TRACTABLE MCMC FOR PRIVATE LEARNING WITH PURE AND GAUSSIAN DIFFERENTIAL PRIVACY

Yingyu Lin[1*], **Yi-An Ma**[1*], **Yu-Xiang Wang**[2*], **Rachel Redberg**[2], **Zhiqi Bu**[3]
[1]UC San Diego, [2]UC Santa Barbara, [3]Amazon AI
{yil208,yianma}@ucsd.edu {yuxiangw,rredberg}@cs.ucsb.edu zhiqibu@amazon.com

ABSTRACT

Posterior sampling, i.e., exponential mechanism to sample from the posterior distribution, provides ε-*pure* differential privacy (DP) guarantees and does not suffer from potentially unbounded privacy breach introduced by (ε, δ)-approximate DP. In practice, however, one needs to apply approximate sampling methods such as Markov chain Monte Carlo (MCMC), thus re-introducing the unappealing δ-approximation error into the privacy guarantees. To bridge this gap, we propose the Approximate SAample Perturbation (abbr. ASAP) algorithm which perturbs an MCMC sample with noise proportional to its Wasserstein-infinity (W_∞) distance from a reference distribution that satisfies *pure* DP or *pure* Gaussian DP (i.e., $\delta = 0$). We then leverage the Metropolis-Hastings algorithm to generate the sample and prove that the algorithm converges in W_∞ distance. We show that by combining our new techniques with a localization step, we obtain the first nearly linear-time algorithm that achieves the optimal rates in the DP-ERM problem with strongly convex and smooth losses.

1 INTRODUCTION

The strongest form of differential privacy (DP) is ε-*pure* DP, which provides an inviolable bound of $\varepsilon \geq 0$ on an algorithm's privacy loss. The posterior sampling mechanism associated with the loss functions (Wang et al., 2015; Dimitrakakis et al., 2017; Gopi et al., 2022) is known to achieve near-optimal privacy-utility tradeoff for learning under pure DP — in theory. In practice, sampling from a general posterior distribution is typically either intractable or inefficient. Both issues can be solved by approximating the posterior, e.g. via MCMC sampling. But of course there is no free lunch: the approximation error introduced by this approach now degrades the privacy guarantee to the weaker notion of (ε, δ)-*approximate* DP, under which we risk a catastrophic privacy breach with some small probability δ. Our paper bridges the gap between theory and practice — thus alleviating a common growing pain of DP research — by proposing an efficient MCMC-based algorithm that samples from an approximate posterior *while satisfying pure DP*.

Recent strides in research have explored diverse implementations of posterior sampling mechanisms, but a recurring challenge remains — each approach grapples with either a compromise on the DP guarantee or extended computational runtimes. As demonstrated by Wang et al. (2015); Gopi et al. (2022), applying *approximate* posterior sampling and directly analyzing the sampling error in terms of the total variation (TV) distance leads back to an irremovable failure probability, ultimately reducing the privacy guarantee to approximate DP. In contrast, Seeman et al. (2021) show that it is possible to sample from the *exact* posterior with pure DP. But the runtime of their sampler may be exponentially large since they use a rejection sampling scheme where the probability of an accepted sample is, in the worst case, exponentially small.

This motivates the following question:

Can we obtain *pure* DP guarantees with an efficient MCMC algorithm?

[*]Equal contribution.

In this paper, we answer the question in the affirmative by developing the Approximate Sample Perturbation (ASAP) algorithm.

1.1 Our Contributions

The main results of our paper are threefold.

1. We propose Approximate Sample Perturbation (Algorithm 2), a novel MCMC-type method designed to maintain pure DP and pure Gaussian DP by perturbing an MCMC sample. We substantiate, through Theorem 2, that the pure DP and Gaussian DP guarantees are maintained when the distribution of the MCMC sample closely approximates the exact posterior, quantified in terms of the Wasserstein-infinity (W_∞) distance.

2. We establish a novel generic lemma, Lemma 8, which provides a systematic method for converting Total Variation (TV) distance into Wasserstein-infinity distance. This transformation enables us to uphold pure DP guarantees, even when utilizing MCMC samplers with convergence measured in TV distance. Notably, this lemma holds its own significance and is of independent interest.

3. We introduce the Metropolis adjusted Langevin algorithm (MALA) with constraint (Algorithm 1) that converges in terms of W_∞ distance. When employed in conjunction with a preceding *localization* step, detailed in Algorithm 3, it empowers our ASAP framework to achieve optimal rates in nearly linear time while ensuring pure DP and pure Gaussian DP within the Differential Privacy - Empirical Risk Minimization (DP-ERM) setting, for strongly convex and smooth losses.

1.2 Related Work

Posterior sampling mechanism, i.e., outputting a sample $\hat{\theta} \sim p(\theta) \propto \exp(-\gamma(F(\theta) + \frac{1}{2}\lambda||\theta||_2^2))$ is a popular method for DP-ERM and private Bayesian learning. It was initially analyzed under pure DP (Mir, 2013; Dimitrakakis et al., 2017; Wang et al., 2015), then later under approximate DP (Minami et al., 2016), Rényi DP(Geumlek et al., 2017) and Gaussian DP (Gopi et al., 2022). In each case, it is known that with appropriate choices of γ, λ parameters, it achieves the optimal rate (Bassily et al., 2014) under each privacy definition. However, the computational complexity of the MCMC methods has been a major challenge for this problem. To the best of our knowledge, Bassily et al. (2014); Chourasia et al. (2021); Ryffel et al. (2022); Mangoubi & Vishnoi (2022) are the only known results that obtain DP guarantees using MCMC methods without having $\delta > 0$. Bassily et al. (2014)'s sampler (a variant of Applegate & Kannan (1991)) runs in $O(n^4)$ and requires explicit discretization. Chourasia et al. (2021); Ryffel et al. (2022)'s algorithms run in nearly linear time (for strongly convex DP-ERM) and enjoy pure Rényi DP, but their utility bound is a factor of condition number κ worse than the statistical limit. Mangoubi & Vishnoi (2022)'s algorithm translates the TV distance guarantee to infinity-distance (d_∞) guarantee by incorporating uniform noise and a membership oracle, while their focus is on Lipschitz continuous loss. The idea of perturbing distributions to strengthen the DP guarantee is discussed in Feldman et al. (2018) and its application to Langevin analysis (Altschuler & Talwar, 2023), where additional noise converts W_∞ guarantee to Rényi DP guarantee. Our work is the first that archives the optimal rate for Strongly-Convex DP-ERM under pure-DP and pure-Gaussian DP with a nearly linear time algorithm (for $\kappa = \mathrm{polylog}(n)$).

DP-ERM can also be solved using other methods that do not require MCMC sampling. However, these methods either do not achieve optimal rates (output perturbation (Chaudhuri et al., 2011)) or are computationally less efficient (Noisy SGD (Bassily et al., 2014; Abadi et al., 2016)) or require the model to be (generalized) linear (e.g., objective perturbation (Chaudhuri et al., 2011; Kifer et al., 2012; Redberg et al., 2023)).

2 Problem Setup and Preliminaries

Symbols and notations. Let $\mathcal{X}$ be the space of data points, $\mathcal{X}^* := \cup_{n=0}^{\infty}\mathcal{X}^n$ be the data space, and $D \in \mathcal{X}^*$ be a dataset with an unspecified number of data points. Let the parameter space $\mathcal{U} \subseteq \mathbb{R}^d$ and for each $x \in \mathcal{X}$

and $\theta \in \mathcal{U}$, $\ell_x(\theta)$ denotes the loss function (or negative log-likelihood function). When $D = \{x_1, ..., x_n\}$, we denote $\ell_{x_i}(\theta)$ by $\ell_i(\theta)$ as a short hand. Denote the total loss $\mathcal{L} = \sum_{i=1}^n \ell_i$. For any set S, we denote the set of all probability distributions as Δ^S or $\mathcal{P}_S$, so that a mechanism $\mathcal{M} : \mathcal{X}^* \to \Delta^{\mathcal{U}}$ is a randomized algorithm. We use $\mathcal{M}(D)$ to denote the probability distribution as well as the corresponding random variable returned by the mechanism. For a set S, we denote $\mathrm{Diam}(S) := \sup_{x,y \in S} \|x - y\|_2$, and $\|S\| := \sup_{x \in S} \|x\|_2$.

2.1 Differential Privacy Empirical Risk Minimization (DP-ERM)

Empirical risk minimization (ERM) is a classic learning framework which casts the problem of finding a "good" model into an optimization task. Our goal is to find the parameter θ^* in the parameter space $\mathcal{U} \subseteq \mathbb{R}^d$ which minimizes the empirical risk: $\theta^* = \arg\min_{\theta \in \mathcal{U}} \left(\sum_{i=1}^n \ell_i(\theta)\right)$.

The problem setting of our paper is differentially private empirical risk minimization (DP-ERM): empirical risk minimization under a privacy constraint.

2.2 Differential privacy definitions

Definition 1 (Differential privacy (Dwork et al., 2006; 2014))**.** *Mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if for all neighboring datasets $D \simeq D'$ and for any measurable set $S \subseteq$ Range$(\mathcal{M})$,*

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

When $\delta = 0$, $\mathcal{M}$ satisfies ε-(pure) DP.

Differential privacy (DP) provably bounds the privacy loss of an algorithm. *Approximate* DP ($\delta > 0$) is a practical and popular DP variant which allows a "failure" event — where the privacy loss exceeds the bound — to occur with some probability. Beware: approximate DP does not bound the *severity* of a privacy breach. The privacy loss under a failure event could be arbitrarily large. Avoiding this risk requires *pure* DP ($\delta = 0$), which provides a deterministic bound on the privacy loss.

In comparison to approximate DP, *Gaussian* DP is a more "controlled" relaxation of pure DP which does not have an unbounded failure mode. We can define Gaussian DP via the hockey-stick divergence.

Definition 2 (Hockey-Stick Divergence)**.** *The Hockey-Stick Divergence of distributions P, Q is defined as*

$$H_\alpha(P \| Q) := \mathbb{E}_{o \sim Q}\left[\left(\frac{dP}{dQ}(o) - \alpha\right)_+\right]$$

where $(\cdot)_+ := \max\{0, \cdot\}$ and $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative.

Definition 3 (Gaussian Differential Privacy (Dong et al., 2022))**.** *We say that a mechanism $\mathcal{M}$ satisfies μ-Gaussian differential privacy if for any neighboring dataset D, D',*

$$H_{e^\varepsilon}(\mathcal{M}(D) \| \mathcal{M}(D')) \leq H_{e^\varepsilon}(\mathcal{N}(0,1) \| \mathcal{N}(\mu, 1)) \qquad \forall \varepsilon \in \mathbb{R}.$$

Notice that this is an equivalent definition to the *dual* definition from the hypothesis testing point of view from (Dong et al., 2022, Definition 2.6). The statement naturally provides a "dominating pair" of distributions that are convenient for adaptive composition, amplification by sampling and efficient numerical computation (Zhu et al., 2022).

2.3 Exact posterior sampling: DP and utility guarantees

The primary algorithm we consider is a variant of the classical exponential mechanism (EM) known as posterior sampling. The posterior sampling algorithm instantiates the exponential mechanism by taking the quality score to be a scaled and regularized log-likelihood function with parameter γ, λ, i.e.,

$$\hat{\theta} \sim p(\theta) \propto \exp\left(-\gamma\left(\sum_{i=1}^n \ell_i(\theta) + \lambda\|\theta\|^2\right)\right) \mathbf{1}(\theta \in \Theta). \tag{1}$$

This mechanism enjoys pure DP and Gaussian DP for appropriate choices of γ, λ, and domain Θ.

Lemma 4 (GDP of posterior sampling (Gopi et al., 2022, Theorem 4))**.** *Assume the loss function is G-Lipschitz, posterior sampling mechanism with parameter $\gamma, \lambda > 0$ satisfying $\gamma \leq \mu^2\lambda/G^2$ satisfies μ-GDP.*

Lemma 5 (Pure DP of posterior sampling)**.** *Assume the loss function is G-Lipschitz, posterior sampling mechanism with parameter $\gamma > 0$ (any of λ) in a domain Θ satisfies ε-pure DP if $\gamma \leq \frac{\varepsilon}{G \cdot \mathrm{Diam}(\Theta)}$.*

The above lemma slightly improves existing analysis (Wang et al., 2015; Dimitrakakis et al., 2017) using the *bounded range* analysis from Dong et al. (2020), hence avoids assuming a potentially large bound on the loss in the privacy calculations.

Regarding utility, Gopi et al. (2022) showed that the posterior sampling mechanism achieves a minimax optimal rate for DP-ERM for Lipschitz convex losses. We will discuss these results in the appendix in more detail, but focus on a narrower setting with strongly convex losses in which the rate is faster. We have the following lemma for the utility of posterior sampling.

Lemma 6 (De Klerk & Laurent (2018, Corollary 1))**.** *For a convex function $F(\theta)$, and a convex set $\Theta \subset \mathbb{R}^d$, $\hat{\theta} \sim p(\theta) \propto \exp(-\gamma F(\theta))$ satisfies that $\mathbb{E}[F(\hat{\theta})] \leq \min_{\theta \in \Theta} F(\theta) + \frac{d}{\gamma}$.*

3 Technical Tools

In this section, we introduce tools that make MCMC sampling for *pure* DP possible. These tools are novel and of independent interest as well. Our contributions consist of a new Lemma 8 that converts TV distance to W_∞ distance, and Algorithm 1, an MCMC sampler designed to guarantee convergence in W_∞ distance.

3.1 Overview: Why do we use W_∞ distance?

Suppose p^* is a Gibbs posterior that satisfies pure DP. In practice, MCMC samplers generate an approximate sample $\tilde{\theta}$ from an approximate distribution $\tilde{p}$ with $d_{TV}(\tilde{p}, p^*) < \xi$, where d_{TV} represents the TV distance. Notice that this TV-distance guarantee does not grant pure DP for sampling from $\tilde{p}$. Instead, it provides only (ε, δ)-DP with $\delta = (1 + e^{\varepsilon})\xi > 0$, as elucidated in Proposition 3 of Wang et al. (2015).

To circumvent this challenge, ASAP (Algorithm 2) adopts a two-part approach to address the privacy of $\tilde{\theta}$. First, we decompose $\tilde{\theta}$ into two components:

$$\tilde{\theta} = \theta_* + (\tilde{\theta} - \theta_*), \text{ where } \theta_* \sim p^*.$$

Note that $\theta_* \sim p^*$ satisfies pure DP by existing works Wang et al. (2015). Now, our goal is to ensure that the perturbed difference $(\tilde{\theta} - \theta_*) + \textit{noise}$ exhibits pure DP. This enables us to leverage the composition lemma (Lemma 14) and establish an MCMC sampling methodology that guarantees pure DP.

However, to release $(\tilde{\theta} - \theta_*)$ with pure DP by adding noise, it becomes paramount to establish an upper bound on the sensitivity of this quantity with a certainty of 1. This necessitates bounding the particle-wise distance $\|\tilde{\theta} - \theta_*\|_1$ (or $\|\tilde{\theta} - \theta_*\|_2$ in Gaussian DP) *with a probability of 1*. Achieving this stronger notion of privacy hinges on the W_∞ distance. Refer to Appendix E for a visualization illustrating the W_∞ distance.

Hence, an MCMC sampling approach that guarantees an error bound in Wasserstein-Infinity distance is indispensable. Our work addresses this need through the instantiation of MALA with constraint, a method meticulously tailored to ensure strong Wasserstein-Infinity distance guarantees.

In the following sections, we delve into the details of our conversion lemma and present the MALA algorithm with constraints, demonstrating their efficacy in achieving pure DP through MCMC sampling.

3.2 TV distance to W_∞ distance

Consider two distributions P and Q on the same metric space (Θ, dist). We sample $x \sim P$ and $y \sim Q$. When P and Q are relatively "close" to each other, $\mathrm{dist}(x, y)$ is expected to be small. Motivated by our preceding

discussion of MCMC for pure DP, a natural question arises: Is there a way to define a distance for distributions P and Q, such that it provides an upper bound for $\mathrm{dist}(x, y)$ with probability 1? Wasserstein-infinity distance provides an affirmative response to this question by constructing an appropriate coupling ζ between P and Q, which is defined as follows.

Definition 7. *Let $\Theta \subseteq \mathbb{R}^d$ be a domain equipped with a metric* dist*. Consider two probability measures, P and Q, defined on Θ. The Wasserstein-infinity distance between P and Q with respect to* dist*, denoted as $W_\infty(P, Q)$, is defined as*

$$W_\infty(P,Q) := \inf_{\zeta\in\Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x,y) = \inf_{\zeta\in\Gamma(P,Q)} \{c \geq 0 | \zeta(\{(x,y)\in\Theta^2 | \mathrm{dist}(x,y)\leq c\}) = 1\},$$

where $\Gamma(P, Q)$ is the set of all couplings of P and Q, i.e., the set of all joint probability distributions ζ with marginals P and Q respectively.

Our pursuit for strong privacy guarantees hinges on the W_∞ distance to achieve pure Differential Privacy (DP). However, practical MCMC samplers primarily provide guarantees based on the convergence of Total Variation (TV) distance. Regrettably, relying solely on TV distance would result in a weaker privacy guarantee, approximate (ε, δ)-Differential Privacy with a $\delta > 0$.

To reconcile this discrepancy, we present a versatile lemma that enables the conversion from TV distance to the W_∞ distance.

Lemma 8 (Converting TV distance to W_∞ distance)**.** *Consider two probability measures, P and Q, both supported on a bounded closed 2-norm-ball $\Theta \subseteq \mathbb{R}^d$. On this domain Θ, suppose that the density of Q is lower-bounded by a constant $p_{\min}$. We establish the following results based on the choice of metric for W_∞:*

1. *Suppose W_∞ is metric with the 1-norm, i.e.,* $\mathrm{dist}(x, y) = \|x - y\|_1$.

$$\textit{If} \quad d_{TV}(P,Q) < p_{\min}\cdot\frac{\pi^{d/2}}{2^{d+1}\cdot\Gamma(\frac{d}{2}+1)\cdot d^{d/2}}\Delta^d, \quad \textit{then} \quad W_\infty(P,Q)\leq\Delta.$$

2. *Suppose W_∞ is metric with the 2-norm, i.e.,* $\mathrm{dist}(x, y) = \|x - y\|_2$.

$$\textit{If} \quad d_{TV}(P,Q) < p_{\min}\cdot\frac{\pi^{d/2}}{2^{d+1}\cdot\Gamma(\frac{d}{2}+1)}\Delta^d, \quad \textit{then} \quad W_\infty(P,Q)\leq\Delta.$$

In both cases, Γ represents the gamma function.

The results naturally extend to the case where Θ is a bounded open 2-norm ball.

This lemma provides a powerful tool to bridge the gap between TV distance and W_∞ distance, ensuring that we can maintain pure DP guarantees even when dealing with TV distance-based MCMC samplers.

Note that to convert from TV distance to W_∞ distance, additional assumptions about the probability measures are required. We introduce a supporting lemma to ensure the conditions for the conversion procedure are met. This lemma establishes a lower bound on the density of distributions that are log-concave and log-Lipschitz continuous or smooth.

Lemma 9. *Consider a distribution with density $p(\theta) = \frac{1}{Z}e^{-\gamma J(\theta)}$ supported on a bounded convex domain Θ with 2-norm diameter R, where $Z = \int_\Theta e^{-\gamma J(\theta)}\mathrm{d}\theta$ is the normalization constant.*

Assuming that J is convex and $\widetilde{G}$-Lipschitz continuous, we establish the following bound on the density:

$$\inf_{\theta\in\Theta} p(\theta) \geq e^{-\gamma\widetilde{G}R}\cdot\frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}(R/2)^d}.$$

Alternatively, if J is convex and β-Lipschitz smooth, the density bound holds for any $\tilde{\theta} \in \Theta$:

$$\inf_{\theta \in \Theta} p(\theta) \geq e^{-\gamma\left(R\|\nabla J(\tilde{\theta})\| + \beta R^2/2\right)} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}(R/2)^d}.$$

These lemmas together provide a comprehensive framework for translating TV distance to W_∞ distance while maintaining crucial assumptions about the underlying probability measures.

By incorporating these results into our privacy framework, we can navigate the challenges of MCMC samplers, ensuring *pure* differential privacy guarantees even when operating in a TV distance-based setting.

3.3 MALA WITH CONSTRAINT

From our previous discussion, we know that to maintain the pure and Gaussian-DP guarantee, the approximate and the target distribution need to be close in the W_∞ distance. From Lemma 8, that translates into an *exponentially* small error in terms of the TV distance. Hence we apply MALA, an asymptotically unbiased sampler, which converges *exponentially* fast in terms of the accuracy in TV distance. Lemma 8 also underscores the necessity of sampling within a bounded set. Consequently, we reject the samples outside the specified domain, as outlined in Algorithm 1. To ensure the algorithm terminates in a finite time, we set a maximum restart number $\tau_{\max}$. We also provide the expected runtime, indicating that the algorithm halts in early trials.

Notations. Let $J(\theta) = \sum_{i=1}^n \ell_i(\theta)$ be the objective function, where ℓ_i's satisfy G-Lipschitz continuity. Let $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} J(\theta)$ be the minimizer. Let π and p^* be the unconstrained and constrained Gibbs distribution respectively with density $\pi(\theta) \propto e^{-\gamma J(\theta)}$, and $p^*(\theta) \propto e^{-\gamma J(\theta)}\mathbf{1}(\theta \in \Theta)$. Let $\kappa = \beta/\alpha$ be the condition number, where β and α are defined in Assumption 1 and 2 respectively.

Assumption 1 (Smoothness). Function J is $n\beta$-Lipschitz smooth.

Assumption 2 (Strong-Convexity). Function J is $n\alpha$-strongly convex.

Assumption 3 (Domain bound). The domain Θ is a convex set such that $\mathbb{B}(\theta^*, R_1) \subset \Theta$, for $R_1 \geq 8\sqrt{\frac{d}{\gamma n \alpha}}$.

Theorem 1 (Mixing time in TV-distance)**.** *Consider Algorithm 1 with initial distribution $p^0 \sim \mathcal{N}(0, \frac{1}{\gamma n \beta}\mathbb{I})$, step size $h^k = \mathbf{\Theta}\left(\min\left\{\kappa^{-1/2}(\gamma n \beta)^{-1}(d\ln\kappa + \ln 1/\xi)^{-1/2}, \frac{1}{\gamma n \beta d}\right\}\right)$, number of iterations $K = \mathbf{\Theta}\left((d\ln\kappa + \ln 1/\xi)\max\left\{\kappa^{3/2}\sqrt{d\ln\kappa + \ln 1/\xi}, d\kappa\right\}\right)$, and maximum number of restarts $\tau_{\max} = \mathbf{\Theta}(\ln 1/\xi)$. Under Assumptions 1, 2, and 3, this algorithm converges to ξ-accuracy in $d_{TV}(p_{\text{out}}, p^*)$, where p_{out} is the output distribution. The expected number of gradient queries to J is given by:*

$$\mathbf{\Theta}\left((d\ln\kappa + \ln 1/\xi)\max\left\{\kappa^{3/2}\sqrt{d\ln\kappa + \ln 1/\xi}, d\kappa\right\}\right).$$

Corollary 10 (Mixing time in W_∞ distance)**.** *Consider the W_∞ distance with respect to the 1-norm. Let the domain Θ be a ball of radius B. Under the same assumptions and conditions as Theorem 1 with choosing $\ln(1/\xi) = \mathbf{\Theta}(\gamma n \beta B^2 + d\ln B + d\ln d + d\ln(1/\Delta))$, Algorithm 1 converges to Δ-accuracy in $W_\infty(p_{\text{out}}, p^*)$. The expected number of gradient queries to J is given by:*

$$\mathbf{\Theta}\left(\left(d\ln(\kappa B d) + d\ln(1/\Delta) + \gamma n \beta B^2\right)\max\left\{\kappa^{3/2}\sqrt{d\ln(\kappa B d) + d\ln(1/\Delta) + \gamma n \beta B^2}, d\kappa\right\}\right).$$

If B, γ is defined as Table 2, then the expected queries is

$$\mathbf{\Theta}\left(d\left(\kappa\ln(d/\rho) + \ln\left(\frac{G\ln(d/\rho)}{\alpha n \varepsilon \Delta}\right)\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(d/\rho) + \ln\left(\frac{G\ln(d/\rho)}{\alpha n \varepsilon \Delta}\right)\right)}, d\kappa\right\}\right).$$

Each gradient query to the empirical risk J translates to n queries to the individual losses ℓ_i's.

Remark 11. *We note that the dependence on d is $\widetilde{\Omega}(d^2)$, and the number of queries to the individual losses ℓ_i's is $\widetilde{\Omega}(n)$.*

In comparison, if we employ vanilla rejection sampling (Casella et al., 2004) with a proposal distribution of $\mathcal{N}(0, \frac{1}{\gamma n \beta}\mathbb{I})$, then the convergence time scales as $e^{\gamma n(\beta-\alpha)R_1^2} \geq e^{64(\kappa-1)d}$, where R_1 is defined according to Assumption 3. This factor scales exponentially with the condition number κ and the dimension d.

Algorithm 1 Metropolis-adjusted Langevin algorithm (MALA) with constraint

Input: Stepsizes $\{h^k\}$, number of (inner-loop) iterations K, objective function J, domain Θ, parameter γ, maximum number of restarts $\tau_{\max}$.
for $t = 0, 1, 2, \ldots, \tau_{\max} - 1$ **do**
  Sample θ^0 according to distribution p^0
  for $k = 0, 1, 2, \ldots, K-1$ **do**
    Sample $\theta^{k+1} \sim \mathcal{N}\left(\theta^k - h^k\gamma\nabla J(\theta^k), 2h^k\mathrm{I}\right)$
    Sample $u^{k+1} \sim \mathcal{U}[0,1]$, denote $p(\cdot|\theta)$ as the density of $\mathcal{N}\left(\theta - h^k\gamma\nabla J(\theta), 2h^k\mathrm{I}\right)$
    if $\dfrac{p\left(\theta^k|\theta^{k+1}\right)\pi(\theta^k)}{p\left(\theta^{k+1}|\theta^k\right)\pi\left(\theta^{k+1}\right)} < u^{k+1}$ **then**
      $\theta^{k+1} \leftarrow \theta^k$
    end if
  end for
  if $\theta^K \in \Theta$ **then** ▷ Accept the sample
    Return θ^K and halt
  end if
end for
Return an arbitrary $\theta \in \Theta$

4 Main Results: Approximate Sample Perturbation (ASAP)

This section introduces Approximate Sample Perturbation, along with the end-to-end localized ASAP.

4.1 Approximate Sample Perturbation (ASAP)

ASAP (Algorithm 2) is designed to maintain pure DP and pure Gaussian DP by smoothing out an MCMC sample. Theorem 2 establishes the pure DP and pure GDP guarantee for the ASAP algorithm.

To clarify, consider a reference randomized mechanism $\mathcal{M}$ alongside another randomized mechanism $\widetilde{\mathcal{M}}$. Assume that $\mathcal{M}$ is assured of pure DP or pure GDP, whereas $\widetilde{\mathcal{M}}$ lacks DP guarantee. According to Theorem 2, if distributions $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ have a bounded disparity in W_∞ distance, then it is feasible to achieve pure DP (or GDP) by generating samples from $\widetilde{\mathcal{M}}$ and appending noise scaled in proportion to the W_∞ bound.

Algorithm 2 Approximate SAmple Perturbation (ASAP)

Input: Dataset D, reference randomized mechanism $\mathcal{M} : \mathcal{X}^* \to \mathcal{P}_{\mathcal{U}}$. W_∞ error Δ, a black box sampler $\widetilde{\mathcal{M}}$ such that $W_\infty(\widetilde{\mathcal{M}}(D), \mathcal{M}(D)) \leq \Delta$. Privacy parameter ε' if pure DP (μ' if Gaussian DP).
Run the MCMC sampler: $\tilde{\theta} \sim \widetilde{\mathcal{M}}(D)$.
Return $\hat{\theta} = \tilde{\theta} + \mathcal{N}(0, \frac{4\Delta^2}{\mu'^2}\mathrm{I}_d)$ if GDP (or $\hat{\theta} = \tilde{\theta} + \mathbf{Z}$ with $Z_i \overset{i.i.d.}{\sim} \mathrm{Lap}(\frac{2\Delta}{\varepsilon'})$, $i = 1, ..., d$ if pure DP.)

Theorem 2 (DP Guarantees for Approximate SAmple Perturbation)**.** *Let $\mathcal{M}, \widetilde{\mathcal{M}}$ be randomized algorithms with output space $\Theta \subset \mathbb{R}^d$ that satisfy the following for any input dataset D:*

$$W_\infty(\mathcal{M}(D), \widetilde{\mathcal{M}}(D)) \leq \Delta.$$

1. *Suppose we define W_∞ with the 2-norm. If $\mathcal{M}$ satisfies μ-GDP, then the procedure $\widetilde{\mathcal{M}}(D) + \mathcal{N}(0, \frac{4\Delta^2}{(\mu')^2} I_d)$ satisfies $\sqrt{\mu^2 + \mu'^2}$-GDP.*

2. *Suppose we define W_∞ with the 1-norm. If $\mathcal{M}$ satisfies ε-DP, then the procedure $\widetilde{\mathcal{M}}(D) + \mathbf{Z}$ with $Z_i \sim \mathrm{Lap}(\frac{2\Delta}{\varepsilon'})$ i.i.d. for $i = 1, ..., d$, satisfies $(\varepsilon + \varepsilon')$-DP.*

As it is made clear in the proof and in Section 3.1, a W_∞ distance guarantee is essential for us to obtain pure-DP and pure-GDP with ASAP. Other weaker distance metrics such as TV-distance and W_2-distance are not useful for this purpose in general. Interestingly, as we establish in Lemma 8, when the domain is a bounded ball and the density is bounded away from zero, we can convert a weak TV-distance bound into a strong W_∞ distance bound, which allows us to apply Algorithm 2 in a wide range of applications.

4.2 Localized ASAP and the End-to-End Guarantees

Lemma 8 implies that we should sample from a bounded 2-norm ball. So before sampling, we should first localize to a bounded ball centered at the initial point θ_0, i.e., $\Theta = \mathbb{B}(\theta_0, B) = \{\theta \mid \|\theta - \theta_0\|_2 \leq B\}$. We then run ASAP on this bounded ball; we call this "Localized-ASAP".

Algorithm 3 End-to-End Localized ASAP

1: Input: Parameters $\gamma, \lambda = 0, B$ as given in Table 2. Wasserstein-infinity error Δ. Dataset D, individual loss ℓ_i's satisfying G Lipschitz continuity, β-smoothness, and α-strong convexity. Privacy parameter μ' for GDP (or ε' for pure DP) for ASAP. Denote $J(\theta) = \sum_{i=1}^n \ell_i(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|^2$.

2: Run the localization Algorithm 4, and assign its output to θ_0.

3: Call ASAP (Algorithm 2) with inputs

$$\begin{cases} \Delta, D, \gamma, \text{privacy parameter } \varepsilon' \text{ if pure DP } (\mu' \text{ if Gaussian DP}) \\ \widetilde{\mathcal{M}} \leftarrow \text{Algorithm 1 on the domain } \Theta = \{\theta | \|\theta - \theta_0\|_2 \leq B\}, \\ \mathcal{M} \leftarrow \text{Reference distribution with density } p_{\mathcal{M}(D)}(\theta) \propto e^{-\gamma J(\theta)} \mathbf{1}\{\|\theta - \theta_0\| \leq B\}, \end{cases}$$

and assign the output to $\hat{\theta}$.

4: Return $\hat{\theta}$.

The computational efficiency of the sampler depends on the choices of B and γ, as well as assumptions on the loss functions. The quality of the solution from the localized ASAP depends on the initialization θ_0, and the associated choice of B. We instantiate these parameters concretely in Appendix B.

We provide the following privacy, utility, and computational guarantees for the end-to-end Algorithm 3.

Theorem 3 (Guarantees for localized-ASAP)**.** *Let $\rho \in (0, 1)$ and set $\theta_0, B, \gamma, \lambda$ as Table 2. Let $\hat{\theta}$ be the output of Algorithm 3. Set the pure-DP (or Gaussian DP) parameters for the output perturbation, approximate-MCMC sampler, and the ASAP to be ε (or μ). Then*

1. *$\hat{\theta}$ satisfies 3ε-pure DP (or $\sqrt{3}\mu$-Gaussian DP).*

2. *In pure DP case, the empirical (total) risk satisfies*

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq \mathcal{O}\left(\frac{d^2 G^2 (\ln d + \ln(1/\rho) + \rho d \kappa)}{\alpha n \varepsilon^2}\right).$$

In the Gaussian DP case, the empirical risk satisfies $\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq \mathcal{O}\left(\frac{G^2(d+\rho\kappa(d+\ln 1/\rho))}{\alpha n \mu^2}\right)$.

3. *The expected overall runtime time of the algorithm is*

$$\mathbf{\Theta}\left(nd\left(\kappa\ln(d/\rho)+\ln n\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(d/\rho)+\ln n\right)}, d\kappa\right\}\left(1+\rho d\left(\kappa\ln(d/\rho)+\ln n\right)\right)\right).$$

Remark 12. *The suboptimality bound matches the information-theoretic limit for this problem in (Bassily et al., 2014) up to a constant factor. Notably, taking* $\rho = \mathcal{O}(1/(d\kappa))$*, we obtain the empirical risk bound* $\mathcal{O}\left(\frac{d^2G^2(\ln d+\ln\kappa)}{\alpha n\varepsilon^2}\right)$*. This reduces the* κ *dependence from* $\mathcal{O}(\kappa)$ *to* $\mathcal{O}(\log\kappa)$*. To our knowledge, this is the first* $\tilde{\mathcal{O}}(n)$ *time algorithm (using incremental gradient oracle, assume* $\kappa = \mathrm{polylog}(n)$*) for DP-ERM (Lipschitz, smooth, strong convex losses) that achieves the optimal rates under pure DP. The closest is the algorithm from (Bassily et al., 2014), which runs in* $\mathcal{O}(n^4)$ *time. To avoid confusion, a nearly linear-time algorithm for DP-SCO in the smooth / strongly convex setting that achieves optimal rates exists (Feldman et al., 2020). But to the best of our knowledge, the problem remains open for DP-ERM until this paper.*

Remark 13 (Improvement under Gaussian DP)**.** *Taking* $\rho = \mathcal{O}(1/\kappa)$*, the empirical risk bound for GDP case is* $\mathcal{O}(\frac{(d+\ln\kappa)G^2}{\alpha n\mu^2})$*. By the conversion from pure DP to Gaussian DP (Lemma 15), the bound for pure DP directly implies a DP-ERM learner under Gaussian DP with suboptimality* $\tilde{\mathcal{O}}(\frac{d^2G^2}{\alpha n\mu^2})$*. However, we obtain an improved dimension dependence and achieve the bound of* $\tilde{\mathcal{O}}(\frac{dG^2}{\alpha n\mu^2})$ *under GDP with the same runtime.*

Comparing to the existing work. To our knowledge, no existing pure-DP or pure-Gaussian DP learner achieves the optimal strongly convex rate in nearly linear time. The closest is Noisy Gradient Descent, which runs in $\widetilde{\mathcal{O}}(n)$ time for strongly convex and smooth problems (runs in $\kappa\log n$ iterations) but is a factor of $\kappa\log n$ worse in the excess empirical risk it obtains. An alternative parameter regime that runs Noisy Gradient Descent for $\mathcal{O}(n^2)$ iterations (i.e., $\mathcal{O}(n^3)$ time) achieves the optimal rate (without additional κ or $\log n$ dependence) albeit much slower and does not take advantage of the smoothness. We include more information about this in Appendix K.

Other works either operate in a different setting or do not apply to all problems that we consider. For example, objective perturbation with approximate minima perturbation Iyengar et al. (2019) runs in $\widetilde{\mathcal{O}}(n)$ time but applies only to generalized linear losses. Feldman et al. (2020) obtain nearly linear time private learner using a "privacy amplification by iteration" technique, but it works under the DP Stochastic Convex Optimization setting, which is different from DP-ERM that we are considering. The same algorithm for DP-ERM requires $\widetilde{\mathcal{O}}(n^2)$ time still. In addition, (Feldman et al., 2020) focused on (ε, δ)-DP too. The sampler from (Gopi et al., 2022) (without combining with Localized-ASAP) runs in $\widetilde{\mathcal{O}}(\frac{n}{\alpha}\log(d/\delta))$ time but becomes vacuous when we want $\delta = 0$ for either pure-DP or pure Gaussian DP.

5 Conclusion

Our proposed sampler Approximate SAample Perturbation (abbr. ASAP) perturbs an MCMC sample with noise proportional to its Wasserstein-∞ distance from a reference distribution that satisfies *pure* DP or *pure* Gaussian DP (i.e., $\delta = 0$). We show that our sampler obtains the first nearly linear-time end-to-end algorithm that achieves the optimal rate in the DP-ERM problem with strongly convex and smooth losses. The new techniques we developed might be of independent interest elsewhere that rely on approximate sampling.
Limitations. While the posterior sampling mechanism is known to achieve optimal rates in the general convex Lipschitz loss cases under pure-DP and pure-Gausssian DP (Gopi et al., 2022), our techniques are insufficient for obtaining optimal rates with faster computation in these broader families of DP-ERM problems than what is known in the literature. It remains an intriguing open problem to characterize the optimal computational complexity when we do not have either smoothness or strong convexity.

ACKNOWLEDGEMENTS

The research is partially supported by the NSF awards: SCALE MoDL-2134209, CCF-2112665 (TILOS), as well as CNS-2048091 (CAREER). It is also supported by the U.S. Department of Energy, the Office of Science, the Facebook Research Award, a Google Research Scholar Award, as well as CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics.

YL acknowledges Geelon So and Yuxing Huang for discussing the proofs in the paper. YW appreciates early conversations with Eric Vigoda, Zongchen Chen, and Daniel Štefankovič at a 2022 Summer School on MCMC in UCSB. Finally, the authors thank Jinshuo Dong for communicating Lemma 15 to us.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Jason Altschuler and Kunal Talwar. Resolving the mixing time of the langevin algorithm to its stationary distribution for log-concave sampling. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2509–2510. PMLR, 2023.

David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pp. 156–163, 1991.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *arXiv preprint arXiv:2206.07136*, 2022.

George Casella, Christian P Robert, and Martin T Wells. Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, pp. 342–347, 2004.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34:14771–14781, 2021.

Etienne De Klerk and Monique Laurent. Comparison of lasserre's measure-based bounds for polynomial optimization to bounds obtained by simulated annealing. *Mathematics of Operations Research*, 43(4): 1317–1325, 2018.

Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikateirni Mitrokotsa, and Benjamin IP Rubinstein. Differential privacy for bayesian inference through posterior sampling. *Journal of machine learning research*, 18(11):1–39, 2017.

Jinshuo Dong, David Durfee, and Ryan Rogers. Optimal differential privacy composition for exponential mechanisms. In *International Conference on Machine Learning*, pp. 2597–2606. PMLR, 2020.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20:1–42, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532. IEEE, 2018.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Joseph Geumlek, Shuang Song, and Kamalika Chaudhuri. Renyi differential privacy mechanisms for posterior sampling. *Advances in Neural Information Processing Systems*, 30, 2017.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, pp. 1948–1989. PMLR, 2022.

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)*, pp. 299–316. IEEE, 2019.

Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Global linear convergence of newton's method without strong-convexity or lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116:20881–20885, 2019.

Oren Mangoubi and Nisheeth Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees. *Advances in Neural Information Processing Systems*, 35:12633–12646, 2022.

Eric Mazumdar, Aldo Pacchiano, Yi-An Ma, Michael Jordan, and Peter Bartlett. On approximate Thompson sampling with Langevin algorithms. *International Conference on Machine Learning*, pp. 6797–6807, 2020.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. *Advances in Neural Information Processing Systems*, 29, 2016.

Darakhshan J Mir. Information-theoretic foundations of differential privacy. In *Foundations and Practice of Security: 5th International Symposium, FPS 2012, Montreal, QC, Canada, October 25-26, 2012, Revised Selected Papers 5*, pp. 374–381. Springer, 2013.

Rachel Redberg, Antti Koskela, and Yu-Xiang Wang. Improving the privacy and practicality of objective perturbation for differentially private linear learners. *Advances in Neural Information Processing Systems*, 2023.

Théo Ryffel, Francis Bach, and David Pointcheval. Differential privacy guarantees for stochastic gradient langevin dynamics. *arXiv preprint arXiv:2201.11980*, 2022.

Jeremy Seeman, Matthew Reimherr, and Aleksandra Slavković. Exact privacy guarantees for markov chain implementations of the exponential mechanism with artificial atoms. *Advances in Neural Information Processing Systems*, 34:13125–13136, 2021.

Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pp. 2493–2502. PMLR, 2015.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pp. 4782–4817. PMLR, 2022.

Appendix

Table of Contents

A Additional Preliminaries **14**

B Localization Techniques **14**

C Parameters in Localized-ASAP **15**

D Experiments **16**
D.1 Theoretical lower bounds 16
D.2 Empirical Risks on Real Datasets 16

E Supplementary Discussion on Wasserstein-Infinity Distance **18**

F Proof the Conversion Lemma 8 **19**
F.1 A Key Lemma: Another Characterization of W_∞ Distance 19
F.2 Proof of Lemma 8 20

G Proofs of Theorem 1 and Corollary 10 **22**

H Proof of Theorem 2 **26**

I Proof of Theorem 3 **26**

J Deferred Proofs of the Supporting Lemmas and Corollaries **30**
J.1 Proof of Lemma 9 and Corollary 20 30
J.2 Proof of Lemma 18 and Lemma 19 31
J.3 Proof of Lemma 14 32
J.4 Proof of Lemma 5 34
J.5 Proof of Lemma 15 35
J.6 Proof of Lemma 22 35

K Facts about Noisy Gradient Descent **38**

A Additional Preliminaries

Lemma 14 (Adaptive Composition)**.** *Let $\mathcal{M}_1 : \mathcal{X}^* \to \Delta^{\Theta_1}$ satisfies ε_1-DP (or μ_1-GDP) and $\mathcal{M}_2 : \Theta_1 \times \mathcal{X}^* \to \Delta^{\Theta_2}$: satisfies ε_2-DP (or μ_2-GDP) with probability 1 when its first input $\theta_1 \sim \mathcal{M}_1(D)$ for any dataset $D \in \mathcal{X}^*$, then $\mathcal{M}_1 \times \mathcal{M}_2$ satisfies $(\varepsilon_1 + \varepsilon_2)$-DP (or $\sqrt{\mu_1^2 + \mu_2^2}$-GDP).*

The composition theorem traditionally requires $\mathcal{M}_2$ to satisfy DP (or GDP) for all $\theta_2 \in \Theta_2$. What is stated above is slightly weaker in the sense that we can ignore a measure 0 set. We defer the proof to Appendix J.3.

Lemma 15 (Conversion between pure DP and GDP)**.** *With Φ denoting the cumulative distribution function of the standard normal, an ε-pure DP mechanism also satisfies μ-GDP with*

$$\mu = 2\Phi^{-1}\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}\right).$$

We defer the proof to Appendix J.5.

Definition 16 (Exponential Mechanism McSherry & Talwar (2007))**.** *For a quality score u, the exponential mechanism $\mathcal{M}_u : \mathcal{X} \to \Delta^{\Theta}$ samples an outcome $\theta \in \Theta$ with probability proportional to $\exp\left(\frac{\varepsilon u(x,\theta)}{2\Delta u}\right)$, and is ε-DP, where Δu is the global sensitivity.*

B Localization Techniques

The choice for the algorithm for providing θ_0 and the associated B parameter in Algorithm 3 is delicate. We construct an appropriate algorithm to find θ_0 in this section. In particular, we use approximate output perturbation.

For the general Lipschitz loss setting, output perturbation does not quite work as it does not even achieve minimax rate. But if individual loss function ℓ satisfies α-strongly convexity, Bassily et al. (2014) showed that one can obtain a localization using output perturbation that qualifies such that the downstream exponential mechanism on the localized set is optimal. The key observation is that under strong convexity, the sensitivity of $\theta^* = \arg\min_\theta \sum_{i=1}^n \ell_i(\theta)$ can be shown to be $G/(\alpha n)$ and output perturbation satisfies pure-DP, also it satisfies (with high probability)

$$\|\theta_0 - \theta^*\| \le \frac{2Gd\log(\cdot)}{\alpha n \varepsilon}.$$

This order suffices for the posterior sampling to achieve the optimal rate.

We provide the following Approximate Output Perturbation algorithm, wherein we introduce perturbations to the approximate optimizer.

Algorithm 4 Approximate Output Perturbation

1: **Input:** individual losses $\{\ell_i\}_{i=1}^n$ satisfying G-Lipschitz continuity and β-smoothness; strong convexity parameter α. Privacy parameter μ for GDP (or ε for pure DP).
2: Denote $\mathcal{L}(\theta) := \sum_{i=1}^n \ell_i(\theta)$, and $\theta^* := \arg\min_\theta \mathcal{L}(\theta)$. Set $\tilde{\Delta} := \frac{2\tau}{n} + \frac{2G}{\alpha n}$.
3: Solve for θ_{opt} that satisfies $\|\theta_{\text{opt}} - \theta^*\|_2 \le \frac{\tau}{n}$. ▷ Instantiation: Gradient Descent or Newton Descent.
4: Output $\theta_0 = \theta_{\text{opt}} + \boldsymbol{Z}$, where $\boldsymbol{Z}_i \sim \text{Lap}(\frac{\sqrt{d}\tilde{\Delta}}{\varepsilon})$ for ε-pure DP ($\boldsymbol{Z}_i \sim \mathcal{N}(0, \frac{\tilde{\Delta}^2}{\mu^2})$ for μ-GDP).

Remark 17. *For step 3, we make the optimization oracle a black-box algorithm that enjoys a linear convergence rate. Instantiation of the optimization oracle could be the Gradient Descent or the Newton Descent, which enjoy a linear convergence rate under the assumptions of strong convexity and Lipschitz*

smoothness (or c-stable Hessian by the Theorem 2 of Karimireddy et al. (2018)). Note that $\|\tilde{\theta} - \theta^*\|_2 \leq \frac{\tau}{n}$ is not a stopping criteria of the optimization algorithm. Instead, it is a property that $\tilde{\theta}$ would fulfill under the stopping criteria the additional assumption on $\mathcal{L}$.

The following lemma provides the bound for the sensitivity of the minimizer θ^*, which helps bound the sensitivity of the approximate minimizer $\tilde{\theta}$.

Lemma 18. *Let $D \in \mathcal{X}^*$. Let $D' \in \mathcal{X}^*$ be a neighbor such that $\mathcal{L}_{D'} - \mathcal{L}_D = \pm \ell_x$ for some $x \in \mathcal{X}$, then*

$$\|\theta^*(D) - \theta^*(D')\|_2 \leq \frac{2G}{\alpha n}.$$

With Lemma 18, we can bound the sensitivity of $\tilde{\theta}$, thus providing the following differential privacy guarantees.

Lemma 19. *Algorithm 4 satisfies ε-differential privacy (or μ-Gaussian DP).*

We defer the proofs of Lemma 18 and 19 to Appendix J.2.

C Parameters in Localized-ASAP

The choice of γ and λ. Larger γ provides a better utility guarantee but a weaker privacy guarantee. We choose γ by the following fact so that the posterior sampling mechanism $\theta \sim p^*(\theta)$ satisfies ε-pure DP or μ-GDP.

Fact 1 (Optimal rates for strongly convex problems)**.** *The optimal (expected) excess empirical risks for G-Lipschitz continuous and α-strongly convex (individual) losses is $\frac{d^2G^2}{\alpha n \varepsilon^2}$ and $\frac{dG^2}{\alpha n \mu^2}$ for ε-DP and μ-GDP respectively. For posterior sampling, choosing*

$$\gamma = \frac{\mu^2 \alpha n}{G^2}, \quad \lambda = 0$$

yields the optimal rate under μ-GDP. Unfortunately, there are no parameter choices for γ, λ that can yield the optimal rate under pure-DP, unless Θ is already localized such that it satisfies that $\mathrm{Diam}(\Theta) \leq \frac{dG}{\alpha n \varepsilon}$. Then the choice of $\gamma = \frac{\varepsilon}{G \cdot \mathrm{Diam}(\Theta)}$ works.

The choice of B (the radius of the localized domain). For desirable the utility and computational guarantees in sampling, we hope that B is large enough such that the mode θ^* and the mean $\bar{\theta}$ are in the localized domain Θ, i.e., $\theta^* \in \Theta$ and $\bar{\theta} \in \Theta$, where $\theta^* := \arg\min_{\theta \in \mathbb{R}^d} J(\theta)$, and $\bar{\theta} := \mathbb{E}_{\theta \sim \pi(\theta) \propto e^{-\gamma J(\theta)}} \theta$. Moreover, Assumption 3 requires $\mathbb{B}(\theta^*, R_1) \subset \Theta$ for efficient constrained MALA, where $R_1 \geq 8\sqrt{\frac{d}{\gamma \alpha n}}$. However, by Corollary 20 and the discussion after Theorem 1, we know that an excessively large B should be avoided. This is because a large B leads to a reduced $p_{\min}$, which in turn necessitates a greater number of steps for constrained MALA to achieve the desired accuracy. Also, note that by Fact 1, for ε-pure DP guarantee, we set $\gamma = \frac{\varepsilon}{2GB}$, which is dependent of B.

In pure DP case, suppose θ_0 is the output of Alogrithm 4, then we have $\theta_0 = \theta_{\mathrm{opt}} + \boldsymbol{Z}$, where $Z_i \sim \mathrm{Lap}(2\sqrt{d}\left(\tau + \frac{G}{\alpha}\right)/n\varepsilon)$, and $\|\theta_{\mathrm{opt}} - \theta^*\|_2 \leq \frac{\tau}{n}$. Therefore we have

$$\|\theta_0 - \theta^*\| \leq \|\theta_0 - \theta_{\mathrm{opt}}\|_2 + \|\theta_{\mathrm{opt}} - \theta^*\|_2 = \|\boldsymbol{Z}\|_2 + \|\tilde{\theta} - \theta^*\|_2 \leq \frac{2d\left(G + \tau\alpha\right)\ln(d/\rho)}{n\alpha\varepsilon} + \frac{\tau}{n},$$

with probability $1 - \rho$. In the last inequality, we apply the fact that for $Y_i \overset{\text{iid}}{\sim} \mathrm{Lap}(\lambda)$, we have

$$\mathbb{P}\left(\|\boldsymbol{Y}\|_2 \geq \sqrt{d}\lambda \ln(d/\rho)\right) = \mathbb{P}\left(\sum_{i=1}^d Y_i^2 \geq d(\lambda \ln(d/\rho))^2\right) \leq \sum_{i=1}^d \mathbb{P}\left(|Y_i| \geq \lambda \ln(d/\rho)\right) \leq \rho.$$

In order to fulfill the condition $\mathbb{B}(\theta^*, R_1) \subset \Theta$, where $R_1 = 8\sqrt{\frac{d}{\gamma\alpha n}}$, note that $\gamma = \frac{\varepsilon}{2GB}$, the following inequality should hold:

$$B - 8\sqrt{\frac{2dGB}{\varepsilon\alpha n}} \geq \frac{2d\,(G+\tau\alpha)\ln(d/\rho)}{n\alpha\varepsilon} + \frac{\tau}{n}. \tag{2}$$

The above inequality is satisfied with $B \geq \frac{8(32dG+(\alpha\tau+G)d\ln(d/\rho)+\tau\varepsilon\alpha)}{\alpha n\varepsilon}$. For the sake of simplicity, we let τ be sufficiently small and set $B = \frac{C_1 dG\ln(d/\rho)}{\alpha n\varepsilon}$.

In GDP case, similarly, we have $\theta_0 = \theta_{\text{opt}} + \boldsymbol{Z}$, where $Z_i \sim \mathcal{N}(0, 4(\tau + \frac{G}{\alpha})^2/n^2\mu^2)$, and $\|\theta_{\text{opt}} - \theta^*\|_2 \leq \frac{\tau}{n}$. Therefore we have

$$\|\theta_0 - \theta^*\| \leq \|\theta_0 - \theta_{\text{opt}}\|_2 + \|\theta_{\text{opt}} - \theta^*\|_2 \leq \|\boldsymbol{Z}\|_2 + \frac{\tau}{n} \leq \frac{2\sqrt{2}(\tau\alpha + G)(\sqrt{d} + \sqrt{\ln(1/\rho)})}{\alpha n\mu} + \frac{\tau}{n}, \tag{3}$$

with probability $1-\rho$. In the last inequality where we use the tail bound for chi-square distributions by Lemma 1 in (Laurent & Massart, 2000). Again we need $B - R_1 \geq \|\theta_0 - \theta^*\|$, where $R_1 = 8\sqrt{\frac{d}{\gamma\alpha n}}$ and $\gamma = \frac{\mu^2\alpha n}{G^2}$. This is satisfied with $B \geq \frac{2\sqrt{2}(\tau\alpha+G)(\sqrt{d}+\sqrt{\ln(1/\rho)})+8G\sqrt{d}+\tau\alpha\mu}{\alpha n\mu}$. Again, we let τ be sufficiently small and set $B = \frac{C_2 G(\sqrt{d}+\sqrt{\ln(1/\rho)})}{\alpha n\mu}$

Calculation of $p_{\min}$ Lemma 9 implies that the choices of $p_{\min}$ in Algorithm 3 are valid lower bounds of the density function. We have the following corollary.

Corollary 20. *Consider probability with density*

$$p(\theta) \propto \exp\left(-\gamma\left(\sum_{i=1}^n \ell_i(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|^2\right)\right)\mathbf{1}\{\|\theta - \theta_0\| \leq B\}.$$

1. *If ℓ_i is G-Lipschitz for all i, then* $\min_\theta p(\theta) \geq e^{-\gamma(2nGB+2\lambda B^2)}\frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}B^d}$.
2. *If ℓ_i is convex and β-smooth for all i, then for all $\tilde{\theta} \in \Theta$, we have*

$$\min_\theta p(\theta) \geq e^{-\gamma\left(2B\|\nabla J(\tilde{\theta})\|+2(n\beta+\lambda)B^2\right)} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}B^d}.$$

Applying Corollary 20, we calculate $p_{\min}$ as follows: Take any θ' such that $\|\theta' - \theta_0\|_2 \leq B$. Compute

$$p_{\min} \leftarrow e^{-\gamma\min\left\{2nGB+2\lambda B^2,\, 2\|\nabla L_\lambda(\theta')\|B+2(n\beta+\lambda)B^2\right\}}\frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}B^d}.$$

D EXPERIMENTS

D.1 THEORETICAL LOWER BOUNDS

We visualize the excess empirical risks in Figure 1and demonstrate that ε-(pure) DP can outperform (ε, δ)-DP in theory, especially when the sample size n is large.

D.2 EMPIRICAL RISKS ON REAL DATASETS

Setup. We experiment on two real datasets: Red Wine Quality and White Wine Quality from UCI repository (`https://archive.ics.uci.edu/dataset/186/wine+quality`). The Red Wine Quality dataset contains 1599 samples with 11 features. We standardize the data and learn a regression task via $\ell_i(\theta) = \frac{1}{2}(x_i^T\theta - y_i)^2 + \frac{\alpha}{2}\|\theta\|^2$, which is a strongly convex problem under $\alpha = 100$ and $\rho = 0.01$. The

Table 1: Summary of our results for the αn-strong convex DP-ERM problem with G-Lipschitz and β-smooth losses and that the total loss satisfies αn-strong convexity. The utility measures the expected excess empirical (total) risk. The $\mathrm{polylog}(n)$ factors are ignored from the computation. Observe that we are the first algorithms that achieve the nearly linear time computation under pure-DP and pure-Gaussian DP while still obtaining the optimal excess empirical risk.

	Privacy	Excess Empirical risk	Computation
NoisyGD (see Appendix K)	μ-GDP	$dG^2/\alpha n\mu^2$	n^3
NoisySGD (Bassily et al., 2014)	(ε, δ)-DP	$dG^2 \log(1/\delta)/\alpha n\varepsilon^2$	n^2
PosteriorSample (Gopi et al., 2022)	δ-approx μ-GDP	$dG^2/\alpha n\mu^2$	n/α
ExponentialMechanism (Bassily et al., 2014)	ε-DP	$d^2G^2/\alpha n\varepsilon^2$	n^4
Localized-PS-ASAP (This paper)	μ-GDP	$dG^2/\alpha n\mu^2$	n/α^2 or $n/\alpha^{5/2}$
Localized-EM-ASAP (This paper)	ε-DP	$d^2G^2/\alpha n\varepsilon^2$	n/α^2 or $n/\alpha^{5/2}$

Table 2: The choice of $\gamma, B, \lambda, \theta_0, \Delta$ when instantiating Algorithm 3 for pure DP or Gaussian DP learning.The choices of γ ensure the DP and GDP guarantee respectively and the choice of B ensures sufficient localization such that the sampling Algorithm 1 run in $\tilde{O}(n)$ time (condition on the high-prob event that θ^* is inside the localized set Θ. Polynomial dependence in d and other parameters are hidden in $\tilde{O}$). For pure-DP, the choice of B has the additional purpose of achieving the optimal rate for DP-ERM — $\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) = O(\frac{d^2G^2}{\alpha n\varepsilon^2})$. For GDP, the choice of B only affects computation. The optimal rate $\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) = O(\frac{dG^2}{\alpha n\mu^2})$ is always attained.

	γ	B	λ	θ_0	Δ
ε-Pure DP	$\frac{\varepsilon}{2GB}$	$\frac{C_1 Gd\ln(d/\rho)}{\alpha n\varepsilon}$	0	Algorithm 4	$\frac{dG\ln(d/\rho)}{2n^2\alpha\varepsilon}$
μ-GDP	$\frac{\mu^2\alpha n}{G^2}$	$\frac{C_2 G(\sqrt{d}+\sqrt{\ln(1/\rho)})}{\alpha n\mu}$	0	Algorithm 4	$\frac{\sqrt{d}G}{\sqrt{2}n^2\alpha\mu}$

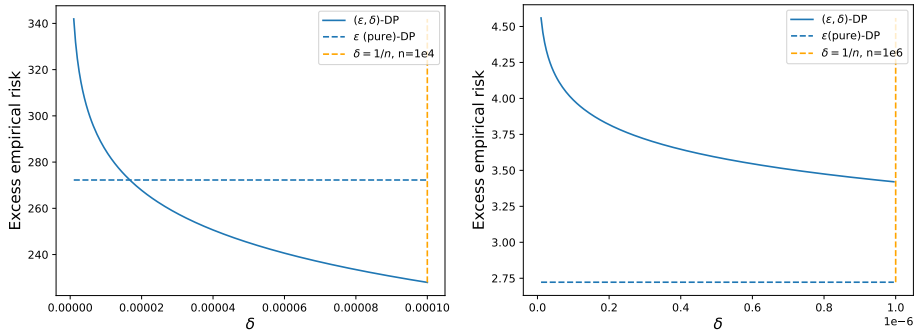

Figure 1: Excess empirical risks from Table 1 for strongly convex losses. Here $d = 11, G = 300, \alpha = 4, \varepsilon = 1$. Left $n = 1e4$. Right $n = 1e6$.

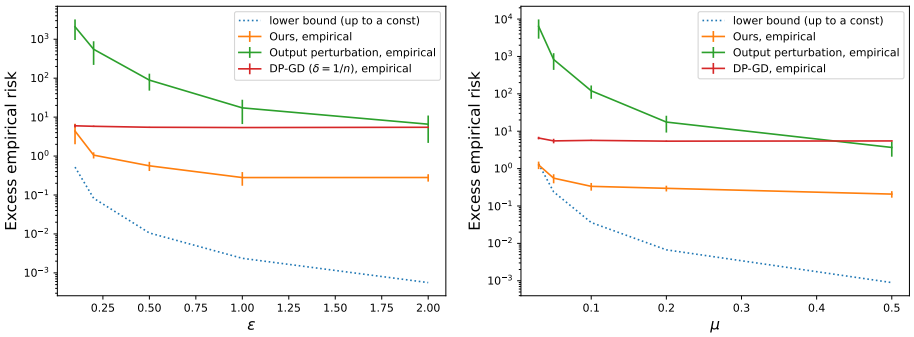

Figure 2: Excess empirical risks for strongly convex losses on Wine Quality – Red dataset.

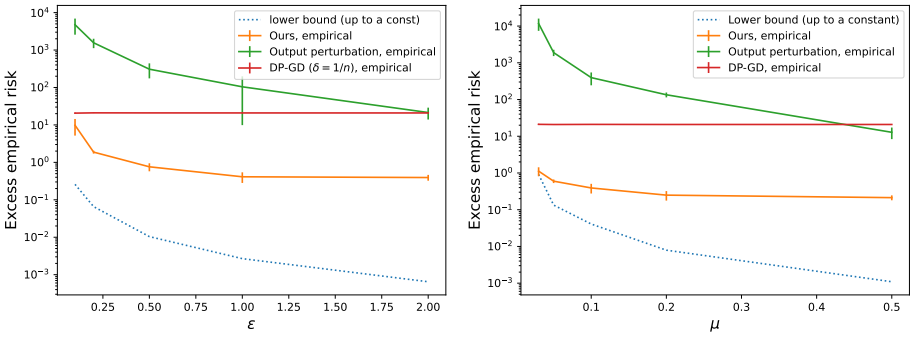

Figure 3: Excess empirical risks for strongly convex losses on Wine Quality – White dataset.

White Wine Quality dataset contains 4898 samples with 11 features. We standardize the data and learn a regression task via $\ell_i = \frac{1}{2}(x_i^T\theta - y_i)^2 + \frac{\alpha}{2}\|\theta\|^2$, which is a strongly convex problem under $\alpha = 32$ and $\rho = 0.01$.

Results. We compare with the theoretical lower bounds in Table 1 (up to a constant), the output perturbation in Algorithm 4, and DP-GD with per-sample gradient clipping (no need to set the clipping threshold) by Bu et al. (2022). The experiments are conducted under both ε-DP and μ-GDP. We consistently observe that our algorithms outperform the existing ones.

E Supplementary Discussion on Wasserstein-Infinity Distance

Building on the insights from Section 3.1, we provide Figure 4 to enhance intuition on W_∞ distance. The visualization demonstrates that the optimal coupling is highly concentrated near the line $y = x$ and sparser in other areas compared to the independent coupling. Notably, Feldman et al. (2018) also explore this "norm-aware" statistical distance and establish the shift reduction lemma (Lemma 20, (Feldman et al., 2018)).

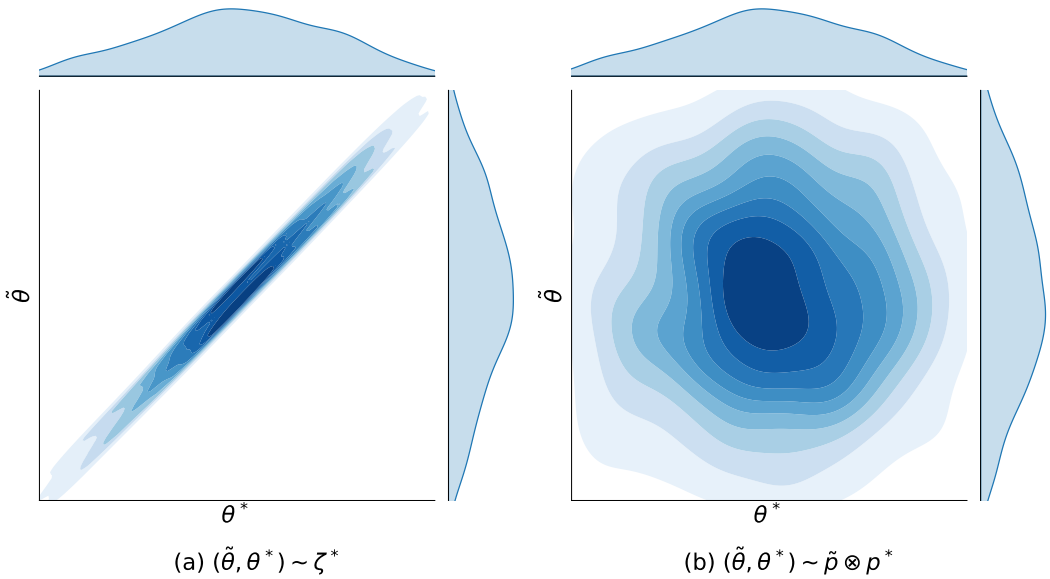

Figure 4: Two examples illustrating the couplings of $\tilde{p}$ and p^*. In figure (a), when $(\tilde{\theta}, \theta^*)$ follows the optimal coupling of $W_\infty(\tilde{p}, p^*)$, $(\tilde{\theta}, \theta^*)$ is confined within the band $|\tilde{\theta} - \theta^*| \leq \Delta$ with a probability of 1. Conversely, in figure (b), when $\tilde{\theta}$ and θ^* are independent, $|\tilde{\theta} - \theta^*|$ can take relatively large values with positive probability.

F Proof the Conversion Lemma 8

F.1 A Key Lemma: Another Characterization of W_∞ Distance

To establish Lemma 8, we commence by furnishing an equivalent characterization for the W_∞ distance. We also demonstrate the attainability of the infimum in Lemma 22. Our exposition begins with the introduction of the concept of tightness.

Definition 21 (Tightness)**.** *A probability measure P on Θ is called tight if for any $\epsilon > 0$, there exists a compact set $K \subset \Theta$, such that*

$$P(\Theta \setminus K) \leq \epsilon,$$

Lemma 22. *Let (Θ, dist) be a complete separable metric space. Let P and Q be two tight probability measures on Θ. The Wasserstein-∞ distance has the following characterization:*

$$W_\infty(P, Q) = \inf\{r > 0 : P(U) \leq Q(U^r), \textit{ for all open subsets } U \subset \Theta\},$$

where the r-expansion of U is denoted by $U^r := \{x \in \Theta : \mathrm{dist}(x, U) \leq r\}$.

Furthermore, the infimum in the Definition 7 can be attained, i.e., there exists $\zeta^* \in \Gamma(P, Q)$ such that

$$W_\infty(P, Q) = \operatorname*{ess\,sup}_{(x,y) \in (\Theta \times \Theta, \zeta^*)} \mathrm{dist}(x, y) = \inf_{\zeta \in \Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y) \in (\Theta \times \Theta, \zeta)} \mathrm{dist}(x, y).$$

The proof of this lemma is presented in Appendix J.6.

F.2 PROOF OF LEMMA 8

Proof of Lemma 8 First note that $\|x-y\|_1 \le \sqrt{d}\|x-y\|_2$, for all x, y. Therefore, it suffices to prove the lemma when dist is the ℓ_2 metric.

Fix Δ and set $\xi = p_{\min} \cdot \frac{\pi^{d/2}}{2^{d+1}\cdot\Gamma(d/2+1)}\Delta^d$, so that $d_{\mathrm{TV}}(P,Q) < \xi$.

To prove the result that $W_\infty \le \Delta$, we use the equivalent definition of W_∞ in Lemma 22. By this definition, to prove $W_\infty(P,Q) \le \Delta$, it suffices to show that

$$P(A) \le Q(A^\Delta), \text{ for all open set } A \subseteq \Theta,$$

where we re-define $A^\Delta := \{x \in \mathbb{R}^d \mid \mathrm{dist}(x, A) \le \Delta\}$. Note that by this definition, A^Δ might extend beyond Θ. However, we still have $Q(A^\Delta) = Q(A^\Delta \cap \Theta)$, since Q is supported on Θ.

Note that if $Q(A^\Delta) = 1$, it is obvious that $P(A) \le Q(A^\Delta)$. When A is an empty set, the proof is trivial. So we only consider nonempty open set $A \subseteq \Theta$ with $Q(A^\Delta) < 1$.

Note that for an arbitrary open set $A \subseteq \Theta$, we have

$$P(A) \le Q(A) + d_{TV}(P,Q) < Q(A) + \xi.$$

Thus to prove $P(A) \le Q(A^\Delta)$, it suffices to prove $Q(A) + \xi \le Q(A^\Delta)$, i.e., $Q(A^\Delta \setminus A) \ge \xi$.

To prove $Q(A^\Delta \setminus A) \ge \xi$, we construct an ℓ_2-ball $U \subset \mathbb{R}^d$ of radius $\Delta/2$ satisfying the properties:

1. U is contained in the set $A^\Delta \setminus A$, i.e., $U \subset A^\Delta \setminus A$, and
2. $Q(U) = Q(U \cap \Theta) \ge \xi$

Notice that we only consider nonempty open set $A \subseteq \Theta$ with $Q(A^\Delta) < 1$, we construct U as follows.

Intuition and Overview of the construction of U. The construction of the ball U involves determining two pivotal points, y_Δ and y_A, located at the "boundaries" of A^Δ and A, respectively. We then define U as $U := \mathbb{B}_{\text{dist}}\left(\frac{y_\Delta + y_A}{2}, \frac{\mathrm{dist}(y_\Delta, y_A)}{2}\right)$. Note that by the definition of A^Δ, $\mathrm{dist}(y_\Delta, y_A) \ge \Delta$. To guarantee Property 1 (i.e., $U \subset A^\Delta \setminus A$), y_A is chosen such that $\mathrm{dist}(y_\Delta, y_A) = \Delta$. Additionally, to ensure Property 2 (i.e., $Q(U) \ge \xi$), since $Q(U) = Q(U \cap \Theta)$, our aim is to maximize the "size" of $U \cap \Theta$. Therefore, we specifically require y_Δ to belong to Θ, while noting that A^Δ might extend beyond Θ. The detailed selection process of y_Δ and y_A, and the subsequent construction of U, are presented below.

Existence and Construction of y_Δ. To find the above discussed $y_\Delta \in \Theta$, we first show that:

$$\text{If } Q(A^\Delta) < 1, \text{ then } \partial(A^\Delta) \cap \Theta \ne \emptyset,$$

where we re-define the boundary of A^Δ as $\partial(A^\Delta) := \{x \in \mathbb{R} \mid \mathrm{dist}(x, A) = \Delta\}$.

We prove it by contradiction. If instead $\partial(A^\Delta) \cap \Theta = \emptyset$, then for all $x \in \Theta$, $\mathrm{dist}(x, A) \ne \Delta$. Due to the continuity of dist and the convexity of Θ, we know that only one of these two statements holds:

- $\mathrm{dist}(x, A) < \Delta$, for all $x \in \Theta$.
- $\mathrm{dist}(x, A) > \Delta$, for all $x \in \Theta$.

Since $\emptyset \ne A \subseteq \Theta$, there exist $x' \in A \in \Theta$, such that $\mathrm{dist}(x', A) = 0$. Therefore, the first statement holds. Thus $\Theta \subseteq A^\Delta$, which contradicts to $Q(A^\Delta) < 1$.

Therefore, $\partial(A^\Delta) \cap \Theta \ne \emptyset$. Then there exists $y_\Delta \in \partial(A^\Delta) \cap \Theta$. Thus there exist $y_\Delta \in \Theta$, such that $\mathrm{dist}(y_\Delta, A) = \Delta$.

Construction of y_A. Note that by the definition of $\mathrm{dist}(\cdot, A)$ and that $\mathrm{dist}(y_\Delta, A) = \Delta$, there exist $y_A \in \overline{A}$ (the closure of A), such that $\mathrm{dist}(y_\Delta, y_A) = \Delta$.

Construction of U. Let $y_0 = \frac{y_\Delta + y_A}{2}$ and let U be the closed ball $U = \mathbb{B}(y_0, \Delta/2) := \{x \in \mathbb{R}^d \mid \mathrm{dist}(x, y_0) \leq \Delta/2\}$. We now prove that $U \subseteq A^\Delta \setminus A$ and that $Q(U) \geq \xi$, and then the proof is concluded.

1. To prove $U \subset A^\Delta \setminus A$, let $x \in U$. We show that $x \in A^\Delta$ and $x \notin A$. Since $x, y_\Delta, y_A \in U$ and U is a ball with diameter Δ, we have

$$\mathrm{dist}(x, y_\Delta) \leq \Delta, \text{ and } \mathrm{dist}(x, y_A) \leq \Delta \tag{4}$$

   - ($x \notin A$). If $x \in A$, since A is an open set and $\mathrm{dist}(y_\Delta, A) = \Delta$, we have that $\mathrm{dist}(x, y_\Delta) > \Delta$, which contradicts to (4). Therefore $x \notin A$.
   - ($x \in A^\Delta$). Note that $y_A \in \bar{A}$, thus $\mathrm{dist}(x, A) \leq \mathrm{dist}(x, y_A) \overset{(4)}{\leq} \Delta$, implying that $x \in A^\Delta$.

2. To prove $Q(U) \geq \xi$, note that we assume the domain Θ to be an ℓ_2 ball. Since $y_A, y_\Delta \in \Theta$, by Lemma 24, we know that $\mathrm{Vol}(U \cap \Theta) \geq \frac{1}{2}\mathrm{Vol}(U)$. Thus

$$Q(U) = Q(U \cap \Theta) \geq p_{\min}\mathrm{Vol}(U \cap \Theta) \geq \frac{1}{2}p_{\min}\mathrm{Vol}(U) = \xi,$$

   which completes the proof.

Remark 23. *It is worth noting that Lemma 8 and Lemma 24 extend naturally to open balls. The necessary modifications in the proof involve changing certain inequalities: "$\leq$" becomes "$<$", and "$>$" becomes "$\geq$."*

■

Lemma 24. *Suppose $\theta \in \mathbb{R}^d$, and $R > 0$. If there exists $a \in \mathbb{R}^d$ such that $a \in \mathbb{B}_{\ell_2}(\theta, R)$, and that $-a \in \mathbb{B}_{\ell_2}(\theta, R)$. Then*

$$\frac{\mathrm{Vol}\left(\mathbb{B}_{\ell_2}(0, \|a\|_2) \cap \mathbb{B}_{\ell_2}(\theta, R)\right)}{\mathrm{Vol}\left(\mathbb{B}_{\ell_2}(0, \|a\|_2)\right)} \geq \frac{1}{2}.$$

Proof Since $a, -a \in \mathbb{B}_{\ell_2}(\theta, R)$, we have

$$\|a - \theta\|_2^2 \leq R^2, \text{ and } \|-a - \theta\|_2^2 \leq R^2,$$

which implies

$$\|a\|_2^2 - 2\langle a, \theta\rangle + \|\theta\|_2^2 \leq R^2, \text{ and } \|a\|_2^2 + 2\langle a, \theta\rangle + \|\theta\|_2^2 \leq R^2.$$

Thus, $\|a\|_2^2 + \|\theta\|_2^2 \leq R^2$.

Denote $S = \mathbb{B}_{\ell_2}(0, \|a\|_2) \setminus \mathbb{B}_{\ell_2}(\theta, R)$. We first prove that $-S := \{-x \mid x \in S\} \subseteq \mathbb{B}_{\ell_2}(\theta, R)$. For an arbitrary $x \in S$, since $x \in \mathbb{B}_{\ell_2}(0, \|a\|_2)$, we have $\|x\|_2 \leq \|a\|_2$, thus $\|x\|_2^2 + \|\theta\|_2^2 \leq \|a\|_2^2 + \|\theta\|_2^2 \leq R^2$. On the other hand, since $x \notin \mathbb{B}_{\ell_2}(\theta, R)$, we have $\|x - \theta\|_2^2 > R^2$. Thus $\|-x - \theta\|_2^2 = 2\left(\|x\|_2^2 + \|\theta\|_2^2\right) - \|x - \theta\|_2^2 < R^2$, which implies $-x \in \mathbb{B}_{\ell_2}(\theta, R)$. Therefore $-S \subseteq \mathbb{B}_{\ell_2}(\theta, R)$.

Since $S = \mathbb{B}_{\ell_2}(0, \|a\|_2) \setminus \mathbb{B}_{\ell_2}(\theta, R)$ and $-S \subseteq \mathbb{B}_{\ell_2}(\theta, R)$, we know that $S \cap (-S) = \emptyset$. Therefore,

$$2\mathrm{Vol}(S) = \mathrm{Vol}(S) + \mathrm{Vol}(-S) = \mathrm{Vol}\left(S \cup (-S)\right) \leq \mathrm{Vol}\left(\mathbb{B}_{\ell_2}(0, \|a\|_2)\right).$$

Therefore

$$\mathrm{Vol}\left(\mathbb{B}_{\ell_2}(0, \|a\|_2) \cap \mathbb{B}_{\ell_2}(\theta, R)\right) = \mathrm{Vol}\left(\mathbb{B}_{\ell_2}(0, \|a\|_2)\right) - \mathrm{Vol}(S) \geq \frac{1}{2}\mathrm{Vol}\left(\mathbb{B}_{\ell_2}(0, \|a\|_2)\right).$$

■

G PROOFS OF THEOREM 1 AND COROLLARY 10

In this section, we provide the omitted proofs and supplementary facts for the Metropolis-adjusted Langevin algorithm (MALA) with constraint in Section 3.3.

We first provide the definition of mixing time. Given an error tolerance $\xi \in (0,1)$ and an initial distribution p_0, define the ξ-mixing time in total variation distance as

$$\tau\left(\xi, p^0, p^*\right) = \min\{k \mid \|p^k - p^*\|_{\mathrm{TV}} \leq \xi\}$$

For simplicity, we denote $L = n\beta$ and $m = n\alpha$ in this proof.

Proof of Theorem 1 We separate the proof of Theorem 1 into three parts via splitting of Algorithm 1 into the inner loop, the outer loop, and the case where the outer loop count reaches the maximum. We note that in the inner loop, Algorithm 1 is performing the MALA algorithm on the unconstrained space $\mathbb{R}^d$. In the outer loop, it performs a rejection sampling step that takes in independent samples from MALA and keeps rejecting samples outside of the domain Θ until one sample falls inside Θ. When the sample keeps falling outside Θ and the outer loop count reaches the maximum, the algorithm stops by outputting an arbitrary point $\theta \in \Theta$. We clarify the notations as follows.

Notations. Let π and p^* be the unconstrained and constrained Gibbs distribution respectively with density $\pi(\theta) \propto e^{-\gamma J(\theta)}$, and $p^*(\theta) \propto e^{-\gamma J(\theta)}\mathbf{1}(\theta \in \Theta)$. Let $\hat{p}^K$ be the distribution of the unconstrained MCMC sample after K inner-loop iterations. Let $p^K(\theta) \propto \hat{p}^K(\theta)\mathbb{1}\{\theta \in \Theta\}$ be the distribution of the accepted MCMC sample. Let p_{out} be the output distribution of Algorithm 1. Let p_{Fail} be the distribution within domain Θ, which is the output distribution when none of the $\tau_{\max}$ trials belong to Θ.

Overview. We first bound the error in the inner loop – the TV distance between the unconstrained MCMC distribution and the unconstrained Gibbs distribution: $d_{\mathrm{TV}}(\hat{p}^K, \pi)$. Next, we bound the error of the accepted sample – the TV distance between the constrained MCMC distribution and the target distribution: $d_{\mathrm{TV}}(p^K, p^*)$. Finally, we bound the TV distance between the output distribution and the target distribution: $d_{\mathrm{TV}}(p_{\mathrm{out}}, p^*)$.

For the inner-loop part, we cite the previous results. Consider the MALA algorithm on $\mathbb{R}^d$ with a step size of

$$h = \mathbf{\Theta}\left(\min\left\{\kappa^{-1/2}(\gamma L)^{-1}(\ln \beta_0/\xi)^{-1/2}, \frac{1}{\gamma L d}\right\}\right),$$

and an initial distribution μ^0 satisfying β_0-warmness assumption that $\sup_{A\subset\mathbb{R}^d}\left(\frac{\mu^0(A)}{\pi(A)}\right) \leq \beta_0$. Then by Theorem 1 of Dwivedi et al. (2019), we know that the MALA algorithm with the number of iterations

$$K = \mathbf{\Theta}\left(\ln\left(\frac{\beta_0}{\xi}\right)\max\left\{\kappa^{3/2}\sqrt{\ln \beta_0/\xi}, d\kappa\right\}\right),$$

converges to ξ accuracy in terms of the TV distance: $d_{TV}(\hat{p}^K, \pi) \leq \xi$. By Lemma 8 of Ma et al. (2019), we also know that on the unconstrained space, the Gaussian distribution $\mathcal{N}(0, \frac{1}{\gamma L}\mathbb{I})$ is $(2\kappa)^{d/2}$ warm with respect to the posterior distribution π with condition number κ. Plugging in this warmness constant, we obtain that

$$K = \mathbf{\Theta}\left((d\ln\kappa + \ln 1/\xi)\max\left\{\kappa^{3/2}\sqrt{d\ln\kappa + \ln 1/\xi}, d\kappa\right\}\right),$$

Next, for the outer-loop (rejection sampling) part, we note that if the unconstrained $\hat{p}^K$ and π satisfies $d_{TV}\left(\hat{p}^K(\theta), \pi(\theta)\right) \leq \xi$, then we also obtain a bound for $d_{TV}\left(p^K(\theta), p^*(\theta)\right)$, the TV distance between the constrained distributions $p^K(\theta) \propto \hat{p}^K(\theta)\mathbb{1}\{\theta \in \Theta\}$ and $p^*(\theta) \propto \pi(\theta)\mathbb{1}\{\theta \in \Theta\}$. Normalizing the

density gives that $p^*(\theta) = \frac{\pi(\theta)\mathbb{1}\{\theta\in\Theta\}}{\int \pi(\theta)\mathbb{1}\{\theta\in\Theta\}\mathrm{d}\theta}$. Applying the concentration argument above provides the bound: $\int \pi(\theta)\mathbb{1}\{\theta\in\Theta\}\mathrm{d}\theta \geq 1 - 2\exp\left(-\frac{\gamma m R_1^2}{32d}\right) \geq \frac{1}{2}$, where the last inequality follows by plugging in our choice of R_1. From the definition of the TV-distance, we know that

$$\begin{aligned}
&d_{TV}\left(p^K(\theta), p^*(\theta)\right) \\
&= \int_{\theta\in\Theta} |p^K(\theta) - p^*(\theta)|\mathrm{d}\theta \\
&= \int_{\theta\in\Theta} \left|\frac{\hat{p}^K(\theta)}{\int_{\theta\in\Theta}\hat{p}^K(\theta)\mathrm{d}\theta} - \frac{p^*(\theta)}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta}\right|\mathrm{d}\theta \\
&\leq \int_{\theta\in\Theta} \left|\frac{\hat{p}^K(\theta)}{\int_{\theta\in\Theta}\hat{p}^K(\theta)\mathrm{d}\theta} - \frac{\hat{p}^K(\theta)}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta}\right|\mathrm{d}\theta + \int_{\theta\in\Theta} \left|\frac{\hat{p}^K(\theta)}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta} - \frac{p^*(\theta)}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta}\right|\mathrm{d}\theta \\
&= \left|1 - \frac{\int_{\theta\in\Theta}\hat{p}^K(\theta)\mathrm{d}\theta}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta}\right| + \frac{1}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta}\int_{\theta\in\Theta}|\hat{p}^K(\theta) - \pi(\theta)|\mathrm{d}\theta \\
&\overset{\text{(i)}}{\leq} 2\frac{\xi}{\int_{\theta\in\Theta}\pi(\theta)\mathrm{d}\theta} \\
&\leq 4\xi,
\end{aligned}$$

where (i) follows from the definition of the TV-distance.

Replacing ξ by $\xi/8$, we have $d_{\mathrm{TV}}(p^K, p^*) \leq \xi/2$.

Then, we bound the "failure" probability – the probability that the outer-loop count reaches $\tau_{\max} = \ln(2/\xi)$, while the MCMC sample still falls outside Θ. For the rejection sampling part, we know that for each individual trial, the rejection probability equals the probability mass outside of Θ. Invoking Lemma 25 below, we obtain that for $\hat{\theta} \sim \pi$,

$$\mathbb{P}\left(\left\|\hat{\theta} - \mathbb{E}_{\theta\sim p(\theta)}[\theta]\right\| \geq r\right) \leq 2e^{-\frac{r^2\gamma m}{8d}},$$

and that for θ^* being the mode of π,

$$\left\|\mathbb{E}_{\theta\sim p(\theta)}[\theta] - \theta^*\right\|^2 \leq \frac{3}{\gamma m}.$$

On the other hand, $\mathbb{B}(\theta^*, R_1) \subset \Theta$. Since $R_1 \geq 8\sqrt{\frac{d}{\gamma m}} \geq 2\sqrt{\frac{3}{\gamma m}}$, we have $\mathbb{B}(\mathbb{E}_{\theta\sim p(\theta)}[\theta], R_1/2) \subset \mathbb{B}(\theta^*, R_1) \subset \Theta$. Hence, applying the concentration argument above for $\hat{\theta} \sim \pi$,

$$\mathbb{P}\left(\hat{\theta} \notin \Theta\right) \leq 2e^{-\frac{\gamma m R_1^2}{32d}}.$$

Combining with the definition of the TV distance, we know that for $\hat{\theta}^K \sim \hat{p}^K$ obtained by running the MALA algorithm (on $\mathbb{R}^d$) for K steps,

$$\mathbb{P}\left(\hat{\theta}^K \notin \Theta\right) \leq \mathbb{P}\left(\hat{\theta} \notin \Theta\right) + d_{TV}(\hat{p}^K, \pi) \leq 2e^{-\frac{\gamma m R_1^2}{32d}} + \xi.$$

Hence with $\tau_{\max}$ independent trials, the probability that none of the $\tau_{\max}$ trials belong to Θ is

$$\begin{aligned}
\mathbb{P}\left(\text{None of the } \tau_{\max} \text{ trials belong to } \Theta\right) = \left(\mathbb{P}\left(\hat{\theta}^K \notin \Theta\right)\right)^{\tau_{\max}} &\leq \left(2e^{-\frac{\gamma m R_1^2}{32d}} + \xi\right)^{\tau_{\max}} \\
&\leq \left(2e^{-2} + \xi\right)^{\tau_{\max}} \\
&\leq e^{-\tau_{\max}},
\end{aligned}$$

where we plug in $R_1 \geq 8\sqrt{\frac{d}{\gamma m}}$, and without loss of generality assume $\xi < 1/16 < e^{-1} - 2e^{-2}$. Therefore, by setting $\tau_{\max} = \ln(2/\xi)$, we have $\mathbb{P}\left(\text{None of the } \tau_{\max} \text{ trials belong to } \Theta\right) \leq \xi/2$.

Finally, we bound the TV distance between the output distribution and the target distribution: $d_{\mathrm{TV}}(p_{\mathrm{out}}, p^*)$. Denote the event $H_{\mathrm{Fail}} = \{\text{None of the } \tau_{\max} \text{ trials belong to } \Theta\}$, and let $H_{\mathrm{Success}} = H_{\mathrm{Fail}}^c$ be its complement. We know from the above analysis that $\mathbb{P}(H_{\mathrm{Fail}}) \leq \xi/2$.

Let $\theta_{\mathrm{out}} \sim p_{\mathrm{out}}$. Notice that

$$\theta_{\mathrm{out}} | H_{\mathrm{Success}} \sim p^K, \text{ and } \theta_{\mathrm{out}} | H_{\mathrm{Fail}} \sim p_{\mathrm{Fail}}.$$

We thus have $p_{\mathrm{out}} = (1 - \mathbb{P}(H_{\mathrm{Fail}}))p^K + \mathbb{P}(H_{\mathrm{Fail}})p_{\mathrm{Fail}}$. Therefore, for

$$\begin{aligned}
d_{\mathrm{TV}}(p_{\mathrm{out}}, p^*) &= \sup_A |p_{\mathrm{out}}(A) - p^*(A)| \\
&= \sup_A \left|(1 - \mathbb{P}(H_{\mathrm{Fail}}))p^K(A) + \mathbb{P}(H_{\mathrm{Fail}})p_{\mathrm{Fail}}(A) - p^*(A)\right| \\
&\leq \sup_A \left|p^K(A) - p^*(A)\right| + \sup_A \left|-\mathbb{P}(H_{\mathrm{Fail}})p^K(A) + \mathbb{P}(H_{\mathrm{Fail}})p_{\mathrm{Fail}}(A)\right| \\
&= d_{\mathrm{TV}}(p^K, p^*) + \mathbb{P}(H_{\mathrm{Fail}})d_{TV}(p^K, p_{\mathrm{Fail}}) \\
&\leq \xi/2 + \xi/2 = \xi,
\end{aligned}$$

where we apply $d_{\mathrm{TV}}(p^K, p^*) \leq \xi/2$, $\mathbb{P}(H_{\mathrm{Fail}}) \leq \xi/2$, and $d_{TV}(p^K, p_{\mathrm{Fail}}) \leq 1$ in the last inequality.

Therefore, we prove that $d_{\mathrm{TV}}(p_{\mathrm{out}}, p^*) \leq \xi$.

At last, we calculate the expected total number of iterations. Let τ_{halt} be the current count of the outer loop when the algorithm stops. Then τ_{halt} follows the finite geometric distribution with the probability of success in each trial equals $\mathbb{P}\left(\hat{\theta}^K \in \Theta\right) \geq 1 - 2e^{-2} - \xi$, and the maximum number of trials $\tau_{\max}$. Therefore, $\mathbb{E}(\tau_{\mathrm{halt}}) \leq \frac{1}{1-2e^{-2}-\xi} \leq \frac{1}{1-2e^{-2}-1/8} < 2$, when $\xi < 1/16 < 1/8$.

Therefore, the expected number of total iterations is given by

$$K \cdot \mathbb{E}(\tau_{\mathrm{halt}}) = \mathbf{\Theta}\left((d \ln \kappa + \ln 1/\xi) \max\left\{\kappa^{3/2}\sqrt{d \ln \kappa + \ln 1/\xi}, d\kappa\right\}\right),$$

which completes the proof. ■

The universal constants regarding the number of iterations can be found in the proof of Lemma 7 in Ma et al. (2019).

Lemma 25. *In $\mathbb{R}^d$, if a distribution p is γm-strongly log-concave, then*

$$\left\|\mathbb{E}_{\theta \sim p(\theta)}[\theta] - \theta^*\right\| \leq \sqrt{\frac{3}{\gamma m}},$$

where θ^* denotes the mode of the distribution $p(\theta)$. We also have that for $\hat{\theta} \sim p$,

$$\mathbb{P}\left(\left\|\hat{\theta} - \mathbb{E}_{\theta \sim p(\theta)}[\theta]\right\| \geq r\right) \leq 2e^{-\frac{r^2 \gamma m}{8d}}.$$

Proof We prove this Lemma following the proof of Lemma 10 in Mazumdar et al. (2020).

For $\left\|\mathbb{E}_{\theta \sim p(\theta)}[\theta] - \theta^*\right\|$, we use the fact that $p(\theta)$ is γm-strongly log-concave and consequently unimodal, and that

$$\left\|\mathbb{E}_{\theta \sim p(\theta)}[\theta] - \theta^*\right\|^2 \leq \frac{3}{\gamma m}.$$

For $\left\|\hat{\theta} - \mathbb{E}_{\theta\sim p(\theta)}[\theta]\right\|$, we note that $p(\theta)$ being γm-strongly log-concave implies that the random variable $\theta \sim p$ is a sub-Gaussian random vector with parameter $\sigma_v^2 = \frac{1}{2\gamma m}$. Consequently, it is a norm-sub-Gaussian random variable with constant $\sigma_n^2 = \frac{4d}{\gamma m}$. Plugging into the definition of norm-sub-Gaussian random variable, we obtain that for $\hat{\theta} \sim \pi$,

$$\mathbb{P}\left(\left\|\hat{\theta} - \mathbb{E}_{\theta\sim p(\theta)}[\theta]\right\| \geq r\right) \leq 2e^{-\frac{r^2\gamma m}{8d}}.$$

■

Proof of Corollary 10 Applying Lemma 8, to obtain Δ-accuracy in W_∞ distance, we want that

$$\xi < p_{\min} \cdot \frac{\pi^{d/2}}{2^{d+1}\cdot\Gamma(\frac{d}{2}+1)d^{d/2}}\Delta^d,$$

where $p_{\min} \leq \min_\theta p^*(\theta)$ and $p^*(\theta) \propto e^{-\gamma J(\theta)}\mathbf{1}(\theta \in \Theta)$ as defined in Section 3.3.

Since $\theta^* \in \Theta$ by Assumption 3, applying Corollary 20, we obtain that

$$\min_\theta p^*(\theta) \geq e^{-2\gamma n\beta B^2} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}B^d}.$$

We then set

$$\begin{aligned}\xi &= p_{\min} \cdot \frac{\pi^{d/2}}{2^{d+1}\cdot\Gamma(\frac{d}{2}+1)d^{d/2}}\Delta^d && (p_{\min} = e^{-2\gamma n\beta B^2}\cdot\tfrac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}B^d})\\ &= e^{-2\gamma n\beta B^2}\cdot\frac{1}{2^{d+1}\cdot d^{d/2}B^d}\Delta^d\end{aligned}$$

Therefore, we have

$$\ln 1/\xi \sim \mathcal{O}\left(\gamma n\beta B^2 + d\log d + d\log B + d\log(1/\Delta)\right).$$

Therefore, applying Theorem 1, we obtain that Algorithm 1 converges to Δ-accuracy in $W_\infty(p_{\text{out}}, p^*)$ with the following expected number of gradient queries to J:

$$\mathbf{\Theta}\left(\left(d\ln(\kappa Bd) + d\ln(1/\Delta) + \gamma n\beta B^2\right)\max\left\{\kappa^{3/2}\sqrt{d\ln(\kappa Bd) + d\ln(1/\Delta) + \gamma n\beta B^2}, d\kappa\right\}\right).$$

Plugging B, γ as given in Table 2, we obtain

$$\begin{aligned}\xi &= e^{-2\gamma n\beta B^2}\cdot\frac{1}{2^{d+1}\cdot d^{d/2}B^d}\Delta^d\\ &= e^{-\frac{\varepsilon n\beta B}{G}}\cdot\frac{1}{2^{d+1}\cdot d^{d/2}B^d}\Delta^d && (\gamma = \tfrac{\varepsilon}{2GB})\\ &= e^{-C_1\kappa d\ln(d/\rho)}\cdot\frac{1}{2^{d+1}\cdot d^{d/2}}\left(\frac{\alpha n\varepsilon}{C_1 Gd\ln(d/\rho)}\right)^d\Delta^d && (B = \tfrac{C_1 Gd\ln(d/\rho)}{\alpha n\varepsilon},\ \kappa = \tfrac{\beta}{\alpha})\end{aligned}$$

Therefore, we have

$$\begin{aligned}\ln 1/\xi &\sim \mathcal{O}\left(\kappa d\ln(d/\rho) + d\ln d + d\ln\left(\frac{Gd\ln(d/\rho)}{\alpha n\varepsilon}\right) + d\ln\frac{1}{\Delta}\right)\\ &\sim \mathcal{O}\left(\kappa d\ln(d/\rho) + d\ln\left(\frac{Gd\ln(d/\rho)}{\alpha n\varepsilon\Delta}\right)\right) && (\kappa = \tfrac{\beta}{\alpha} > 1,\ \rho < 1)\end{aligned}$$

Therefore, applying Theorem 1, we obtain that Algorithm 1 converges to Δ-accuracy in $W_\infty(p_{\text{out}}, p^*)$ with the following expected number of gradient queries to J:

$$\mathbf{\Theta}\left(d\left(\ln\kappa + \kappa\ln(d/\rho) + \ln\left(\frac{Gd\ln(d/\rho)}{\alpha n\varepsilon\Delta}\right)\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(d/\rho) + \ln\left(\frac{Gd\ln(d/\rho)}{\alpha n\varepsilon\Delta}\right)\right)}, d\kappa\right\}\right)$$
$$\sim \mathbf{\Theta}\left(d\left(\kappa\ln(d/\rho) + \ln\left(\frac{G\ln(d/\rho)}{\alpha n\varepsilon\Delta}\right)\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(d/\rho) + \ln\left(\frac{G\ln(d/\rho)}{\alpha n\varepsilon\Delta}\right)\right)}, d\kappa\right\}\right)$$

■

H Proof of Theorem 2

We first provide the following lemma that converts the distance between probability measures to the distance between random variables.

Lemma 26. *Let P, Q be two distributions defined on Θ such that $W_\infty(P, Q) \le \Delta$ under a metric* $\mathrm{dist} : \Theta \times \Theta \to \mathbb{R}_+$. *Let $X \sim P$, there exists a random variable $Y \sim Q$ such that* $\mathbb{P}[\mathrm{dist}(X, Y) \le \Delta] = 1$.

The proof of this lemma is deferred to the end of this section.

Proof of Theorem 2 Let $x \sim \widetilde{\mathcal{M}}(D)$, by Lemma 26 and the condition on W_∞ distance, there is a coupling ζ such that random variable $y \sim \mathcal{M}(D)$, $(x, y) \sim \zeta$, and $\mathrm{dist}(x, y) \le \Delta$ with probability 1 (with respect to ζ). The distribution of x is equivalent to the following two-step procedure

(1) $y \sim \mathcal{M}(D)$

(2) $x \sim \zeta(\cdot|y)$

The second step can be further viewed as first sampling $u \sim \mathrm{Uniform}([0, 1])$ and then mapping u to x deterministically by a data-dependent function $x = g(u)$ where g is completely determined by $\zeta(\cdot|y)$ via the inverse integral transform. g depends on D and y through ζ.

Define query $f_{u,y}(D) = g_\zeta(u) - y$ and on a neighbor dataset D', $f_{u,y}(D') = g_{\zeta'}(u) - y$ where ζ' is the resulting coupling under D'. With probability 1 under ζ, $\|g_\zeta(u) - y\| \le \Delta$. Similarly with probability 1, under ζ', $\|g_{\zeta'}(u) - y\| \le \Delta$.

By a union bound, with probability 1 (under $\zeta(\cdot|y) \times \zeta'(\cdot|y)$), the global sensitivity of this query $f_{u,y}$ satisfies

$$\|f_{u,y}(D) - f_{u,y}(D')\| \le \|g_\zeta(u) - y\| + \|g_{\zeta'}(u) - y\| \le 2\Delta.$$

Recall that ASAP returns $\tilde{x} = x + Z = y + (x - y) + Z$. In the above, $x - y = f_{u,y}(D)$. By choosing Z appropriately as the Gaussian mechanism (or Laplace mechanism) and the adaptive composition theorem (of GDP and pure-DP), we establish the two stated claims. ■

Proof of Lemma 26 By the attainability of infimum for W_∞ stated in Lemma 22 and that $W_\infty(P, Q) \le \Delta$, we know that there exists a joint distribution $\zeta(x, y)$, such that the marginals in x and y follow P and Q respectively, and that $\operatorname{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x, y) = W_\infty(P, Q) \le \Delta$. For given $X \sim P$, we define a conditional distribution $\zeta(\cdot|x)$, then taking $Y \sim \zeta(\cdot|x)$ statisfies $\mathbb{P}[\mathrm{dist}(X, Y) \le \Delta] = 1$.

■

I Proof of Theorem 3

Proof of Theorem 3 We divide the proof into three parts: privacy, accuracy, and computation. We first present the proof for pure DP case.

Privacy. The privacy analysis follows from the adaptive composition (Lemma 14) of output perturbation ε-DP and ASAP (2ε-DP, Theorem 2).

Accuracy. Let E denote the event of $\|\theta_0 - \theta_{\text{opt}}\|_1 \leq \frac{2d(G+\tau\alpha)\ln(d/\rho)}{n\alpha\varepsilon}$, where $\theta_0, \theta_{\text{opt}}$ is defined in Algorithm 4 for pure DP. Then by Lemma 17 of (Chaudhuri et al., 2011), $\mathbb{P}(E) \geq 1-\rho$. By the choice of B from (2), we know that under event E, we have $\mathbb{B}(\theta^*, R_1) \subset \Theta$, where $R_1 = 8\sqrt{\frac{d}{\gamma\alpha n}}$, and thus Assumption 3 holds.

Denote $\tilde{\theta}$ the output of MALA with constraint (Algorithm 1), and denote p^* the exact posterior.

We then divide the risk into two parts

$$\mathbb{E}\left[\mathcal{L}(\hat{\theta})\right] - \mathcal{L}(\theta^*) = \mathbb{E}\left[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta})\right] + \mathbb{E}\left[\mathcal{L}(\tilde{\theta})\right] - \mathcal{L}(\theta^*). \tag{5}$$

For the first part, note that $\hat{\theta} = \tilde{\theta} + \mathbf{Z}$, where $Z_i \overset{\text{i.i.d.}}{\sim} \text{Lap}(\Delta/\varepsilon),\ i = 1, ..., d$, thus we have that

$$\mathbb{E}\left[\|\hat{\theta} - \tilde{\theta}\|_2\right] \overset{\text{Jensen's inequality}}{\leq} \sqrt{\mathbb{E}\left[\|\hat{\theta} - \tilde{\theta}\|_2^2\right]} = \sqrt{\mathbb{E}\left[\sum\nolimits_{i=1}^{d} Z_i^2\right]} = \sqrt{2d}\frac{\Delta}{\varepsilon}$$

With the nG-Lipschitz continuity of $\mathcal{L}$, we have

$$\mathbb{E}\left[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta})\right] \leq nG\mathbb{E}\left[\|\hat{\theta} - \tilde{\theta}\|_2\right] \leq \frac{\sqrt{2d}nG\Delta}{\varepsilon} \tag{6}$$

For $\mathbb{E}\left[\mathcal{L}(\tilde{\theta})\right] - \mathcal{L}(\theta^*)$, the second part of the right-hand side of the equation (5), we consider two cases:

1. Under the event E. By Corollary 10, we have $W_\infty(\tilde{p}, p^*) \leq \Delta$. Since $W_\infty(\tilde{p}, p^*) \leq \Delta$, applying Lemma 22, there exist a coupling of $\tilde{p}$ and p^*, denoted as ζ^*, such that $\operatorname{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)}\|x - y\|_1 \leq \Delta$, thus $\operatorname{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)}\|x-y\|_2 \leq \Delta$. Take $\theta_{\text{post}}|\tilde{\theta} \sim \zeta(\cdot|\tilde{\theta})$, then we know $\theta_{\text{post}} \sim p^*$.

   We divide

$$\mathbb{E}\left[\mathcal{L}(\tilde{\theta})|E\right] - \mathcal{L}(\theta^*) = \mathbb{E}\left[\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta_{\text{post}})|E\right] + \left(\mathbb{E}\left[\mathcal{L}(\theta_{\text{post}})|E\right] - \mathcal{L}(\theta^*)\right) \tag{7}$$

   With the nG-Lipschitz continuity of $\mathcal{L}$, we have

$$\mathbb{E}_{(\tilde{\theta},\theta_{\text{post}})\sim\zeta}\left[\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta_{\text{post}})|E\right] \leq nG\mathbb{E}_{(\tilde{\theta},\theta_{\text{post}})\sim\zeta}\left[\|\tilde{\theta} - \theta_{\text{post}}\|_2|E\right] \leq nG\Delta, \tag{8}$$

   which is dominated by (6). By Lemma 6, we have

$$\mathbb{E}[\mathcal{L}(\theta_{\text{post}})|E] - \mathcal{L}(\theta^*) \leq \frac{d}{\gamma}. \tag{9}$$

2. Under the complementary event E^c. For simplicity, we set τ in Algorithm 4 to be sufficiently small and exclude this factor. By the $n\beta$-Lipschitz smooth of $\mathcal{L}$ and the first-order condition, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\tilde{\theta})|E^c\right] - \mathcal{L}(\theta^*) \leq \frac{n\beta}{2}\mathbb{E}\left[\|\tilde{\theta} - \theta^*\|_2^2|E^c\right] &\leq \frac{n\beta}{2}\mathbb{E}\left[\left(\|\tilde{\theta} - \theta_0\|_2 + \|\theta_0 - \theta^*\|_2\right)^2|E^c\right] \\ &\leq \frac{n\beta}{2}\mathbb{E}\left[\left(B + \|\theta_0 - \theta^*\|_1\right)^2|E^c\right]\end{aligned} \tag{10}$$

By Lemma 27, we have

$$\mathbb{E}\left[\|\theta_0-\theta^*\|_1^2|E^c\right] \leq \mathcal{O}\left(d(d+1)\left(\frac{\sqrt{d}G}{\alpha n\varepsilon}\right)^2+B^2+(d+1)B\frac{\sqrt{d}G}{\alpha n\varepsilon}\right),$$

and

$$\mathbb{E}\left[\|\theta_0-\theta^*\|_1|E^c\right] \leq \mathcal{O}\left(d\frac{\sqrt{d}G}{\alpha n\varepsilon}+B\right).$$

Plugging these into (10), plugging in $\kappa=\beta/\alpha$, we obtain that

$$\mathbb{E}[\mathcal{L}(\tilde{\theta})|E^c]-\mathcal{L}(\theta^*) \leq \mathcal{O}\left(\frac{d^3G^2\kappa}{\alpha n\varepsilon^2}\right). \tag{11}$$

Therefore, adding (6), (8), and (9), as well as adding (6), and (11), respectively, we obtain

$$\mathbb{E}\left[\mathcal{L}(\hat{\theta})|E\right]-\mathcal{L}(\theta^*) \leq \frac{\sqrt{2d}nG\Delta}{\varepsilon}+nG\Delta+\frac{d}{\gamma}=\mathcal{O}\left(\frac{d^2G^2\ln(d/\rho)}{\alpha n\varepsilon^2}\right), \text{ and} \tag{12}$$

$$\mathbb{E}\left[\mathcal{L}(\hat{\theta})|E^c\right]-\mathcal{L}(\theta^*) \leq \frac{\sqrt{2d}nG\Delta}{\varepsilon}+\mathcal{O}\left(\frac{d^3G^2\kappa}{\alpha n\varepsilon^2}\right)=\mathcal{O}\left(\frac{d^2G^2\left(\ln(d/\rho)+d\kappa\right)}{\alpha n\varepsilon^2}\right), \tag{13}$$

where we instantiated our choice of $\gamma=\frac{\varepsilon}{G\mathrm{Diam}(\Theta_{\text{local}})}$ with $\mathrm{Diam}(\Theta_{\text{local}})=2B=\frac{2CdG\log(d/\rho)}{\alpha n\varepsilon}$, and $\Delta=\frac{dG\log(d/\rho)}{2n^2\alpha\varepsilon}\leq\frac{d^{\frac{3}{2}}G\log(d/\rho)}{2n^2\alpha\varepsilon}$.

With (12) and (13), we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\hat{\theta})\right]-\mathcal{L}(\theta^*) &= \mathbb{P}(E)\cdot\mathbb{E}\left[\mathcal{L}(\hat{\theta})|E\right]+(1-\mathbb{P}(E))\cdot\mathbb{E}\left[\mathcal{L}(\hat{\theta})|E^c\right]-\mathcal{L}(\theta^*)\\ &\leq \mathcal{O}\left(\frac{d^2G^2\ln(d/\rho)}{\alpha n\varepsilon^2}\right)+\rho\cdot\mathcal{O}\left(\frac{d^2G^2\left(\ln(d/\rho)+d\kappa\right)}{\alpha n\varepsilon^2}\right)\\ &= \mathcal{O}\left(\frac{d^2G^2\left(\ln d+\ln(1/\rho)+\rho d\kappa\right)}{\alpha n\varepsilon^2}\right).\end{aligned}$$

Taking $\rho=\mathcal{O}\left(1/(d\kappa)\right)$, we obtain

$$\mathbb{E}\left[\mathcal{L}(\hat{\theta})\right]-\mathcal{L}(\theta^*)=\mathcal{O}\left(\frac{d^2G^2\left(\ln d+\ln\kappa\right)}{\alpha n\varepsilon^2}\right).$$

Computation. The optimization oracle (e.g., Newton's method or Gradient Descent) runs in $\mathcal{O}(n\log n)$ time for getting a solution satisfying $\|\theta_0-\theta^*\|\leq 1/n^2$. For the sampling step, following the proof of Corollary 10, the expected computation complexity is bounded by

$$K\cdot\left((1-\rho)\mathbb{E}\left[\tau_{\text{halt}}|E\right]+\rho\tau_{\max}\right).$$

Applying Corollary 10 with taking proper $B=\frac{CdG\log(d/\rho)}{\alpha n\varepsilon}$, we have $\mathbb{E}\left[\tau_{\text{halt}}|E\right]\sim\mathcal{O}(1)$.

Plugging in $\Delta=\frac{Gd\log(d/\rho)}{2n^2\alpha\varepsilon}$ as given in Table 2, we have

$$K\sim\mathbf{\Theta}\left(d\left(\kappa\ln(d/\rho)+\ln n\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(d/\rho)+\ln n\right)},d\kappa\right\}\right).$$

Therefore, the expected total computation complexity is given by

$$\mathbf{\Theta}\left(nd\left(\kappa\ln(d/\rho)+\ln n\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(d/\rho)+\ln n\right)},d\kappa\right\}\left(1+\rho d\left(\kappa\ln(d/\rho)+\ln n\right)\right)\right).$$

Gaussain DP case. The analysis of Gaussian DP closely parallels the proof of pure DP case except for the slight difference in the accuracy and computation analysis. Here, we present a simplified proof.

Let E denote the event of $\|\theta_0 - \theta_{\text{opt}}\|_2 \leq \frac{2\sqrt{2}(\tau\alpha+G)(\sqrt{d}+\sqrt{\ln(1/\rho)})}{\alpha n\mu}$, where $\theta_0, \theta_{\text{opt}}$ is defined in Algorithm 4 for Gaussian DP. Since $\theta_0 = \theta_{\text{opt}} + \boldsymbol{Z}$, where $Z_i \sim \mathcal{N}(0, 4(\tau + \frac{G}{\alpha})^2/n^2\mu^2)$, applying the tail bound for chi-square distributions by Lemma 1 of (Laurent & Massart, 2000), we have $\mathbb{P}(E) \geq 1 - \rho$. By the choice of B from (3), we know that under event E, Assumption 3 holds. For simplicity, we set τ in Algorithm 4 to be sufficiently small and exclude this factor.

Under this event E, by Lemma 6, similar to the pure DP case, we have

$$\mathbb{E}[\mathcal{L}(\hat{\theta})|E] - \mathcal{L}(\theta^*) \leq \frac{d}{\gamma} + \mathcal{O}\left(\frac{nG\sqrt{d}\Delta}{\mu}\right) = \mathcal{O}\left(\frac{dG^2}{\alpha n\mu^2}\right),$$

where we instantiated our choice of $\gamma = \frac{\mu^2\alpha n}{G^2}$, and $\Delta = \frac{\sqrt{d}G}{\sqrt{2}n^2\alpha\mu}$.

On the other hand, under E^c, applying Lemma 27 by taking $k = d/2$, and $\sigma = 8(\frac{\Delta}{\mu})^2$, we have

$$\mathbb{E}\left[\|\theta_0 - \theta_{\text{opt}}\|_2^2 | E^c\right] \leq \mathcal{O}\left(\frac{G^2(d + \ln(1/\rho))}{(\alpha n\mu)^2}\right).$$

Thus by the $n\beta$-Lipschitz smooth of $\mathcal{L}$ and the first-order condition, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\tilde{\theta})|E^c\right] - \mathcal{L}(\theta^*) \leq \frac{n\beta}{2}\mathbb{E}\left[\|\tilde{\theta} - \theta^*\|_2^2 | E^c\right] &\leq \frac{n\beta}{2}\mathbb{E}\left[\left(\|\tilde{\theta} - \theta_0\|_2 + \|\theta_0 - \theta^*\|_2\right)^2 | E^c\right] \\ &\leq \mathcal{O}\left(\frac{\kappa G^2(d + \ln(1/\rho))}{\alpha n\mu^2}\right).\end{aligned} \tag{14}$$

Therefore, by combining the two cases, we have

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq \mathcal{O}\left(\frac{G^2\left(d + \rho\kappa d + \rho\kappa\ln(1/\rho)\right)}{\alpha n\mu^2}\right).$$

Taking $\rho = \mathcal{O}(1/\kappa)$, we obtain that

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq \mathcal{O}\left(\frac{G^2(d + \ln\kappa)}{\alpha n\mu^2}\right).$$

Applying Lemma 8, the expected total computation complexity of the Gaussian DP case is given by

$$\boldsymbol{\Theta}\left(nd\left(\ln\kappa + \kappa\ln(1/\rho) + \ln n\right)\max\left\{\kappa^{3/2}\sqrt{d\left(\kappa\ln(1/\rho) + \ln n\right)}, d\kappa\right\}\left(1 + \rho d\left(\kappa\ln(1/\rho) + \ln n\right)\right)\right),$$

with a slight reduction of a $\ln d$ factor. ■

Lemma 27. *Let X be a random variable drawn from the Gamma distribution $\Gamma(k, \sigma)$, with the density*

$$f(x; k, \sigma) = \frac{x^{k-1}e^{-\frac{x}{\sigma}}}{\sigma^k\Gamma(k)}, \quad \textit{for } x > 0, k > 0, \textit{ and } \sigma > 0.$$

Suppose $k \geq 1$, then for any $a > 0$, we have

$$\mathbb{E}(X|X > a) \leq k\sigma + a, \quad \textit{and} \quad \mathbb{E}(X^2|X > a) \leq k(k+1)\sigma^2 + a^2 + (k+1)a\sigma.$$

Proof We have

$$\mathbb{E}(X|X>a) = \frac{\int_a^\infty x f(x;k,\sigma)dx}{\int_a^\infty f(x;k,\sigma)dx} = \frac{\int_a^\infty x^k e^{-\frac{x}{\sigma}}dx}{\int_a^\infty x^{k-1}e^{-\frac{x}{\sigma}}dx}$$
$$= \sigma \frac{\int_{a/\sigma}^\infty x^k e^{-x}dx}{\int_{a/\sigma}^\infty x^{k-1}e^{-x}dx}$$
$$= \sigma \frac{\Gamma(k+1, a/\sigma)}{\Gamma(k, a/\sigma)},$$

where $\Gamma(k,t)$ is the incomplete gamma function defined as $\Gamma(k,t) = \int_t^\infty x^{k-1}e^{-x}\mathrm{d}x$.

Applying integral by parts, we have

$$\Gamma(k+1,t) = k\Gamma(k,t) + t^k e^{-t}.$$

We also have, for $k \geq 1$,

$$\Gamma(k,t) = (k-1)\Gamma(k-1,t) + t^{k-1}e^{-t} \geq t^{k-1}e^{-t}.$$

Therefore, for $k \geq 1$,

$$\frac{\Gamma(k+1,t)}{\Gamma(k,t)} = k + \frac{t^k e^{-t}}{\Gamma(k,t)} \leq k + \frac{t^k e^{-t}}{t^{k-1}e^{-t}} = k + t.$$

Taking $t = a/\sigma$, we have

$$\mathbb{E}(X|X>a) \leq \sigma \frac{\Gamma(k+1, a/\sigma)}{\Gamma(k, a/\sigma)} \leq k\sigma + a.$$

For $\mathbb{E}(X^2|X>a)$, we have

$$\mathbb{E}(X^2|X>a) = \frac{\int_a^\infty x^2 f(x;k,\sigma)dx}{\int_a^\infty f(x;k,\sigma)dx} = \frac{\int_a^\infty x^{k+1} e^{-\frac{x}{\sigma}}dx}{\int_a^\infty x^{k-1}e^{-\frac{x}{\sigma}}dx}$$
$$= \sigma^2 \frac{\int_{a/\sigma}^\infty x^{k+1} e^{-x}dx}{\int_{a/\sigma}^\infty x^{k-1}e^{-x}dx}$$
$$= \sigma^2 \frac{\Gamma(k+2, a/\sigma)}{\Gamma(k, a/\sigma)}$$
$$= \sigma^2 \frac{(k+1)\Gamma(k+1, a/\sigma) + (a/\sigma)^{k+1}e^{-a/\sigma}}{\Gamma(k, a/\sigma)}$$
$$\leq \sigma^2 \left((k+1)(k+a/\sigma) + \frac{(a/\sigma)^{k+1}e^{-a/\sigma}}{(a/\sigma)^{k-1}e^{-a/\sigma}} \right)$$
$$= \sigma^2 k(k+1) + a^2 + a\sigma(k+1).$$

■

J Deferred Proofs of the Supporting Lemmas and Corollaries

J.1 Proof of Lemma 9 and Corollary 20

Proof of Lemma 9 Denote θ^* as the minimizer of J on Θ. Then

$$Z = \int_\Theta e^{-\gamma J(\theta)}\mathrm{d}\theta \leq e^{-\gamma J(\theta^*)} \int_\Theta \mathrm{d}\theta = e^{-\gamma J(\theta^*)} \cdot \frac{\pi^{d/2}(R/2)^d}{\Gamma(d/2+1)}.$$

On the other hand, by the Lipschitz continuity assumption, $J(\theta) - J(\theta^*) \leq G \cdot \|\theta - \theta^*\| \leq GR$. Hence

$$\frac{1}{Z} e^{-\gamma J(\theta)} \geq e^{-\gamma G R} \cdot \frac{\Gamma(d/2+1)}{\pi^{d/2}(R/2)^d}.$$

If J instead satisfies the β-Lipschitz smoothness and convextiy assumptions, then for all $\tilde{\theta} \in \Theta$

$$\begin{aligned} J(\theta) - J(\theta^*) &= J(\theta) - J(\tilde{\theta}) + J(\tilde{\theta}) - J(\theta^*) \\ &\leq \langle J(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{\beta}{2}\|\theta - \tilde{\theta}\|^2 - \langle \nabla J(\tilde{\theta}), \theta^* - \tilde{\theta} \rangle \\ &= \langle J(\tilde{\theta}), \theta - \theta^* \rangle + \frac{\beta}{2}\|\theta - \tilde{\theta}\|^2 \\ &\leq \|J(\tilde{\theta})\| R + \frac{1}{2}\beta R^2 \end{aligned}$$

Hence

$$\inf_{\theta \in \Theta} p(\theta) \geq e^{-\gamma\left(R\|\nabla J(\tilde{\theta})\| + \beta R^2/2\right)} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}(R/2)^d}, \ \forall \tilde{\beta} \in \Theta.$$

Furthermore, if θ^* is the global minimizer, i.e., $\nabla J(\theta^*) = 0$, then

$$\inf_{\theta \in \Theta} p(\theta) \geq e^{-\gamma \beta R^2/2} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}(R/2)^d}, \ \forall \tilde{\theta} \in \Theta.$$

■

Proof of Corollary 20 It suffices to invoke Lemma 9 by setting $J(\theta) := \sum_{i=1}^n \ell_i(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|^2$ and $\Theta := \{\theta | \|\theta - \theta_0\| \leq B\}$, notice that the diameter of Θ is $2B$. ■

J.2 Proof of Lemma 18 and Lemma 19

Proof of Lemma 18 By the α-strong convexity of $\mathcal{L}_D$ and first-order optimality conditions for $\theta^*(D)$

$$\begin{aligned} \mathcal{L}_D(\theta^*(D')) &\geq \mathcal{L}_D(\theta^*(D)) + \langle \theta^*(D') - \theta^*(D), \nabla \mathcal{L}_D(\theta^*(D)) \rangle + \frac{\alpha n}{2}\|\theta^*(D) - \theta^*(D')\|^2 \\ &\geq \mathcal{L}_D(\theta^*(D)) + \frac{\alpha n}{2}\|\theta^*(D) - \theta^*(D')\|^2. \end{aligned}$$

On the other side, by the α-strong convexity of $\mathcal{L}_{D'}$ and first-order optimality conditions for $\theta^*(D')$

$$\begin{aligned} \mathcal{L}_{D'}(\theta^*(D)) &\geq \mathcal{L}_{D'}(\theta^*(D')) + \langle \theta^*(D) - \theta^*(D'), \nabla \mathcal{L}_{D'}(\theta^*(D')) \rangle + \frac{\alpha n}{2}\|\theta^*(D') - \theta^*(D)\|^2 \\ &\geq \mathcal{L}_{D'}(\theta^*(D')) + \frac{\alpha n}{2}\|\theta^*(D) - \theta^*(D')\|^2. \end{aligned}$$

Add the two inequalities we get

$$\begin{aligned} \alpha n \|\theta^*(D) - \theta^*(D')\|^2 &\leq \mathcal{L}_D(\theta^*(D')) - \mathcal{L}_{D'}(\theta^*(D')) + \mathcal{L}_{D'}(\theta^*(D)) - \mathcal{L}_D(\theta^*(D)) \\ &\leq |\ell_x(\theta^*(D')) - \ell_x(\theta^*(D))| \leq G\|\theta^*(D) - \theta^*(D')\|. \end{aligned}$$

The proof is complete by dividing both sides by $\|\theta^*(D) - \theta^*(D')\|$. ■

Proof of Lemma 19 It suffices to show the sensitivity of $\tilde{\theta}(D)$ is bounded by $\tilde{\Delta}$. Applying Lemma 18,the sensitivity of $\tilde{\theta}(D)$ is bounded by

$$\begin{aligned}\max_{D\simeq D'} ||\tilde{\theta}(D)-\tilde{\theta}(D')||_2 &= \max_{D\simeq D'} ||\tilde{\theta}(D)-\theta^*(D)+\theta^*(D)-\theta^*(D')+\theta^*(D')-\tilde{\theta}(D')||_2 \\ &\leq ||\tilde{\theta}(D)-\theta^*(D)||_2 + \max_{D\simeq D'} ||\theta^*(D)-\theta^*(D')||_2 + ||\theta^*(D')-\tilde{\theta}(D')||_2 \\ &\leq \frac{2\tau}{n}+\frac{2G}{\alpha n} = \tilde{\Delta}.\end{aligned}$$

■

J.3 Proof of Lemma 14

Proof of Lemma 14 First, observe that pure-DP (with any $\varepsilon < \infty$) by definition implies absolute continuity. Let μ, ν be the two measures induced by a pure-DP mechanism on dataset D, D' respectively. DP implies that for any measurable set S, $\mu(S) \leq e^{\varepsilon}\nu(S)$. This inequality implies that if $\nu(S) = 0$ then $\mu(S) = 0$, which verifies the definition of absolute continuity, i.e., $\mu \ll \nu$.

By our assumption, $\mathcal{M}_1$ satisfies DP, thus $P_{\mathcal{M}_1(D)}$ absolutely continuous w.r.t. $P_{\mathcal{M}_1(D')}$. Similarly, $\mathcal{M}_2(o_1, D)$ satisfies DP for all o_1 except when o_1 belongs to a measure 0 set, thus with probability 1, $P_{\mathcal{M}_2(O_1,D)}$ is absolutely continuous w.r.t. $P_{\mathcal{M}_2(O_1,D')}$. It follows that the "density" function (technically, Radon-Nikodym derivative) $\frac{dP_{\mathcal{M}_1(D)}}{dP_{\mathcal{M}_1(D')}}$ exists and $\frac{dP_{\mathcal{M}_2(O_1,D)}}{dP_{\mathcal{M}_2(O_1,D')}}$ exists almost surely. In addition, by taking $S = \{\frac{dP_{\mathcal{M}_1(D)}}{dP_{\mathcal{M}_1(D')}} > e^{\varepsilon}\}$ the DP definition, by a proof by contradiction*, we have

$$\mathbb{P}_{O_1\sim\mathcal{M}_1(D')}[\tfrac{dP_{\mathcal{M}_1(D)}}{dP_{\mathcal{M}_1(D')}}(O_1) \leq e^{\varepsilon_1}] = 1 \tag{15}$$

and

$$\mathbb{P}_{O_2\sim\mathcal{M}_2(o_1,D')}[\tfrac{dP_{\mathcal{M}_2(o_1,D)}}{dP_{\mathcal{M}_2(o_1,D')}}(O_2) \leq e^{\varepsilon_2}] = 1 \tag{16}$$

almost surely under $o_1 \sim \mathcal{M}_1(D')$.

Let $O_1 \sim \mathcal{M}_1(D)$ and $O_2 \sim \mathcal{M}_2(O_1, D)$. Similarly, let $\tilde{O}_1 \sim \mathcal{M}_1(D')$ and $\tilde{O}_2 \sim \mathcal{M}_2(\tilde{O}_2, D')$. Consider any measurable set $S \subset \Theta_1 \times \Theta_2$, by the Lebesgue integral

$$\begin{aligned}&\mathbb{P}[(O_1, O_2) \in S] \\ =&\int\int \mathbf{1}((u_1,u_2)\in S) dP_{\mathcal{M}_2(u_1,D)}(u_2) dP_{\mathcal{M}_1(D)}(u_1) \\ =&\int\int \mathbf{1}((u_1,u_2)\in S) \tfrac{dP_{\mathcal{M}_2(u_1,D)}}{dP_{\mathcal{M}_2(u_1,D')}}(u_2)\cdot dP_{\mathcal{M}_2(u_1,D')}(u_2)\cdot \tfrac{dP_{\mathcal{M}_1(D)}}{dP_{\mathcal{M}_1(D')}}(u_1)\cdot dP_{\mathcal{M}_1(D')}(u_1) \\ \leq&\int\int \mathbf{1}((u_1,u_2)\in S) e^{\varepsilon_2}\cdot dP_{\mathcal{M}_2(u_1,D')}(u_2)\cdot e^{\varepsilon_1}\cdot dP_{\mathcal{M}_1(D')}(u_1) \\ =&e^{\varepsilon_1+\varepsilon_2}\mathbb{P}[(\tilde{O}_1,\tilde{O}_2)\in S].\end{aligned}$$

The inequality above follows from equation 15 and equation 16.

For Gaussian DP, let P_1 and Q_1 be the probability measures of $M_1(D)$ and $M_1(D')$ respectively, and $P_2(\cdot|x = o)$ and $Q_2(\cdot|x = o)$ be the probability measures of $M_2(o, D)$ and $M_2(o, D')$ respectively. Let P and Q be the probability measures of $M(D)$ and $M(D')$.

*Assume S is not measure 0. By definition of DP $\mathbb{P}[\mathcal{M}_1(D') \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}_1(D') \in S]$ which contradicts with the definition of S unless S has measure 0.

We first show that P is absolutely continuous w.r.t Q. Noting that by the definition of Gaussian DP and Hockey-stick divergence, P_1 is absolutely continuous w.r.t. Q_1 and $P_2(\cdot|x=o)$ is continuous w.r.t. $Q_2(\cdot|x=o)$ with probability 1. Then, for arbitrary Q-measureable set A, we have that

$$\begin{aligned}
P(A) &= \int\int \mathbf{1}\{(x,y)\in A\} dP(x,y) \\
&= \int\int \mathbf{1}\{(x,y)\in A\} dP_2(y|x) dP_1(x) \\
&= \int\int \mathbf{1}\{(x,y)\in A\} dP_2(y|x) dP_1(x) \\
&= \int\int \mathbf{1}\{(x,y)\in A\} \frac{dP_2}{dQ_2}(y|x) dQ_2(y|x) \frac{dP_1}{dQ_1}(x) dQ_1(x) \\
&= \int\int \mathbf{1}\{(x,y)\in A\} \frac{dP_2}{dQ_2}(y|x) \frac{dP_1}{dQ_1}(x) dQ(x,y)
\end{aligned}$$

So $Q(A)=0$ implies $P(A)=0$, thus P is absolutely continuous w.r.t Q. Moreover, from the above equity, we know that the Radon-Nikodym derivative $\frac{dP}{dQ}(x,y) = \frac{dP_2}{dQ_2}(y|x)\cdot\frac{dP_1}{dQ_1}(x)$.

Since P is absolutely continuous w.r.t Q, their Hockey-stick distance can be defined. We then have for $\alpha > 0$,

$$\begin{aligned}
&H_\alpha(M(D)\|M(D')) \\
&= H_\alpha(P\|Q) \\
&= \int\int \left[\frac{dP}{dQ}(x,y) - \alpha\right]_+ dQ(x,y) \\
&= \int\int \mathbf{1}\left\{\frac{dP}{dQ}(x,y) \geq \alpha\right\}\left(\frac{dP}{dQ}(x,y) - \alpha\right) dQ(x,y) \\
&= \int\int \mathbf{1}\left\{\frac{dP_2}{dQ_2}(y|x)\cdot\frac{dP_1}{dQ_1}(x) \geq \alpha\right\}\left(\frac{dP_2}{dQ_2}(y|x)\cdot\frac{dP_1}{dQ_1}(x) - \alpha\right) dQ_2(y|x) dQ_1(x) \\
&= \int\int \mathbf{1}\left\{\frac{dP_2}{dQ_2}(y|x)\cdot\frac{dP_1}{dQ_1}(x) \geq \alpha\right\}\mathbf{1}\left\{\frac{dP_1}{dQ_1}(x) > 0\right\}\left(\frac{dP_2}{dQ_2}(y|x)\cdot\frac{dP_1}{dQ_1}(x) - \alpha\right) dQ_2(y|x) dQ_1(x) \\
&= \int\int \mathbf{1}\left\{\frac{dP_2}{dQ_2}(y|x)\cdot\frac{dP_1}{dQ_1}(x) \geq \alpha\left(\frac{dP_1}{dQ_1}(x)\right)^{-1}\right\}\left(\frac{dP_2}{dQ_2}(y|x) - \alpha\left(\frac{dP_1}{dQ_1}(x)\right)^{-1}\right) dQ_2(y|x) \\
&\cdot \mathbf{1}\left\{\frac{dP_1}{dQ_1}(x) > 0\right\}\frac{dP_1}{dQ_1}(x) dQ_1(x) \\
&= \int H_{\alpha\left(\frac{dP_1}{dQ_1}(x)\right)^{-1}}\left(dP_2(\cdot|x)\|dQ_2(\cdot|x)\right)\mathbf{1}\left\{\frac{dP_1}{dQ_1}(x) > 0\right\}\frac{dP_1}{dQ_1}(x) dQ_1(x) \\
&\leq \int H_{\alpha\left(\frac{dP_1}{dQ_1}(x)\right)^{-1}}\left(\mathcal{N}(0,1)\|\mathcal{N}(\mu_2,1)\right)\mathbf{1}\left\{\frac{dP_1}{dQ_1}(x) > 0\right\}\frac{dP_1}{dQ_1}(x) dQ_1(x) \\
&= \int\int \left[\frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z) - \alpha\left(\frac{dP_1}{dQ_1}(x)\right)^{-1}\right]_+ dP_{\mathcal{N}(\mu_2,1)}(z)\,\mathbf{1}\left\{\frac{dP_1}{dQ_1}(x) > 0\right\}\frac{dP_1}{dQ_1}(x) dQ_1(x) \\
&= \int\int \left[\frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z)\frac{dP_1}{dQ_1}(x) - \alpha\right]_+ dP_{\mathcal{N}(\mu_2,1)}(z) dQ_1(x) \\
&= H_\alpha\left(P_1\times N(0,1)\parallel Q_1\times\mathcal{N}(\mu_2,1)\right)
\end{aligned}$$

Continuing this argument, we have

$$
\begin{aligned}
&H_\alpha\left(P_1 \times N(0,1) \parallel Q_1 \times \mathcal{N}(\mu_2,1)\right) \\
&= \int\int \left[\frac{dP_1}{dQ_1}(x) - \alpha \cdot \frac{dP_{\mathcal{N}(\mu_2,1)}}{dP_{\mathcal{N}(0,1)}}(z)\right]_+ \frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z) dQ_1(x) dP_{\mathcal{N}(\mu_2,1)}(z) \\
&= \int H_{\alpha \cdot \frac{dP_{\mathcal{N}(\mu_2,1)}}{dP_{\mathcal{N}(0,1)}}(z)}\left(P_1 \| Q_1\right) \frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z) dP_{\mathcal{N}(\mu_2,1)}(z) \\
&\leq \int H_{\alpha \cdot \frac{dP_{\mathcal{N}(\mu_2,1)}}{dP_{\mathcal{N}(0,1)}}(z)}\left(\mathcal{N}(0,1) \| \mathcal{N}(\mu_1,1)\right) \frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z) dP_{\mathcal{N}(\mu_2,1)}(z) \\
&= \int\int \left[\frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_1,1)}}(w) - \alpha \cdot \frac{dP_{\mathcal{N}(\mu_2,1)}}{dP_{\mathcal{N}(0,1)}}(z)\right]_+ \frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z) dP_{\mathcal{N}(\mu_1,1)}(w) dP_{\mathcal{N}(\mu_2,1)}(z) \\
&= \int\int \left[\frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_1,1)}}(w) \frac{dP_{\mathcal{N}(0,1)}}{dP_{\mathcal{N}(\mu_2,1)}}(z) - \alpha\right]_+ dP_{\mathcal{N}(\mu_1,1)}(w) dP_{\mathcal{N}(\mu_2,1)}(z) \\
&= H_\alpha\left(\mathcal{N}(0,1) \times N(0,1) \parallel \mathcal{N}(\mu_1,1) \times \mathcal{N}(\mu_2,1)\right)
\end{aligned}
$$

By taking $s = (\mu_1 w + \mu_2 z)/\sqrt{\mu_1^2 + \mu_2^2},\ t = (\mu_1 w - \mu_2 z)/\sqrt{\mu_1^2 + \mu_2^2}$, we have,

$$
\begin{aligned}
&H_\alpha\left(\mathcal{N}(0,1) \times N(0,1) \parallel \mathcal{N}(\mu_1,1) \times \mathcal{N}(\mu_2,1)\right) \\
&= \int\int \left[\frac{\exp(-(w^2+z^2))}{\exp(-((w-\mu_1)^2 + (z-\mu_2)^2))} - \alpha\right]_+ \frac{1}{2\pi} \exp(-((w-\mu_1)^2 + (z-\mu_2)^2)) dw dz \\
&= \int\int \left[\exp(-2(\mu_1 w + \mu_2 z) + \mu_1^2 + \mu_2^2) - \alpha\right]_+ \frac{1}{2\pi} \exp(-(w^2 + z^2 - 2(\mu_1 w + \mu_2 z) + \mu_1^2 + \mu_2^2)) dw dz \\
&= \int\int \left[\exp(-2s\sqrt{\mu_1^2+\mu_2^2} + \mu_1^2 + \mu_2^2) - \alpha\right]_+ \frac{1}{2\pi} \exp(-(s^2 + t^2 - 2s\sqrt{\mu_1^2+\mu_2^2} + \mu_1^2 + \mu_2^2)) dt ds \\
&= \int\int \left[\exp(-2s\sqrt{\mu_1^2+\mu_2^2} + \mu_1^2 + \mu_2^2) - \alpha\right]_+ \frac{1}{\sqrt{2\pi}} \exp(-(s^2 - 2s\sqrt{\mu_1^2+\mu_2^2} + \mu_1^2 + \mu_2^2)) ds \\
&= H_\alpha\left(\mathcal{N}(0,1) \parallel \mathcal{N}(\sqrt{\mu_1^2+\mu_2^2}, 1)\right)
\end{aligned}
$$

The proof is completed. ■

J.4 Proof of Lemma 5

Proof of Lemma 5 (Dong et al., 2020) showed that exponential sampling with utility function $q(D,\theta)$ that satisfies a property called *bounded range* is differentially private even if the sensitivity is not bounded (unlike the original exponential mechanism).

$$
\mathrm{range}(q) := \sup_{D, D' \text{are neighbors}} (\max_{\theta\in\Theta} - \min_{\theta\in\Theta})[q(D,\theta) - q(D',\theta)].
$$

In our problem, q is the (regularized) sum of loss functions, and the difference in q between two neighbor datasets created by adding or removing a datapoint is simply the loss of one data point. $q(D,\theta) - q(D',\theta) = \pm\ell(\theta)$. It is easy to see that the $\mathrm{range}(q) \leq G\mathrm{Diam}(\Theta)$. By (Dong et al., 2020), we get that choosing $\gamma \leq \frac{\varepsilon}{G\mathrm{Diam}(\Theta)}$ gives ε-DP. ■

J.5 Proof of Lemma 15

Proof Denote

$$f_{\varepsilon,0}(x) = \max\{0, 1 - e^{\varepsilon}x, e^{-\varepsilon}(1-x)\}, \text{ for } 0 \le x \le 1,$$

and

$$G_{\mu}(x) = \Phi\left(\Phi^{-1}(1-x) - \mu\right), \text{ for } 0 \le x \le 1.$$

By Definition 3, Proposition 3 and Definition 4 of Dong et al. (2022), it suffices to show that

$$f_{\varepsilon,0}(x) \ge G_{\mu}(x), \text{ for all } 0 \le x \le 1.$$

By the concavity of G_{μ} and the piece-wise linearity of $f_{\varepsilon,0}$, it suffices to show that $G_{\mu}(x_0) \le f_{\varepsilon,0}(x_0)$, with $x_0 = \frac{1}{1+e^{\varepsilon}}$ satisfying $1 - e^{\varepsilon}x_0 = e^{-\varepsilon}(1-x_0) = x_0$. Taking $\mu = 2\Phi^{-1}\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}\right)$, we have $G_{\mu}(x_0) = f_{\varepsilon,0}(x_0)$, which finishes the proof. ■

J.6 Proof of Lemma 22

Proof of Lemma 22 Denote $\|\mathrm{dist}(\cdot,\cdot)\|_{L_\infty(\Theta^2,\zeta)} = \operatorname{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x,y)$.

To prove the lemma, it suffices to show these two inequalities both hold:

$$\inf_{\zeta\in\Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x,y) \le \inf\{\alpha > 0 : P(U) \le Q(U^{\alpha}),\ \forall \textit{ open } U \subset \Theta\} \tag{17}$$

$$\inf_{\zeta\in\Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x,y) \ge \inf\{\alpha > 0 : P(U) \le Q(U^{\alpha}),\ \forall \textit{ open } U \subset \Theta\} \tag{18}$$

For the sake of clarity, we denote

$$W_{\mathrm{LHS}} = \inf_{\zeta\in\Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x,y) = \inf_{\zeta\in\Gamma(P,Q)} \|\mathrm{dist}(\cdot,\cdot)\|_{L_\infty(\Theta^2,\zeta)}, \text{ and}$$

$$W_{\mathrm{RHS}} = \inf\{\alpha > 0 : P(U) \le Q(U^{\alpha}),\ \forall \textit{ open } U \subset \Theta\}.$$

We first prove (17). To prove (17), it suffices to show that for any $\alpha > W_{\mathrm{RHS}}$, the relationship $\alpha \ge W_{\mathrm{LHS}}$ inherently holds.

In particular, we leverage the following Strassen's theorem and prove that both constants β and ε can be driven to zero.

Lemma 28 (Strassen's Theorem, Strassen (1965))**.** *Suppose that* (Θ, dist) *is a separable metric space and* $\alpha, \beta > 0$*. Suppose the laws P and Q are such that, for all open sets* $U \subset \Theta$*,*

$$P(U) \le Q(U^{\alpha}) + \beta$$

where $U^{\alpha} = \{x \in \Theta : \mathrm{dist}(x, U) \le \alpha\}$*.*

Then for any $\varepsilon > 0$ *there exist a law* ζ *on* $\Theta \times \Theta$ *with marginals P and Q, such that*

$$\zeta(\mathrm{dist}(x,y) > \alpha + \varepsilon) \le \beta + \varepsilon. \tag{19}$$

Take an arbitrary $\alpha > W_{\mathrm{RHS}}$. Our goal is to prove $\alpha \ge W_{\mathrm{LHS}}$. By the definition of infimum, we have

$$P(U) \le Q(U^{\alpha}),\ \forall \textit{ open } U \subset \Theta,$$

For arbitrary β, ε, by Strassen's theorem, (plugging in (19)) there exist $\zeta_{\beta,\varepsilon} \in \Gamma(P,Q)$ such that

$$\zeta_{\beta,\varepsilon}\left(\mathrm{dist}(x,y) > \alpha + \varepsilon\right) \le \beta + \varepsilon.$$

We are going to choose β, ε to be sufficiently small (both go to zero) and take the limits using Prohorov's Theorem and Portmanteau theorem.

Taking $\beta = \varepsilon = \frac{1}{n}$, by Strassen's theorem, there exists $\zeta_n \in \Gamma(P, Q)$ such that

$$\zeta_n \left(\mathrm{dist}(x, y) > \alpha + \frac{1}{n} \right) \leq \frac{2}{n}.$$

Since P, Q are tight, for any $\epsilon > 0$, there exists two compact sets $K_1, K_2 \subset \Theta$, such that

$$P(\Theta \setminus K_1) \leq \epsilon, \; Q(\Theta \setminus K_2) \leq \epsilon$$

Note that $\zeta_n \in \Gamma(P, Q)$, we have (since $\mathbf{1}\{(x, y) \notin K_1 \times K_2\} \leq \mathbf{1}\{x \notin K_1\} + \mathbf{1}\{y \notin K_2\}$)

$$\zeta_n \left((\Theta \times \Theta) \setminus (K_1 \times K_2) \right) \leq \zeta_n \left((\Theta \setminus K_1) \times \Theta \right) + \zeta_n \left(\Theta \times (\Theta \setminus K_2) \right) \leq 2\epsilon$$

thus $\{\zeta_n\}_{n=1}^{\infty}$ is a tight sequence of probability measures.

Since $\{\zeta_n\}_{n=1}^{\infty}$ is tight, by Prohorov's Theorem, there exists a weakly convergent subsequence $\zeta_{n(k)} \Rightarrow \zeta$. We have $\zeta \in \Gamma(P, Q)$ because $\zeta(A \times \Theta) = \lim_{k \to \infty} \zeta_{n(k)}(A \times \Theta) = P(A)$, and that $\zeta(\Theta \times A) = \lim_{k \to \infty} \zeta_{n(k)}(\Theta \times A) = Q(A)$.

For arbitrary $\delta > 0$, the set $\{(x, y) \in \Theta \times \Theta : \mathrm{dist}(x, y) > \alpha + \delta\}$ is a open set in $\Theta \times \Theta$. By the Portmanteau Theorem,

$$\begin{aligned}
\zeta \left(\mathrm{dist}(x, y) > \alpha + \delta \right) &\leq \liminf_{k \to \infty} \zeta_{n(k)} \left(\mathrm{dist}(x, y) > \alpha + \delta \right) \\
&\leq \liminf_{k \to \infty} \zeta_{n(k)} \left(\mathrm{dist}(x, y) > \alpha + \frac{1}{n(k)} \right) \\
&\leq \liminf_{k \to \infty} \left(\frac{2}{n(k)} \right) = 0,
\end{aligned}$$

where the first inequality is by Portmanteau Theorem. The second inequality holds because there exist k_0 such that $\delta > \frac{1}{n(k_0)}$. The third inequality holds due to the construction of ζ_n.

Since δ is arbitrary, (taking $\delta = \frac{1}{n}$ and by Fatou's Lemma,) we have that

$$\zeta \left(\mathrm{dist}(x, y) > \alpha \right) = 0. \tag{20}$$

Therefore, $\alpha \geq \operatorname{ess\,sup}_{(x,y) \in (\Theta \times \Theta, \zeta)} \mathrm{dist}(x, y) = W_{\mathrm{LHS}}$. Thus (17) is proved.

Next, we show that (18) holds. Similarly, to prove (17), it suffices to show that for any $t > W_{\mathrm{LHS}}$, the relationship $t \geq W_{\mathrm{RHS}}$ inherently holds.

Take an arbitrary $t > W_{\mathrm{LHS}}$, our goal is then to prove that $t \geq W_{\mathrm{RHS}}$. By the definition of infimum, there exists a $\zeta \in \Gamma(P, Q)$, such that

$$\operatorname*{ess\,sup}_{(x,y) \in (\Theta \times \Theta, \zeta)} \mathrm{dist}(x, y) < t.$$

Therefore, $\zeta \left(\{(x, y) \in \Theta^2 : \mathrm{dist}(x, y) > t\} \right) = 0$, which translate into

$$\int \mathbf{1}\{\mathrm{dist}(x, y) > t\} \mathrm{d}\zeta(x, y) = 0. \tag{21}$$

Thus, we obtain that

$$
\begin{aligned}
Q(U^t) &= Q\left(\{y \in \Theta : \mathrm{dist}(y, U) \le t\}\right) \\
&= \int \mathbf{1}\{x \in \Theta\}\mathbf{1}\{\mathrm{dist}(y, U) \le t\}\mathrm{d}\zeta(x, y) \\
&\ge \int \mathbf{1}\{x \in U\}\mathbf{1}\{\mathrm{dist}(y, U) \le t\}\mathrm{d}\zeta(x, y) \\
&= \int \mathbf{1}\{x \in U\}\mathbf{1}\{y \in \Theta\}\mathrm{d}\zeta(x, y) - \int \mathbf{1}\{x \in U\}\mathbf{1}\{\mathrm{dist}(y, U) > t\}\mathrm{d}\zeta(x, y) \\
&\qquad\qquad \text{(If } x \in U \text{ and } \mathrm{dist}(y, U) > t\text{, then } \mathrm{dist}(x, y) > t.\text{)} \\
&\ge \int \mathbf{1}\{x \in U\}\mathbf{1}\{y \in \Theta\}\mathrm{d}\zeta(x, y) - \int \mathbf{1}\{\mathrm{dist}(x, y) > t\}\mathrm{d}\zeta(x, y) \qquad \text{(Plug in (21))} \\
&= \int \mathbf{1}\{x \in U\}\mathbf{1}\{y \in \Theta\}\mathrm{d}\zeta(x, y) \\
&= P(U).
\end{aligned}
$$

Since $Q(U^t) \ge P(U)$, we have

$$t \in \{r > 0 : P(U) \le Q(U^r), \forall \textit{ open } U \subset \Theta\}.$$

Thus

$$t \ge \inf\{\alpha > 0 : P(U) \le Q(U^\alpha),\ \forall \textit{ open } U \subset \Theta\} = W_{\mathrm{RHS}}.$$

Therefore, (18) holds.

Since (17) and (18) hold, we have

$$\inf_{\zeta \in \Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y) \in (\Theta \times \Theta, \zeta)} \mathrm{dist}(x, y) = \inf\{\alpha > 0 : P(U) \le Q(U^\alpha),\ \forall \textit{ open } U \subset \Theta\},$$

which demonstrates the equivalence of the two definitions.

We now prove the attainability of the infimum in Definition 7 by repeating the above construction of ζ in the first part of our proof. We provide a simplified proof as follows.

The above construction of ζ before (20) tells us that for an arbitrary $\alpha > W_\infty(P, Q)$, there exists a $\zeta \in \Gamma(P, Q)$ such that $\zeta\left(\mathrm{dist}(x, y) > \alpha\right) = 0$. Taking $\alpha_m = W_\infty(P, Q) + \frac{1}{m}$, there exists a sequence $\{\zeta_m\} \subset \Gamma(P, Q)$ such that $\zeta_m\left(\mathrm{dist}(x, y) > \alpha + \frac{1}{m}\right) = 0$. Since $\{\zeta_m\} \subset \Gamma(P, Q)$, likewise, we obtain that $\{\zeta_m\}$ is tight. By Prohorov's Theorem, there exists a weakly convergent subsequence $\zeta_{m(k)} \Rightarrow \zeta^*$. Similarly, $\zeta^* \in \Gamma(P, Q)$. By the same token, for arbitrary $\delta > 0$, we have

$$
\begin{aligned}
\zeta^*\left(\mathrm{dist}(x, y) > W_\infty(P, Q) + \delta\right) &\le \liminf_{k \to \infty} \zeta_{m(k)}\left(\mathrm{dist}(x, y) > W_\infty(P, Q) + \delta\right) \\
&\le \liminf_{k \to \infty} \zeta_{m(k)}\left(\mathrm{dist}(x, y) > W_\infty(P, Q) + \frac{1}{m(k)}\right) \\
&= 0,
\end{aligned}
$$

where the first inequality is by the Portmanteau Theorem. The second inequality holds because there exist k_0 such that $\delta > \frac{1}{m(k_0)}$. The third inequality holds due to the construction of ζ_m.

Since δ is arbitrary, (taking $\delta = \frac{1}{n}$ and by Fatou's Lemma,) we have that

$$\zeta^*\left(\mathrm{dist}(x, y) > W_\infty(P, Q)\right) = 0.$$

That is,

$$\zeta^* \left(\mathrm{dist}(x, y) \leq W_\infty(P, Q)\right) = 1.$$

Therefore,

$$W_\infty(P, Q) \geq \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta^*)} \mathrm{dist}(x, y).$$

On the other hand, by the definition of W_∞, since $\zeta^* \in \Gamma(P, Q)$, we have

$$W_\infty(P, Q) \leq \inf_{\zeta\in\Gamma(P,Q)} \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta)} \mathrm{dist}(x, y) \leq \operatorname*{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta^*)} \mathrm{dist}(x, y).$$

Therefore, $W_\infty(P, Q) = \operatorname{ess\,sup}_{(x,y)\in(\Theta\times\Theta,\zeta^*)} \mathrm{dist}(x, y)$, for $\zeta^* \in \Gamma(P, Q)$, which proves the attainability of the infimum in the definition of W_∞. ■

K Facts about Noisy Gradient Descent

Noisy gradient descent is an alternative algorithm that can be used to obtain pure-DP or pure-Gaussian DP under the same assumptions we have. While it uses full batch gradients, its analysis relies on the theory of stochastic gradient descent since the full gradients in each iteration are perturbed by either the Laplace mechanism or the Gaussian mechanism. We explicitly write out the guarantees of noisy gradient descent in this section so as to substantiate our discussion related to our computational guarantee of localized-ASAP with MALA. We focus the discussion on Gaussian DP in the αn-strongly convex setting.

Theorem 4 (Noisy Gradient Descent for Lipschitz and strongly Convex Losses)**.** *Assume the loss function $\ell(\theta, (x, y))$ is G-Lipschitz for any data point (x, y). Assume $\sum_i \ell_i(\theta)$ is αn-strongly convex on Θ. Consider the following (projected) noisy gradient descent algorithm that initializes at θ_0 and update the parameter by T rounds using*

$$\theta_t = \mathrm{Proj}_\Theta(\theta_{t-1} - \eta_t(\sum_{i=1}^n \nabla\ell_i(\theta_{t-1}) + \mathcal{N}(0, \frac{TG^2}{2\mu^2} I_d)))$$

with $\eta_t = \frac{1}{\lambda t}$. Let θ^*_λ be the minimizer of the regularized ERM problem and θ^* to be the minimizer of the unregularized ERM problem.

1. *This algorithm that releases the whole trajectory $\theta_1, ..., \theta_T$ satisfies satisfies μ-GDP.*
2. *It also satisfies that*

$$\mathbb{E}\left[\mathcal{L}\left(\sum_{t=1}^T \frac{2t}{T(T+1)}\theta_t\right)\right] - \mathcal{L}(\theta^*_\lambda) \leq \frac{4n^2G^2}{\alpha n(T+1)} + \frac{dG^2}{2\alpha n\mu^2}.$$

3. *If $T \geq \frac{8n^2\mu^2}{d}$, then it achieves an excess empirical risk of $\frac{dG^2}{\alpha n\mu^2}$.*

Proof Observe that the G-Lipschitz loss says that the global sensitivity of the gradient is G. The privacy analysis follows from the composition of Gaussian mechanism via Gaussian DP for T rounds (Dong et al., 2022). This releases the entire sequence of parameters. The weighted average of the parameters over time is post-processing. The second statement follows from Section 3.2 of (Lacoste-Julien et al., 2012) by choosing the noise level and other parameters appropriately. The last statement is corollary by choosing T to be large. ■

In the above, the excess empirical risk achieves the optimal rates but requires the algorithm to run for $O(n^2)$ iterations. In the smooth case, one can obtain faster convergence but at the cost of the resulting excess empirical risk.

Theorem 5 (Noisy Gradient Descent for smooth and strongly convex Losses)**.** *Consider the same algorithm and assume all assumptions from Theorem 4. In addition, assume individual loss functions ℓ_x are β-smooth.*

1. *This algorithm that releases the whole trajectory $\theta_1, ..., \theta_T$ satisfies satisfies μ-GDP.*
2. *Choose a constant learning rate $\eta \leq \frac{1}{n\beta}$, run for T iterations*

$$\mathbb{E}\left[\mathcal{L}\left(\theta_T\right)\right] - \mathcal{L}(\theta_\lambda^*) \leq \frac{n\beta}{2}(1-\eta\alpha n)^T \mathrm{Diam}(\Theta)^2 + \frac{\eta n \beta d T G^2}{n\alpha\mu^2}$$

3. *Choose $\eta = 1/(n\beta)$ and $T = O(\frac{\beta}{\alpha}\log n)$, the excess risk is*

$$\mathbb{E}\left[\mathcal{L}\left(\theta_T\right)\right] - \mathcal{L}(\theta_\lambda^*) \leq O(\frac{\beta d G^2 \log n}{n\alpha^2\mu^2}).$$

Proof From Theorem 5.7 of (Garrigos & Gower, 2023) we can derive the convergence in various regimes. ■

It is not hard to see that there isn't a good choice of the learning rate parameter η based on the bound above that can get rid of the additional $\frac{\beta \log n}{\alpha}$ factor.