right\}$$

if distribution of $\theta^{\text{priv}} \leftarrow \mathcal{A}$ is λ -far in total variation distance (TVD) from a distribution π such that

1. **Uniform sampling:** The marginal distribution of π over \mathcal{G} is uniform.
2. **Maximality condition:** For any $\theta \in \mathcal{G}$ and $\theta' \notin \mathcal{G}$, the distribution π assigns probability density to θ which is greater than or equal to that of θ' .
3. **Excess empirical risk guarantee:** $\mathbb{E}_{\mathcal{A}} [\mathcal{L}(\theta^{\text{priv}}; D)] \leq \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi + \gamma$.

The *maximality* condition rules out trivial and uninteresting solutions. It ensures that (ignoring privacy constraints), it is possible to combine the Rashomon sampler with rejection sampling to efficiently get an uniform sample from \mathcal{G} . In particular, we show in Theorem B.6 that, if $\psi = \omega(p\gamma)$, then the Gibbs sampler has $1 - o(1)$ probability of hitting the Rashomon set for the values of β stated in Theorem 1.1. We presented Definition 1.4 with respect to empirical accuracy guarantee for the ease of presentation. One can also define *excess population risk guarantee*:

$$\mathbb{E}_{\mathcal{A}, d \sim \mathcal{D}} [\ell(\theta^{\text{priv}}; d)] \leq \psi + \min_{\theta \in \mathcal{C}} \mathbb{E}_{d \sim \mathcal{D}} [\ell(\theta; d)] + \gamma',$$

where γ' is the *population level slack* and \mathcal{D} is the distribution from which the data samples in the data set D are drawn i.i.d.

⁵Predictive multiplicity refers to models that achieve statistically-indistinguishable performance on a test set assign wildly different predictions to an input sample. Therefore, if we naively pick a model from a Rashomon set, it can have highly disparate impact on the predictions on an individual test sample resulting in unfair and potentially individual level harm [Smi20, CH22].

In Theorems 2.1 and 2.3, we show that our LD based Gibbs sampler from Theorem 1.1 is a $(0, \psi, p/\beta)$ -Rashomon sampler. Furthermore, in Theorem 2.6, we show that the excess population risk (i.e., $\psi + \gamma'$) for the Rashomon set sampler, is bounded by $\frac{p}{\beta} + O(\psi + L\sqrt{\psi/m} + L^2/mn)$. We give a summary of the bounds in Table 1.

We next discuss the use case of predictive multiplicity eluded above. Consider a classification problem with c classes and Δ_c be the probability simplex. Let (y, \mathbf{x}) be a test sample and let $f(\mathbf{x}; \theta) \in \Delta_c$ be the prediction function that provides a probability distribution across the c classes. The objective is to estimate the variance, $\text{Var}(f(\mathbf{x}, \theta))$, where $\theta \sim_{\text{unif}} \mathcal{G}$. One can get its approximate estimate by sampling $\{\theta_1, \dots, \theta_k\} \sim_{\text{iid}} \mathcal{G}$, and estimating the standard deviation of $\{f(\mathbf{x}, \theta_1), \dots, f(\mathbf{x}, \theta_k)\}$. While there is a privacy cost in sampling k models, [BH18, Section 6.4] shows that the variance estimation comes at *no additional privacy cost* when compared to that of outputting a single model. Given the standard deviation, one can then decide on the class to predict for \mathbf{x} , either via further randomization, or other strategy, including more sophisticated measures like *Rashomon capacity* [HC22]. We leave exploring their DP variants for future research.

Organization: For the ease of presentation and owing to the generality of the Rashomon set sampling problem, we state all our main results in that context in Section 2. In Section 3 we provide the details for the discretization of the LD algorithm. In Section 4 we provide DP-ERM/DP-SCO results obtained by setting $\psi = 0$ for the Rashomon set sampler. Finally, in Section 5 we end with discussions and future directions. We enumerate our notations in Table 3.

1.4 Omitted results from v3

Most results in the previous version of this paper on arXiv (v3) appear in this version or are subsumed by results in this version. For ease of presentation, a few results in v3 were not carried over. The following is a complete list of the omitted results:

- Section 3.3 of v3, which gives a tighter analysis of the empirical loss guarantees for the continuous exponential mechanism on non-convex losses than the one in Bassily et al. [BST14] by removing the “small ball” assumption.
- Section 5 and Appendix F of v3 analyze the DP-ERM/SCO guarantees of finite-time LD under approximate DP by a gradient descent-like analysis (as opposed to the results in this version, which use the analysis of the exponential mechanism).
- Section 7 of v3 (i) gives intuition for why the sum of step sizes in DP-SGD can be quite small (ii) shows that at some time when DP-SGD/LD achieve the asymptotically optimal ERM bound, their output distribution is total variation distance $1 - o(1)$ from their stationary distribution. Note that (ii) in v3 is shown for a loss which is not strongly convex, i.e. (ii) does not contradict the results in this paper which rederive the optimal ERM bound by showing the chain mixes for a strongly convex loss.
- Section 8 of v3 shows a bound on the last-iterate Rényi divergence between running DP-LD on two adjacent databases that does not go to infinity as t goes to infinity. This is somewhat subsumed by the analysis in section 2, which shows a qualitatively similar statement in terms of approximate DP instead of Rényi DP.

2 Rashomon Set Sampling

In this section, we provide the privacy analysis (Sections 2.1 and 2.4)⁶, and the utility analysis (Sections 2.2 and 2.3) for the Rashomon set sampler based on the Gibbs distribution proportional to $\exp(-\beta \tilde{\mathcal{L}}(\theta; D))$, where

$$\tilde{\mathcal{L}}(\theta; D) := \max\{\mathcal{L}(\theta; D), \psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D)\}.$$

⁶The analysis in Section 2.1 is via analyzing the privacy of the path of the LD that converges to the Gibbs distribution, and an alternate one in Section 2.4 via directly analyzing the privacy of the Gibbs distribution.

Both the privacy and the utility guarantee primarily depend on the choice of the inverse temperature β . Under (ϵ, δ) -DP, we can operate with higher values of β than in the case of ϵ -DP (see Theorems 2.1 and 2.3): this translates to a better utility for certain choices of ψ . In Table 1 we provide a summary of the empirical/population risk guarantees for our Rashomon samplers satisfying a given privacy constraint.

In our algorithmic version of the Gibbs distribution, we instantiate it with the LD described in Figure 1. The LD that is used to instantiate the Gibbs distribution in the ϵ -DP case unfortunately requires running the algorithm for $t \rightarrow \infty$ to reach within ϵ of the stationary distribution in terms of ℓ_∞ -distance⁷. However, if we are willing to tolerate (ϵ, δ) -DP guarantee, then the LD can be run in time $\approx \frac{\log(1/\delta)}{\beta m}$, where m is the strong convexity parameter. A standard discretization of the LD we use in this section (via SGLD), that makes the algorithm run on a finite precision machine, can be found in Section 3. All the missing proofs of this section appear in Appendix B.

Privacy guarantee	ϵ -DP	(ϵ, δ) -DP
Excess empirical risk (γ)	$\frac{Lp}{\epsilon n}$	$\frac{L^2 p \log(1/\delta)}{m \epsilon^2 n^2}$
Excess population risk	$\frac{Lp}{\epsilon n} + \psi + L \sqrt{\frac{\psi}{m} + \frac{L^2}{mn}}$	$\frac{L^2 p \log(1/\delta)}{m \epsilon^2 n^2} + \psi + L \sqrt{\frac{\psi}{m} + \frac{L^2}{mn}}$

Table 1: Summary of our Rashomon sampler guarantees. In all results, λ (the sampling error) is 0. In bounds where the parameters appear, we assume L -Lipschitzness, m -strong convexity, and M -smoothness within \mathcal{C} .

2.1 Privacy Guarantees and Running Time

The privacy in ϵ -DP case follows from the fact that $\max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\}$ cannot change by more than $L \|\mathcal{C}\|_2 / n$ when we change one data point because each $\ell \in [0, L \|\mathcal{C}\|_2]$.

Theorem 2.1 (ϵ -DP sampler; Theorem 3.1 of [BST14]). *Suppose we have a constraint set \mathcal{C} and a loss function $\ell(\theta; d)$ such that for all d , $\ell(\cdot; d) \in \mathbb{R}^+$, and is L -Lipschitz within \mathcal{C} . Then, sampling θ^{priv} from the Gibbs distribution $\exp(-\beta \max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\})$ is ϵ -differentially private for $\beta = O(\frac{\epsilon n}{L \|\mathcal{C}\|_2})$ and all ψ .*

(ϵ, δ) -DP Sampler: If $\ell(\theta; \cdot)$ is m -strongly convex and our goal is (ϵ, δ) -DP, we can use larger values of β . However, for the settings when $\psi > 0$, we would additionally require $\ell(\theta; \cdot)$ to be M -smooth. We first need a ‘‘continuous’’ composition theorem that bounds the Rényi divergence between two instances of LDs run on adjacent databases:

Lemma 2.2. *Let θ_0, θ'_0 have the same distribution Θ_0 , θ_t be the solution to (2) given θ_0 and data set D (and correspondingly θ'_0 and θ'_t for a data set D'). Let $\Theta_{[0,t]}$ ($\Theta'_{[0,t]}$) be the distribution of the trajectory of LD $\{\theta_{t'}\}_{t' \in [0,t]}$ ($\{\theta'_{t'}\}_{t' \in [0,t]}$, respectively). Suppose we have that $\|\nabla \mathcal{L}(\theta; D) - \nabla \mathcal{L}(\theta; D')\|_2 \leq \Delta$ for all θ . Then $\forall \alpha \geq 1$:*

$$R_\alpha(\Theta_{[0,t]}, \Theta'_{[0,t]}) \leq \frac{\alpha \beta^2 \Delta^2 t}{4}.$$

The idea behind the proof is to use a bound on the divergence between Gaussians and RDP composition to provide a bound on the divergence between the projected noisy gradient descents on datasets D and D' . Then, taking the limit as the step size in gradient descent goes to 0 and applying Fatou’s lemma (Lemma A.14), we get the bound above. A full proof is deferred to Appendix B.1. Rényi divergence bounds imply (ϵ, δ) -DP privacy guarantees (Fact A.4), which we use to prove the following theorem in Appendix B.2:

⁷[BST14, MV21] provides rejection sampling based polytime algorithms, but lack the generalization properties of LD.

Theorem 2.3 ((ϵ, δ) -DP sampler). *Suppose we have a constraint set \mathcal{C} and a loss function $\ell(\theta; d)$ such that for all d , $\ell(\cdot; d) \in \mathbb{R}^+$ is m -strongly convex, M -smooth, and is L -Lipschitz within \mathcal{C} . For $t = \tilde{O}_\delta\left(\frac{1}{\beta m}\right)$ and an appropriate choice of θ_0 , let Θ_t be the distribution over θ_t given by running (2) on $\max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\}$. Then Θ_t is within total variation distance δ of the Gibbs distribution on $\max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\}$, and is (ϵ, δ) -DP if any of the following holds:*

- (i) $\psi \in \left(\frac{2L^2}{Mn^2}, \frac{L\|\mathcal{C}\|_2}{2}\right]$, and $\beta = \tilde{\Theta}\left(\frac{\epsilon^2(m/M)}{\log(L\|\mathcal{C}\|_2/\delta^2)\log(1/\delta)} \cdot \frac{1}{\psi}\right)$,
- (ii) $\psi \leq \frac{2L^2}{Mn^2}$ and $\beta = \tilde{\Theta}\left(\frac{\epsilon^2 n^2 m}{L^2 \log(L\|\mathcal{C}\|_2/\delta^2)\log(1/\delta)}\right)$.

Furthermore, the privacy holds even if we release the entire trajectory $\{\Theta_{t'}\}_{t' \in [0, t]}$, and even without convexity.

If ψ has a dependence $o(1/n)$ on n , this gives a better dependence of β on n than Theorem 2.1. For privacy, we first bound the sensitivity of the thresholded loss by $O(\max\{L/n, \sqrt{M\psi}\})$. We then appeal to Lemma 2.2 and the translation from Rényi divergence bounds to (ϵ, δ) -DP bounds (Fact A.4). Since Lemma 2.2 does not require convexity and allows for releasing the entire trajectory, the same is true for the privacy guarantee of Theorem 2.3. The sampling guarantee is given by showing that the Gibbs distribution satisfies log-Sobolev inequality (LSI; see Definition A.19), a measure of concentration. This implies convergence of LD to the Gibbs distribution by results in [VW19]. A few comments are in order about the theorem.

Direct analysis of Gibbs distribution Note that by triangle inequality, Theorem 2.3 also implies that, under strong convexity, the Gibbs distribution is $(\epsilon, 2\delta)$ -DP. We can also analyze the Gibbs distribution directly: Since the Gibbs distribution satisfies LSI, one can obtain an isoperimetric inequality for the probability measure via [Led99]. Using this isoperimetric inequality and the coupling between Gibbs distributions that we later state in Theorem 2.5, one can obtain a bound on β that improves Theorem 2.3 by log factors, giving tight bounds in the case $\psi = O\left(\frac{L^2}{Mn^2}\right)$. However, this privacy proof relies heavily on the convexity and does not give an “anytime” private sampler like our LD-based proof. For the sake of completeness, we provide a proof of this result in Section 2.4 (see Theorem 2.7 for a precise statement).

Dependence of β on ψ We believe the relation $\beta \propto \min\{1/\psi, n^2\}$ in Theorem 2.3 is necessary. To see this, consider mean estimation on an all zero databases D and on neighboring D' that contain just a single 1. Fix $\epsilon = 1$ for simplicity. The Gibbs samplers for these datasets have densities proportional to $\exp(-\beta \min\{\|x\|_2^2/2, \psi\})$ and $\exp(-\beta \min\{\|x - 1/n\|_2^2/2, \psi\})$, respectively. For the case $\psi = 0$, this is just a Gaussian mechanism with sensitivity $1/n$ and variance $1/\beta$, and one way to argue this mechanism is differentially private is to provide a tail bound on x , which gives a tail bound on the privacy loss. For $\psi > 0$, the unnormalized density decreases by a factor of $\exp(-\Omega(\beta\psi))$ in the range $[-\sqrt{2\psi}, \sqrt{2\psi}]$ when we apply thresholding to the loss function. In turn, if most of the probability mass is in this interval, a tail bound on events outside this interval can get worse by a factor of $\exp(-\Omega(\beta\psi))$. Suppose we use $\beta \approx n^2$, which gives $(1, \delta)$ -DP for $\psi = 0$. Then the interval $[-\sqrt{2\psi}, \sqrt{2\psi}]$ contains all points within $\sqrt{\psi/\beta} \approx n\sqrt{\psi}$ standard deviations of the mean. So this interval contains most of the probability mass of the mechanism when $\psi = \tilde{\Omega}(1/n^2)$. This roughly matches the “transition point” in Theorem 2.3. In order for a tail bound that holds w.p. $1 - \delta$ on the Gaussian mechanism to be non-vacuous after thresholding, we need $\beta\psi = O(\log(1/\delta))$. This roughly matches our choice of β after the “transition point.”

Running time In the following discussion, we will assume $L = \|\mathcal{C}\|_2 = \Theta(1)$ for brevity. The LD for the ϵ -DP sampler (mentioned in Theorem 2.1) runs for time $t \rightarrow \infty$ to obtain the privacy/utility trade-off obtained by Theorems 2.1 and 2.4. However, if we are willing to tolerate (ϵ, δ) -DP, then assuming the loss function $\mathcal{L}(\theta; D)$ is m -strongly convex, one can obtain asymptotically the same privacy/utility trade-off, and run for time $t_\beta = \tilde{O}_\delta\left(\max\left\{\frac{1}{\beta m}, \frac{\psi}{m}\right\}\right)$. This follows from Lemma B.5. Setting $\beta = \epsilon n$ from

Theorem 2.1, we have $t_\beta = \tilde{O}_\delta \left(\max \left\{ \frac{1}{m\epsilon n}, \frac{\psi}{m} \right\} \right)$. The (ϵ, δ) -DP sampler from Theorem 2.3 also runs in time t_β , where β is chosen based on Theorem 2.3. Hence, to obtain the best (ϵ, δ) -DP Rashomon sampler, one needs $t_\beta = \tilde{O}_\delta \left(\max \left\{ \frac{1}{m\epsilon^2 n^2}, \frac{\epsilon^2 M \psi}{m} \right\} \right)$. In Section 3, we discuss how one can obtain an approximate sampler using only discrete noisy gradient steps.

Probability of hitting the Rashomon Set The Gibbs measure induced by the LD algorithm i.e.,

$$\exp(-\beta \max\{\mathcal{L}(\theta; D), \psi + \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)\})$$

trivially satisfies the maximality condition for the Rashomon sampler. In Theorem B.6, we show that the ϵ -DP Rashomon set sampler hits the set \mathcal{G} with probability at least $1 - o(1)$, if $\psi = \tilde{\omega}(p^2/(\epsilon n))$. (Here, we assumed all other parameters to be constant.) Although all our Rashomon set samplers are forced to provide non-zero probability mass on the Rashomon set due to the maximality condition, it is unclear how to obtain a similar guarantee as Theorem B.6 for our (ϵ, δ) -DP sampler. We leave it as an open question.

2.2 Excess Empirical Risk Guarantees

The excess empirical risk guarantee follows from Theorem 3.2 in [BST14]:

Theorem 2.4. *Assume each loss function is convex. Then sampling θ^{priv} from the Gibbs distribution*

$$\exp(-\beta \max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\})$$

is a $(0, \psi, p/\beta)$ -Rashomon sampler.

Using Theorem 2.4, to get the best risk bound it suffices to find the largest value of β that preserves DP under given assumptions on the loss function. Theorem 2.3 suggest that when $\psi = 0$, the setting of β that is required for (ϵ, δ) -DP is independent of the smoothness parameter M , and, hence can also be applied to non-smooth functions. Combining Theorems 2.1, 2.3, and 2.4 we get the existence of the following three Rashomon samplers:

- $\left(0, \psi, \tilde{O} \left(p \cdot \frac{L\|\mathcal{C}\|_2}{\epsilon n} \right)\right)$ -sampler for all ψ and is ϵ -DP.
- $\left(0, \psi, \tilde{O} \left(p\psi \cdot \frac{M \log(L\|\mathcal{C}\|_2/\delta^2) \log(1/\delta)}{m\epsilon^2} \right)\right)$ -sampler when $\psi \in \left(\frac{2L^2}{Mn^2}, \frac{L\|\mathcal{C}\|_2}{2} \right]$ and is (ϵ, δ) -DP.
- $\left(0, \psi, \tilde{O} \left(p \cdot \frac{L^2 \log(L\|\mathcal{C}\|_2/\delta^2) \log(1/\delta)}{m\epsilon^2 n^2} \right)\right)$ -sampler when $\psi \leq \frac{2L^2}{Mn^2}$ and is (ϵ, δ) -DP.

In the (ϵ, δ) -DP setting if instead of sampling from the Gibbs distribution, we operate with the LD, then in the Rashomon sampler we set $\lambda = \delta$ instead of $\lambda = 0$ as done until now. The loss guarantee worsens by at most $O(\delta L^2/m)$ if we use LD instead of the Gibbs sampler, so the earlier bounds remain unchanged for $\delta = O(1/\epsilon^2 n^2)$.

Ignoring the polylogarithmic terms, if $\psi = \tilde{\Omega} \left(\frac{\epsilon L\|\mathcal{C}\|_2}{(M/m)} \cdot \frac{1}{n} \right)$, then our privacy analysis of the ϵ -DP Rashomon sampler gives a better bound on β than our analysis of the (ϵ, δ) -DP sampler, and vice versa when $\psi = \tilde{O} \left(\frac{\epsilon L\|\mathcal{C}\|_2}{(M/m)} \cdot \frac{1}{n} \right)$. As previously mentioned, we believe that the weaker bound on β for higher values of ψ in our analysis of the (ϵ, δ) -DP sampler is fundamental to the problem.

2.3 Excess Population Risk Guarantees

For the same Rashomon samplers, we can derive bounds on their population risks under strong convexity. We give a bound on uniform stability (see Definition A.15) of the Gibbs distribution:

Theorem 2.5. Suppose we have \mathcal{C} and $\ell(\cdot; d) \in \mathbb{R}^+$ such that, for all d , $\ell(\cdot; d)$ is m -strongly convex, and is L -Lipschitz within \mathcal{C} . Then sampling θ^{priv} from the Gibbs distribution proportional to

$$\exp(-\beta \max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\})$$

satisfies $\left(4L\sqrt{\frac{2\psi}{m}} + \frac{2L^2}{mn}\right)$ -uniform stability.

Proof. Let θ_t, θ'_t be the solutions to running (2) from the same initialization on $\mathcal{L}(\theta; D)$ and $\mathcal{L}(\theta; D')$, respectively. Similarly, let $\tilde{\theta}_t$ be the solutions to running eq. (2) on $\max\left\{\psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D), \mathcal{L}(\theta; D)\right\}$ and $\tilde{\theta}'_t$ be the solutions to running eq. (2) on $\max\left\{\psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D'), \mathcal{L}(\theta; D')\right\}$. Let W_∞ denote the the ∞ -Wasserstein distance. Then we will show that for all t , $W_\infty(\tilde{\theta}_t, \tilde{\theta}'_t) \leq 4\sqrt{\frac{2\psi}{m}} + \frac{2L}{mn}$. Taking the limit as t goes to infinity, we get a ∞ -Wasserstein distance bound between the Gibbs samplers. The theorem now follows by using L -Lipschitzness.

We now show the desired bound on $W_\infty(\tilde{\theta}_t, \tilde{\theta}'_t)$. By the triangle inequality, we have

$$W_\infty(\tilde{\theta}_t, \tilde{\theta}'_t) \leq W_\infty(\tilde{\theta}_t, \theta_t) + W_\infty(\theta_t, \theta'_t) + W_\infty(\theta'_t, \tilde{\theta}'_t).$$

So, it suffices to prove that for all t ,

$$W_\infty(\tilde{\theta}_t, \theta_t) \leq 2\sqrt{\frac{2\psi}{m}}, \quad W_\infty(\theta'_t, \tilde{\theta}'_t) \leq 2\sqrt{\frac{2\psi}{m}}, \quad \text{and} \quad W_\infty(\theta_t, \theta'_t) \leq \frac{2L}{mn}.$$

We first prove the desired bound on $W_\infty(\tilde{\theta}_t, \theta_t)$, the bound on $W_\infty(\theta'_t, \tilde{\theta}'_t)$ follows identically.

Bounding $W_\infty(\tilde{\theta}_t, \theta_t)$: We show that conditioned on the value of a *shared* Brownian motion W_t , $\|\theta_t - \tilde{\theta}_t\|_2 \leq 2\sqrt{\frac{2\psi}{m}}$ holds deterministically, which implies the desired Wasserstein distance bound. We split $[0, \infty)$ into intervals of maximal length for which one of the following three holds throughout each interval:

- (i) $\tilde{\theta}_t \notin \mathcal{G}$,
- (ii) $\tilde{\theta}_t \in \mathcal{G}, \theta_t \in \mathcal{G}$, and
- (iii) $\tilde{\theta}_t \in \mathcal{G}, \theta_t \notin \mathcal{G}$.

In case (ii), by strong convexity throughout the interval we have

$$\|\theta_t - \tilde{\theta}_t\|_2 \leq \|\mathcal{G}\|_2 \leq 2\sqrt{\frac{2\psi}{m}}.$$

So it suffices to show that in cases (i) and (iii), $\|\theta_t - \tilde{\theta}_t\|_2$ is non-increasing throughout the interval. Then the desired Wasserstein distance bound holds by induction, since initially $\theta_0 = \tilde{\theta}_0$.

In case (i), since \mathcal{L} is convex, projection is contractive (which implies $\frac{d\|\theta_t - \tilde{\theta}_t\|_2^2}{dt}$ can only increase if we use (1) instead of (2)), and that we are using a shared Brownian motion W_t , we have

$$\frac{1}{2} \cdot \frac{d}{dt} \|\theta_t - \tilde{\theta}_t\|_2^2 \leq \left\langle \frac{d\theta_t}{dt} - \frac{d\tilde{\theta}_t}{dt}, \theta_t - \tilde{\theta}_t \right\rangle = \beta \langle -(\nabla \mathcal{L}(\theta_t; D) - \nabla \mathcal{L}(\tilde{\theta}_t; D)), \theta_t - \tilde{\theta}_t \rangle \leq 0.$$

Similarly, in case (iii), by convexity and since $\tilde{\theta}_t$ is in the Rashomon set and θ_t is not (or $\mathcal{L}(\tilde{\theta}_t; D) \leq \mathcal{L}(\theta_t; D)$), we have

$$\frac{1}{2} \cdot \frac{d \|\theta_t - \tilde{\theta}_t\|_2^2}{dt} \leq \left\langle \frac{d\theta_t}{dt} - \frac{d\tilde{\theta}_t}{dt}, \theta_t - \tilde{\theta}_t \right\rangle = \beta \langle \nabla \mathcal{L}(\theta_t; D), \tilde{\theta}_t - \theta_t \rangle \leq \beta (\mathcal{L}(\tilde{\theta}_t; D) - \mathcal{L}(\theta_t; D)) \leq 0.$$

Bounding $W_\infty(\theta_t, \theta'_t)$: We again show that conditioned on the value of a *shared* Brownian motion W_t , $\|\theta_t - \theta'_t\|_2 \leq \frac{2L}{mn}$ holds deterministically. We again use the fact that projection is contractive, so we can consider using (1) instead of (2). Then by m -strong convexity and L -Lipschitzness:

$$\begin{aligned} \frac{1}{2} \cdot \frac{d \|\theta_t - \theta'_t\|_2^2}{dt} &\leq \left\langle \frac{d\theta_t}{dt} - \frac{d\theta'_t}{dt}, \theta_t - \theta'_t \right\rangle = -\beta \langle \nabla \mathcal{L}(\theta_t; D) - \nabla(\mathcal{L}(\theta'_t, D')), \theta_t - \theta'_t \rangle \\ &= -\beta (\langle \nabla \mathcal{L}(\theta_t; D) - \nabla(\mathcal{L}(\theta'_t, D)), \theta_t - \theta'_t \rangle + \langle \nabla \mathcal{L}(\theta'_t; D) - \nabla(\mathcal{L}(\theta'_t, D')), \theta_t - \theta'_t \rangle) \\ &\leq \beta \left(-m \|\theta_t - \theta'_t\|_2^2 + \frac{2L}{n} \|\theta_t - \theta'_t\|_2 \right). \end{aligned} \quad (3)$$

Then, if $\|\theta_t - \theta'_t\|_2 > \frac{2L}{mn}$, then we have (3) < 0 . That is, we have $\frac{d \|\theta_t - \theta'_t\|_2^2}{dt} < 0$. This implies the desired Wasserstein distance bound completing the proof of Theorem 2.5. \square

Remark 1. We note that in the second part of the proof, one can instead take the L^2/mn -uniform stability of (noisy) gradient descent on strongly convex losses proved in [HRS16], and then take the limit as $\eta \rightarrow 0, T \rightarrow \infty$ to conclude the same uniform stability bound holds for the Gibbs distribution, rather than appeal to the derivative of the distance between θ_t and θ'_t .

It is straightforward to see that a $(0, \psi, \gamma)$ -Rashomon sampler has expected excess empirical risk at most $\psi + \gamma$. Then, an algorithm's excess population risk is at most its excess empirical risk plus its uniform stability (see Lemma A.16), giving us the following result:

Theorem 2.6. Assume that each of the individual loss function $\ell(\theta; \cdot) \in \mathbb{R}^+$ is m -strongly convex and L -Lipschitz within the constraint set \mathcal{C} . Then

- There exists an ε -DP rashomon sampler with threshold ψ and excess population risk

$$O \left(\frac{L \|\mathcal{C}\|_2 p}{\varepsilon n} + \psi + L \sqrt{\frac{\psi}{m} + \frac{L^2}{mn}} \right).$$

- There exists an (ε, δ) -DP rashomon sampler with threshold ψ and excess population risk (assuming M -smoothness)

$$\tilde{O} \left(p \max \left\{ \psi, \frac{L^2}{Mn^2} \right\} \cdot \frac{M \log(L \|\mathcal{C}\|_2 / \delta^2) \log(1/\delta)}{m\varepsilon^2} + \psi + L \sqrt{\frac{\psi}{m} + \frac{L^2}{mn}} \right).$$

2.4 Interlude: Proof of Privacy for the Gibbs Sampler via Isoperimetry

Theorem 2.7. Suppose ℓ is m -strongly convex, M smooth, and L -Lipschitz within \mathcal{C} . Let

$$\mathcal{L}'(\theta; D) := \max\{\mathcal{L}(\theta; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\}.$$

Then for $\beta = O \left(\frac{\min\{\varepsilon^2, \varepsilon\} m}{\max\{L^2/n^2, M\psi\} \log(1/\delta)} \right)$, sampling from the distribution with density proportional to $\exp(-\beta \mathcal{L}'(\theta; D)) \cdot \mathbb{1}(\theta \in \mathcal{C})$ satisfies (ε, δ) -DP.

Proof. Let P be the probability measure for the distribution induced by D and Q be the same for D' . Let $g(\theta) = \log(P(\theta)/Q(\theta))$. Then g is a $6\beta \max\{L/n, \sqrt{M\psi}\}$ -Lipschitz function. It suffices to show that $\Pr_{\theta \sim P}[g(\theta) > \varepsilon] \leq \delta$. We first bound $E_{\theta \sim P}[g(\theta)]$ as a function of β . This is simply the KL divergence between P, Q , which must be non-negative. By symmetry, $E_{\theta \sim Q}[g(\theta)]$ is non-positive. In addition, we can bound the difference between these two expressions: in the proof of Theorem 2.5, we showed that the ∞ -Wasserstein distance between P and Q is at most $4\sqrt{\frac{2\psi}{m}} + \frac{2L}{mn}$. Then, by Lipschitzness of g ,

$$\begin{aligned} E_{\theta \sim P}[g(\theta)] &\leq E_{\theta \sim Q}[g(\theta)] + 6\beta \max\{L/n, \sqrt{M\psi}\} \cdot \left(4\sqrt{\frac{2\psi}{m}} + \frac{2L}{mn}\right) \\ &\leq 6\beta \max\{L/n, \sqrt{M\psi}\} \cdot \left(4\sqrt{\frac{2\psi}{m}} + \frac{2L}{mn}\right) \end{aligned} \quad (4)$$

$$\leq 36\sqrt{2}\beta \max\left\{\frac{L^2}{mn^2}, \sqrt{\frac{M}{m}}\psi\right\}. \quad (5)$$

We next give a high probability bound on g as a function of β . P satisfies LSI with constant $\beta m \exp(-\beta\psi)$ by Lemma B.4. Then by Proposition 2.3 in [Led99], plugging in our LSI constant and Lipschitzness bound for g we have:

$$\Pr_{\theta \sim P}[g(\theta) > E_{\theta \sim P}[g(\theta)] + \beta r] \leq \exp\left(-\frac{\beta m \exp(-\beta\psi)r^2}{18 \max\{L^2/n^2, M\psi/2\}}\right) \quad (6)$$

Setting

$$r = \frac{3\sqrt{2} \max\{L/n, \sqrt{M\psi/2}\}}{\sqrt{\beta m} \exp(-\beta\psi/2)} \sqrt{\log(1/\delta)},$$

the right hand side in eq. (6) becomes δ . Setting

$$\beta \leq \frac{\min\{\varepsilon^2, \varepsilon\}m}{c \max\{L^2/n^2, M\psi\} \log(1/\delta)},$$

where c is a sufficiently large constant, gives that the upper bound in (5) is at most $\varepsilon/2$ and $\beta r \leq \varepsilon/2$. Here, we use that for this choice of $\beta, \beta\psi \leq 1$. Plugging in to (6) we get $\Pr_{\theta \sim P}[g(\theta) > \varepsilon] \leq \delta$ exactly as desired. This completes the proof of Theorem 2.7. \square

2.5 Comparison with Contemporary Work of [GLL22]

All our results for DP-ERM and DP-SCO, and their relation with Langevin diffusion is contemporary to [GLL22] (which is also formally acknowledged by [GLL22]). Our results specific to Rashomon sets (i.e., with setting $\psi > 0$) and SGLD are subsequent to their work. In the following, we discuss the technical differences between two works, and highlight the settings in which one might be better over the other.

On the privacy aspect Since [GLL22] gives Gaussian DP guarantees, which do not translate to ε -DP guarantees, we will restrict the discussion specific to (ε, δ) -DP. Unlike [GLL22], our privacy guarantee (for LD as well as SGLD) is independent of convexity. This is highly desirable for broader applicability in settings where the loss function may not be globally convex, but has local convexity properties (e.g., losses emerging from deep learning settings [IPG⁺18, KMN⁺16].) Our privacy holds for the entire path of the optimization. In contrast, [GLL22] only guarantee privacy for the final model release. In particular, this implies that we can stop and output the model at any point of the optimization trajectory. While this might not yield optimal model, the privacy is never compromised. The ability to release the path has been used in subsequent works [SGM⁺22, RTT⁺23] for uncertainty quantification.

On the proof technique. In Theorem 2.7, we demonstrated that the proof technique of [GLL22] can be extended to obtain privacy for the Gibbs distribution used in Rashomon set sampling problem (i.e., when $\psi > 0$). However, unlike Lemma 2.2, the proof of Theorem 2.7 requires an isoperimetric inequality based on LSI [Led99], and bounding the LSI constant for the Rashomon set sampler. In addition, while both our work and [GLL22] shows uniform stability property for the Gibbs distribution, our proof is arguably simpler: Theorem 2.5 uses only well-known properties of gradient descent and avoids tools such as the Talagrand transportation inequality. Since the proof techniques of [GLL22] and ours are completely independent, we believe that both the techniques will find adoption in subsequent works on DP optimization.

On run time In the DP-SCO setting, the gradient oracle complexity of [GLL22] is better than our SGLD discretization of the Langevin diffusion. In particular, they achieve oracle complexity with logarithmic dependence on p and the total variation distance error δ , whereas we have a dependence p/δ^2 on these two parameters. Furthermore, we require exact gradient oracles, whereas they only require an unbiased function oracle.

3 Discrete Approximation of the Langevin Diffusion and SGLD

While exactly sampling from the Gibbs distribution may be computationally intractable, under some additional assumptions, a number of papers study polynomial-time algorithms for approximate sampling from the Gibbs distribution with various metric of approximation, such as Renyi divergence [VW19, GT20, EHZ21, CEL⁺21] and ∞ -divergence [BST14, MV21]. For sampling from the distribution $\exp(-\beta\mathcal{L})$, a popular approximate sampler is the Stochastic Gradient Langevin Dynamics (SGLD). The SGLD approximates a finite-time solution to (1) via the discrete updates: $\theta_{t+1} = \theta_t - \eta\beta\nabla\mathcal{L}(\theta_t; D) + \xi_t$, where $\xi_t \sim \mathcal{N}(0, 2\eta I_p)$. This update can be seen as equivalent to eq. (1) if we use $\nabla\mathcal{L}(\theta_{\eta\lfloor t/\eta\rfloor}; D)$ instead of $\nabla\mathcal{L}(\theta_t; D)$, i.e., instead of continuously, update the gradient drift term in eq. (1) every η time.

Many works study SGLD as an approximate sampler, and show that, for sufficiently small η and large T , θ_T is an approximate sample from the stationary distribution $\exp(-\beta\mathcal{L})$ of (1). For an appropriate definition of approximation, such as total variation distance, privacy of the stationary distribution then implies privacy of SGLD. SGLD also converges in polynomial time for stronger notions of convergence such as the Rényi divergence, but since divergences are “one-directional” bounds, whereas privacy requires “bi-directional” bounds, these results combined with the privacy of the stationary distribution do not necessarily ensure privacy of SGLD. One could also use the result of [AT22c], which shows SGLD is an approximate sampler for the *discrete chain’s* stationary distribution. However, one would then need to show a bound on the bias due to the discretization that preserves privacy and utility.

Instead, we appeal to the privacy of the noisy gradient steps taken since SGLD is just a reparameterization of DP-SGD. Using the composition theorem for Rényi divergences (Fact A.8) and translation from Rényi divergence bounds to (ϵ, δ) -DP (Fact A.4), we get the following.

Lemma 3.1. *If*

$$\|\nabla\mathcal{L}(\theta; D) - \nabla\mathcal{L}(\theta; D')\|_2 \leq \Delta \quad \text{and} \quad \beta \leq \frac{2\epsilon}{\Delta\sqrt{T\eta\log(1/\delta)}},$$

then outputting θ_T sampled from SGLD is (ϵ, δ) -DP.

This statement matches Theorem 2.2 as $\eta \rightarrow 0$. For the utility guarantee of SGLD as an approximate sampler, we use results from [CEL⁺21], though these require smoothness. The thresholded loss does not satisfy smoothness for $\psi > 0$. In Appendix B.4, we show that using standard smoothing techniques, one can still get a discrete approximate sampler for the Gibbs distribution of a smoothed version of the thresholded loss, which implies the following sampler:

	Assumption	ε -DP	(ε, δ) -DP
DP-ERM	Convex	$\frac{Lp}{\varepsilon n}$	$\tilde{O}_\delta \left(\frac{L\sqrt{p}}{\varepsilon n} \right)$
	SC	$\frac{L^2(p^2+p \log n)}{m\varepsilon^2 n^2}$	$\tilde{O}_\delta \left(\frac{L^2 p}{m\varepsilon^2 n^2} \right)$
DP-SCO	Convex	$\frac{L}{\sqrt{n}} + \frac{Lp}{\varepsilon n}$	$\frac{L}{\sqrt{n}} + \tilde{O}_\delta \left(\frac{L\sqrt{p}}{\varepsilon n} \right)$
	SC	$\frac{L^2}{mn} + \frac{L^2 p^2 \log n}{m\varepsilon^2 n^2}$	$\frac{L^2}{mn} + \tilde{O}_\delta \left(\frac{L^2 p}{m\varepsilon^2 n^2} \right)$

Table 2: Summary of results that can be (re-)derived using the Rashomon sampler. The bounds marked in blue were not known even via different algorithms, and all other bounds are tight. Convex: Convex bounded Lipschitz losses, SC: Convex with $\nabla^2 \ell(\theta; \cdot) \succcurlyeq m\mathbb{I}$.

Theorem 3.2. Suppose $\|\nabla \mathcal{L}(\theta; D) - \nabla \mathcal{L}(\theta; D')\|_2 \leq \Delta$, and that each individual loss function ℓ is m -strongly convex and M -smooth. Let $0 \leq \lambda \leq \sqrt{\frac{\psi}{Mp}}$ and Q be the (unconstrained) Gibbs distribution of $\beta \tilde{\mathcal{L}}'$, where

$$\tilde{\mathcal{L}}' := \mathbb{E}_{\xi \sim N(0, \lambda^2 \mathbb{I}_p)} \left[\min\{\mathcal{L}(\theta + \xi; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\} \right].$$

Then for

$$\beta = \tilde{O} \left(\frac{\varepsilon^2 m}{\max\{\Delta^2, M\psi\} \log^2(1/\delta)} \right), T = \tilde{\Omega} \left(\frac{p(M^2 + \frac{M\psi}{\lambda^2}) \max\{\Delta^4, M^2\psi^2\}}{\varepsilon^4 m^4 \delta^2} \right),$$

there exists an algorithm using T iterations that satisfies (ε, δ) -DP and returns a sample from a distribution whose TVD from Q is δ . Furthermore, the privacy guarantee holds even without convexity. In particular, if $\psi = 0$, for $\lambda = 0$ the bounds are $\beta = \tilde{O} \left(\frac{\varepsilon^2 m}{\Delta^2 \log^2(1/\delta)} \right)$, $T = \tilde{\Omega} \left(\frac{pM^2 \Delta^4}{\varepsilon^4 m^4 \delta^2} \right)$.

In Theorem 3.2, we do not know if one can exactly compute the values of $\nabla \tilde{\mathcal{L}}'(\theta)$, but one can make approximate oracle calls to the gradients of $\tilde{\mathcal{L}}'$ via Monte Carlo sampling. That is, one samples ξ_1, \dots, ξ_k , and then uses $\frac{1}{k} \sum_{i \in [k]} \nabla(\min\{\mathcal{L}(\theta + \xi; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\})$ as an estimate of $\nabla \tilde{\mathcal{L}}'(\theta)$. It is easy to check that using Monte Carlo sampling instead of exact gradients does not affect our (worst-case) privacy analysis. Of course, in practice, the assumptions in Theorem 3.2 may not hold anyway, and using the convolved loss function and using the unperturbed loss function will lead to similar outcomes for small λ , and also have the same privacy guarantees.

4 Optimal DP-ERM/SCO Bounds from Rashomon Samplers

In this section we show that just using the Rashomon sampler with $\psi = 0$ as a primitive is enough to derive near-optimal bounds for DP-ERM/SCO in all settings. Our results are summarized in Table 2.

4.1 Pure DP, Convex Losses

For L -Lipschitz convex losses on \mathcal{C} , the best possible excess empirical risk under ε -DP is $O \left(\frac{L\|\mathcal{C}\|_2 p}{\varepsilon n} \right)$. This bound is achieved by the exponential mechanism as shown in [BST14], which is exactly what the Gibbs distribution is for $\psi = 0$.

The best possible excess population risk under ε -DP is $O \left(\frac{L\|\mathcal{C}\|_2 p}{\varepsilon n} + \frac{L\|\mathcal{C}\|_2}{\sqrt{n}} \right)$. We can achieve this via Theorem 2.6 by setting $\psi = 0$ and adding the regularizer $\frac{L}{2\|\mathcal{C}\|_2 \sqrt{n}} \|\theta - \theta_0\|_2^2$, where θ_0 is an arbitrary point

in \mathcal{C} , to the loss function to give the algorithm uniform stability. The excess population risk of the Gibbs distribution with respect to the regularized loss is $O\left(\frac{L\|\mathcal{C}\|_2^p}{\varepsilon n} + \frac{L\|\mathcal{C}\|_2}{\sqrt{n}}\right)$ by Theorem 2.6, and the regularized and unregularized loss differ by at most $O\left(\frac{L\|\mathcal{C}\|_2}{\sqrt{n}}\right)$ everywhere in \mathcal{C} . Putting it all together, we get the following:

Theorem 4.1. *For convex, L -Lipschitz losses over \mathcal{C} , there exists an ε -DP algorithm with excess population risk $O\left(\frac{L\|\mathcal{C}\|_2^p}{\varepsilon n} + \frac{L\|\mathcal{C}\|_2}{\sqrt{n}}\right)$.*

4.2 Pure DP and Strongly Convex Losses

For L -Lipschitz, m -strongly convex losses on \mathcal{C} , the best possible excess empirical risk under ε -DP is $O\left(\frac{L^2 p^2}{\varepsilon^2 n^2 m}\right)$. Unfortunately, the guarantee given by Theorem 2.4 is worse by a quadratic factor. In [BST14], the optimal excess empirical risk is achieved (up to log factors) by first choosing a smaller ball using the Laplace mechanism, and then running the Gibbs sampler on this smaller ball. We show the best-known bound can be achieved using the Gibbs sampler only as a primitive, which does not allow us to use the Laplace mechanism. We propose the iterated exponential mechanism, given as Algorithm 1, with the following guarantee:

Algorithm 1 Iterated Exponential Mechanism

Require: Loss function \mathcal{L} , constraint set \mathcal{C}_0 , Lipschitz constant L , strong convexity parameter m , number of iterations k , privacy parameter sequence $\{\varepsilon_i\}_{i=1}^k$, flag and data set D of n samples.

- 1: **for** $i = 1$ to k **do**
 - 2: Sample θ_i from \mathcal{C}_{i-1} with probability proportional to $\exp\left(-\frac{\varepsilon_i n}{2L\|\mathcal{C}_{i-1}\|_2} \mathcal{L}(\theta; D)\right)$.
 - 3: **If** flag = 1 **then**
 - 4: $\mathcal{C}_i \leftarrow \left\{ \theta \in \mathcal{C}_{i-1} : \|\theta - \theta_i\|_2 \leq \sqrt{\frac{cL(p+3\log n)\|\mathcal{C}_{i-1}\|_2}{m\varepsilon_i n}} \right\}$.
 - 5: **else**
 - 6: $\mathcal{C}_i \leftarrow \left\{ \theta \in \mathcal{C}_{i-1} : \|\theta - \theta_i\|_2 \leq \sqrt{\frac{cL(p+3\log n)\|\mathcal{C}_{i-1}\|_2}{m\varepsilon_i n}} + \frac{cL\sqrt{\log n}}{m\sqrt{n}} \right\}$.
 - 7: **end**
 - 8: **return** θ_k
-

Theorem 4.2. *Assume each of the individual loss function in $\mathcal{L}(\theta; D)$ is L -Lipschitz within the constraint set \mathcal{C}_0 . For any ε , if we instantiate Algorithm 1 with $k = 1 + \lceil \log \log\left(\frac{\varepsilon n}{(p+\log n)}\right) \rceil$ and $\varepsilon_i = \varepsilon/2^{k-i+1}$, then Algorithm 1 is ε -differentially private.*

Additionally, if the loss function $\mathcal{L}(\theta; D)$ is m -strongly convex and the constraint set \mathcal{C}_0 is convex and flag = 1, then over the randomness of the algorithm, the output θ_k of Algorithm 1 has excess empirical risk:

$$O\left(\frac{L^2(p^2 + p \log n)}{\varepsilon^2 n^2 m}\right).$$

The theorem follows by solving a recurrence for $\|\mathcal{C}_i\|_2$ to bound the diameter of the final set \mathcal{C}_{k-1} . Then we show that the minimizer over \mathcal{C}_0 is also in \mathcal{C}_{k-1} with high probability. Therefore, the analysis of the exponential mechanism on \mathcal{C}_{k-1} gives the theorem. We note that in addition to only using the Gibbs sampler as a primitive, we improve the $p^2 \log n$ in [BST14] result to $p^2 + p \log n$. To bound the excess loss, we first need the following lemma, which shows that with high probability we choose a series of \mathcal{C}_i that all contain the optimal θ for \mathcal{C}_0 . It follows from a tail bound on the excess loss of the exponential mechanism, and using m -strong convexity to translate this into a distance bound.

Lemma 4.3. Let $\mathcal{L}(\cdot; D)$ be an m -strongly convex function. Suppose we sample θ from the convex constraint set \mathcal{C} with probability proportional to $\exp\left(-\frac{\varepsilon n}{2L\|\mathcal{C}\|_2}\mathcal{L}(\theta; D)\right)$. Let $\theta^* = \arg \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$. Then for any $t \geq 0$ and for some sufficiently large constant c we have

$$\Pr \left[\|\theta - \theta^*\|_2 \leq \sqrt{\frac{cL(p+t)\|\mathcal{C}\|_2}{m\varepsilon n}} \right] \geq 1 - 2^{-t}.$$

Proof. By e.g. the proof of [BST14, Theorem III.2], we know that for some sufficiently large constant c :

$$\Pr \left[\mathcal{L}(\theta; D) - \mathcal{L}(\theta^*; D) \leq \frac{cL\|\mathcal{C}\|_2}{2\varepsilon n}(p+t) \right] \geq 1 - 2^{-t}. \quad (7)$$

We now show that the claim holds conditioned on this event. By optimality of θ^* and convexity of \mathcal{C} , we know

$$\langle \nabla \mathcal{L}(\theta^*; D), \theta - \theta^* \rangle \geq 0. \quad (8)$$

So, by m -strong convexity, we have

$$\begin{aligned} \frac{cL\|\mathcal{C}\|_2}{2\varepsilon n}(p+t) &\stackrel{(7)}{\geq} \mathcal{L}(\theta; D) - \mathcal{L}(\theta^*; D) \\ &\geq \langle \nabla \mathcal{L}(\theta^*; D), \theta - \theta^* \rangle + \frac{m}{2} \|\theta - \theta^*\|_2^2 \\ &\stackrel{(8)}{\geq} \frac{m}{2} \|\theta - \theta^*\|_2^2. \end{aligned}$$

Rearranging gives the claim in Lemma 4.3. \square

Given Lemma 4.3, we can now prove Theorem 4.2.

Proof of Theorem 4.2. The privacy guarantee is immediate from the privacy guarantee of the exponential mechanism, composition, and the fact that for this choice of ε_i, k , we have $\sum_{i=1}^k \varepsilon_i < \varepsilon$.

Setting $t = 3 \log n$ in Lemma 4.3, in iteration i , letting $\theta_i^* = \arg \min_{\theta \in \mathcal{C}_{i-1}} \mathcal{L}(\theta; D)$, we have that with probability $1 - 2^{-t} = 1 - \frac{1}{n^3}$, $\theta_i^* \in \mathcal{C}_i$, and thus $\theta_i^* = \theta_{i+1}^*$. Then by a union bound, we have that with probability $1 - \frac{k}{n^3} \geq 1 - \frac{\log \log(\varepsilon n)}{n^3}$, $\theta_1^* \in \mathcal{C}_{k-1}$ (equivalently, $\theta_1^* = \theta_2^* = \dots = \theta_k^*$). When this event fails to happen, our excess loss is at most $L\|\mathcal{C}_0\|_2$, and in turn the contribution of this event failing to hold to the expected excess loss is $O\left(\frac{L\|\mathcal{C}_0\|_2 \log \log(\varepsilon n)}{n^3}\right)$, which is asymptotically less than our desired excess loss bound. So it suffices to provide the desired expected excess loss bound conditioned on this event. By the analysis of the exponential mechanism, conditioned on this event, we have that

$$\mathbb{E}_{\theta_k} [\mathcal{L}(\theta_k; D)] - \mathcal{L}(\theta_1^*; D) = O\left(\frac{Lp\|\mathcal{C}_{k-1}\|_2}{\varepsilon_k n}\right) = O\left(\frac{Lp\|\mathcal{C}_{k-1}\|_2}{\varepsilon n}\right). \quad (9)$$

Note that $\mathcal{L}(\theta_1^*; D) = \min_{\theta \in \mathcal{C}_0} \mathcal{L}(\theta; D)$ by definition, so it now suffices to bound $\|\mathcal{C}_{k-1}\|_2$ by $O\left(\frac{L(p+\log n)}{m\varepsilon n}\right)$. To do this, we have the recurrence relation:

$$\|\mathcal{C}_i\|_2 \leq 2\sqrt{\frac{cL(p+3\log n)\|\mathcal{C}_{i-1}\|_2}{m\varepsilon_i n}}.$$

Solving the recurrence relation for \mathcal{C}_{k-1} , we get:

$$\begin{aligned}\|\mathcal{C}_{k-1}\|_2 &\leq \left(\frac{4cL(p+3\log n)}{mn}\right)^{1-2^{-(k-1)}} \cdot (\|\mathcal{C}_0\|_2)^{2^{-(k-1)}} \cdot \prod_{i=1}^{k-1} \varepsilon_i^{-2^{-(k-i)}} \\ &= \left(\frac{4cL(p+3\log n)}{m\varepsilon n}\right)^{1-2^{-(k-1)}} \cdot (\|\mathcal{C}_0\|_2)^{2^{-(k-1)}} \cdot \prod_{i=1}^{k-1} (2^{(k-i+1)})^{2^{-(k-i)}}.\end{aligned}\tag{10}$$

We claim the following:

$$\|\mathcal{C}_0\|_2 \leq \frac{2L}{m}.\tag{11}$$

Let θ_{global} be the minimizer of $\mathcal{L}(\theta; D)$ over all of \mathbb{R}^p . By triangle inequality, there exists a point θ in \mathcal{C}_0 which is at distance at least $\|\mathcal{C}_0\|_2/2$ far from θ_{global} . By m -strong convexity, this implies that the gradient at θ has ℓ_2 -norm at least $m\|\mathcal{C}_0\|_2/2$. Now, by Lipschitzness over \mathcal{C}_0 , we know that the gradient at θ has ℓ_2 -norm at most L . This gives us eq. (11).

Using eq. (11), we can simplify eq. (10) to

$$\|\mathcal{C}_{k-1}\|_2 \leq \frac{2L}{m} \cdot \left(\frac{2c(p+3\log n)}{\varepsilon n}\right)^{1-2^{-(k-1)}} \cdot \prod_{i=1}^{k-1} (2^{(k-i+1)})^{2^{-(k-i)}}.$$

We have:

$$\log_2 \left(\prod_{i=1}^{k-1} (2^{(k-i+1)})^{2^{-(k-i)}} \right) = \sum_{i=1}^{k-1} (k-i+1)2^{-(k-i)} \leq \sum_{j=1}^{\infty} (j+1)2^{-j} = 3.$$

In other words, $\prod_{i=1}^{k-1} (2^{(k-i+1)})^{2^{-(k-i)}}$ is at most 8, regardless of the value of k . Now, using the fact that $m^{1/\log m} = O(1)$ is a constant, our final upper bound on $\|\mathcal{C}_{k-1}\|_2$ is:

$$\|\mathcal{C}_{k-1}\|_2 = O\left(\frac{L}{m} \cdot \left(\frac{(p+\log n)}{\varepsilon n}\right)^{1-2^{-(k-1)}}\right) = O\left(\frac{L(p+\log n)}{m\varepsilon n}\right).$$

Plugging in eq. (9) gives us Theorem 4.2. □

The best possible population risk bound is $O\left(\frac{L^2 p^2}{\varepsilon^2 n^2 m} + \frac{L^2}{mn}\right)$. In order for Algorithm 1 to achieve this bound (up to log factors) we make a slight modification: we choose the radius of each ball defined by the algorithm such that the population minimizer, rather than the empirical minimizer, is in \mathcal{C}_{k-1} with high probability. Then, we can apply uniform stability of the Gibbs sampler on strongly convex losses to the exponential mechanism run on \mathcal{C}_{k-1} to get the following DP-SCO bound:

Theorem 4.4. *Let θ_k be the output of Algorithm 1 when $\text{flag} = 0$. Then θ_k has excess population risk*

$$O\left(\frac{L^2 p^2 \log n}{m\varepsilon^2 n^2} + \frac{L^2}{mn}\right).$$

We first show that the empirical minimizer is close to the population minimizer with high probability:

Lemma 4.5. Let ℓ be a m -strongly convex function and \mathcal{C} be a convex set such that for any d, θ ,

$$\|\nabla\ell(\theta; d) - \mathbb{E}_{d \sim \mathcal{D}}[\nabla\ell(\theta; d)]\|_2 \leq \Delta,$$

and let $\theta^* := \arg \min_{\theta \in \mathcal{C}} \mathbb{E}_{d \sim \mathcal{D}}[\ell(\theta; d)]$ and $\theta^{\text{emp}} := \arg \min_{\theta \in \mathcal{C}} \ell(\theta; D)$. Then for $D \sim \mathcal{D}^n$, with probability $1 - \gamma$, we have:

$$\|\theta^{\text{emp}} - \theta^*\|_2 = O\left(\frac{\Delta}{m} \sqrt{\frac{\log(1/\gamma)}{n}}\right)$$

Proof. Consider a function $\tilde{\ell}$ which has gradient $\nabla\tilde{\ell}(\theta) = \nabla\ell(\Pi_{\mathcal{C}}(\theta)) + m(\theta - \Pi_{\mathcal{C}}(\theta))$. For any D , the empirical minimizer of $\tilde{\ell}$ over \mathbb{R}^p is equal to

$$\tilde{\theta}^{\text{emp}} := \theta^{\text{emp}} - \frac{1}{m} \cdot \nabla\ell(\theta^{\text{emp}}; D),$$

and the population minimizer of $\mathbb{E}_{d \sim \mathcal{D}}[\tilde{\ell}(\theta; d)]$ is

$$\tilde{\theta}^* := \theta^* - \frac{1}{m} \cdot \mathbb{E}_{d \sim \mathcal{D}}[\nabla\ell(\theta^*; d)].$$

By optimality of $\theta^{\text{emp}}, \theta^*$ and convexity of \mathcal{C} ,

$$\langle \nabla\ell(\theta^{\text{emp}}; D), \theta^* - \theta^{\text{emp}} \rangle \geq 0 \quad \text{and} \quad \langle \mathbb{E}_{d \sim \mathcal{D}}[\nabla\ell(\theta^*; D)], \theta^{\text{emp}} - \theta^* \rangle \geq 0,$$

This implies that $\|\tilde{\theta}^{\text{emp}} - \tilde{\theta}^*\|_2 \geq \|\theta^{\text{emp}} - \theta^*\|_2$. In addition, since projection to convex sets is a non-expansive map, $\tilde{\ell}$ is m -strongly convex if ℓ is, and for any d, θ we have

$$\|\nabla\ell(\theta; d) - \mathbb{E}_{d \sim \mathcal{D}}[\nabla\ell(\theta; d)]\|_2 = \|\nabla\tilde{\ell}(\theta; d) - \mathbb{E}_{d \sim \mathcal{D}}[\nabla\tilde{\ell}(\theta; d)]\|_2 \leq \Delta.$$

This holds for any D . Therefore, if we prove the lemma for $\tilde{\ell}$ and \mathbb{R}^p , then this would imply that the lemma holds for ℓ and \mathcal{C} . So it suffices to show the lemma for $\mathcal{C} = \mathbb{R}^p$.

If $\mathcal{C} = \mathbb{R}^p$ then $\mathbb{E}_{d \sim \mathcal{D}}[\nabla\ell(\theta; d)] = \mathbf{0}$. Now, by the assumptions in the lemma and a vector Azuma inequality [Hay03], we have $\|\nabla\ell(\theta^*; D)\|_2 = O\left(\frac{\Delta\sqrt{\log(1/\gamma)}}{\sqrt{n}}\right)$ with probability $1 - \gamma$ over D . Furthermore, we know $\nabla\ell(\theta^{\text{emp}}; D) = \mathbf{0}$ by strong convexity and since $\mathcal{C} = \mathbb{R}^p$. Then by strong convexity, we have

$$\|\theta^* - \theta^{\text{emp}}\|_2 \leq \frac{\|\nabla\ell(\theta^*; D) - \nabla\ell(\theta^{\text{emp}}; D)\|_2}{m} = \frac{\|\nabla\ell(\theta^*; D)\|_2}{m} = O\left(\frac{\Delta}{m} \sqrt{\frac{\log(1/\gamma)}{n}}\right)$$

with probability $1 - \gamma$ as desired in the statement of Lemma 4.5. \square

Given Lemma 4.5, if we want to ensure the *population* minimizer rather than empirical minimizer remains in the sets we choose in Algorithm 1, we just need to choose a slightly larger ball. From this modification and uniform stability, we get our DP-SCO bound:

Proof of Theorem 4.4. Note that by L -Lipschitzness of ℓ in \mathcal{C} , we have

$$\|\nabla\ell(\theta; d) - \mathbb{E}_{d \sim \mathcal{D}}[\nabla\ell(\theta; d)]\|_2 \leq 2L.$$

By Lemma 4.5, Lemma 4.3, and a triangle inequality, we have that the population minimizer of ℓ in \mathcal{C}_i is in \mathcal{C}_{i+1} for each i with probability $1 - 2/n^3$. Then by a union bound, we have that the population minimizer is in \mathcal{C}_{k-1} . When this event fails to hold, our excess population risk is $O(L \|\mathcal{C}_0\|_2)$ and so the contribution

of this event to the expected excess loss is $O\left(\frac{L\|\mathcal{C}_0\|_2 \log \log(\varepsilon n)}{n^3}\right) = O\left(\frac{L^2 \log \log(\varepsilon n)}{mn^3}\right)$, which is asymptotically less than our desired bound. So it suffices to provide the desired expected excess loss bound conditioned on this event. We can bound the radius of \mathcal{C}_{k-1} similarly to the proof of Theorem 4.2, by noting that:

$$\|\mathcal{C}_i\|_2 \leq 2 \cdot \max \left\{ \sqrt{\frac{cL(p+3\log n) \|\mathcal{C}_{i-1}\|_2}{m\varepsilon_i n}}, \frac{cL\sqrt{\log n}}{m\sqrt{n}} \right\}$$

Then, rolling out the recursion, we have similarly to the proof of Theorem 4.2:

$$\|\mathcal{C}_{k-1}\|_2 = O\left(\frac{L(p+\log n)}{m\varepsilon n} + \frac{L\sqrt{\log n}}{m\sqrt{n}}\right).$$

Now, combining Theorem A.11 and the uniform stability bound of Theorem 2.5 (for $\psi = 0$), we get that the expected excess population risk of θ_k compared to the population minimizer over \mathcal{C}_{k-1} is:

$$\begin{aligned} O\left(\frac{Lp\|\mathcal{C}_{k-1}\|_2}{\varepsilon n} + \frac{L^2}{mn}\right) &= O\left(\frac{L^2}{mn} \cdot \left(\frac{(p^2 + \log^2 n)}{\varepsilon^2 n} + \frac{p\sqrt{\log n}}{\varepsilon\sqrt{n}} + 1\right)\right) \\ &= O\left(\frac{L^2(p^2 \log n + \log^2 n)}{m\varepsilon^2 n^2} + \frac{L^2}{mn}\right). \end{aligned}$$

In the final equality, we use the fact that $\frac{p}{\varepsilon} \sqrt{\frac{\log n}{n}} \leq \max\left\{\frac{p^2 \log n}{\varepsilon^2 n}, 1\right\}$. We conclude by noting that conditioned on the event the population minimizer over \mathcal{C}_0 is contained in \mathcal{C}_{k-1} , θ_k has this same excess population risk bound compared to the population minimizer over \mathcal{C}_0 , completing the proof of Theorem 4.4. \square

4.3 Approximate DP and Strongly Convex Losses

The results in Theorem 2.4, 2.7 and 2.6 with $\psi = 0$ combined immediately give that the Gibbs sampler achieves the optimal bounds of $O\left(\frac{L^2 p \log(1/\delta)}{\varepsilon^2 n^2 m}\right)$ and $O\left(\frac{L^2 p \log(1/\delta)}{\varepsilon^2 n^2 m} + \frac{L^2}{mn}\right)$ for excess empirical and population risk respectively in this setting. We note that one could also use Theorem 2.3 instead of Theorem 2.7 and obtain bounds that are within logarithmic factors of the optimal bounds, with a finite-time object.

4.4 Approximate DP and Convex Losses

Similarly to pure-DP, we can adapt our results in the strongly convex setting to the convex setting by adding a regularizer. For a near-optimal empirical guarantee, we use the regularizer $\frac{L\sqrt{p \log(1/\delta)}}{2\|\mathcal{C}\|_2 \varepsilon n} \|\theta - \theta_0\|_2^2$. The regularized and unregularized losses differ by at most $O\left(\frac{L\|\mathcal{C}\|_2 \sqrt{p \log(1/\delta)}}{\varepsilon n}\right)$ everywhere in \mathcal{C} , and the empirical excess loss bound we get by plugging in the strong convexity parameter $\frac{L\sqrt{p \log(1/\delta)}}{\|\mathcal{C}\|_2 \varepsilon n}$ into the bound for strongly convex losses is $O\left(\frac{L\|\mathcal{C}\|_2 \sqrt{p \log(1/\delta)}}{\varepsilon n}\right)$, giving nearly the optimal bound of $O\left(\frac{L\|\mathcal{C}\|_2 \sqrt{p \log(1/\delta)}}{\varepsilon n}\right)$.

For population risk, we use the regularizer $\frac{L}{2\|\mathcal{C}\|_2} \cdot \max\left\{\frac{\sqrt{p \log(1/\delta)}}{\varepsilon n}, \frac{1}{\sqrt{n}}\right\} \|\theta - \theta_0\|_2^2$. By a similar argument, this gives the optimal bound of $O\left(\frac{L\|\mathcal{C}\|_2 \sqrt{p \log(1/\delta)}}{\varepsilon n} + \frac{L\|\mathcal{C}\|_2}{\sqrt{n}}\right)$.

5 Discussion and Future Directions

In this work we demonstrated the power of Langevin diffusion (LD) by simultaneously obtaining tight guarantees for DP-ERM, DP-SCO, and obtaining the first private uniform sampling algorithms from Rashomon sets. Furthermore, we demonstrated that, via SGLD, it is possible to maintain the same privacy/utility trade-offs while allowing the algorithm to be implemented on a finite precision machine. We believe the idea of using a LD to analyze the privacy of the mechanism that samples from the Gibbs distribution, and using a LD to analyze the uniform stability of this mechanism has wider applicability in the DP literature. We leave it for future exploration. Furthermore, the gradient complexity of our SGLD algorithm is inferior to that of [GLL22]. It is an important open question if it is at all possible to close this gap while not relying on the convexity of the loss function as in our case. This would have a significant real world implications where we aim to ensure privacy and also train learning models that are inherently non-convex. Finally, we believe that exploring other applications of Rashomon sets would facilitate wider adoption of machine learning models that are differentially private.

Acknowledgements

We would like to thank Walid Krichene, Dvijotham Krishnamurthy, Ryan McKenna, Sewoong Oh, Adam Smith, and Thomas Steinke for their helpful comments.

References

- [ACG⁺16] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS'16)*, pages 308–318, 2016.
- [AD20a] Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117, 2020.
- [AD20b] Hilal Asi and John C Duchi. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020.
- [ALD21] Hilal Asi, Daniel Asher Nathan Levy, and John Duchi. Adapting to function difficulty and growth conditions in private optimization. In *Advances in Neural Information Processing Systems*, 2021.
- [AT22a] Jason M Altschuler and Kunal Talwar. Concentration of the langevin algorithm’s stationary distribution. *arXiv preprint arXiv:2212.12629*, 2022.
- [AT22b] Jason M Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *arXiv preprint arXiv:2205.13710*, 2022.
- [AT22c] Jason M Altschuler and Kunal Talwar. Resolving the mixing time of the langevin algorithm to its stationary distribution for log-concave sampling. *arXiv preprint arXiv:2210.08448*, 2022.
- [BBM05] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.
- [BE85] Dominique Bakry and Michel Émery. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985.
- [BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- [BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete and Computational Geometry*, 59, 06 2018.
- [BFGT20] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*, 2020.
- [BFTT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pages 11279–11288, 2019.
- [BH18] Thomas Brawner and James Honaker. Bootstrap inference and differential privacy: Standard errors for free. 2018.
- [Bre01] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- [BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*, pages 464–473, 2014.
- [BUV18] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018.

- [CB18] Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 07–09 Apr 2018.
- [CCAY⁺18] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.
- [CEL⁺21] Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of langevin monte carlo from poincaré to log-sobolev. *arXiv preprint arXiv:2112.12662*, 2021.
- [CH22] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, pages 1–18, 2022.
- [CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.
- [CRC21] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pages 2144–2155. PMLR, 2021.
- [CRK21] Beau Coker, Cynthia Rudin, and Gary King. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, 67(10):6174–6197, 2021.
- [CYS21] Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. In *Advances in Neural Information Processing Systems*, 2021.
- [Dal17] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(3):651–676, 2017.
- [DKM⁺06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT*, pages 486–503, 2006.
- [DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.
- [DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.
- [DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pages 265–284, 2006.
- [DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [DRD20] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.
- [EHZ21] Murat A. Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. Convergence of langevin monte carlo in chi-squared and renyi divergence, 2021.

- [FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proc. of the Fifty-Second ACM Symp. on Theory of Computing (STOC'20)*, 2020.
- [FMTT18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *59th Annual IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 521–532, 2018.
- [FRD19] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.
- [GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. *arXiv preprint arXiv:2203.00263*, 2022.
- [GT20] Arun Ganesh and Kunal Talwar. Faster differentially private samplers via rényi divergence analysis of discretized langevin mcmc. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7222–7233. Curran Associates, Inc., 2020.
- [Hay03] Thomas P. Hayes. A large-deviation inequality for vector-valued martingales. 2003.
- [HC22] Hsiang Hsu and Flavio P Calmon. Rashomon capacity: Measuring predictive multiplicity in probabilistic classification. In *NeurIPS*, 2022.
- [HKMN23] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. *STOC*, 2023.
- [HRS16] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, page 1225–1234. JMLR.org, 2016.
- [HS87] Richard Holley and Daniel W. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.
- [HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714, 2010.
- [INS⁺19] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.
- [IPG⁺18] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.
- [KLL21] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth erm and sco in sub-quadratic steps. *Advances in Neural Information Processing Systems*, 34, 2021.
- [KM16] Alexander Kolesnikov and Emanuel Milman. Riemannian metrics on convex sets with applications to poincaré and log-sobolev inequalities. *Calculus of Variations and Partial Differential Equations*, 55, 06 2016.
- [KMN⁺16] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

- [KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1, 2012.
- [Led99] Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.
- [Led01] Michel Ledoux. Logarithmic sobolev inequalities for unbounded spin systems revisited. *Séminaire de Probabilités XXXV*, pages 167–194, 2001.
- [LLRB16] Benjamin Letham, Portia A Letham, Cynthia Rudin, and Edward P Browne. Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063110, 2016.
- [MAD19] David Madras, James Atwood, and Alex D’Amour. Detecting underspecification with local ensembles. *arXiv e-prints*, pages arXiv–1910, 2019.
- [MB10] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- [Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.
- [MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07)*, pages 94–103. IEEE, 2007.
- [MV21] Oren Mangoubi and Nisheeth K Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees. In *Advances in Neural Information Processing Systems*, 2021.
- [MV22] Oren Mangoubi and Nisheeth K Vishnoi. Re-analyze gauss: Bounds for private matrix approximation via dyson brownian motion. *arXiv preprint arXiv:2211.06418*, 2022.
- [NR17] Daniel Nevo and Ya’acov Ritov. Identifying a minimal class of models for high-dimensional data. *The Journal of Machine Learning Research*, 18(1):797–825, 2017.
- [RBP22] Théo Ryffel, Francis Bach, and David Pointcheval. Differential privacy guarantees for stochastic gradient langevin dynamics. *arXiv preprint arXiv:2201.11980*, 2022.
- [RCC⁺22] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.
- [RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1674–1703, 2017.
- [RTT⁺23] Stephan Rabanser, Anvith Thudi, Abhradeep Thakurta, Krishnamurthy Dvijotham, and Nicolas Papernot. Training private models that know what they don’t know. *arXiv preprint arXiv:2305.18393*, 2023.
- [Sch19] André Schlichting. Poincaré and log-sobolev inequalities for mixtures. *Entropy*, 21(1):89, jan 2019.
- [SCS13] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

- [SGM⁺22] Virat Shejwalkar, Arun Ganesh, Rajiv Mathews, Om Thakkar, and Abhradeep Thakurta. Recycling scraps: Improving private learning by leveraging intermediate checkpoints. *arXiv preprint arXiv:2210.01864*, 2022.
- [Smi20] Helen Smith. Algorithmic bias: should students pay the price? *AI & society*, 35(4):1077–1078, 2020.
- [SRP22] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.
- [SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.
- [STT20] Shuang Song, Om Thakkar, and Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783*, 2020.
- [STU17] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.
- [SU15] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *arXiv preprint arXiv:1501.06095*, 2015.
- [SU17] Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 552–563. IEEE, 2017.
- [Tan79] Hiroshi Tanaka. Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Mathematical Journal*, 9(1):163 – 177, 1979.
- [TR13] Theja Tulabandhula and Cynthia Rudin. Machine learning with operational costs. 2013.
- [TR14a] Theja Tulabandhula and Cynthia Rudin. On combining machine learning with decision making. *Machine learning*, 97(1):33–64, 2014.
- [TR14b] Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014.
- [vH14] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014.
- [VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [WCX19] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535. PMLR, 2019.
- [WLK⁺17] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey F. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In Semih Salihoglu, Wenchao Zhou, Rada Chirkova, Jun Yang, and Dan Suciuc, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD*, 2017.
- [WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Notation	
$D = \{d_1, \dots, d_n\}$	data set
\mathcal{D}	data distribution
τ	domain set of data
$\mathcal{C} \subset \mathbb{R}^p$	convex set/parameter space
ℓ	loss function
\mathcal{L}	empirical loss function
Risk_{ERM}	excess empirical risk
Risk_{SCO}	excess population risk
m	strong convexity parameter
M	smoothness parameter
L	Lipschitz constant
θ^{priv}	private model output
θ^*	optimal model
θ^{emp}	model for ERM
β	inverse temperature
W_t	standard Brownian motion
$R_\alpha(\cdot, \cdot)$	Rényi divergence of order α
T	time
$A \succeq 0$	A is positive semidefinite
$A \succeq B$	$A - B$ is positive semidefinite
\mathbb{I}_p	$p \times p$ identity matrix
(ϵ, δ)	Privacy parameters
γ	Empirical loss slack
γ'	Population loss slack
ψ	Threshold of Rashomon set
λ	Sampling error wrt TVD

Table 3: Notation Table

A Notation and Preliminaries

In this section, we give a brief exposition of the concepts and results used in the rest of the paper. In Table 3 we provide a summary of the notation used in the paper.

Rényi divergence and differential privacy. Rényi divergence is the generalization of KL divergence to higher order and satisfies many useful properties [vH14]. More formally,

Definition A.1 (Rényi Divergence). For $0 < \alpha < \infty$, $\alpha \neq 1$ and distributions P, Q , such that $\text{supp}(P) = \text{supp}(Q)$ the α -Rényi divergence between P and Q is

$$R_\alpha(P, Q) = \frac{1}{\alpha - 1} \ln \int_{\text{supp}(Q)} \frac{P(x)^\alpha}{Q(x)^{\alpha-1}} dx = \frac{1}{\alpha - 1} \ln \mathbb{E}_{x \sim Q} \left[\frac{P(x)^\alpha}{Q(x)^\alpha} \right].$$

The α -Rényi divergence for $\alpha = 1$ (resp. ∞) is defined by taking the limit of $R_\alpha(P, Q)$ as α approaches 1 (resp. ∞) and equals the KL divergence (resp. max divergence).

We next define differential privacy, our choice of the notion of data privacy. Central to the notion of differential privacy is the definition of *adjacent* or *neighboring* datasets. Two datasets D and D' are called adjacent if they differ in exactly one data point.

Definition A.2 (Approximate Differential privacy [DMNS06]). A randomized mechanism $\mathcal{M} : \mathcal{D}^n \rightarrow \mathcal{R}$ is said to have (ϵ, δ) -differential privacy, or (ϵ, δ) -DP for short, if for any adjacent $D, D' \in \mathcal{D}^n$ and measurable subset $S \subset \mathcal{R}$, it holds that

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S] + \delta.$$

When $\delta = 0$, it is known as pure differential privacy, and we denote it by ϵ -DP.

Definition A.3 (Rényi Differential privacy [Mir17]). A randomized mechanism $\mathcal{M} : \mathcal{D}^n \rightarrow \mathcal{R}$ is said to have (α, ϵ) -Rényi differential privacy, or (α, ϵ) -RDP for short, if for any adjacent $D, D' \in \mathcal{D}^n$ it holds that

$$R_\alpha(\mathcal{M}(D), \mathcal{M}(D')) \leq \epsilon.$$

It is easy to see that ϵ -DP is merely (∞, ϵ) -RDP. Similarly, the following fact relates (ϵ, δ) -DP to (α, ϵ) -RDP:

Fact A.4 (Proposition 3 in [Mir17]). If \mathcal{M} satisfies (α, ϵ) -RDP, then \mathcal{M} is $(\epsilon + \frac{\log 1/\delta}{\alpha-1}, \delta)$ -differentially private for any $0 < \delta < 1$.

Rényi divergences satisfy a number of other useful properties, which we list here.

Fact A.5 (Monotonicity [vH14, Theorem 3]). For any distributions P, Q and $0 \leq \alpha_1 \leq \alpha_2$ we have $R_{\alpha_1}(P, Q) \leq R_{\alpha_2}(P, Q)$.

Fact A.6 (Post-Processing [vH14, Theorem 9]). For any sample spaces \mathcal{X}, \mathcal{Y} , distributions P, Q over \mathcal{X} , and any function $f : \mathcal{X} \rightarrow \mathcal{Y}$ we have $R_\alpha(f(P), f(Q)) \leq R_\alpha(P, Q)$.

Lemma A.7 (Gaussian dichotomy [vH14, Example 3]). Let $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2 \times \dots$ and $\mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2 \times \dots$, where \mathcal{P}_i and \mathcal{Q}_i are unit variance Gaussian distributions with mean μ_i and ν_i , respectively. Then

$$R_\alpha(\mathcal{P}_i, \mathcal{Q}_i) = \frac{\alpha}{2}(\mu_i - \nu_i)^2,$$

and by additivity for $\alpha > 0$,

$$R_\alpha(\mathcal{P}, \mathcal{Q}) = \frac{\alpha}{2} \sum_{i=1}^{\infty} (\mu_i - \nu_i)^2.$$

As a corollary, we have:

$$R_\alpha(N(0, \sigma^2 \mathbb{I}_p), N(\mathbf{x}, \sigma^2 \mathbb{I}_p)) \leq \frac{\alpha \|\mathbf{x}\|_2^2}{2\sigma^2}.$$

Fact A.8 (Adaptive Composition Theorem [Mir17, Proposition 1]). Let $\mathcal{X}_0, \mathcal{X}_1, \dots, \mathcal{X}_k$ be arbitrary sample spaces. For each $i \in [k]$, let $f_i, f'_i : \Delta(\mathcal{X}_{i-1}) \rightarrow \Delta(\mathcal{X}_i)$ be maps from distributions over \mathcal{X}_{i-1} to distributions over \mathcal{X}_i such that for any distribution X_{i-1} over \mathcal{X}_{i-1} , $R_\alpha(f_i(X_{i-1}), f'_i(X_{i-1})) \leq \epsilon_i$. Then, for $F, F' : \Delta(\mathcal{X}_0) \rightarrow \Delta(\mathcal{X}_k)$ defined as $F(\cdot) = f_k(f_{k-1}(\dots f_1(\cdot) \dots))$ and $F'(\cdot) = f'_k(f'_{k-1}(\dots f'_1(\cdot) \dots))$ we have $R_\alpha(F(X_0), F'(X_0)) \leq \sum_{i=1}^k \epsilon_i$ for any $X_0 \in \Delta(\mathcal{X}_0)$.

Fact A.9 (Weak Triangle Inequality [Mir17, Proposition 11]). For any $\alpha > 1, q > 1$ and distributions $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$ with the same support:

$$R_\alpha(\mathcal{P}_1, \mathcal{P}_3) \leq \frac{\alpha - 1/q}{\alpha - 1} R_{q\alpha}(\mathcal{P}_1, \mathcal{P}_2) + R_{\frac{q\alpha-1}{q-1}}(\mathcal{P}_2, \mathcal{P}_3).$$

We discuss two differentially private mechanisms for optimization in this paper. The first one is the *exponential mechanism*. [MT07]. Given some arbitrary domain \mathcal{D} and range \mathfrak{R} , the exponential mechanism is defined with respect to some loss function, $\ell : \mathcal{D} \times \mathfrak{R} \rightarrow \mathbb{R}$.

Definition A.10 (Exponential mechanism [MT07]). Given a privacy parameter ε , the range \mathfrak{X} and a loss function $\ell : \mathfrak{D} \times \mathfrak{X} \rightarrow \mathbb{R}$, the exponential mechanism samples a single element from \mathfrak{X} based on the probability distribution

$$\pi_D(r) = \frac{e^{-\varepsilon\ell(D,r)/2\Delta_\ell}}{\sum_{r \in \mathfrak{X}} e^{-\varepsilon\ell(D,r)/2\Delta_\ell}}$$

where Δ_ℓ is the sensitivity of u , defined as $\Delta_\ell := \max_{D \sim D', r \in \mathfrak{X}} |u(D,r) - u(D',r)|$. If \mathfrak{X} is continuous, we instead sample from the distribution with pdf:

$$p_D(r) = \frac{e^{-\varepsilon\ell(D,r)/2\Delta_\ell}}{\int_{r \in \mathfrak{X}} e^{-\varepsilon\ell(D,r)/2\Delta_\ell} dr}.$$

Algorithm 2 Exponential mechanism

Input: Loss function \mathcal{L} , constraint set \mathcal{C} , Lipschitz constant L , number of iterations k , privacy parameter ε , data set D of n -samples.

- 1: Sample and **output** a point θ^{priv} from the constraint set \mathcal{C} w.p. $\propto \exp\left(-\frac{\varepsilon n}{2L\|\mathcal{C}\|_2} \cdot \mathcal{L}(\theta; D)\right)$.
-

Theorem A.11. Assume each of the individual loss function in $\mathcal{L}(\theta; D)$ is L -Lipschitz within the constraint set \mathcal{C} , individual loss function $\ell(\theta; \cdot)$ is convex, and the constraint set \mathcal{C} is convex. Then, Algorithm 2 is ε -differentially private. Furthermore, for θ^{priv} as specified in Algorithm 2, over the randomness of the algorithm,

$$\mathbb{E}_{\theta^{\text{priv}}} [\text{Risk}_{\text{ERM}}(\theta^{\text{priv}})] = O\left(\frac{Lp \cdot \|\mathcal{C}\|_2}{\varepsilon n}\right).$$

Equivalence of Algorithm 2 and Langevin diffusion: The following lemma, which is implied by, e.g. [Tan79, Theorem 4.1], shows that one can implement Algorithm 1 using only solutions to eq. (2); note that this does not necessarily mean solutions to eq. (2) are efficiently sampleable.

Lemma A.12. Let \mathcal{L} be a M -smooth function for some finite M . Then if $\beta_t = \beta$ for all t , then the stationary distribution of (2) has pdf proportional to $\exp(-\beta\mathcal{L}(\theta; D)) \cdot \mathbf{1}(\theta \in \mathcal{C})$.

We recall that one can ensure smoothness by convolving \mathcal{L} (appropriately extended to all of \mathbb{R}^p) with the Gaussian kernel of finite variance [FMTT18, Appendix C]. In particular, since we only need M to be finite, we can take the convolution with the Gaussian kernel $\mathcal{N}(\mathbf{0}, \lambda^2 \mathbb{I}_p)$ for arbitrarily small $\lambda > 0$, and in turn the result of the convolution is L/λ -smooth (which is perhaps arbitrarily large but still finite) and differs from \mathcal{L} by an arbitrarily small amount everywhere in \mathcal{C} .

We use the result by [SU17] for our lower bound proof. We use their equivalent result for empirical mean (see equation (2) in [SU17]) and for privacy parameters (ε, δ) using a standard reduction [BUV18, SU15]⁸:

Theorem A.13. Fix $n, s, k \in \mathbb{N}$. Set $\beta = 1 + \frac{1}{2} \log\left(\frac{s}{8 \max\{2k, 28\}}\right)$. Let $P^1, \dots, P^s \sim \text{Beta}(\beta, \beta)$ and let $X := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be such that $\mathbf{x}_i \in \{0, 1\}^s$ for all $i \in [n]$, $\mathbf{x}_{i,j}$ is independent (conditioned on P) and $\mathbf{E}[\mathbf{x}_{i,j}] = P^j$ for all $i \in [n]$ and $j \in [s]$. Let $\mathcal{M} : (\{0, 1\}^s)^n \rightarrow \{0, 1\}^d$ be $(1, \frac{1}{ns})$ -differentially private. Suppose $\|\mathcal{M}(x)\|_1 = \|\mathcal{M}(x)\|_2^2 = k$ for all X with probability 1 and

$$\mathbf{E}_{\mathcal{M}} \left[\frac{1}{n} \sum_{i=1}^n \sum_{\substack{j \in [s] \\ \mathcal{M}(x)^j=1}} \mathbf{x}_{i,j} \right] \geq \frac{1}{n} \max_{\substack{S \subseteq [d] \\ |S|=k}} \sum_{i=1}^n \sum_{u \in S} \mathbf{x}_{i,u} - \frac{k}{20}. \quad (12)$$

Then $n \in \Omega\left(\sqrt{k} \log\left(\frac{s}{k}\right)\right)$.

⁸[SU17] present their result in the terms of population mean and privacy parameters $(1, \frac{1}{ns})$.

Results from statistics and machine learning: We will sometimes use Fatou’s lemma in our proofs. The form we will use is stated here for convenience:

Lemma A.14 (Fatou’s Lemma). *Let $\{X_i\}$ be a sequence of random variables such that there is some constant c such that for all i , $\Pr[X_i \geq c] = 1$. Then:*

$$\mathbb{E} \left[\liminf_{i \rightarrow \infty} X_i \right] \leq \liminf_{i \rightarrow \infty} \mathbb{E} [X_i].$$

For our SCO bounds, we will use uniform stability. Uniform stability of a learning algorithm is a notion of algorithmic stability introduced to derive high-probability bounds on the generalization error. Formally, it is defined as follows:

Definition A.15 (Uniform stability [BE02]). *A mechanism \mathcal{M} is $\mu(n)$ -uniformly stable with respect to ℓ if for any pair of databases D, D' of size n differing in at most one individual:*

$$\sup_{d \in \tau} [\mathbb{E}_{\mathcal{M}} [\ell(\mathcal{M}(D), d)] - \mathbb{E}_{\mathcal{M}} [\ell(\mathcal{M}(D'), d)]] \leq \mu(n).$$

In this paper, we will need the following result.

Lemma A.16 ([BE02]). *Suppose \mathcal{M} is $\mu(n)$ -uniformly stable. Then:*

$$\mathbb{E}_{D \sim \mathcal{D}^n, \mathcal{M}} [\text{Risk}_{\text{SCO}}(\mathcal{M}(D))] \leq \mathbb{E}_{D \sim \mathcal{D}^n, \mathcal{M}} [\text{Risk}_{\text{ERM}}(\mathcal{M}(D))] + \mu(n).$$

We also use KL divergence and TVD, and the relation between the two.

Definition A.17. *The KL divergence (equivalent to the 1-Rényi divergence) between two distributions P, Q is given by $R_1(P, Q) := \mathbb{E}_{x \sim P} \log \left(\frac{P(x)}{Q(x)} \right)$.*

Lemma A.18 (Pinsker’s inequality). *Let $\text{TVD}(P, Q)$ be the total variation distance between P and Q . Then*

$$\text{TVD}(P, Q) \leq \sqrt{\frac{1}{2} R_1(P, Q)}.$$

Next, we define the LSI constant of a distribution.

Definition A.19. *A distribution P satisfies LSI with constant c if for all smooth functions $g : \mathbb{R}^p \rightarrow \mathbb{R}$ with $\mathbb{E}_{x \sim P}[g(x)^2] < \infty$:*

$$\mathbb{E}_{x \sim P}[g(x)^2 \log(g(x)^2)] - \mathbb{E}_{x \sim P}[g(x)^2] \cdot \mathbb{E}_{x \sim P}[\log(g(x)^2)] \leq \frac{2}{c} \mathbb{E}_{x \sim P}[\|\nabla g(x)\|_2^2].$$

B Proofs from Section 2 and Section 3

B.1 Proof of Lemma 2.2

Proof. For ease of presentation, we will show a divergence bound between Θ_t, Θ'_t which are the distributions of θ_t, θ'_t , and then describe how to modify the proof to show the same bound between $\Theta_{[0,t]}, \Theta'_{[0,t]}$.

Let $\Psi_{D,m,i}$ be a map from (distributions over) \mathbb{R}^p to (distributions over) \mathbb{R}^p that takes the point θ to the distribution $\Pi_{\mathcal{C}} \left(N \left(\theta - \left(\frac{\beta t}{m} \right) \nabla \mathcal{L}(\theta; D), 2 \frac{t}{m} \mathbb{I} \right) \right)$, where $\Pi_{\mathcal{C}}$ is the ℓ_2 -projection into \mathcal{C} . It is well known (see e.g. Lemma A.7) that:

$$R_\alpha(N(0, \sigma^2 \mathbb{I}), N(\mathbf{x}, \sigma^2 \mathbb{I})) \leq \frac{\alpha \|\mathbf{x}\|_2^2}{2\sigma^2}.$$

So by post-processing (Fact A.6) and the Lipschitzness assumption, $R_\alpha(\Psi_{D,m,i}(\theta), \Psi_{D',m,i}(\theta))$ is bounded by

$$\begin{aligned} & R_\alpha \left(N \left(\theta - \left(\frac{\beta t}{m} \right) \nabla \mathcal{L}(\theta; D), \frac{2t}{m} \mathbb{I} \right), N \left(\theta - \left(\frac{\beta t}{m} \right) \nabla \mathcal{L}(\theta; D'), \frac{2t}{m} \mathbb{I} \right) \right) \\ &= R_\alpha \left(N \left(\mathbf{0}, \frac{2t}{m} \mathbb{I} \right), N \left(\left(\frac{\beta t}{m} \right) (\nabla \mathcal{L}(\theta; D) - \nabla \mathcal{L}(\theta; D')), 2 \frac{t}{m} \mathbb{I} \right) \right) \\ &\leq \frac{\alpha \Delta^2}{4} \cdot \frac{\left(\frac{\beta t}{m} \right)^2}{t/m}. \end{aligned}$$

Let $\Psi_{D,m}$ denote the composition $\Psi_{D,m,m} \circ \Psi_{D,m,m-1} \circ \dots \circ \Psi_{D,m,1}$. By Fact A.8, we have

$$R_\alpha(\Psi_{D,m}(\Theta_0), \Psi_{D',m}(\Theta_0)) \leq \sum_{i=1}^m \max_{\theta} \{ R_\alpha(\Psi_{D,m,i}(\theta), \Psi_{D',m,i}(\theta)) \}.$$

Plugging in the bound on $R_\alpha(\Psi_{D,m,i}(\theta), \Psi_{D',m,i}(\theta))$, we get

$$R_\alpha(\Psi_{D,m}(\Theta_0), \Psi_{D',m}(\Theta_0)) \leq \frac{\alpha \Delta^2}{4} \cdot \frac{m}{t} \sum_{i=1}^m \left(\frac{\beta t}{m} \right)^2 = \frac{\alpha \beta^2 \Delta^2 t}{4}$$

Note that $\Theta_t = \lim_{m \rightarrow \infty} \Psi_{D,m}(\Theta_0)$, and $\Theta'_t = \lim_{m \rightarrow \infty} \Psi_{D',m}(\Theta_0)$. Since $\exp((\alpha - 1)R_\alpha(\mathcal{P}, \mathcal{Q}))$ is a monotone function of $R_\alpha(\mathcal{P}, \mathcal{Q})$ and is the expectation of a positive random variable, by Fatou's lemma we have:

$$\begin{aligned} R_\alpha(\Theta_t, \Theta'_t) &\leq \lim_{m \rightarrow \infty} R_\alpha(\Psi_{D,m}(\Theta_0), \Psi_{D',m}(\Theta_0)) \\ &\leq \frac{\alpha \beta^2 \Delta^2 t}{4}. \end{aligned}$$

This gives the bound on $R_\alpha(\Theta_t, \Theta'_t)$. To obtain the same bound for $R_\alpha(\Theta_{[0,t]}, \Theta'_{[0,t]})$, we modify $\Psi_{D,m,i}$ so that instead of receiving $\Theta_{(i-1)t/m}$ and outputting $\Theta_{it/m}$, it receives the joint distribution $\{\Theta_{jt/m}\}_{0 \leq j \leq i-1}$ and outputs $\{\Theta_{jt/m}\}_{0 \leq j \leq i}$ by appending the (also jointly distributed) variable

$$\Theta_{it/m} = \Pi_{\mathcal{C}} \left(N \left(\theta - \left(\frac{\beta t}{m} \right) \nabla \mathcal{L}(\Theta_{(i-1)t/m}; D), 2 \frac{t}{m} \mathbb{I} \right) \right)$$

That is, we update $\Psi_{D,m,i}$ so it outputs the distributions of all iterates seen so far instead of just the distribution of the last iterate; the limiting value of the joint distribution $\{\Theta_{jt/m}\}_{0 \leq j \leq i}$ is then $\Theta_{[0,t]}$ according to eq. (2), and the same divergence bound holds. \square

B.2 Proof of Theorem 2.3

We first need the following results, which let us analyze the LSI constant of the Gibbs distribution of interest easily. These results were originally stated for unconstrained domains defined over the space of real numbers, a proof for which can also be found in [Led01] using the theory of semigroup (see Corollary 1.4, 1.6 and Lemma 1.2). For the general convex set, \mathcal{C} , we refer the readers to Theorem 2.1 in [KM16].

Lemma B.1 (Proposition 3 and Corollaire 2 in [BE85]). *Let P be the distribution with density proportional to $\exp(-f(x)) \cdot \mathbb{1}(x \in \mathcal{C})$ for convex \mathcal{C} . If f is m -strongly convex, then P satisfies LSI with constant m .*

Lemma B.2 (Page 1184 in [HS87]). *Let f, f' be two functions such that*

$$\sup_{\theta \in \mathcal{C}} |f(\theta) - f'(\theta)| \leq \Delta.$$

Suppose the distribution with density proportional to $\mathbb{1}(\theta \in \mathcal{C}) \cdot \exp(-f)$ satisfies LSI with constant c . Then the distribution with density proportional to $\mathbb{1}(\theta \in \mathcal{C}) \cdot \exp(-f')$ satisfies LSI with constant $c \cdot \exp(-\Delta)$.

The following result shows convergence of (2) under LSI.

Lemma B.3 (Theorem 4 of [VW19]). *Suppose Q satisfies LSI with constant c . Then let P_0 be any initial distribution over θ_0 , and P_t be the distribution over θ_t given by running (1) on $-\log q$, where q is the density of Q , for $\beta = 1$. Then $R_1(P_t, Q) \leq \exp(-2ct) \cdot R_1(P_0, Q)$.*

Using these results, we first prove an LSI constant for the Gibbs distribution.

Lemma B.4. *Suppose we have non-negative $\ell(\theta; d)$ such that ℓ is m -strongly convex wrt θ for all d , and let Q be the distribution with density proportional to*

$$\mathbb{1}(\theta \in \mathcal{C}) \cdot \exp\left(-\beta \max\left\{\psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D), \mathcal{L}(\theta; D)\right\}\right).$$

Then Q satisfies LSI with constant c , where

$$c := \begin{cases} \frac{m}{e^\psi} & \text{for } \beta\psi > 1 \\ \beta m \exp(-\beta\psi) & \text{for } \beta\psi \leq 1 \end{cases}.$$

Proof. For $a \in [0, 1]$, consider the distribution with density proportional to $e^{-f(\theta)} \cdot \mathbb{1}(\theta \in \mathcal{C})$ for

$$f(\theta) := \beta \left(a \cdot \mathcal{L}(\theta; D) + (1 - a) \cdot \max\left\{\psi + \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D), \mathcal{L}(\theta; D)\right\} \right). \quad (13)$$

Since \mathcal{L} is m -strongly convex, the function $\max\{\psi + \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D), \mathcal{L}(\theta; D)\}$ is a convex function. Therefore, $f(\theta)$ defined in eq. (13) is an $(a\beta m)$ -strongly convex function. In other words, this distribution satisfies LSI with constant $a\beta m$ using Lemma B.1.

Now the distribution Q has density proportional to $\exp(-f)$, where

$$f' := \max\{\psi + \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D), \mathcal{L}(\theta; D)\}.$$

When $\mathcal{L}(\theta; D) > \psi$, $f = f'$. Otherwise, since ℓ is non-negative, we have that f differs from f' by at most $a\beta\psi$ everywhere. Then by lemma B.2, Q satisfies LSI with constant $a\beta m \exp(-a\beta\psi)$.

Now, we maximize this bound over $a \in [0, 1]$. If $\beta\psi \leq 1$, then this bound is maximized at $a = 1$ and we get an LSI constant of $\beta m \exp(-\beta\psi)$. Otherwise, this bound is maximized at $a = 1/\beta\psi$ and we get an LSI constant of $\frac{m}{e^\psi}$. This completes the proof of Lemma B.4. \square

Given the LSI constant of Rashomon sampler, we can use Lemma B.3 to upper bound the time we need to run DP-LD.

Lemma B.5. *Under the assumptions of Theorem 2.3, let*

$$\mathcal{L}'(\theta; D) := \max\left\{\psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D), \mathcal{L}(\theta; D)\right\}.$$

Let Q be the distribution with density proportional to $\exp(-\beta\mathcal{L}')$. Let P_0 be the distribution over θ_0 that is uniform over \mathcal{C} . Let P_t be the resulting distribution over θ_t given by running (2) on \mathcal{L}' . Then assuming $\psi \leq L \|\mathcal{C}\|_2$:

- If $\beta\psi > 1$, for $t = \frac{e\psi}{2m} \log\left(\frac{2\beta L\|\mathcal{C}\|_2}{\delta^2}\right)$, we have $\text{TVD}(P_t, Q) \leq \delta/2$.
- If $\beta\psi \leq 1$, for $t = \frac{e}{2\beta m} \log\left(\frac{2\beta L\|\mathcal{C}\|_2}{\delta^2}\right)$, we have $\text{TVD}(P_t, Q) \leq \delta/2$.

Proof. We wish to apply Lemma B.3, which was originally stated in the unconstrained case for (1). One can see that it applies to running (2) on \mathcal{L}' by the following argument: Consider extending \mathcal{L}' to \mathbb{R}^p by defining $\mathcal{L}'(\theta') = \mathcal{L}'(\Pi_{\mathcal{C}}(\theta')) + c\|\theta' - \Pi_{\mathcal{C}}(\theta')\|_2$ for $\theta' \notin \mathcal{C}$. As c goes to infinity, the Gibbs distribution induced by the extended loss in the unconstrained setting approaches the Gibbs distribution induced by the loss in the constrained setting, and (1) approaches (2).

We have:

$$\begin{aligned}
R_1(P_0, Q) &\leq \max_{\theta \in \text{supp}(P_0)} \log(P_0(\theta)/Q(\theta)) \\
&\leq \log\left(\frac{1}{\int_{\theta \in \mathcal{C}} 1 d\theta} \cdot \frac{\int_{\theta \in \mathcal{C}} \exp(-\beta \max\{\mathcal{L}(\theta^*; D) + \psi, \mathcal{L}(\theta; D)\}) d\theta}{\min_{\theta \in \mathcal{C}} \exp(-\beta \max\{\mathcal{L}(\theta^*; D) + \psi, \mathcal{L}(\theta; D)\})}\right) \\
&\leq \log\left(\frac{\exp(-\beta(\mathcal{L}(\theta^*; D) + \psi))}{\exp(-\beta(L\|\mathcal{C}\|_2 + \psi))}\right) \\
&\leq 2\beta L\|\mathcal{C}\|_2.
\end{aligned} \tag{14}$$

Here, we use the fact that each ℓ has minimum 0. Recall that

$$\mathcal{L}'(\theta; D) := \max\left\{\psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D), \mathcal{L}(\theta; D)\right\}.$$

Now running (2) on $\beta\mathcal{L}'$, replacing β in (2) with 1, is equivalent to running (2) on \mathcal{L}' . Therefore, applying Lemma B.3 and Lemma B.4:

- If $\beta\psi > 1$, for $t = \frac{e\psi}{2m} \log\left(\frac{4\beta L\|\mathcal{C}\|_2}{\delta^2}\right)$, we have:

$$R_1(P_t, Q) \leq R_1(P_0, Q) \cdot \exp\left(-\frac{2m}{e\psi}t\right) \leq 2\beta L\|\mathcal{C}\|_2 \cdot \exp\left(-\frac{2m}{e\psi}t\right) = \delta^2/2,$$

where the first inequality is using Lemma B.3 and Lemma B.4 and the second inequality is due to eq. (14). The first bullet now follows using Pinsker's inequality (Lemma A.18).

- If $\beta\psi \leq 1$, note that $\beta m \exp(-\beta\psi) \geq \beta m/e$. So for $t = \frac{e}{2\beta m} \log\left(\frac{4\beta L\|\mathcal{C}\|_2}{\delta^2}\right)$, we have:

$$R_1(P_t, Q) \leq R_1(P_0, Q) \cdot \exp\left(-\frac{2m}{e\psi}t\right) \leq 2\beta L\|\mathcal{C}\|_2 \cdot \exp\left(-\frac{2\beta m}{e}t\right) = \delta^2/2,$$

where the first inequality is using Lemma B.3 and Lemma B.4 and the second inequality is due to eq. (14). The second bullet now follows using Pinsker's inequality (Lemma A.18).

This completes the proof of Lemma B.5. \square

In Lemma B.5, we used a uniform distribution on \mathcal{C} as our initialization (i.e., P_0 was uniform distribution over \mathcal{C}). In the case where \mathcal{C} is *unconstrained* and the losses satisfy $\|\nabla \ell(\theta; d) - \nabla \ell(\theta; d')\|_2 \leq L$ instead of Lipschitzness within \mathcal{C} , we can obtain the same bound by letting \mathcal{C} be the convex hull of the minimizers of all per-example loss functions $\ell(\theta; d)$ and using a uniform distribution on \mathcal{C} as our initial distribution.

We now complete the proof of Theorem 2.3.

Proof. Let $\mathbf{v}_D(\theta) = \partial_{\theta} \max\{\mathcal{L}(\theta; D), \mathcal{L}(\theta^*; D) + \psi\}$ be the sub-differential of the score function used in the Gibbs distribution. Using the Lipschitz property and the smoothness assumption on $\ell(\theta; d)$, for any two neighboring data sets D and D' , we have

$$\|\mathbf{v}_D(\theta) - \mathbf{v}_D(\theta^*)\|_2 \leq 3 \cdot \max \left\{ \frac{L}{n}, \sqrt{\frac{M\psi}{2}} \right\}. \quad (15)$$

In particular, the Lipschitzness bounds the sensitivity between the gradients of $\mathcal{L}(\theta; D)$ and $\mathcal{L}(\theta; D')$ by L/n . Using smoothness, for D , the sensitivity between the gradients of $\mathcal{L}(\theta; D)$ and

$\max \left\{ \psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D), \mathcal{L}(\theta; D) \right\}$ is bounded by $\sqrt{M\psi/2}$. Similarly, we have the sensitivity between the gradients of $\mathcal{L}(\theta; D')$ and $\max \left\{ \psi + \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D'), \mathcal{L}(\theta; D') \right\}$ bounded by $\sqrt{M\psi/2}$.

By Lemma 2.2, Fact A.4, and sensitivity bound in eq. (15), in order for $\{\Theta_{t'}\}_{t' \in [0, t]}$ to satisfy (ϵ, δ) -DP, it suffices if

$$\beta \leq \frac{\epsilon}{6\sqrt{2} \cdot \max \left\{ \frac{L}{n}, \sqrt{\frac{M\psi}{2}} \right\} \sqrt{t \log(1/\delta)}}. \quad (16)$$

Suppose $\beta\psi \leq 1$. Recall that, in Lemma B.5, $t = \frac{\epsilon}{2\beta m} \log\left(\frac{4\beta L \|\mathcal{C}\|_2}{\delta^2}\right)$. Plugging in this value of t in eq. (16) and observing that $\psi \leq \frac{L\|\mathcal{C}\|_2}{2}$, it suffices to ensure (ϵ, δ) -DP if

$$\beta = \tilde{O} \left(\frac{\epsilon^2 n^2 m}{\max \{L, n\sqrt{M\psi/2}\}^2 \log(L \|\mathcal{C}\|_2 / \delta^2) \log(1/\delta)} \right). \quad (17)$$

We now consider two cases (on top of the constraint that $\psi \leq L\|\mathcal{C}\|_2/2$):

- (i) $\psi > \frac{2L^2}{Mn^2}$, and
- (ii) $\psi \leq \frac{2L^2}{Mn^2}$.

In the case (i), in order to satisfy $\beta\psi \leq 1$ it is sufficient to set $\beta = \tilde{O} \left(\frac{\epsilon^2 (m/M)}{\log((L\|\mathcal{C}\|_2 - \psi)/\delta^2) \log(1/\delta)} \cdot \frac{1}{\psi} \right)$. In the second, case it sufficient to set $\beta = \tilde{O} \left(\frac{\epsilon^2 n^2 m}{L^2 \log((L\|\mathcal{C}\|_2 - \psi)/\delta^2) \log(1/\delta)} \right)$.

We will not analyze the setting when $\beta\psi > 1$. From Lemma B.5, it is not hard to observe that it will not provide any better conditions on ψ and β to ensure (ϵ, δ) -DP.

So, we have that for the choices of ψ, β, t in Theorem 2.3, outputting $\{\Theta_{t'}\}_{t' \in [0, t]}$ satisfies (ϵ, δ) -DP. Furthermore, by Lemma B.5, this gives the desired bound on total variation distance completing the proof of Theorem 2.3. \square

B.3 Probability of Hitting the Rashomon Set

Theorem B.6. *Let θ^{priv} be the model output by a Rashomon sampler. Let $\Pr[\mathcal{L}(\theta^{\text{priv}}; D) \leq \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) + \psi + \gamma] \geq 1 - \phi$. Then for Rashomon set \mathcal{G} , $\Pr[\theta^{\text{priv}} \in \mathcal{G}] \geq 1 - \phi - \frac{p\gamma}{\psi}$.*

Proof. Consider the differential conic region Ω centered at the true minimizer $\theta^* = \arg \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$. Let

A be the cone from this region with height r , and B be the one with height $(r + \Delta)$, with r and Δ chosen such that θ in the boundaries of the cones A and B satisfy $\mathcal{L}(\theta; D) = \psi + \min_{\theta} \mathcal{L}(\theta; D)$ and $\mathcal{L}(\theta; D) = \psi + \gamma + \min_{\theta} \mathcal{L}(\theta; D)$ respectively. By law of total probability, it suffices to show that conditioned on sampling θ in any such cone, the desired probability bound holds. By the maximality condition on the Rashomon sampler,

$$\Pr[\theta^{\text{priv}} \in \mathcal{G} | \theta \in \Omega] \geq 1 - \phi - \frac{\text{Vol}(A)}{\text{Vol}(B)} = 1 - \phi - \left(1 - \frac{\Delta}{r + \Delta}\right)^p \geq 1 - \phi - \frac{p\Delta}{r}.$$

Note that within Ω , we have $\mathcal{L}(\theta; D) = f(\|\theta - \theta^*\|_2)$ where f is a convex function. By convexity, we have:

$$f(r) - f(0) \leq f'(r) \cdot r \rightarrow f'(r) \geq \frac{f(r) - f(0)}{r} = \psi/r.$$

$$\gamma = f(r + \Delta) - f(r) \geq f'(r) \cdot \Delta \geq \frac{\psi\Delta}{r} \implies \Delta/r \leq \gamma/\psi.$$

This completes the proof of Theorem B.6. \square

B.4 Proof of Theorem 3.2

We start by showing some properties of the convolved loss function:

Theorem B.7. *Let \mathcal{L} be a convex and M -smooth loss, let $\mathcal{L}'(\theta) = \max\{\mathcal{L}(\theta), \mathcal{L}(\theta^*) + \psi\}$ and let $\tilde{\mathcal{L}}'(\theta) := \mathbb{E}_{\xi \sim N(0, \lambda^2 \mathbb{I}_p)} [\mathcal{L}'(\theta + \xi)]$. Then:*

1. $\max_{\theta} |\tilde{\mathcal{L}}'(\theta) - \mathcal{L}'(\theta)| \leq 2\psi + Mp\lambda^2$.
2. $\tilde{\mathcal{L}}'(\theta)$ is $M + \frac{\sqrt{\psi M/2}}{\lambda}$ -smooth.

Proof. Let $\tilde{\mathcal{L}}(\theta) := \mathbb{E}_{\xi \sim N(0, \lambda^2 \mathbb{I}_p)} [\mathcal{L}(\theta + \xi)]$. Then:

$$\max_{\theta} |\tilde{\mathcal{L}}'(\theta) - \mathcal{L}'(\theta)| \leq 2\psi + \max_{\theta} |\tilde{\mathcal{L}}(\theta) - \mathcal{L}(\theta)|.$$

We bound the second term for all θ simultaneously as follows:

$$\begin{aligned} |\tilde{\mathcal{L}}(\theta) - \mathcal{L}(\theta)| &= \left| \mathbb{E}_{\xi \sim N(0, \lambda^2 \mathbb{I}_p)} [\mathcal{L}(\theta + \xi) - \mathcal{L}(\theta)] \right| \\ &\stackrel{(*)}{=} \left| \mathbb{E}_{\xi \sim N(0, \lambda^2 \mathbb{I}_p)} [\mathcal{L}(\theta + \xi) - \mathcal{L}(\theta) - \langle \nabla \mathcal{L}(\theta), \xi \rangle] \right| \\ &\stackrel{(**)}{\leq} \mathbb{E}_{\xi \sim N(0, \lambda^2 \mathbb{I}_p)} [M \|\xi\|_2^2] = Mp\lambda^2. \end{aligned}$$

In equality (*), we use the fact that ξ is mean 0, and in inequality (**), we use the fact that $0 \leq \mathcal{L}(\theta + \xi) - \mathcal{L}(\theta) - \langle \nabla \mathcal{L}(\theta), \xi \rangle \leq M \|\xi\|_2^2$ by convexity and M -smoothness of \mathcal{L} . This gives the first part of the theorem.

For the second part, fix any θ_1, θ_2 . The smoothness parameter of $\tilde{\mathcal{L}}'$ is at most the smoothness parameter of $\tilde{\mathcal{L}}$ (which is at most M , since \mathcal{L} is M -smooth), plus the smoothness of $\tilde{\mathcal{L}}' - \tilde{\mathcal{L}}$.

Now $\tilde{\mathcal{L}}' - \tilde{\mathcal{L}}$ is the convolution of the Gaussian kernel with the function $\mathcal{L}' - \mathcal{L}$, which is 0 outside of Rashomon set, \mathcal{G} and equal to $\mathcal{L}(\theta^*) + \psi - \mathcal{L}(\theta)$ inside the Rashomon set, \mathcal{G} . By M -smoothness of \mathcal{L} , $\mathcal{L}' - \mathcal{L}$ is globally $\sqrt{\frac{\psi M}{2}}$ -Lipschitz. Then, by Theorem 33 of [FMTT18], $\tilde{\mathcal{L}}' - \tilde{\mathcal{L}}$ is $\frac{1}{\lambda} \sqrt{\frac{\psi M}{2}}$ -smooth, giving the second part of the theorem.

This completes the proof of Theorem B.7. \square

Similarly to Lemma B.4 we have the following:

Lemma B.8. *Let $\tilde{\mathcal{L}}'$ be defined as in Theorem B.7, for some $\lambda \geq 0$. Then the distribution with density proportional to $\exp(-\beta \cdot \tilde{\mathcal{L}}')$ satisfies LSI with constant c for*

$$c := \beta m \exp(-\beta(3\psi + Mp\lambda^2)).$$

Recall the SGLD equation:

$$\theta_{t+1} = \theta_t - \eta\beta\nabla\mathcal{L}(\theta_t; D) + \xi_t, \quad \xi_t \sim \mathcal{N}(0, 2\eta I_p). \quad (18)$$

We can now apply the following result of [CEL⁺21]:

Theorem B.9 (Theorem 4 of [CEL⁺21]). *For any T and η , let Θ_T be the distribution of $\theta_{T\eta}$ given by the solution to (1), and let Θ'_T be the distribution of θ_T given by (18). Suppose \mathcal{L} is M -smooth and satisfies LSI (Definition A.19) with constant c . Then for any $\kappa > 0, \alpha \geq 2$,*

$$\eta = O\left(\frac{c\kappa}{p\alpha M^2} \cdot \min\left\{\frac{1}{\log(\alpha)}, \frac{p}{\alpha\kappa}\right\}\right), \quad \text{and} \quad T = \Omega\left(\frac{p\alpha^2 M^2}{c^2\kappa} r \cdot \max\left\{\log(\alpha), \frac{\alpha\kappa}{p}\right\}\right),$$

we have $R_\alpha(\Theta'_T, \Theta_\infty) \leq \kappa$.

We can now prove a general result in terms of Rényi divergence:

Theorem B.10. *Suppose $\|\nabla\mathcal{L}(\theta; D) - \nabla\mathcal{L}(\theta; D')\|_2 \leq \Delta$, and that each individual loss function ℓ is m -strongly convex and M -smooth. Let $0 \leq \lambda \leq \sqrt{\frac{\psi}{Mp}}$ and let*

$$\tilde{\mathcal{L}} := \mathbb{E}_{\xi \sim N(0, \lambda^2 I_p)} \left[\min\{\mathcal{L}(\theta + \xi; D), \min_{\theta^* \in \mathcal{C}} \mathcal{L}(\theta^*; D) + \psi\} \right].$$

Let Q be the (unconstrained) Gibbs distribution of $\beta \cdot \tilde{\mathcal{L}}$. Let Θ'_T be the solution to (18) starting from the distribution Θ'_0 , run on the loss $\tilde{\mathcal{L}}$. Fix any $r > 0$. Then for

$$\beta = O\left(\frac{\varepsilon^2 m}{\max\{\Delta^2, M\psi\} \alpha \log(R_\alpha(\Theta'_0, Q)/\kappa) \log(1/\delta)}\right),$$

$$T = \Omega\left(\frac{p\alpha^4 \log^3(R_\alpha(\Theta'_0, Q)/\kappa) (M^2 + \frac{M\psi}{\lambda^2}) \max\{\Delta^4, M^2\psi^2\} \log^2(1/\delta)}{\varepsilon^4 m^4 \kappa} \cdot \max\left\{\log(\alpha), \frac{\alpha\kappa}{p}\right\}\right)$$

and an appropriate choice of η , outputting a sample from Θ'_T is (ε, δ) -DP. Furthermore,

$$R_\alpha(\Theta'_T, Q) \leq \kappa.$$

Proof. In Theorem B.9 we need:

$$T\eta = \Theta\left(\frac{\alpha}{c} \log\left(\frac{R_\alpha(\Theta'_0, Q)}{\kappa}\right)\right)$$

Plugging in Lemmas 3.1 and B.8 for outputting Θ'_T from (18) to be (ε, δ) -DP it suffices if:

$$\beta = O\left(\frac{\varepsilon^2 m \exp(-\beta(3\psi + Mp\lambda^2))}{\max\{\Delta^2, M\psi\} \alpha \log(R_\alpha(\Theta'_0, Q)/\kappa) \log(1/\delta)}\right).$$

If we assume $\lambda \leq \sqrt{\frac{\psi}{Mp}}$, for $\varepsilon \leq 1$ this condition implies $\beta = O(\frac{1}{\psi + Mp\lambda^2})$. So this can be simplified to:

$$\beta = O\left(\frac{\varepsilon^2 m}{\max\{\Delta^2, M\psi\} \alpha \log(R_\alpha(\Theta'_0, Q)/\kappa) \log(1/\delta)}\right).$$

Now we can plug in this value of β , the smoothness bound from Theorem B.7, Lemma B.8, and our assumption on λ into the number of iterations T to get an iteration complexity requirement of:

$$T = \Omega\left(\frac{p\alpha^4 (M^2 + \frac{M\psi}{\lambda^2}) \max\{\Delta^4, M^2\psi^2\} \log^2(1/\delta)}{\varepsilon^4 m^4 \kappa} \log^3(R_\alpha(\Theta'_0, Q)/\kappa) \max\{\log(\alpha), \frac{\alpha\kappa}{p}\}\right).$$

This completes the proof of Theorem B.10. \square

In order to bound the initial divergence, we use Θ'_0 that is a normal distribution centered at a point in the convex hull of minimizers of ℓ , i.e. the convex hull of $\{\arg \min_{\theta} \ell(\theta; d) : d \in \tau\}$.

Lemma B.11. *Suppose $\|\nabla \mathcal{L}(\theta; D) - \nabla \mathcal{L}(\theta; D')\|_2 \leq \Delta$, and let θ_0 be an arbitrary point in the convex hull of $\{\arg \min_{\theta} \ell(\theta; d) | d \in \tau\}$. Let P be the Gibbs distribution on $\beta \cdot \tilde{\mathcal{L}}'$ as defined in Theorem B.7. Then for $\alpha \geq 2$:*

$$R_{\alpha} \left(N \left(\theta_0, \frac{1}{\beta M} \mathbb{I}_p \right), P \right) = O \left(\frac{\alpha \Delta \beta^2 M^2}{m^2} + p \ln \left(\frac{M}{m} \right) + \beta \psi \right).$$

Proof. By e.g. Lemma 16 in [GT20], if θ^* is the true minimizer of \mathcal{L} and P_1 is the Gibbs distribution on $\beta \cdot \mathcal{L}$, for all $\alpha \geq 1$:

$$R_{\alpha} \left(N \left(\theta^*, \frac{1}{\beta M} \mathbb{I}_p \right), P_1 \right) \leq \frac{p \ln \left(\frac{M}{m} \right)}{2}.$$

By Theorem B.7, $\beta \cdot \mathcal{L}$ and $\beta \cdot \tilde{\mathcal{L}}'$ differ by at most $4\beta\psi$ everywhere if $\lambda \leq \sqrt{\frac{\psi}{Mp}}$. Then:

$$R_{\infty}(P_1, P) \leq 8\beta\psi.$$

So by the approximate triangle inequality for Rényi divergences (Fact A.9):

$$R_{\alpha} \left(N \left(\theta^*, \frac{1}{\beta M} \mathbb{I}_p \right), P \right) \leq \frac{p \ln \left(\frac{M}{m} \right)}{2} + 8\beta\psi$$

Finally, by the assumption on $\nabla \mathcal{L}$, all points in $\{\arg \min_{\theta} \ell(\theta; d) | d \in \tau\}$ are distance at most Δ/m apart by strong convexity. Then by Lemma A.7:

$$R_{\alpha} \left(N \left(\theta_0, \frac{1}{\beta M} \mathbb{I}_p \right), N \left(\theta^*, \frac{1}{\beta M} \mathbb{I}_p \right) \right) \leq \frac{\alpha \Delta \beta^2 M^2}{2m^2}$$

Applying Fact A.9 again gives Lemma B.11. □

Now by Pinsker's inequality and monotonicity of Rényi divergences, we can use Lemma B.11, $\alpha = 2$, and $\kappa = \sqrt{\delta/2}$ to get a total variation distance bound of δ to the stationary distribution in Theorem B.10, proving Theorem 3.2.

C Lower Bound on DP-ERM for Non-Convex Losses

In this section, we show the following lower bound on the excess empirical risk for 1-Lipschitz non-convex loss functions. The lower bound implies that there is no advantage, in terms of the dependence on dimensions (p), to move from ε -DP to (ε, δ) -DP.

Theorem C.1. *Let $\varepsilon \leq 1$, $2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, and $B(\mathbf{0}, 1)$ be a unit Euclidean ball centered at origin. Then there exists 1-Lipschitz non-convex function $\mathcal{L} : B(\mathbf{0}, 1) \times \mathcal{X} \rightarrow \mathbb{R}$ and a dataset⁹ $D = \{d_1, \dots, d_n\}$ such that for every $p \in \mathbb{N}$, there is no (ε, δ) -differentially private algorithm \mathcal{A} that outputs θ^{priv} such that*

$$\mathbb{E} \left[\mathcal{L}(\theta^{\text{priv}}; D) - \min_{\theta \in B_p(1)} \mathcal{L}(\theta; D) \right] = o \left(\frac{p \log(1/\delta)}{n\varepsilon} \right), \quad (19)$$

⁹The dataset, $D = \{d_1, \dots, d_n\}$ is such that $d_i \in \{0, 1\}^s$ for all $i \in [n]$, $d_{i,j}$ is independent (conditioned on P) and $\mathbb{E}[d_{i,j}] = P^j$ for all $i \in [n]$ and $j \in [s]$. Here \mathcal{P} is the distribution that is defined in Theorem A.13.

Proof. We first perform two translations of Theorem A.13: first from $(1, \frac{1}{ns})$ to (ε, δ) from [SU15] and then from sample complexity to a result stated in the terms of accuracy bound. A direct corollary of Theorem A.13 with $k = 1$ is as follows: for every $s \in \mathbb{N}$, no (ε, δ) -differentially private algorithm on input X satisfying the premise of Theorem A.13 outputs an index $j \in [s]$ such that

$$\mathbf{E}_{\mathcal{M}} \left[\frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i,j} \right] - \max_{u \in [s]} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i,u} = o \left(\frac{1}{n\varepsilon} \log(s) \log(1/\delta) \right), \quad (20)$$

where $\varepsilon \leq 1$ and $2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$.

Using this lower bound on top-selection, we give our lower bound by defining an appropriate non-convex loss function. In particular, we define a packing over the p -dimensional Euclidean ball such that there is an bijective mapping between the centers of the packing and $[s]$. Then the function attains the minimum at the center of packing which corresponds to the coordinate $j \in [s]$ with maximum frequency. Since the size of the α -net is $\approx 1/\alpha^p$ and there is a bijective mapping, this gives a lower bound using eq. (20).

Let $B(\mathbf{0}, 1)$ be the p -dimensional Euclidean ball centered at origin and let $\alpha \in (0, 1/2)$ be a constant. Consider an α -packing with centers $C = \{\mathbf{c}_1, \mathbf{c}_2, \dots\}$. It is known that the size of such packing, $N(\alpha)$ is $(\frac{1}{\alpha})^p \leq N(\alpha) \leq (\frac{3}{\alpha})^p$. Let $s = N(\alpha)$. Further, let $f : B(\mathbf{0}, 1) \rightarrow \{1, \dots, s\}$ be an injective function defined as follows:

$$f(\theta) = \left\{ j : \mathbf{c}_j = \arg \min_{\mathbf{c} \in C} \|\theta - \mathbf{c}\|_2 \right\}.$$

In particular, f is the function that maps a point on the unit ball to its closest point in C .

We now define our loss function as follows:

$$\mathcal{L}(\theta; D) := \frac{1}{n} \sum_{d_i \in D} \ell(\theta; d_i) \text{ where } \ell(\theta; d_i) = \min_{\mathbf{c}_j \in C} \left(\frac{\|\theta - \mathbf{c}_j\|}{\alpha} - 1 \right) d_{i,j}. \quad (21)$$

For Lipschitz property, note that each loss function is $1/\alpha$ -Lipschitz because the gradient when it is defined is just $\frac{\theta - \mathbf{c}_j}{\alpha \|\theta - \mathbf{c}_j\|_2}$. We prove it formally.

Consider any θ, θ' in $B(\mathbf{0}, 1)$ and a data point $d_i \in D$. We wish to show $|\ell(\theta; d_i) - \ell(\theta'; d_i)| \leq \frac{1}{\alpha} \|\theta - \theta'\|_2$. We can split the line segment from θ to θ' into a sequence of line segments

$$(\theta_0, \theta_1), (\theta_1, \theta_2), \dots, (\theta_{k-1}, \theta_k),$$

where $\theta_0 = \theta, \theta_k = \theta'$, such that for any line segment (θ_m, θ_{m+1}) , θ_m and θ_{m+1} share a minimizer in C of $\left(\frac{\|\theta - \mathbf{c}_j\|_2}{\alpha} \right) d_{i,j}$.¹⁰

It now suffices to show $|\ell(\theta_m; d_i) - \ell(\theta_{m+1}; d_i)| \leq \frac{1}{\alpha} \|\theta_m - \theta_{m+1}\|_2$ for each m , since we then have:

$$|\ell(\theta; d_i) - \ell(\theta'; d_i)| \leq \sum_{m=0}^{k-1} |\ell(\theta_m; d_i) - \ell(\theta_{m+1}; d_i)| \leq \frac{1}{\alpha} \sum_{m=0}^{k-1} \|\theta_m - \theta_{m+1}\|_2 = \frac{1}{\alpha} \|\theta - \theta'\|_2.$$

Let \mathbf{c}_j be a shared minimizer of $\left(\frac{\|\theta - \mathbf{c}_j\|_2}{\alpha} \right) d_{i,j}$ for θ_m and θ_{m+1} . If $d_{i,j} = 0$, then trivially $|\ell(\theta_m; d_i) - \ell(\theta_{m+1}; d_i)| \leq \frac{1}{\alpha} \|\theta_m - \theta_{m+1}\|_2$. Otherwise $d_{i,j} = 1$ and by triangle inequality, we have:

¹⁰In particular, for each \mathbf{c}_j let B_j be the set of points in $B(\mathbf{0}, 1)$ such that \mathbf{c}_j is a minimizer of $\left(\frac{\|\theta - \mathbf{c}_j\|_2}{\alpha} \right) d_{i,j}$. We can split the line segment from θ to θ' at each point where it enters or leaves some B_j to get this sequence of line segments, and by this construction each line segment's endpoints are both in B_j for some j .

$$|\ell(\theta_m; d_i) - \ell(\theta_{m+1}; d_i)| = \left| \frac{\|\theta_m - \mathbf{c}_j\|_2}{\alpha} - \frac{\|\theta_{m+1} - \mathbf{c}_j\|_2}{\alpha} \right| \leq \frac{1}{\alpha} \|\theta_m - \theta_{m+1}\|_2.$$

Now let us suppose there is an (ϵ, δ) -differentially private algorithm \mathcal{A} that on input a non-convex function \mathcal{L} and n data points $\{d_1, \dots, d_n\}$, outputs a θ^{priv} such that

$$\mathbb{E}_{\mathcal{A}} \left[\mathcal{L}(\theta^{\text{priv}}; D) \right] - \min_{\theta \in B(1)} \mathcal{L}(\theta; D) = o\left(\frac{p \log(1/\delta)}{n\epsilon}\right), \quad (22)$$

where $D = \{d_1, \dots, d_n\}$.

We will construct an algorithm that uses \mathcal{A} as subroutine and solve top-selection problem with an error $o(\log(s))$, contracting the lower bound of Theorem A.13.

Algorithm \mathcal{B} :

- On input $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, invokes \mathcal{A} on the function defined by eq. (21) and data points X to get θ^{priv} as output.
- Output $f(\theta^{\text{priv}})$.

Since the last step is post-processing, \mathcal{B} is (ϵ, δ) -differentially private. We now show that if \mathcal{A} outputs a θ^{priv} satisfying eq. (22), then $j := f(\theta^{\text{priv}})$ satisfies eq. (20) leading to a contradiction.

First note that, for any $\mathbf{c} \in C$ and all $\theta \in \mathbb{B}_p(\mathbf{c}, \alpha)$ such that $\|\theta - \mathbf{c}\|_2 \leq \frac{\alpha}{2}$,

$$\mathcal{L}(\mathbf{c}; D) = -\frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i, f(\mathbf{c})} \leq \mathcal{L}(\theta; D).$$

Therefore,

$$\mathcal{L}(\theta^*; X) := \min_{\mathbf{c} \in C} \mathcal{L}(\mathbf{c}, X) = \min_{\mathbf{c} \in C} \left(-\frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i, f(\mathbf{c})} \right)$$

This implies that

$$f(\theta^*) = \arg \max_{1 \leq j \leq s} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i, j},$$

which is exactly the top-selection problem. Therefore, eq. (22) implies eq. (20) because $p \log\left(\frac{1}{\alpha}\right) \leq \log(s) \leq p \log\left(\frac{3}{\alpha}\right)$ and $\alpha \in (0, 1/2)$ is a constant. □