Journal of Machine Learning Research 26 (2025) 1-33 Submitted 1/25; Revised 10/25; Published 11/25

Mixing Times and Privacy Analysis for the Projected Langevin Algorithm under a Modulus of Continuity

Mario Bravo MARIO.BRAVO.G@USACH.CL
Departamento de Administración
Facultad de Administración y Economía
Universidad de Santiago de Chile
Av. Lib. Bdo. O'Higgins 3363, Santiago, Chile

Juan Pablo Flores-Mella JPFLORES1@UC.CL
Facultad de Matemáticas
Pontificia Universidad Católica de Chile
Vicuña Mackenna 4860, Santiago, Chile

Cristóbal Guzmán CRGUZMANP@UC.CL
Institute for Mathematical and Computational Engineering,
Facultad de Matemáticas and Escuela de Ingeniería
Pontificia Universidad Católica de Chile
Vicuña Mackenna 4860, Santiago, Chile

Editor: Jianfeng Lu

Abstract

We study the mixing time of the projected Langevin algorithm (LA) and the privacy curve of noisy Stochastic Gradient Descent (SGD), beyond nonexpansive iterations. Specifically, we derive new mixing time bounds for the projected LA which are, in some important cases, dimension-free and poly-logarithmic on the accuracy, closely matching the existing results in the smooth convex case. Additionally, we establish new upper bounds for the privacy curve of the subsampled noisy SGD algorithm. These bounds show a crucial dependency on the regularity of gradients, and are useful for a wide range of convex losses beyond the smooth case. Our analysis relies on a suitable extension of the Privacy Amplification by Iteration (PABI) framework (Feldman et al., 2018; Altschuler and Talwar, 2022, 2023) to noisy iterations whose gradient map is not necessarily nonexpansive. This extension is achieved by designing an optimization problem which accounts for the best possible Rényi divergence bound obtained by an application of PABI, where the tractability of the problem is crucially related to the modulus of continuity of the associated gradient mapping. We show that, in several interesting cases –namely the nonsmooth convex, weakly smooth and (strongly) dissipative– such optimization problem can be solved exactly and explicitly, yielding the tightest possible PABI-based bounds.

Keywords: Privacy Amplification by Iteration, Langevin Algorithm, Noisy Stochastic Gradient Descent, Differential Privacy

1. Introduction

Sampling from a log-concave distribution π (i.e., $\pi \propto e^{-f}$, where f is a convex potential) is a fundamental algorithmic problem and a basic building block for problems such as volume estimation (Kannan et al., 1997), optimization (Kalai and Vempala, 2006), Bayesian statistics (Welling and Teh, 2011), machine learning (Ho et al., 2020), and differential privacy (McSherry and Talwar, 2007). There is a wide variety of algorithms designed to solve this problem, with the Langevin algorithm (LA) as one of the prominent examples. The idea is to consider the Euler-Maruyama discretization of the Langevin diffusion

$$dL_t = -\nabla f(L_t)dt + \sqrt{2}dW_t \quad (t \geq 0),$$

where W_t is the d-dimensional Brownian motion. It is well-known that under mild assumptions, the diffusion has $\pi \propto e^{-f}$ as its unique stationary distribution –referred to as the *target distribution.* The rationale is that by discretizing the diffusion with a small step $\eta > 0$, we can use the Markov chain

$$X_{t+1} = X_t - \eta\nabla f(X_t) + \sqrt{2\eta}\xi_t \quad (t \in \mathbb{N}_0), \tag{LA}$$

where $(\xi_t)_t$ are i.i.d. standard d-dimensional Gaussians, to (approximately) simulate π.

Recently, Altschuler and Talwar (2022, 2023) have made major progress on understanding the procedure defined by

$$X_{t+1} = \Pi_{\mathcal{X}}[X_t - \eta\nabla f(X_t) + \sigma\xi_t] \quad (t \in \mathbb{N}_0), \tag{PLA}$$

where $(\xi_t)_t$ are i.i.d. d-dimensional Gaussians, $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact and convex set and Π is the projection operator. Their analysis focuses particularly on two cases: (i) when $\sigma = \sqrt{2\eta}$, referred to as the *Projected Langevin Algorithm*, first introduced in Bubeck et al. (2018), for which they establish mixing times, and (ii) when $\sigma = O(\eta)$, corresponding to *Noisy Stochastic Gradient Descent*, for which they investigate the privacy curve.

The thrust of their analysis is based on a technique known as *Privacy Amplification by Iteration* (henceforth, PABI) (Feldman et al., 2018), which leverages the nonexpansive properties of the gradient step in the smooth convex setting to gradually and recursively control the Rényi divergence of iterates, either under different initializations (used for mixing time arguments) or potentials (used for privacy arguments). Given the power of this technique and its potential to aid in understanding both privacy and sampling, we consider it important to study the PABI technique beyond the smooth convex scenario. We highlight that among the cases that we study are the convex and L-Lipschitz, which encompasses functions that can be nondifferentiable, the convex and (p, M)-weakly smooth, which interpolates between the Lipschitz and the smooth one, and the smooth strongly dissipative case. See Table 1 for a more precise summary.

1.1 Our Results

In this work, we conduct a study of the PABI technique beyond the case of nonexpansive iterations, together with some consequences for the mixing time and privacy analysis of this algorithm.

Extension of PABI for general mappings in terms of the modulus of continuity. We start by providing an extension of the PABI technique to iterations beyond the nonexpansive case. In order to do this, we quantify the regularity of the underlying mapping by its *modulus of continuity.* More precisely, for a vector valued map $\Phi : \mathbb{R}^d \to \mathbb{R}^d$, the nondecreasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a modulus of continuity of Φ if $\|\Phi(x) - \Phi(y)\| \leq \varphi(\|x - y\|)$ for all x, y. An interesting feature of this extension is that we can even address discontinuous mappings (characterized by their moduli of continuity being discontinuous at the origin). For instance, this extension is crucial for studying PABI under convex Lipschitz potentials, which are only subdifferentiable.

PABI works by gradually interpolating between a worst-case distance bound (quantified by the ∞-Wasserstein distance) and a Rényi divergence bound. This interpolation is performed by using the *shifted Rényi divergence*, which is an infimal convolution between the convex indicator of a ∞-Wasserstein ball (with radius given by the shift) and the Rényi divergence. By using a shift-reduction property of Gaussian noise addition, in conjunction with the nonexpansiveness of the gradient mapping, one can gradually reduce the shifts on the shifted divergences at the expense of an increase in the upper bound. This process concludes with zero shift, i.e. an upper bound on the Rényi divergence. Notice in particular that the shifts applied at the different steps are tunable parameters, which in the case of nonexpansive mappings are easily optimized by uniform shifts. In our case, the modulus of continuity leads to a nonconvex optimization problem in terms of the tuning parameters. Remarkably, when the modulus of continuity of the iteration is of the form $\varphi(\delta) = \sqrt{c\delta^2 + h}$, where $c, h \geq 0$ are two parameters[1], this optimization problem has a unique optimal solution with a closed-form expression.

The list below highlights important contexts where this type of modulus of continuity arises. Table 1 provides expressions for these moduli and also the bounds on the Rényi divergence of the final iterate, obtained from the application of Theorem 12 to each case. To the best of our knowledge, all bounds in Table 1 are new. We regard this as our main technical contribution.

Let $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a function where $\mathcal{X}$ is a closed convex set. We say that

- f is convex if for all $0 \leq \lambda \leq 1$, and $x, y \in \mathcal{X}$, $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y)$.
- f is (λ, κ)-strongly dissipative[2] if there exist $\lambda, \kappa > 0$ such that for all $x, y \in \mathcal{X}$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq -\lambda + \kappa \|x - y\|^2$.
- f is L-Lipschitz if there exists $L > 0$ such that for all $x, y \in \mathcal{X}$, $|f(x) - f(y)| \leq L \|x - y\|$.
- f is (p, M)-weakly smooth (or have p-Hölder continuous gradient) if there exist $M > 0$ and $0 \leq p \leq 1$ such that for all $x, y \in \mathcal{X}$, $\|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|^p$.
- f is β-smooth if there exist $\beta > 0$ such that for all $x, y \in \mathcal{X}$, $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$.

See Section 2 for further details.

Mixing times. We show that the PABI technique yields a polynomial upper bound on the mixing time in total variation distance of the projected Langevin algorithm in the convex and nonsmooth case, including cases where the potential is only subdifferentiable.

[1] It is understood that for any function the modulus of continuity is such that $\varphi(0) = 0$. Hence, we will use formulae as above to refer to φ for strictly positive values. With this in mind, it is clear that $\varphi(\delta) = \sqrt{c\delta^2 + h}$ is discontinuous at zero if and only if $h > 0$.

[2] The denomination of strongly dissipative is not standard in the literature. We introduce it to distinguish it from the more standard notion of dissipativity, where y is a fixed vector.

Assumptions on f	Rényi divergence of order α	c	h
Convex, L-Lipschitz	$\frac{\alpha}{2\sigma^2}\left(\frac{D^2}{T}+h\sum_{t=1}^{T}\frac{1}{t}\right)$	1	$(2\eta L)^2$
Convex, (p,M)-w.s.	$\frac{\alpha}{2\sigma^2}\left(\frac{D^2}{T}+h\sum_{t=1}^{T}\frac{1}{t}\right)$	1	$\left(2\eta^{\frac{1}{1-p}}\sqrt{\frac{1-p}{1+p}}\left(\frac{M}{2}\right)^{\frac{1}{1-p}}\right)^2$
(λ,κ)-Str. Dissip, β-smooth	$\frac{\alpha}{2\sigma^2}\left(\frac{D^2c^T(1-c)}{(1-c^T)}+h\ln\left(\left(\frac{1-c^T}{1-c}\right)e\right)\right)$	$1-2\eta\kappa+\eta^2\beta^2$	$2\eta\lambda$

Table 1: Summary of the Rényi divergence bounds between the last iterates under two different initializations of (PLA). For all rows, the corresponding moduli of continuity can be bounded by $\varphi(\delta)=\sqrt{c\delta^2+h}$, with c,h given in the corresponding columns. Here D is the diameter of $\mathcal{X}$, T is the number of iterations in the algorithm and σ^2 is the (coordinate-wise) variance for Gaussian noise. For more information, see the list below Definition 4.

We also provide a mixing time bound for the strongly dissipative case, which is logarithmic in the diameter, but exponential in the parameter λ. The following are informal versions of the Theorems, whose complete statements can be found in Section 4.

Theorem 1 (Abridged version of Theorem 19) *Let $\mathcal{X}\subseteq\mathbb{R}^d$ be a convex, compact set with diameter $D>0$ and suppose that $f:\mathcal{X}\to\mathbb{R}$ is a convex and (p,M)-weakly smooth function, with $0\le p\le 1$. There exists a constant Θ such that if $1/\eta\ge\Theta$, then for all $\varepsilon>0$,*

$$T_{mix,TV}(\varepsilon)\le\left\lceil\frac{D^2}{\eta}\right\rceil\cdot\lceil\log_2(1/\varepsilon)\rceil.$$

We remark that Θ depends polylogarithmicaly in the diameter D and polynomially in M,p; furthermore, p modulates these dependencies, making them near-quadratic on M when $p=0$ (Lipschitz case) and linear on M when $p=1$ (smooth case). For an explicit expression, see (5). Also, we stress that our upper bound operate under a slightly more restrictive stepsize constraint, but otherwise the bound on mixing time is identical to that of the smooth one (Altschuler and Talwar, 2023), for all $0\le p\le 1$. This includes the whole interpolation from the convex and nonsmooth ($p=0$) to the convex and smooth case ($p=1$).

Theorem 2 (Abridged version of Theorem 20) *Let $\mathcal{X}\subseteq\mathbb{R}^d$ be a convex, compact set with diameter $D>0$ and suppose that $f:\mathcal{X}\to\mathbb{R}$ is a (λ,κ)-strongly dissipative and β-smooth function. If $c=1-2\eta\kappa+\eta^2\beta^2<1$, then for all $\varepsilon>0$,*

$$T_{mix,TV}(\varepsilon)=O\left(\log_{1/c}\left(1+\frac{D^2(1-c)}{4\eta}\right)\cdot\left(\frac{e}{1-c}\right)^{\lambda/2}\log_2(1/\varepsilon)\right).$$

Privacy curve. We also study the impact of our PABI results for the privacy curve of noisy SGD, in the convex setting. For any nontrivial Hölder gradient regularity, we have that the privacy curve caps in a similar fashion to that proven in (Altschuler and Talwar, 2022), except for the addition of an extra term (denoted by V in the corresponding section) that depends on η (the step) and the Hölder regularity of the gradient. Particularly, we prove that for L-Lipschitz and (p, M)-weakly smooth losses, and under mild restrictions over the Rényi divergence parameter $\alpha \geq 1$, variance σ^2, and number of iterations T; noisy SGD satisfies the following Rényi DP bound.

Theorem 3 (Abridged version of Theorem 22) *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex and compact set with diameter $D > 0$. There exists $\overline{T} > 0$ and a function V such that for $T > \overline{T}$, datasize $n \in \mathbb{N}$, expected batch size b, stepsize $\eta > 0$ and initialization $x_0 \in \mathcal{X}$, the last iteration satisfies (α, ε)-RDP for*

$$\varepsilon \leq \frac{16\alpha L^2}{n^2\sigma^2} \min\Big\{T, 2\overline{T} + V(D, M, \overline{T}, \eta, p)\Big\}.$$

This shows that for convex and (p, M)-weakly smooth losses, the privacy curve caps in a similar fashion to that of the smooth and convex ones, except for an additive term V. For details, see Section 5. In particular, see Figure 1 for a plot comparing the privacy curves. However, our conclusions for the nondifferentiable case fall short: it is not possible to obtain any nontrivial privacy amplification, even when the sample size tends to infinity. Note however that due to the optimality of our PABI optimization problem, and the tighness of the modulus of continuity for the gradient mapping we use, these pessimistic results exhibit the inherent limits of PABI in the nonsmooth convex setting.

1.2 Related Work

A substantial part of the studies for the Langevin algorithm have focused on the strongly convex and smooth potential setting (e.g. Dalalyan 2017a,b; Durmus and Moulines 2019). Most of the research here has focused on approximation bounds (e.g. in Wasserstein or total variation distance) between the last iterate of LA with the target distribution, π. These arguments are based on using the stationarity of the target distribution under the difussion, together with a coupling between the discrete and continuous Langevin dynamics, to control the distance. These results lead to bounds that blow up with respect to the time (of the discrete chain), and therefore these are inherently finite-iteration statements. This body of work primarily targets approximation to the stationary distribution, π, of the diffusion, typically in unconstrained Euclidean settings. By contrast, the PABI approach undertaken in Altschuler and Talwar (2023), provides a mixing time bound of the Projected Langevin algorithm to its own stationary distribution. Although this distribution may not be the target distribution π, for some purposes this approximation in unnecessary: this is the case e.g. in differentially-private optimization, where the goal is bounding the empirical or population risk, together with a divergence bound (e.g. Rényi) between two outputs using datasets that only differ in a single example. For references on the capabilities of noisy iterative methods for this problem, we refer the reader to (Bassily et al., 2014, 2020).

Far less work has been devoted to the case of nonsmooth convex potentials; in fact, several works consider the Langevin algorithm for nonsmooth potentials as far less understood (e.g. Pereyra 2016; Chatterji et al. 2020; Mitra and Wibisono 2024). A first natural approach to reduce the nonsmooth setting to a smooth one is using a convolution-type smoothing. This includes the case of proximal algorithms (e.g. Pereyra 2016; Durmus et al.

2018; Wibisono 2019), or randomized smoothing (e.g. Chatterji et al. 2020). Regarding the former, while it can be preferable to use proximal smoothing due to its stability properties, these methods require proximal mapping computations, which are tractable only for very structured cases. Regarding the latter, even if randomized smoothing is easily implementable, the smoothness of the resulting functions have polynomial dependence on the dimension, which leads to sample complexity results which are quite expensive. All these works focus on the approximation to the diffusion's stationary distribution, without any projection.

To the best of our knowledge, the only existing work that analyzes the (Projected) Langevin algorithm in the convex Lipschitz case is Lehec (2023). This work extends the coupling techniques used in the smooth case (Dalalyan, 2017a), observing that the monotonicity of the gradient suffices for nonexpansiveness of the Langevin difussion, and using discrete/continuous time couplings establishes approximation results in the Wasserstein metric. Lehec (2023) covers both constrained (compactly supported) and unconstrained domains, illustrating how the same coupling techniques can be adapted to different geometric settings.

We emphasize that almost all of these results are independent and incomparable to ours, as they do not imply mixing-time bounds. On the other hand, our results do not focus on the approximation to the target distribution. Additionally, upon completing this paper, we learned of related work by Johnston et al. (2025), which establishes approximation results to π in the Wasserstein-1 and Wasserstein-2 metrics for potentials that are semiconvex on a ball and strongly convex outside it. Their analysis also accommodates discontinuous gradients by employing subgradient techniques. The results are of a different nature from ours, due to the differing assumptions and approximation criteria.

In the case of differential privacy, classical analyses of differentially private iterative algorithms assume that all iterates are published, leading to unbounded growth in privacy parameters with the number of iterations. It is often the case however that only the last iteration is published. Recent works have explored this setting. Among these, Chourasia et al. (2021) and Ye and Shokri (2022) show that, in the smooth and strongly convex case, the Rényi Differential Privacy (RDP) of variants of noisy SGD approaches a constant bound exponentially quickly. Altschuler and Talwar (2022) show, using PABI, that in the smooth and (strongly) convex settings the RDP of projected noisy SGD stops growing after a certain number of iterations. Asoodeh and Diaz (2023) prove a convergent upper bound for the privacy of DP-SGD, even in nonconvex settings, using Hockey-Stick divergence.

We further note that several other recent studies have addressed related questions, highlighting the critical role of modulus of continuity in the privacy analysis of the last iterate of noisy SGD, underscoring the importance of this line of investigation. The first, Kong and Ribero (2024), investigates the last iteration of noisy SGD without assuming random sampling of the data set, employing clipped gradients. Their analysis diverges from ours for several reasons: they operate under distinct hypotheses, leverage an affine modulus of continuity, and directly incorporate sensitivity into their methodology. The second paper, Chien and Li (2024), explores scenarios where the gradients exhibit Hölder continuity, allowing the objective function to be non-convex. This assumption leads to a modulus of continuity that differs from ours and yields distinct bounds. Moreover, they cannot analytically resolve their shift optimization problem. Their work also employs a variation of the PABI mechanism, which differs slightly from the diameter-aware version introduced by Altschuler and Talwar (2022).

1.3 Organization of the Paper

This paper is organized as follows. Section 2 presents the necessary background results for the reading of the paper. Section 3 provides the extension of PABI to the modulus of continuity setting. Section 4 presents mixing times in total variation for convex and Lipschitz, convex and (p, M)-weakly smooth and (λ, κ)-strongly dissipative and β-smooth settings. Section 5 presents a privacy analysis for the last iteration of noisy SGD for convex and (p, M)-weakly smooth functions.

We also add several appendices to complement the presentation. Appendix A presents a brief summary of useful results that may be used for quick inspection. Appendix B shows that the problem we solve for the PABI extension is nonconvex, while Appendix C deals with the existence of a stationary distribution of (PLA) with $\sigma^2 = 2\eta$, when the potential, f, is convex and L-Lipschitz (possibly nondifferentiable).

2. Preliminaries

2.1 Vector Spaces and Convex Functions

We work over the standard Euclidean space $(\mathbb{R}^d, \|\cdot\|)$ (that is $\|\cdot\| = \|\cdot\|_2$ is the ℓ_2-norm). If $\mathcal{X}$ is a closed convex set, we denote by $\Pi_{\mathcal{X}} : \mathbb{R}^d \mapsto \mathcal{X}$ the Euclidean projection operator, which we recall is nonexpansive. We denote by $I_{d\times d}$ the d-dimensional identity matrix.

We now introduce the modulus of continuity, which serves as a measure of regularity of functions, and it is the main property used in the PABI technique we will introduce later.

Definition 4 (Modulus of continuity) *Let $\Phi : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ be a map. We say that a nondecreasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a modulus of continuity of Φ if*

$$\|\Phi(x) - \Phi(y)\| \leq \varphi(\|x - y\|) \quad \forall x, y \in \mathcal{X}\,.$$

Note that $\lim_{t\to 0^+} \varphi(t) = 0$ implies Φ is continuous and that we can always assume that $\varphi(0) = 0$.

For a function $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$, given $\eta > 0$, let $\Phi(x) = x - \eta\nabla f(x)$ be the gradient mapping. We have that

1. If f is convex and L-Lipschitz, then $\varphi(\delta) = \sqrt{\delta^2 + (2\eta L)^2}$ (Bassily et al., 2020, Lemma 3.1).

2. If f is convex and (p, M)-weakly smooth, then $\varphi(\delta) = \sqrt{\delta^2 + \left(2\eta^{\frac{1}{1-p}}\sqrt{\frac{1-p}{1+p}}\left(\frac{M}{2}\right)^{\frac{1}{1-p}}\right)^2}$ (Lei and Ying, 2020, Lemma D.3).

   Note that we recover the modulus of continuity of the Lipschitz case when $p = 0$. The only difference is in the Lipschitz constant that is now $(M/2)$. Thus, a $(0, 2L)$-weakly smooth function has the same modulus of continuity as a L-Lipschitz one.

3. If f is (λ,κ)-strongly dissipative and β-smooth, then $\varphi(\delta)=\sqrt{(1 - 2\eta\kappa + \eta^2\beta^2)\delta^2 + 2\eta\lambda}$.

We were unable to find a reference providing a modulus of continuity bound in the strongly dissipative case; therefore, we include a brief explanation for completeness. Suppose $\|x-y\| \leq \delta$. Then

$$\begin{aligned}\|x-\eta\nabla f(x)-(y-\eta\nabla f(y))\|^2 &= \|x-y\|^2 - 2\eta\langle\nabla f(x)-\nabla f(y), x-y\rangle + \eta^2\,\|\nabla f(x)-\nabla f(y)\|^2\\ &\leq \delta^2 - 2\eta\langle\nabla f(x)-\nabla f(y), x-y\rangle + \eta^2\beta^2\delta^2\\ &\leq (1-2\eta\kappa+\eta^2\beta^2)\delta^2 + 2\eta\lambda,\end{aligned}$$

where in the first line we expand the square; in the second line we use the β-smoothness of f and the bound on $\|x-y\|$; in the third line we use the (λ,κ)-strong dissipativity of f.

2.2 Information Theory and Probability Divergences

In the following we use $\mathcal{P}(\mathcal{X})$ to denote the set of all probability measures supported in $\mathcal{X}$ and $\mathcal{B}(\mathcal{X})$ to denote the Borel σ-algebra of $\mathcal{X}$.

Definition 5 (Rényi Divergence) *Let $\alpha \in (1,+\infty)$ and $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures on $\mathbb{R}^d$. We define the Rényi divergence of order α between μ,ν as:*

$$R_\alpha(\mu||\nu) = \begin{cases}\frac{1}{\alpha-1}\ln\left(\int_{\mathbb{R}^d}\left(\frac{d\mu}{d\nu}(x)\right)^\alpha \nu(dx)\right) & \textit{if } \mu \ll \nu\\ +\infty & \textit{otherwise.}\end{cases}$$

It is well-known (e.g. Van Erven and Harremos (2014, Theorem 5)) that if there exists $\beta > 1$ such that $R_\beta(\mu||\nu) < \infty$, then $\mathrm{KL}(\mu||\nu) = \lim_{\alpha\to 1^+} R_\alpha(\mu||\nu)$.

In a slight abuse of notation, we write Rényi divergences applied to random variables meaning the divergence of the respective distributions.

Definition 6 (Coupling and ∞-Wasserstein Distance) *Let $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. We say that $\gamma \in \mathcal{P}(\mathbb{R}^d)$ is a coupling of μ and ν if for all $A \in \mathcal{B}(\mathbb{R}^d)$*

$$\gamma(A\times\mathbb{R}^d) = \mu(A) \qquad \textit{and} \qquad \gamma(\mathbb{R}^d\times A) = \nu(A).$$

We denote by $\Gamma(\mu,\nu)$ the set of all couplings between μ and ν.

We say that a pair of random variables is a coupling of μ and ν if its joint distribution is in $\Gamma(\mu,\nu)$; i.e. $X \sim \mu$ and $X' \sim \nu$.

Finally, we define the ∞-Wasserstein distance between μ and ν as

$$W_\infty(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \operatorname*{ess\,sup}_{(x,y)\sim\gamma} \|x-y\|\,.$$

A key tool for analyzing is the so called *Shifted Rényi Divergence*, which helps to interpolate between a W_∞ guarantee –which holds trivially in the compact setting– and a Rényi divergence one.

Definition 7 (Shifted Rényi Divergence) *Let $\alpha \in [1,+\infty)$, $\delta \geq 0$, and $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. We define the δ-shifted Rényi divergence of order α as:*

$$R_\alpha^{(\delta)}(\mu||\nu) = \inf_{\mu':W_\infty(\mu,\mu')\leq\delta} R_\alpha(\mu'||\nu).$$

Two useful properties of the shifted Rényi divergence are $R_\alpha^{(0)}(\mu||\nu) = R_\alpha(\mu||\nu)$, and that $W_\infty(\mu,\nu) \leq \delta$, implies $R_\alpha^{(\delta)}(\mu||\nu) = 0$. These properties account for the final and initial bounds on PABI. A key property for PABI is the shift-reduction property of Gaussian noise addition (Feldman et al., 2018, Lemma 20).

Lemma 8 (Shift-reduction) *For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $a, \delta \geq 0$,*

$$R_\alpha^{(\delta)}\left(\mu * \mathcal{N}\left(0, \sigma^2 I_{d\times d}\right) || \nu * \mathcal{N}\left(0, \sigma^2 I_{d\times d}\right)\right) \leq R_\alpha^{(\delta+a)}\left(\mu || \nu\right) + \frac{\alpha a^2}{2\sigma^2}.$$

3. Privacy Amplification by Iteration Under a Modulus of Continuity

We start by providing a simple extension of the PABI framework for nonexpansive iterations to the case of general maps under a modulus of continuity assumption. As discussed in the introduction, we want to study iterations of the form

$$X_{t+1} = \Pi_{\mathcal{X}}[\Phi_t(X_t) + \xi_t], \qquad \xi_t \sim \mathcal{N}(0, \sigma_t^2 I_{d\times d}).$$

In what follows, we denote by φ_t the modulus of continuity of the map Φ_t. We are interested in bounding the Rényi divergence of two different trajectories of this algorithm, either under different initializations (to obtain mixing time results) or under different maps Φ_t, Φ_t' (to prove privacy results). The following two results are an adaptation of Feldman et al. (2018, Lemma 21) and Feldman et al. (2018, Theorem 22), respectively.

Lemma 9 (Coupling under modulus of continuity) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\delta \geq 0$ and $\alpha \in [1, +\infty)$. Let also $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ be a map with modulus of continuity φ. Then*

$$R_\alpha^{(\varphi(\delta))}\left(\Phi_\# \mu || \Phi_\# \nu\right) \leq R_\alpha^{(\delta)}\left(\mu || \nu\right),$$

where $\Phi_\#\mu$ and $\Phi_\#\nu$ denote the pushforward measure of μ and ν through Φ, respectively.

Proof Let μ' be such that $W_\infty(\mu, \mu') \leq \delta$ and $R_\alpha(\mu'||\nu) = R_\alpha^{(\delta)}(\mu||\nu)$. Let (X, X') be a coupling of (μ, μ') such that $\|X - X'\| \leq \delta$ a.s.
Then $\|\Phi(X) - \Phi(X')\| \leq \varphi(\|X - X'\|) \leq \varphi(\delta)$ a.s. Also, by the data-processing inequality (Proposition 28),

$$R_\alpha(\Phi_\#\mu'||\Phi_\#\nu) \leq R_\alpha(\mu'||\nu) = R_\alpha^{(\delta)}(\mu||\nu).$$

Therefore, since $(\Phi(X), \Phi(X'))$ is a coupling of $(\Phi_\#\mu, \Phi_\#\mu')$,

$$R_\alpha^{(\varphi(\delta))}(\Phi_\#\mu||\Phi_\#\nu) \leq R_\alpha(\Phi_\#\mu'||\Phi_\#\nu) \leq R_\alpha^{(\delta)}(\mu||\nu).$$

■

Lemma 10 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex set and $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ be a map with modulus of continuity φ. Also, let X, X' be two random variables in $\mathbb{R}^d$ and $\xi \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$ be centered Gaussian noise with variance σ^2. Let $Y = \Pi_{\mathcal{X}}\left[\Phi(X) + \xi\right]$ and $Y' = \Pi_{\mathcal{X}}\left[\Phi(X') + \xi\right]$, and let $\delta \geq 0$. Then, for any $0 < a \leq \varphi(\delta)$,*

$$R_\alpha^{(\varphi(\delta)-a)}(Y||Y') \leq R_\alpha^{(\delta)}(X||X') + \frac{\alpha a^2}{2\sigma^2}.$$

Proof By a succesive application of Lemmas 9, 8 and 9 again, we have that:

$$\begin{aligned} R_\alpha^{(\delta)}(X||X') &\geq R_\alpha^{(\varphi(\delta))}(\Phi(X)||\Phi(X')) \\ &\geq R_\alpha^{(\varphi(\delta)-a)}(\Phi(X)+\xi||\Phi(X')+\xi) - \frac{\alpha a^2}{2\sigma^2} \\ &\geq R_\alpha^{(\varphi(\delta)-a)}(\Pi_{\mathcal{X}}\left[\Phi(X)+\xi\right]||\Pi_{\mathcal{X}}\left[\Phi(X')+\xi\right]) - \frac{\alpha a^2}{2\sigma^2}, \end{aligned}$$

where in the last step we used the nonexpansiveness of $\Pi_{\mathcal{X}}$. ■

We introduce now a simplifying notation for the PABI induction. Given $T \in \mathbb{N}$, $D > 0$ and $\mathbf{a} = (a_t)_{t=1}^T$ be a sequence of nonnegative reals. We define $\varphi_{[0:0]}(D, \mathbf{a}) := \varphi_0(D) - a_1$, and

$$\varphi_{[0:t]}(D, \mathbf{a}) := \varphi_t\left(\varphi_{[0:t-1]}(D, \mathbf{a})\right) - a_{t+1}, \quad \text{for} \quad t = 1, \ldots, T-1.$$

Lemma 11 (Privacy amplification by iteration under a modulus of continuity) *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex and compact set with diameter $D > 0$. Let $\mu_0, \mu_0' \in \mathcal{P}(\mathcal{X})$ be two probability measures. For every $t \in \mathbb{N}_0$, let $\Phi_t : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ be a mapping with modulus of continuity φ_t and let $(\xi_t)_{t\in\mathbb{N}_0} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_t^2 I_{d\times d})$. Define $(X_t)_{t\in\mathbb{N}_0}$ and $(X_t')_{t\in\mathbb{N}_0}$ respectively as*

$$\begin{aligned} &X_0 \sim \mu_0 && \textit{and} && X_0' \sim \mu_0' \\ &X_{t+1} = \Pi_{\mathcal{X}}\left[\Phi_t(X_t) + \xi_t\right], && && X_{t+1}' = \Pi_{\mathcal{X}}\left[\Phi_t(X_t') + \xi_t\right]. \end{aligned}$$

Let $\mathbf{a} = (a_t)_{t=1}^T$ be such that $a_t \geq 0$ and $\varphi_{[0:t-1]}(D, \mathbf{a}) \geq 0$ for all $t \leq T$. Then

$$R_\alpha^{\left(\varphi_{[0:T-1]}(D,\mathbf{a})\right)}(X_T||X_T') \leq \frac{\alpha}{2}\sum_{t=1}^T \frac{a_t^2}{\sigma_{t-1}^2}. \tag{1}$$

Proof The proof follows by iteratively applying Lemma 10 starting from $R_\alpha^{(D)}(X_0||X_0') = 0$. ■

We will refer to those sequences $(X_t)_{t\in\mathbb{N}_0}$ having form as in Lemma 11 as *Projected Noisy Iterations with moduli of continuity* $(\varphi_t)_t$.

3.1 The Shifts Optimization Problem

We focus now on the optimization of parameters to obtain the tightest possible PABI bound. In order to bound $R_\alpha(X_T||X_T')$, we need to find a suitable sequence $\mathbf{a} = (a_t)_{t=1}^T$ such that $\varphi_{[0:T-1]}(D, \mathbf{a}) = 0$. This is because, from Lemma 11 and the fact that the 0-shifted Rényi Divergence is the Rényi Divergence, we can obtain a bound for $R_\alpha(X_T||X_T')$. Since there are various *feasible shifts* $\mathbf{a}$ to choose from, our goal is to minimize the right hand side of equation (1). More precisely, for $D > 0$, a sequence $\mathbf{a} = (a_t)_{t=1}^T$ of nonnegative reals is a sequence of feasible shifts if for all $t = 1, \ldots, T$

$$a_t \geq 0, \quad \varphi_{[0:t-1]}(D, \mathbf{a}) \geq 0, \quad \text{and} \quad \varphi_{[0:T-1]}(D, \mathbf{a}) = 0.$$

Note that $\mathbf{a} = (a_t)_{t=1}^T$ is a feasible shift if and only if $a_t = \varphi_{t-1}(u_{t-1}) - u_t$ for all $t = 1, \dots, T$, where $(u_t)_{t=0}^T$ is a sequence of nonnegative numbers that satisfies

$$u_0 = D, \quad u_T = 0, \text{ and } \quad \varphi_{t-1}(u_{t-1}) \geq u_t \quad \forall t = 1, \dots, T. \tag{2}$$

Hence we can restate the problem of finding a sequence of feasible shifts $(a_t)_{t=1}^T$ that minimizes the right-hand-side of (1) by the equivalent problem of finding a sequence of nonnegative real numbers $(u_t)_{t=0}^T$ that satisfies (2) and minimizes

$$\frac{\alpha}{2} \sum_{t=1}^{T} \frac{(\varphi_{t-1}(u_{t-1}) - u_t))^2}{\sigma_{t-1}^2}.$$

Let us call $\mathcal{R}$ the set of parameters $\mathbf{u} = (u_1, \dots, u_{T-1}) \in \mathbb{R}^{T-1}$ that satisfy (2).

Then, in order to obtain the tightest possible PABI upper bound, we consider the problem

$$\min_{\mathbf{u} \in \mathcal{R}} \left[E(\mathbf{u}) := \sum_{t=1}^{T} \frac{(\varphi_{t-1}(u_{t-1}) - u_t)^2}{\sigma_{t-1}^2} \right]. \tag{P}$$

3.2 Solving the Shifts Optimization Problem

We now study the shifts optimization problem under a modulus of continuity assumption that encompasses families of both nonsmooth and smooth potentials.

While under the studied modulus of continuity the objective E has a positive-definite Hessian at every point of $\mathcal{R}$, in general $\mathcal{R}$ is a nonconvex domain, which prevents us from a simple first-order condition characterization of optimality (see Appendix B for more details). We will nevertheless characterize the first-order conditions, and then show by alternative arguments that this is indeed an optimal solution for problem (P).

We present next what we deem as our main result.

Theorem 12 *Let $(c_t)_{t \in \mathbb{N}_0}$ be a sequence of strictly positive real numbers, $(h_t)_{t \in \mathbb{N}_0}$ a sequence of nonnegative real numbers, and $\varphi_t(\delta) = \sqrt{c_t \delta^2 + h_t}$.*

If $(X_t)_{t \in \mathbb{N}_0}$ and $(X'_t)_{t \in \mathbb{N}_0}$ are projected noisy iterations with moduli of continuity $(\varphi_t)_t$, which only differ in their initialization and whose domain, $\mathcal{X}$, has diameter $D > 0$, then

$$R_\alpha(X_T || X'_T) \leq \frac{\alpha}{2} \left(\frac{\Pi_{k=0}^{T-1} c_k D^2}{\sum_{j=0}^{T-1} \sigma_j^2 \Pi_{l=j+1}^{T-1} c_l} + \sum_{t=0}^{T-1} \frac{h_t \Pi_{k=t+1}^{T-1} c_k}{\sum_{j=t}^{T-1} \sigma_j^2 \Pi_{l=j+1}^{T-1} c_l} \right).$$

To prove the Theorem, we need to first characterize the optimal solution of the shifts optimization problem (P). This is done in following Lemma, which provides the unique solution for this problem.

Lemma 13 *Let $(c_t)_{t=0}^{T-1}$ be a sequence of strictly positive numbers, $(h_t)_{t=0}^{T-1}$ be a sequence of nonnegative numbers and let $\varphi_t(\delta) = \sqrt{c_t \delta^2 + h_t}$ be the t-th modulus of continuity. Let also $E : \mathbb{R}^{T-1} \to \mathbb{R}$ be defined as in* (P) *and $\mathbf{u}^* \in \mathbb{R}^{T-1}$ be recursively defined as:*

$$u_t^* = \left(\frac{\sum_{k=t}^{T-1} \Pi_{l=k+1}^{T-1} c_l \sigma_k^2}{\sum_{j=t-1}^{T-1} \Pi_{l=j+1}^{T-1} c_l \sigma_j^2} \right) \varphi_{t-1}(u_{t-1}), \quad \text{for all} \quad t = 1, \dots, T-1.$$

Then $\mathbf{u}^* \in \mathcal{R}$ and is the unique minimizer of E over $\mathbb{R}^{T-1}$.

Proof First, note that $\mathbf{u}^* \in \mathcal{R}$. This follows from the fact that

$$\left(\frac{\sum_{k=t}^{T-1} \Pi_{l=k+1}^{T-1} c_l \sigma_k^2}{\sum_{j=t-1}^{T-1} \Pi_{l=j+1}^{T-1} c_l \sigma_j^2} \right) \leq 1 \quad (\forall t = 1, \dots, T).$$

We break the proof of $\mathbf{u}^*$ being a minimizer into two separate statements: $\mathbf{u}^*$ is the unique stationary point and $\mathbf{u}^*$ is the global minimizer.

$\mathbf{u}^*$ is the unique stationary point of E: Computing the partial derivatives of E and arranging the terms, we get that the stationary conditions for $\mathbf{u}$ are

$$(c_t \sigma_{t-1}^2 + \sigma_t^2) u_t - \sigma_{t-1}^2 \varphi_t'(u_t) u_{t+1} = \sigma_t^2 \varphi_{t-1}(u_{t-1}), \quad \forall t = 1, \dots, T-1. \tag{3}$$

We will show by reverse induction that if $\mathbf{u}$ satisfies the (3), then $\mathbf{u} = \mathbf{u}^*$.

Suppose $\mathbf{u}$ satisfies the stationary conditions. Then, for $t = T-1$, we have that

$$\begin{aligned} u_{T-1} &= \left(\frac{\sigma_{T-1}^2}{c_{T-1} \sigma_{T-2}^2 + \sigma_{T-1}^2} \right) \varphi_{T-2}(u_{T-2}) \\ &= u_{T-1}^* \end{aligned}$$

As induction hypothesis (IH, from now on), suppose that

$$u_{T-s} = u_{T-s}^* = \left(\frac{\sum_{k=T-s}^{T-1} \Pi_{l=k+1}^{T-1} c_l \sigma_k^2}{\sum_{j=T-(s+1)}^{T-1} \Pi_{l=j+1}^{T-1} c_l \sigma_j^2} \right) \varphi_{T-(s+1)}(u_{T-(s+1)}).$$

Plugging IH into (3) for $t = T-(s+1)$ and reordering terms, we get

$$\sigma_{T-(s+1)}^2 \left(\frac{\sum_{k=T-(s+2)}^{T-1} \Pi_{l=k+1}^{T-1} c_l \sigma_k^2}{\sum_{j=T-(s+1)}^{T-1} \Pi_{l=j+1}^{T-1} c_l \sigma_j^2} \right) u_{T-(s+1)} = \sigma_{T-(s+1)}^2 \varphi_{T-(s+2)}(u_{T-(s+2)}).$$

Therefore,

$$u_{T-(s+1)} = \left(\frac{\sum_{k=T-(s+1)}^{T-1} \Pi_{l=k+1}^{T-1} c_l \sigma_k^2}{\sum_{j=T-(s+2)}^{T-1} \Pi_{l=j+1}^{T-1} c_l \sigma_j^2} \right) \varphi_{T-(s+2)}(u_{T-(s+2)}),$$

completing the induction.

$\mathbf{u}^*$ is the global minimizer: Since E is continuous and nonnegative, there exists a sequence $(\mathbf{x}^n)_{n \in \mathbb{N}} \subseteq \mathbb{R}^{T-1}$ such that

$$\lim_{n \to \infty} E(\mathbf{x}^n) = \inf_{\mathbf{u} \in \mathbb{R}^{T-1}} E(\mathbf{u}) > -\infty.$$

We will prove by reverse induction that the sequences of coordinates of $(\mathbf{x}^n)_{n \in \mathbb{N}}$ are bounded. To simplify notation, let $x_T^n = 0$ and $x_0^n = D$ for all $n \in \mathbb{N}$. Since

$$E(\mathbf{x}^n) = \sum_{t=1}^{T} \frac{(\varphi_{t-1}(x_{t-1}^n) - x_t^n)^2}{\sigma_{t-1}^2} \geq \frac{\varphi_{T-1}(x_{T-1}^n)^2}{\sigma_{T-1}^2},$$

the sequence $(x^n_{T-1})_{n\in\mathbb{N}}$ must be bounded. Assume now that the sequence $(x^n_{T-s})_{n\in\mathbb{N}}$ is bounded for $s \geq 1$. Then, by the induction hypothesis and the fact that

$$E(\mathbf{x}^n) = \sum_{t=1}^{T} \frac{(\varphi_{t-1}(x^n_{t-1}) - x^n_t)^2}{\sigma^2_{t-1}} \geq \frac{\left(\varphi_{T-(s+1)}(x^n_{T-(s+1)}) - x^n_{T-s}\right)^2}{\sigma^2_{T-(s+1)}},$$

$(x^n_{T-(s+1)})_{n\in\mathbb{N}}$ must also be bounded, which concludes the induction. Since the sequences of coordinates of $(\mathbf{x}^n)_{n\in\mathbb{N}}$ are all bounded, $(\mathbf{x}^n)_{n\in\mathbb{N}}$ is also bounded. Then, by the Bolzano-Weierstrass Theorem, there exists $\mathbf{x}^* \in \mathbb{R}^{T-1}$ that minimizes E, and by first-order conditions $\mathbf{x}^* = \mathbf{u}^*$. ■

Proof *(Proof of Theorem 12)* Note that the optimal solution $(u^*_1, \ldots, u^*_{T-1}) = \mathbf{u}^* \in \mathcal{R}$ determined in the above Lemma, allow to define a sequence of feasible shifts $a^*_t = \varphi_{t-1}(u^*_{t-1}) - u^*_t$ (see Section 3.1). Hence, by Lemma 11,

$$R_\alpha(X_T||X'_T) \leq \frac{\alpha}{2}\sum_{t=1}^{T} \frac{a_t^2}{\sigma^2_{t-1}} = \frac{\alpha}{2}E(\mathbf{u}^*) = \frac{\alpha}{2}\left(\frac{\Pi_{k=0}^{T-1} c_k D^2}{\sum_{j=0}^{T-1}\sigma_j^2 \Pi_{l=j+1}^{T-1} c_l} + \sum_{t=0}^{T-1} \frac{h_t \Pi_{k=t+1}^{T-1} c_k}{\sum_{j=t}^{T-1}\sigma_j^2 \Pi_{l=j+1}^{T-1} c_l}\right),$$

where the last equality follows by a simple evaluation of $E(\mathbf{u}^*)$. ■

Remark 14 *There are moduli of continuity of interesting problems which are not of the form $\varphi(\delta) = \sqrt{c\delta^2 + h}$. One of them comes from the noisy Gradient Descent-Ascent applied to convex/concave saddle-point problems. It can be shown that if the potential and its gradient are L-Lipschitz and β-smooth, respectively, then the modulus of continuity of the iteration is $\varphi(\delta) = \sqrt{\delta^2 + \min\{2\eta L, \eta\beta\delta\}^2}$. We were not able to analyze this modulus. However, it should be noted that it is possible to compare it with that obtained from the convex and Lipschitz potential, which always dominates it. Therefore, one can prove a PABI bound for convex/concave saddle-point problems which is upper bounded by that of the convex Lipschitz case.*

3.3 Consequences of Theorem 12

In this subsection we give explicit bounds for the different modulus of continuity presented in Section 2. In all of the results below, we used constant stepsize, $\eta > 0$, and constant variance of the noise added in each iteration, $\sigma^2 > 0$. The cases $\sigma^2 = 2\eta$ and $\sigma^2 = \eta^2$ are of particular interest for sampling and differential privacy, respectively. Each result is obtained by replacing the respective c_t and h_t in the bound of Theorem 12.

3.3.1 Convex and Lipschitz Potentials, Convex and (p, M)-Weakly Smooth Potentials

Recall that when the potential, f, is convex and L-Lipschitz or convex and (p, M)-weakly smooth, their moduli of continuity associated to the gradient maps are of the form $\varphi(\delta) = \sqrt{\delta^2 + h}$, where $h = (2\eta L)^2$ when the potential is Lipschitz, and $h = \left(2\eta^{\frac{1}{1-p}}\sqrt{\frac{1-p}{1+p}}(M/2)^{\frac{1}{1-p}}\right)^2$

when the potential is (p, M)-weakly smooth. The following Corollary establishes a bound for both cases.

Corollary 15 *Let $(X_t)_{t\in\mathbb{N}_0}$ and $(X'_t)_{t\in\mathbb{N}_0}$ be two projected noisy iterations with modulus of continuity $\varphi(\delta) = \sqrt{\delta^2 + h}$, where $h > 0$, which only differ in their initialization and whose domain, $\mathcal{X}$, has diameter $D > 0$, then*

$$R_\alpha(X_T||X'_T) \le \frac{\alpha}{2\sigma^2}\left(\frac{D^2}{T} + h\sum_{t=1}^{T}\frac{1}{t}\right) \le \frac{\alpha}{2\sigma^2}\left(\frac{D^2}{T} + h\ln(T\cdot e)\right).$$

3.3.2 Strongly Dissipative Potential

Finally, we state the Rényi divergence bound when the potential, f, is (λ, κ)-strongly dissipative and β-smooth.

Corollary 16 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ with diameter $D > 0$, and $(X_t)_{t\in\mathbb{N}_0}$, $(X'_t)_{t\in\mathbb{N}_0}$ be two projected noisy iterations with modulus of continuity $\varphi(\delta) = \sqrt{(1 - 2\eta\kappa + \eta^2\beta^2)\delta^2 + 2\eta\lambda}$, which only differ in their initialization. Then*

$$\begin{aligned} R_\alpha(X_T||X'_T) &\le \frac{\alpha}{2\sigma^2}\left(\frac{D^2c^T(1-c)}{(1-c^T)} + h\sum_{t=0}^{T-1}\frac{c^t}{\sum_{j=0}^{t}c^j}\right) \\ &\le \frac{\alpha}{2\sigma^2}\left(\frac{D^2c^T(1-c)}{(1-c^T)} + h\ln\left(\left(\frac{1-c^T}{1-c}\right)e\right)\right), \end{aligned}$$

where $c = 1 - 2\eta\kappa + \eta^2\beta^2$ and $h = 2\eta\lambda$.

The first inequality is a direct application of Theorem 12, while the second is obtained through an integral estimation of the sum. Indeed, by observing that if a function, F, is nonnegative and decreasing, then $\sum_{t=1}^{T-1} c^t F\left(\sum_{t=0}^{t} c^t\right) \le \int_1^{\sum_{t=0}^{T-1} c^t} F(x)dx$, we get

$$\begin{aligned} 1 + \ln\left(\frac{1-c^{T+1}}{1-c^2}\right) \le \sum_{t=0}^{T-1}\frac{c^t}{\sum_{j=0}^{t}c^j} &= 1 + \sum_{t=1}^{T-1}\frac{c^t}{1+\sum_{j=1}^{t}c^j} \\ &\le 1 + \int_1^{1+\sum_{t=1}^{T-1}c^t}\frac{1}{x}\,dx = 1 + \ln\left(\frac{1-c^T}{1-c}\right), \end{aligned}$$

where we also include a lower bound to show that our integral estimate is nearly tight.

Remark 17 *We would like to point out that our general result allows to recover some settings where the PABI approach has been applied. More precisely, by setting $h \equiv 0$, c and σ constant in Theorem 12, we can recover the contractive, nonexpansive and expansive results obtained in Feldman et al. (2018) and Remark A.4 of Altschuler and Talwar (2023). This corresponds, respectively, to the cases where the potential f is κ-strongly convex and β-smooth, convex and β-smooth and nonconvex and β-smooth.*

4. Mixing Time Bounds for the Projected Langevin Algorithm

In this section we study the mixing time in total variation distance for the Projected Langevin Algorithm (PLA). Since the result only relies on the bound obtained in Section 3, it also holds for (possibly nondifferentiable) potentials f that are convex and L-Lipschitz, by replacing $p = 0$ and $M = 2L$. The problem at hand is not only of academic interest, finding applications in Bayesian inference. An interesting example involves sampling from potentials of the form $\exp(-\|Ax-b\|_2^2 - \|Bx\|_{p+1}^{p+1})$, where $0 \le p \le 1$. This model arises from considering a prior distribution on hypotheses given by a linear transformation B of an ℓ^{p+1}-ball, and its posterior resulting from linear observations, determined by input data matrix A and corresponding output vector b (see Chatterji et al. 2020 for further discussions).

Recall that PLA is (PLA), with $\sigma = \sqrt{2\eta}$. That is

$$X_{t+1} = \Pi_{\mathcal{X}}\left[X_t - \eta\nabla f(X_t) + \sqrt{2\eta}\xi_t\right], \tag{PLA}$$

where $(\xi_t)_{t\in\mathbb{N}_0} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{d\times d})$.

Notice that (PLA) is a homogeneous Markov chain (henceforth HMC), whose only involved map (apart from the noise addition) is the gradient mapping $\Phi = I - \eta\nabla f$.

4.1 Convex and Weakly Smooth Case

Based on our settings of interest, we assume Φ has a modulus of continuity $\varphi(\delta) = \sqrt{\delta^2 + h}$ and that $\sigma^2 = 2\eta$. From Corollary 15 and taking $\alpha = 1$, we obtain the KL bound

$$\mathrm{KL}(X_T||X'_T) \le \underbrace{\frac{D^2}{4\eta T}}_{\textcircled{I}} + \underbrace{\frac{h\ln(T\cdot e)}{4\eta}}_{\textcircled{II}}. \tag{4}$$

Observe that $\textcircled{I}$ in (4) is exactly the bound obtained for the nonexpansive case in Altschuler and Talwar (2023, Proposition 2.10). Consequently, the price required for utilizing moduli of continuity of the form $\varphi(\delta) = \sqrt{\delta^2 + h}$ is encapsulated by the term added in $\textcircled{II}$.

Replacing $h = \left(2\eta^{\frac{1}{1-p}}\sqrt{\frac{1-p}{1+p}}(M/2)^{\frac{1}{1-p}}\right)^2$ in equation (4), we get

$$\mathrm{KL}(X_T||X'_T) \le \frac{D^2}{4\eta T} + \ln(T\cdot e)\left(\eta^{\frac{1+p}{1-p}}\left(\frac{1-p}{1+p}\right)(M/2)^{\frac{2}{1-p}}\right).$$

Since a given potential can be (p, M)-weakly smooth with multiple parameters, the bound also is satisfied with the infimum; that is

$$\mathrm{KL}(X_T||X'_T) \le \frac{D^2}{4\eta T} + \ln(T\cdot e)\cdot\inf\left\{\left(\eta^{\frac{1+p}{1-p}}\left(\tfrac{1-p}{1+p}\right)(M/2)^{\frac{2}{1-p}}\right) : f \text{ is } (p, M)\text{-weakly smooth}\right\}.$$

Notice that $M(p) = \inf\{M > 0 : f \text{ is } (p, M)\text{-weakly smooth}\}$ is a log-convex function with respect to p; therefore, the infimum above may have a nontrivial optimal choice of p. In this regard, it is interesting that we can automatically obtain this adaptivity in terms of p.

We use the obtained KL bound (4) to establish a new mixing time for the projected Langevin algorithm in the weakly smooth and Lipschitz convex cases. We remark that, aside from a slightly more restrictive range of stepsize parameters, the mixing time bound is entirely analogous to that of the smooth convex case (Altschuler and Talwar, 2023). Since it is of practical interest to set the stepsize sufficiently small, the stepsize restriction is not problematic. Note that the result only uses the modulus of continuity of the potential, so it also holds for the L-Lipschitz setting replacing $p = 0$ and $M = 2L$.

We will start providing a total variation mixing time (see Definition 36) to constant error 1/2.

Lemma 18 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex, compact set with diameter $D > 0$ and suppose that $f : \mathcal{X} \to \mathbb{R}$ is a convex and (p, M)-weakly smooth function, with $p \in [0, 1]$ and $M > 0$. Let $(X_t)_{t\in\mathbb{N}_0}$ and $(X'_t)_{t\in\mathbb{N}_0}$ be two HMC generated by* (PLA) *that only differ in their initialization. Let*

$$\Theta = \left(\tfrac{M}{2}\right)^{\left(\frac{2}{1+p}\right)} \left[\left(\tfrac{1-p}{1+p}\right) \max\left\{16 \ln\left(D \left(\tfrac{M}{2}\right)^{\frac{1}{1+p}} e\right), 27\right\}\right]^{\left(\frac{1-p}{1+p}\right)}. \tag{5}$$

If $\frac{1}{\eta} \geq \Theta$ and $T = \left\lceil \frac{D^2}{\eta} \right\rceil$, then $\left\|\mathbb{P}_{X_T} - \mathbb{P}_{X'_T}\right\|_{TV} \leq \frac{1}{2}$.

Proof To simplify notation, let us call $\tilde{\eta} := \eta^{\frac{1+p}{1-p}}$ and $\tilde{M} := \sqrt{\frac{1-p}{1+p}}(M/2)^{\frac{1}{1-p}}$.

By equation (4) and Proposition 26 it is enough to find T and η such that $\frac{D^2}{4\eta T} + \tilde{\eta}\tilde{M}^2 \ln(T \cdot e) \leq 1/2$. One way to get this bound is to derive individually:

$$\frac{D^2}{4\eta T} \leq \frac{1}{4} \qquad \text{and} \qquad \tilde{\eta}\tilde{M}^2 \ln(T \cdot e) \leq \frac{1}{4}.$$

For the first inequality it is enough to take $T = \left\lceil \frac{D^2}{\eta} \right\rceil$. For the second inequality, we plug $\frac{2D^2}{\eta}$ in the place of T, which is a sufficient condition for the inequality when $\eta \leq D^2$. Then, after some simple algebraic manipulation, we get that

$$2 \ln\left(\sqrt{2e} D \tilde{M}^{\frac{1-p}{1+p}}\right) \leq \frac{1}{4\tilde{\eta}\tilde{M}^2} - \left(\frac{1-p}{1+p}\right) \ln\left(\frac{1}{\tilde{\eta}\tilde{M}^2}\right). \tag{6}$$

The right side of (6) can be written as a function of $1/(\tilde{\eta}\tilde{M}^2)$, namely $F_p(x) = \frac{x}{4} - \left(\frac{1-p}{1+p}\right) \ln(x)$. Since $p \in [0, 1]$, F_p can be lower bounded by $F(x) = \frac{x}{4} - \ln(x)$ when $\eta^{-1} \geq \left(\frac{1-p}{1+p}\right)^{\frac{1-p}{1+p}} \left(\frac{M}{2}\right)^{\frac{2}{1+p}}$. It can be shown that $F(x) \geq \frac{x}{8}$ when $x \geq 27$. Thus, a sufficient condition for (6) to hold is to ask for $2 \ln\left(\sqrt{2e} D \tilde{M}^{\frac{1-p}{1+p}}\right) \leq \frac{1}{8\tilde{\eta}\tilde{M}^2}$, subject to $\frac{1}{\tilde{\eta}\tilde{M}^2} \geq 27$. Therefore, a sufficient condition for (6) to hold is:

$$\frac{1}{\tilde{\eta}} \geq \tilde{M}^2 \max\left\{16 \ln\left(\sqrt{2e} D \tilde{M}^{\frac{1-p}{1+p}}\right), 27\right\},$$

which is equivalent to

$$\frac{1}{\eta} \geq \left(\frac{M}{2}\right)^{\left(\frac{2}{1+p}\right)} \left[\left(\frac{1-p}{1+p}\right) \max\left\{16 \ln\left(\sqrt{2e}D\left(\frac{1-p}{1+p}\right)^{\frac{1-p}{2(1+p)}}\left(\frac{M}{2}\right)^{\frac{1}{1+p}}\right), 27\right\}\right]^{\left(\frac{1-p}{1+p}\right)}.$$

Finally, noting that

$$\ln\left(\sqrt{2e}\left(\frac{1-p}{1+p}\right)^{\frac{1-p}{2(1+p)}}\right) \leq 1 \qquad (\forall p \in [0,1]),$$

we obtain the result. ■

We highlight the two extreme cases: $p = 0$ and $p = 1$. When $p = 0$ we are in the Lipschitz case and the restriction over η boils down to $1/\eta \geq (M/2)^2 \max\{16 \ln(D(M/2)e), 27\}$. On the other hand, when $p = 1$ (which can be obtained through a limit) we are in the smooth case and we recover the restriction $1/\eta \geq M/2$ which makes the gradient mapping nonexpansive.

A direct consequence of Lemma 18 is that for weakly smooth potentials:

$$T_{mix,TV}(1/2) \leq \left\lceil \frac{D^2}{\eta} \right\rceil, \tag{7}$$

when its restriction over η is satisfied. This can be proved by letting X_0' follow the stationary distribution of (PLA) (for a proof of existence of stationary distributions of the HMC defined by PLA in the nondifferentiable case see Appendix C). It has been established in Altschuler and Talwar (2023, Theorem 3.2) that (7) is tight up to constants by a constant potential, and thus tightness applies to the above case as well.

Using Lemma 18 and a well-known boosting argument (Proposition 37), we can convert a constant error in total variation into an arbitrary one at a polylogarithmic cost in the accuracy.

Theorem 19 (Mixing for weakly smooth functions) *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex, compact set with diameter $D > 0$ and suppose that $f : \mathcal{X} \to \mathbb{R}$ is a convex and (p, M)-weakly smooth function. If $1/\eta \geq \Theta$, where Θ is as in* (5)*, then, for all $\varepsilon > 0$,*

$$T_{mix,TV}(\varepsilon) \leq \left\lceil \frac{D^2}{\eta} \right\rceil \cdot \lceil \log_2(1/\varepsilon) \rceil.$$

Similarly to (Altschuler and Talwar, 2022), when $f = \sum_{i=1}^n f_i$ and each f_i is (p, M)-weakly smooth, we can replace the use of gradients in (PLA) by stochastic (formed through minibatches) gradients. This change yields the same result as in Theorem 19.

4.2 Smooth and Strongly Dissipative Case

We can also establish a mixing time bound for the β-smooth and (λ, κ)-strongly dissipative case. The analysis is analogous to the one presented before, only that instead of using the

Pinsker inequality to derive a 1/2 bound to which the boosting argument is applied, we use the Bretagnolle-Huber inequality (Proposition 27), which is valid in a broader regime. The boosting argument is still applicable, but at the cost of a worst dependence on the parameters.

Theorem 20 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex, compact set with diameter $D > 0$ and suppose that $f : \mathcal{X} \to \mathbb{R}$ is a (λ, κ)-strongly dissipative and β-smooth function. Let $c = 1 - 2\eta\kappa + \eta^2\beta^2$. Then, for all $\varepsilon > 0$,*

$$T_{mix,TV}(\varepsilon) \leq \left\lceil \log_{1/c}\left(1 + \frac{D^2(1-c)}{4\eta}\right)\right\rceil \cdot \left\lceil 2e\ln(2)\left(\frac{e}{1-c}\right)^{\lambda/2} \log_2(1/\varepsilon)\right\rceil .$$

Proof By Corollary 16, with $\alpha = 1$ and $\sigma^2 = 2\eta$,

$$\mathrm{KL}(X_T || X_T') \leq \frac{D^2}{4\eta} \cdot \frac{c^T(1-c)}{1-c^T} + \frac{\lambda}{2}\ln\left(\left(\frac{1-c^T}{1-c}\right)e\right).$$

Taking $T^* = \left\lceil \log_{1/c}\left(1 + \frac{D^2(1-c)}{4\eta}\right)\right\rceil$, we get

$$\mathrm{KL}(X_{T^*} || X_{T^*}') \leq 1 + \frac{\lambda}{2}\ln\left(\frac{e}{1-c}\right). \tag{8}$$

Using Proposition 27, we convert this KL inequality into a total variation one,

$$\left\|\mathbb{P}_{X_{T^*}} - \mathbb{P}_{X_{T^*}'}\right\|_{\mathrm{TV}} \leq \sqrt{1 - \exp\left(-\mathrm{KL}(X_{T^*} || X_{T^*}')\right)} =: \gamma$$

Let $\varepsilon > 0$. In order to get $\gamma^R \leq \varepsilon$, we need $R \geq \left\lceil \log_{1/\gamma}(1/\varepsilon)\right\rceil$.

Notice that

$$\log_{1/\gamma}(1/\varepsilon) = \frac{\ln(2)\log_2(1/\varepsilon)}{\ln(1/\gamma)} \tag{9}$$

and that by the convexity of $-\ln(1-x)$

$$\begin{aligned} \ln(1/\gamma) &= -\frac{1}{2}\ln\left(1 - e^{-\mathrm{KL}(X_{T^*} || X_{T^*}')}\right) \\ &\geq \frac{1}{2}e^{-\mathrm{KL}(X_{T^*} || X_{T^*}')}. \end{aligned} \tag{10}$$

Hence, by a boosting argument of the total variation (Levin and Peres, 2017, Lemma 4.11), we get $T_{mix,TV}(\varepsilon) \leq T^* \cdot R$. Finally, using the definition of T^*, (8), (9) and (10), we conclude

$$T_{mix,TV}(\varepsilon) \leq \left\lceil \log_{1/c}\left(1 + \frac{D^2(1-c)}{4\eta}\right)\right\rceil \cdot \left\lceil 2e\ln(2)\left(\frac{e}{1-c}\right)^{\lambda/2} \log_2(1/\varepsilon)\right\rceil .$$

■

5. Privacy Analysis of Noisy SGD

In this section, we combine the analysis from Altschuler and Talwar (2022) with the PABI bounds from Section 3 to derive new privacy bounds for noisy SGD's last iteration.

We remind the reader of the definition of differential privacy. We denote a data set by an n-tuple $S = (s_1, \dots, s_n) \in \mathcal{Z}^n$, where $\mathcal{Z}$ is the data space. First, we say that two data sets S, S' are neighbours, denoted by $S \simeq S'$, if they only differ in one of their entries.

Definition 21 (Differential Privacy) *A randomized algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{X}$ is (ε, δ)-differentially private (DP) if for every pair of data sets $S \simeq S'$, and any event $O \subseteq \mathcal{X}$,*

$$\mathbb{P}[\mathcal{A}(S) \in O] \leq \exp(\varepsilon)\mathbb{P}[\mathcal{A}(S') \in O] + \delta.$$

As mentioned above, we will analyze the privacy curve (i.e. a bound of the Rényi divergence of two outputs of the same algorithm when executed on two neighboring data sets) of noisy SGD, a data dependent algorithm that, given a data set $S = (s_1, \dots, s_n)$ and an initializaton $X_0 \in \mathcal{X}$:

1. To update its state at time $t \in \{1, \dots, T\}$:

   (i) Using Poisson sampling, randomly chooses a minibatch $B_t \subseteq \{1, \dots, n\}$ of expected size b (i.e. each index i has probability b/n of being in B_t).

   (ii) Given $\xi_t \sim \mathcal{N}(0, \eta^2\sigma^2 I_{d\times d})$, computes $X_{t+1} = \Pi_{\mathcal{X}}\left[X_t - \frac{\eta}{b}\sum_{i\in B_t} \nabla f(X_t, s_i) + \xi_t\right]$.
2. Return X_T.

We study the case where the loss functions $f(\cdot, s)$ are convex, L-Lipschitz and (p, M)-weakly smooth for every $s \in \mathcal{Z}$, where $p \in [0, 1]$, $M > 0$ and $\mathcal{Z}$ is a data space.

Theorem 22 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex and compact set with diameter $D > 0$. For any number of iterations $T > \overline{T} = \left\lceil \frac{Dn}{4\eta L} \right\rceil$, datasize $n \in \mathbb{N}$, expected batch size $b \leq n$, stepsize $\eta > 0$, initialization $x_0 \in \mathcal{X}$ and noise parameter $\sigma > 8\sqrt{2}L/b$, noisy SGD applied to convex, L-Lipschitz and (p, M)-weakly smooth losses satisfies (α, ε)-RDP for $1 < \alpha \leq \alpha^*\left(\frac{b}{n}, \frac{b\sigma}{2\sqrt{2}L}\right)$, where α^* is defined in Lemma 41, and:*

$$\varepsilon \leq \frac{\alpha}{\sigma^2}\left(\frac{16L^2\overline{T}}{n^2} + \frac{D^2}{\eta^2\overline{T}} + 4\eta^{\frac{2p}{1-p}}\left(\frac{1-p}{1+p}\right)\left(\frac{M}{2}\right)^{\frac{2}{1-p}} \ln\left(\overline{T} \cdot e\right)\right)$$

Proof The proof is an analogue of Altschuler and Talwar (2022, Theorem 3.1). Let $S, S' \in \mathcal{Z}^n$ be two neighbor data sets that differs in the data point corresponding to i^*, that is, $s_i = s'_i$ for all $i \neq i^*$. Run noisy SGD on both data sets, S and S', for T iterations and call the respective trajectories:

$$X_{t+1} = \Pi_{\mathcal{X}}\left[X_t - \frac{\eta}{b}\sum_{i\in B_t} \nabla f(X_t, s_i) + \xi_t\right]$$

$$X'_{t+1} = \Pi_{\mathcal{X}}\left[X'_t - \frac{\eta}{b}\sum_{i\in B_t} \nabla f(X'_t, s'_i) + \xi_t\right].$$

These trajectories start from the same point, in which the noise injection $(\xi_t)_{t=0}^{T-1}$ and minibatch $(B_t)_{t=0}^{T-1}$ are coupled. One can rewrite this expressions as

$$\begin{aligned} X_{t+1} &= \Pi_{\mathcal{X}}\left[X_t - \frac{\eta}{b}\sum_{i\in B_t}\nabla f(X_t, s_i) + Y_t + Z_t\right] \\ X'_{t+1} &= \Pi_{\mathcal{X}}\left[X'_t - \frac{\eta}{b}\sum_{i\in B_t}\nabla f(X'_t, s_i) + Y_t + Z'_t\right], \end{aligned}$$

where $Y_t \sim \mathcal{N}(0, (\eta^2\sigma^2/2)I_{d\times d})$, $Z_t \sim \mathcal{N}(0, (\eta^2\sigma^2/2)I_{d\times d})$ and

$$Z'_t \sim \mathcal{N}\left(\frac{\eta}{b}\left[\nabla f(X'_t, s_{i^*}) - \nabla f(X'_t, {s'}_{i^*})\right]\cdot \mathbb{1}_{\{i^*\in B_t\}},\, (\eta^2\sigma^2/2)I_{d\times d}\right).$$

It is important to remark that the gradients of the convex losses that we are using in both trajectories come from the data set S, not S'. Observe also that the bias term is realized with probability

$$\mathbb{P}\left(i^* \in B_t\right) = \frac{b}{n}. \tag{11}$$

Conditional on the event that $Z_t = Z'_t$ (call z_t its realization):

$$\begin{aligned} X_{t+1} &= \Pi_{\mathcal{X}}\left[\Phi_t(X_t) + Y_t\right] \\ X'_{t+1} &= \Pi_{\mathcal{X}}\left[\Phi_t(X'_t) + Y_t\right], \end{aligned}$$

where

$$\Phi_t(x) := x - \frac{\eta}{b}\sum_{i\in B_t}\nabla f_i(x) + z_t. \tag{12}$$

Note that the modulus of continuity of (12) is upper bounded by the modulus of continuity of the noiseless gradient mapping. This leads to a modulus of continuity $\varphi(\delta) = \sqrt{\delta^2 + h}$ for $h = \left(2\eta^{\frac{1}{1-p}}\sqrt{\frac{1-p}{1+p}}\left(\frac{M}{2}\right)^{\frac{1}{1-p}}\right)^2$.

Conditional on the event $Z_t = Z'_t$ for all $t \geq \tau$, the processes $\{X_t\}_{t\geq\tau}$ and $\{X'_t\}_{t\geq\tau}$ are projected noisy iterations with modulus of continuity φ, where $\tau \in \{0, \ldots, T-1\}$ is a parameter chosen *a posteriori*.

The bound of $R_\alpha(\mathbb{P}_{X_T}\,||\,\mathbb{P}_{X'_T})$ is obtained through Privacy Amplification by Sampling and Privacy Amplification by Iteration (with modulus of continuity):

$$\begin{aligned} R_\alpha\left(\mathbb{P}_{X_T}\,||\,\mathbb{P}_{X'_T}\right) &\leq R_\alpha\left(\mathbb{P}_{X_T, Z_{\tau:T}}\,||\,\mathbb{P}_{X'_T, Z'_{\tau:T}}\right) \\ &\leq \underbrace{R_\alpha\left(\mathbb{P}_{Z_{\tau:T-1}}\,||\,\mathbb{P}_{Z'_{\tau:T-1}}\right)}_{①} + \underbrace{\sup_z R_\alpha\left(\mathbb{P}_{X_T|Z_{\tau:T-1}=z}\,||\,\mathbb{P}_{X'_T|Z'_{\tau:T-1}=z}\right)}_{②}, \end{aligned} \tag{13}$$

where the first line follows from the data-processing inequality (Proposition 28) and the second from strong composition (Proposition 29).

Bounding ① through Privacy Amplification by Sampling:

$$
\begin{aligned}
\textcircled{1} &\le \sum_{t=\tau}^{T-1} \sup_{z_{\tau:t-1}} R_\alpha \left(\mathbb{P}_{Z_t | Z_{\tau:t-1} = z_{\tau:t-1}} \,||\, \mathbb{P}_{Z'_t | Z'_{\tau:t-1} = z_{\tau:t-1}} \right) \\
&= \sum_{t=\tau}^{T-1} R_\alpha \left(\mathcal{N}\left(0, \frac{\eta^2\sigma^2}{2} I_{d\times d}\right) \Big|\Big| \left(1 - \frac{b}{n}\right) \mathcal{N}\left(0, \frac{\eta^2\sigma^2}{2} I_{d\times d}\right) + \frac{b}{n} \mathcal{N}\left(m_t, \frac{\eta^2\sigma^2}{2} I_{d\times d}\right)\right) \\
&\le (T-\tau) S_\alpha \left(\frac{b}{n}, \frac{b\sigma}{2\sqrt{2}L} \right), \qquad (14)
\end{aligned}
$$

where the first line follows from strong composition (Proposition 29). The second line follows by the independence of the $(Z_t)_t$: $Z_t \sim \mathcal{N}\left(0, \frac{\eta^2\sigma^2}{2} I_{d\times d}\right)$ conditioned on $Z_{\tau:t-1} = z_{\tau:t-1}$ for any $z_{\tau:t-1}$; also, by (11) and the independence of the (Z'_t), the law of Z'_t is the mixture of $\mathcal{N}\left(0, \frac{\eta^2\sigma^2}{2} I_{d\times d}\right)$ and $\mathcal{N}\left(m_t, \frac{\eta^2\sigma^2}{2} I_{d\times d}\right)$, where $m_t := \frac{\eta}{b}\left[\nabla f(X'_t, s_{i^*}) - \nabla f(X_t, s_{i^*})\right]$. The last line follows from the fact that $\|m_t\| \le 2\eta L/b$ and the bound in Lemma 40.

Bounding ② through Corollary 15: As we already mentioned, conditional on the event that $Z_t = Z'_t$ for all $t \ge \tau$, the sequences $\{X_t\}_{t\ge\tau}$ and $\{X'_t\}_{t\ge\tau}$ are projected noisy iterations with moduli of continuity $\varphi(\delta) = \sqrt{\delta^2 + h}$. Then, by Corollary 15, using $h = \left(2\eta^{\frac{1}{1-p}} \sqrt{\frac{1-p}{1+p}} \left(\frac{M}{2}\right)^{\frac{1}{1-p}}\right)^2$ for all $t \ge \tau$ and $\sigma_j^2 = \frac{\eta^2\sigma^2}{2}$ for all $j \ge \tau$:

$$
\begin{aligned}
\textcircled{2} &\le \frac{\alpha D^2}{\eta^2\sigma^2(T-\tau)} + \frac{\alpha \left(2\eta^{\frac{1}{1-p}} \sqrt{\frac{1-p}{1+p}} \left(\frac{M}{2}\right)^{\frac{1}{1-p}}\right)^2}{\eta^2\sigma^2} \ln\left((T-\tau)\cdot e\right) \\
&\le \frac{\alpha D^2}{\eta^2\sigma^2(T-\tau)} + \frac{4\alpha\eta^{\frac{2p}{1-p}}}{\sigma^2} \left(\frac{1-p}{1+p}\right) \left(\frac{M}{2}\right)^{\frac{2}{1-p}} \ln\left((T-\tau)\cdot e\right). \qquad (15)
\end{aligned}
$$

Plugging (14) and (15) into (13), we obtain that noisy SGD is (α, ε)-RDP with:

$$
\begin{aligned}
\varepsilon \le \min_{\tau\in\{0,\dots,T-1\}} \Bigg\{ & (T-\tau) S_\alpha \left(\frac{b}{n}, \frac{b\sigma}{2\sqrt{2}L}\right) + \frac{\alpha D^2}{\eta^2\sigma^2(T-\tau)} + \\
& + \frac{4\alpha\eta^{\frac{2p}{1-p}}}{\sigma^2} \left(\frac{1-p}{1+p}\right) \left(\frac{M}{2}\right)^{\frac{2}{1-p}} \ln\left((T-\tau)\cdot e\right) \Bigg\}
\end{aligned}
$$

By Lemma 41, for all $1 < \alpha \le \alpha^* \left(\frac{b}{n}, \frac{b\sigma}{2\sqrt{2}L}\right)$ and $\sigma \ge 8\sqrt{2}L/b$:

$$
S_\alpha \left(\frac{b}{n}, \frac{b\sigma}{2\sqrt{2}L}\right) \le \frac{16\alpha L^2}{n^2\sigma^2}.
$$

Then

$$
\varepsilon \le \frac{\alpha}{\sigma^2} \min_{\tau\in\{0,\dots,T-1\}} \left\{ (T-\tau)\frac{16L^2}{n^2} + \frac{D^2}{\eta^2(T-\tau)} + 4\eta^{\frac{2p}{1-p}} \left(\frac{1-p}{1+p}\right) (M/2)^{\frac{2}{1-p}} \ln\left((T-\tau)\cdot e\right) \right\}.
$$

One can easily optimize the first two terms of the above expression by naming $R = T - \tau$ and differentiating with respect to R. Taking the ceiling of the optimal value for R, one obtains:

$$T - \tau = \overline{T} = \left\lceil \frac{Dn}{4\eta L} \right\rceil,$$

whenever $T \geq \overline{T}$. Therefore,

$$\varepsilon \leq \frac{\alpha}{\sigma^2} \left(\frac{16L^2\overline{T}}{n^2} + \frac{D^2}{\eta^2\overline{T}} + 4\eta^{\frac{2p}{1-p}} \left(\frac{1-p}{1+p}\right) \left(\frac{M}{2}\right)^{\frac{2}{1-p}} \ln\left(\overline{T} \cdot e\right) \right).$$

■

Remark 23 *Note that by only using Lemmas 40 and 41 in conjunction with sequential composition for RDP (Mironov, 2017, Proposition 1) in the privacy analysis of noisy SGD, one obtains that* $\varepsilon \leq \frac{16\alpha L^2}{n^2\sigma^2}T$. *Therefore, from Theorem 22, we conclude that the last iterate of noisy SGD is* (α, ε)*-RDP with*

$$\varepsilon \leq \frac{16\alpha L^2}{n^2\sigma^2} \min\left\{T, 2\overline{T} + \underbrace{\left(\frac{2\overline{T}}{D} \cdot \left(\frac{\eta M}{2}\right)^{\frac{1}{1-p}}\right)^2 \left(\frac{1-p}{1+p}\right) \ln\left(\overline{T} \cdot e\right)}_{V(D,M,\overline{T},\eta,p)}\right\}$$

In comparison to Altschuler and Talwar (2022, Theorem 3.1), which holds for smooth functions, working with Hölder continuous gradients adds an extra term to the privacy bound, which we denote above by $V(D, M, \overline{T}, \eta, p)$. *See Figure 1 for some exemplary plots of the privacy curve. Note that in this figure we omit the graph of the case* $p = 0.8$, *as it is indistinguishable from that of* $p = 1$. *Moreover, it appears that for any* $p \geq 0.7$ *there are no significant differences in the privacy curve bounds with that of the smooth case. On the other hand, in the Lipschitz case* ($p = 0$), *it is never possible to obtain a bound that vanishes with* $n \to \infty$ *since* $V(D, M, \overline{T}, \eta, 0)$ *grows as* $\tilde{O}(n^2)$.

Acknowledgments

We thank Jason Altschuler for valuable feedback on a first version of this manuscript.

Research partially supported by INRIA Associate Teams project, ANID FONDECYT 1210362 grant, ANID FONDECYT 1241805 grant, ANID Anillo ACT210005 grant, and National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

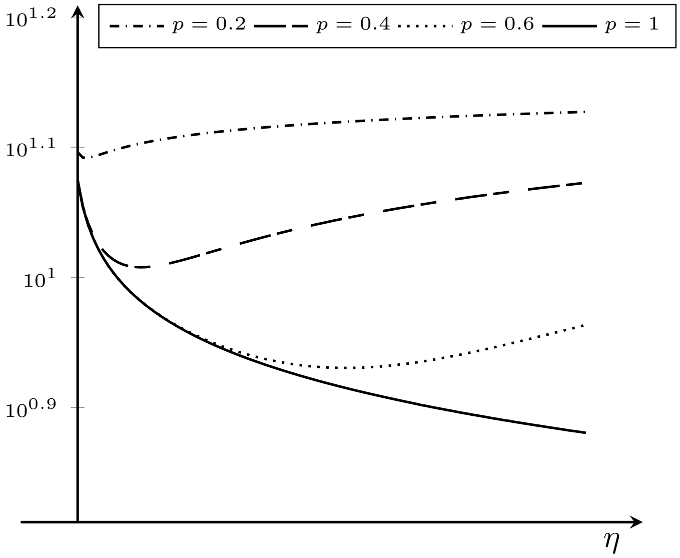

Figure 1: Different values for the bound $2\overline{T} + V(D, M, \overline{T}, \eta, p)$ in logarithmic scale, for $\eta \in [n^{-1}, n^{-1/5}]$, $n = 1000$, $L = 1$, $M = 2$, $D = 1$.

Appendix A. Basic Definitions

A.1 Information Theory and Probabilistic Divergences

Definition 24 (Kullback-Leibler divergence) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. We define the Kullback-Leibler divergence (abbreviated as KL divergence) as:*

$$\mathrm{KL}(\mu||\nu) = \begin{cases} \int_{\mathbb{R}^d} \frac{d\mu}{d\nu}(x) \ln\left(\frac{d\mu}{d\nu}(x)\right) \nu(dx) \text{ if } \mu \ll \nu \\ +\infty \text{ otherwise} \end{cases}$$

It is a well-known fact (Van Erven and Harremos, 2014, Theorem 5) that one can extend, by taking limits, the Rényi divergence to the case $\alpha = 1$ when $R_\beta < \infty$ for some $\beta > 1$. This extreme case results in the KL divergence, i.e. $R_1(\mu||\nu) = \mathrm{KL}(\mu||\nu)$.

Definition 25 (Total Variation distance) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. We define the total variation distance between μ and ν as:*

$$\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mu(A) - \nu(A)|.$$

Note that if a sequence of probability measures $(\nu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ converges in total variation to a measure $\nu \in \mathcal{P}(\mathcal{X})$, then

$$\lim_{n\to\infty} \int_{\mathcal{X}} g(x)\nu_n(dx) = \int_{\mathcal{X}} g(x)\nu(dx)$$

for all measurable and bounded function g with support on $\mathcal{X}$. This is because a measurable and bounded function can be uniformly approximated by a simple functions.

A useful inequality that compares total variation with KL divergence is Pinsker's inequality (Tsybakov, 2009, Lemma 2.5):

Proposition 26 (Pinsker's inequality) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. Then*

$$\|\mu - \nu\|_{TV} \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mu\|\nu)}.$$

Another useful inequality for comparing KL divergence and total variation is Bretagnolle-Huber's inequality (Canonne, 2023, Lemma 3). Although this inequality is worse than Pinsker's when the KL divergence moves between 0 and 2, it has the advantage of being nonvacous when it exceeds this limit.

Proposition 27 (Bretagnolle-Huber inequality) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. Then*

$$\|\mu - \nu\|_{TV} \leq \sqrt{1 - \exp\left(-\mathrm{KL}(\mu||\nu)\right)}.$$

The following is the well-known data-processing inequality (Van Erven and Harremos, 2014, Theorem 9) for Rényi divergence:

Proposition 28 *Let $P : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$ be a measurable map and let $J : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ be the transition operator associated to P (see Definition 32). Then, for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $R_\alpha\left(J\mu||J\nu\right) \leq R_\alpha(\mu||\nu)$.*

Let $X_1, \ldots, X_n$ be a sequence of (possibly random) vectors. We abbreviate $X_1, \ldots, X_k$ by $X_{1:k}$. The following proposition corresponds to Altschuler and Talwar (2022, Lemma 2.9).

Proposition 29 (Strong composition) *Set $\alpha \geq 1$ and let $X_{1:k}$ and $Y_{1:k}$ be two sequences of random variables. Then*

$$R_\alpha(\mathbb{P}_{X_{1:k}} \,\|\, \mathbb{P}_{Y_{1:k}}) \leq \sum\nolimits_{i=1}^{k} \sup_{x_{1:i-1}} R_\alpha\left(\mathbb{P}_{X_i|X_{1:i-1}=x_{1:i-1}} \,\|\, \mathbb{P}_{Y_i|Y_{1:i-1}=x_{1:i-1}}\right).$$

A.2 Mixing Times

We start by introducing the terminology and basic results regarding homogeneous Markov chains (HMC). For more information, we refer the reader to Hairer (2006).

Definition 30 (Homogeneous Markov Chain, Transition Probabilities) *We say that a Markov Chain taking values on a set $\mathcal{X} \subseteq \mathbb{R}^d$, $(X_t)_{t\in\mathbb{N}_0}$, is (time) homogeneous if there exists a measurable map $P : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ such that:*

$$\mathbb{P}\left(X_t \in A | X_{t-1} = x\right) = P(x, A)$$

for every $A \in \mathcal{B}(\mathcal{X})$, almost every $x \in \mathcal{X}$, and every $t \geq 1$. The map P from above is called the transition probabilities of the chain.

We will usually call transition probabilities to all measurable maps $P : \mathcal{X} \to \mathcal{P}(\mathcal{X})$, even if no HMC is specified. This is justified since for every such map there exists an HMC that has it as transition probabilities (see, for example, Hairer 2006, Proposition 2.38).

Proposition 31 (Hairer 2006, Theorem 2.29) *Let $(X_t)_{t\in\mathbb{N}_0}$ be an HMC taking values on $\mathcal{X}$ and with transition probabilities P. Then, for all $t \geq 1$,*

$$\mathbb{P}\left(X_t \in A | X_0 = x\right) = P^t(x, A),$$

where P^t is defined recursively by:

$$P^t(x, A) = \int_{\mathcal{X}} P(z, A)P^{t-1}(x, dz).$$

An easy consequence of the previous proposition is that

$$P^{t+s}(x, A) = \int_{\mathcal{X}} P^t(z, A)P^s(x, dz)$$

for all $t, s \geq 1$.

Definition 32 (Transition Operator) *Given transition probabilities $P : \mathcal{X} \to \mathcal{P}(\mathcal{X})$, we define the transition operator $J : \mathcal{P}(\mathcal{X}) \mapsto \mathcal{P}(\mathcal{X})$ by:*

$$(J\mu)(A) = \int_{\mathcal{X}} P(z, A)\mu(dz),$$

for every $A \in \mathcal{B}(\mathcal{X})$.

Remark 33 *If $(X_t)_{t\in\mathbb{N}_0}$ is an HMC with transition probabilities P and transition operator J such that $X_0 \sim \mu_0$, then the distribution of X_t, for $t \geq 1$, is the one that for all $A \in \mathcal{B}(\mathcal{X})$:*

$$J^t\mu_0(A) = \int_{\mathcal{X}} P^t(z, A)\mu_0(dz),$$

as one can check.

Definition 34 (Invariant measure) *Given a transition operator J, we say that the measure π is an invariant (or stationary) measure of J if*

$$J\pi = \pi.$$

When we we talk about the HMC (instead of its transition operator), we usually call π the **stationary distribution** of $(X_t)_{t\in\mathbb{N}_0}$, instead of the invariant measure (of its transition operator). This is justified by the following Proposition:

Proposition 35 *Let $(X_t)_{t\in\mathbb{N}_0}$ be an HMC and let P and J be its transition probabilities and its transition operator, respectively. If π is the invariant measure of J and $X_0 \sim \pi$, then:*

$$X_t \sim \pi \quad (\forall t \in \mathbb{N}).$$

Definition 36 (Mixing time) *Let $(X_t)_{t\in\mathbb{N}_0}$ be an HMC with transition probabilities P and stationary distribution π. We define the mixing time in total variation up to error $\varepsilon > 0$ of the chain as:*

$$T_{mix,TV}(\varepsilon) := \min\{t \in \mathbb{N} : d(t) \leq \varepsilon\},$$

where

$$d(t) := \sup_{x\in\mathcal{X}} \left\| P^t(x,\cdot) - \pi \right\|_{TV}.$$

Proposition 37 (Levin and Peres 2017, Section 4.5) *Let $(X_t)_{t\in\mathbb{N}_0}$ be an HMC supported on a compact set $\mathcal{X}$ and with stationary distribution π. Let*

$$\bar{d}(t) := \sup_{x,y\in\mathcal{X}} \left\| P^t(x,\cdot) - P^t(y,\cdot) \right\|_{TV}.$$

If T^* is such that:

$$\bar{d}(T^*) \leq \frac{1}{2},$$

then:

$$T_{mix,TV}(\varepsilon) \leq T^* \cdot \lceil \log_2(1/\varepsilon) \rceil .$$

We highlight the distance $\bar{d}$ also holds when the points $x, y \in \mathcal{X}$ are replaced by probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, i.e.

$$\bar{d}(t) = \sup_{\mu,\nu\in\mathcal{P}(\mathcal{X})} \left\| J^t\mu - J^t\nu \right\|_{\mathrm{TV}}.$$

Definition 38 (Dual operator) *We define the dual operator of J, denoted J_* as:*

$$(J_*f)(x) = \mathbb{E}\left[f(X_1)|X_0 = x\right] = \int_{\mathcal{X}} f(z)P(x,dz).$$

It should be noted that, for all bounded and measurable function g and all probability measure $\mu \in \mathcal{P}(\mathcal{X})$, the dual operator satisfies:

$$\int_{\mathcal{X}} (J_*g)(x)\mu(dx) = \int_{\mathcal{X}} g(x)(J\mu)(dx),$$

and that it sends bounded and measurable functions into bounded and measurable functions.

A.3 Differential Privacy

A.3.1 Privacy Amplification by Sampling

Definition 39 (Rényi Divergence of the Sampled Gaussian Mechanism) *Let $\alpha \geq 1$ be a Rényi parameter, $q \in (0,1)$ be a mixture parameter and $\sigma > 0$ be a noise level. Define*

$$S_\alpha(q,\sigma) := R_\alpha\left(\mathcal{N}(0,\sigma^2)||(1-q)\,\mathcal{N}(0,\sigma^2) + q\,\mathcal{N}(1,\sigma^2)\right).$$

Lemma 40 (Altschuler and Talwar 2022, Lemma 2.11) *Let $\alpha \geq 1$ be a Rényi parameter, $q \in (0,1)$ be a mixture parameter, $\sigma > 0$ be a noise level, $d \in \mathbb{N}$ be the dimension and $r > 0$ be a radius. Then:*

$$\sup_{\mu \in \mathcal{P}(B(0,r))} R_\alpha\left(\mathcal{N}(0,\sigma^2 I_{d\times d})||(1-q)\,\mathcal{N}(0,\sigma^2 I_{d\times d}) + q\left(\mathcal{N}(0,\sigma^2 I_{d\times d}) * \mu\right)\right) = S_\alpha(q,\sigma/r),$$

where $B(0,r)$ denotes the Euclidean d-dimensional closed ball centered at the origin and with radius r.

Lemma 41 (Mironov et al. 2019, Theorem 11) *Let $\alpha \geq 1$ be a Rényi parameter, $q \in (0,1/5)$ be a mixture parameter and $\sigma \geq 4$ be a noise level. If $\alpha \leq \alpha^*(q,\sigma)$, then:*

$$S_\alpha(q,\sigma) \leq 2\alpha q^2/\sigma^2,$$

where $\alpha^*(q,\sigma)$ is the largest α satisfying:

$$\begin{aligned} \alpha &\leq \frac{M\sigma^2}{2} - \log(\sigma^2) \\ \textit{and} \quad \alpha &\leq \frac{M^2\sigma^2/2 - \log(5\sigma^2)}{M + \log(q\alpha) + 1/(2\sigma^2)}, \end{aligned}$$

with:

$$M = \log\left(1 + \frac{1}{q(\alpha-1)}\right).$$

Appendix B. Nonconvexity of E for the convex weakly smooth case

Recall that:

$$E(\mathbf{u}) := \sum_{t=1}^{T} \frac{(\varphi_{t-1}(u_{t-1}) - u_t)^2}{\sigma_{t-1}^2}.$$

Let's call g_{t-1} to each of addends of E, except for the fist and the last; i.e. for each $t = 2, \ldots, T-1$, let $g_{t-1}(u_{t-1}, u_t) := (\varphi_{t-1}(u_{t-1}) - u_t)^2$.

Proposition 42 *If $\varphi_t(\delta) = \sqrt{\delta^2 + h_t}$, then the Hessian of g_{t-1} is:*

$$\nabla^2 g_{t-1}(u_{t-1}, u_t) = \begin{pmatrix} 2 - \frac{2h_t u_t}{(u_{t-1}^2 + h_t)^{3/2}} & -\frac{2u_{t-1}}{\sqrt{u_{t-1}^2 + h_t}} \\ -\frac{2u_{t-1}}{\sqrt{u_{t-1}^2 + h_t}} & 2 \end{pmatrix}$$

The determinant and trace of this Hessian are:

$$\begin{aligned} \det \nabla^2 g_{t-1} &= \frac{4h_t\left(\sqrt{u_{t-1}^2 + h_t} - u_t\right)}{(u_{t-1}^2 + h_t)^{3/2}} \\ Tr\nabla^2 g_{t-1} &= 4 - \frac{2h_t u_t}{(u_{t-1}^2 + h_t)^{3/2}}. \end{aligned}$$

Moreover, $\nabla^2 E(\mathbf{u})$ is positive semidefinite when $\mathbf{u} \in \mathcal{R}$.

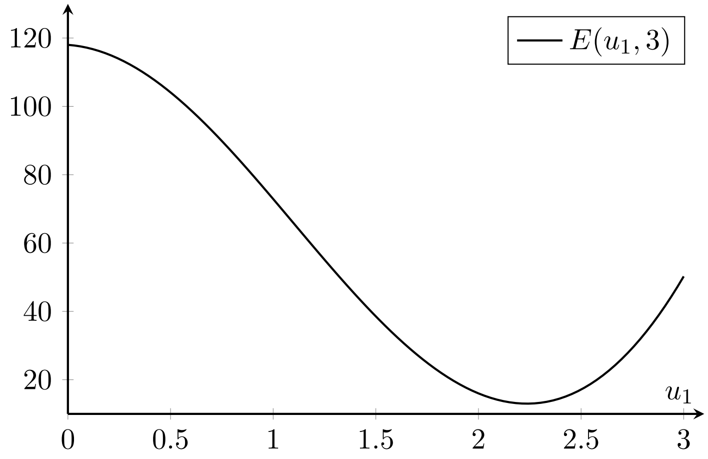

Figure 2: $E(u_1, 3)$ with $D = 1$, $\eta \equiv 1$, $L = 1$, $\sigma_0 = 1$, $\sigma_1 = 0.1$, $\sigma_2 = 1$.

Even though the Hessian of E is positive semidefinite over $\mathcal{R}$, it is easy to see that in the case $\varphi_t(\delta) = \sqrt{\delta^2 + h_t}$, the feasible set $\mathcal{R}$ is nonconvex. This prevents the shifts optimization problem from being convex.

Moreover, as can be seen in Figure 2, E is not convex over $\mathbb{R}^{T-1}$. So even if we prove that $u^* \in \mathcal{R}$, through second order conditions we can only guarantee that it is a local minimum.

Appendix C. Existence of Stationary Distributions for Nondifferentiable Potentials

When the potential f is in $\mathcal{C}^1$, the existence of a stationary distribution follows by standard results, which are based on the Feller condition (see, for example, Hairer 2006, Theorem 4.22 and Corollary 4.18). Since the potentials with Hölder continuous gradients fall in this case, we will focus only in the Lipschitz case.

If one only asks to the potential f to be Lipschitz, it is no longer necessary that it is differentiable. Of course, since our potentials are always convex, f will be subdifferentiable. In order to keep notation simple, we will use $\nabla f(x)$ to denote a subgradient of f in x. We will assume that we have access to an oracle that selects such subgradient and that is consistent with future choices; with this we mean that if the oracle have access to the same point in to different iterations of the algorithm, it will give the same subgradient.

Lemma 43 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex, compact set with diameter $D > 0$ and suppose $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ is a subdifferentiable function. Let P be the transition probabilities of the HMC defined by:*

$$X_{t+1} = \Pi_{\mathcal{X}} \left[X_t - \eta \nabla f(X_t) + \sqrt{2\eta}\xi_t \right],$$

where $\eta > 0$ and $(\xi_t)_{t \in \mathbb{N}_0} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{d \times d})$, then, for every $x \in \mathcal{X}$, the sequence $(P^t(x, \cdot))_{t \in \mathbb{N}}$ is a Cauchy sequence with respect to the total variation norm.

Proof Denote by J the transition operator associated with P. By Lemma 9 applied to the projection (which is nonexpansive) and Lemma 8 with $\alpha = 1$ and $a = D$, we have that for any $\mu, \nu \in \mathcal{P}(\mathcal{X})$:

$$\mathrm{KL}(J\mu || J\nu) \leq \frac{D^2}{4\eta}.$$

Then, by Bretagnolle-Huber inequality (Proposition 27)

$$\bar{d}(1) = \sup_{\mu,\nu \in \mathbb{P}(\mathcal{X})} \|J\mu - J\nu\|_{\mathrm{TV}} \leq \sqrt{1 - \exp\left(-\frac{D^2}{4\eta}\right)} =: \kappa < 1.$$

Let $\varepsilon > 0$ and take $l \geq \left\lceil \frac{\ln(1/\varepsilon)}{\ln(1/\kappa)} \right\rceil$. Then, by the submultiplicativity of $\bar{d}$ (Levin and Peres, 2017, Lemma 4.11), we have that

$$\sup_{\mu,\nu \in \mathcal{P}(\mathcal{X})} \left\| J^l \mu - J^l \nu \right\|_{\mathrm{TV}} = \bar{d}(l) \leq \bar{d}(1)^l \leq \kappa^l < \varepsilon\,.$$

Fix $x \in \mathcal{X}$. Taking $n \geq l$ and $m > j \geq 1$, we have that:

$$\begin{aligned}\left\|P^{n+m}(x,\cdot) - P^{n+j}(x,\cdot)\right\|_{\mathrm{TV}} &= \left\| J^{n-l}\left(J^l P^m(x,\cdot) - J^l P^j(x,\cdot)\right)\right\|_{\mathrm{TV}} \\ &\leq \left\| J^l P^m(x,\cdot) - J^l P^j(x,\cdot)\right\|_{\mathrm{TV}} \\ &< \varepsilon\end{aligned}$$

where the first inequality follows by the data processing inequality and the second by the way we chose l. ■

In order to save space, in the proof of the following theorem we will use sometimes the bracket notation to denote integrals with respect a measure. That is, if $g : \mathcal{X} \to \mathbb{R}$ is an integrable function and $\nu \in \mathcal{P}(\mathcal{X})$:

$$\langle g, \nu \rangle = \int_{\mathcal{X}} g(x)\nu(dx).$$

The following theorem is based on the proof of Hairer 2006, Theorem 4.17.

Theorem 44 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex, compact set with diameter $D > 0$ and suppose $f : \mathcal{X} \to \mathbb{R}$ is a subdifferentiable potential. Then the HMC defined by:*

$$\begin{aligned} X_0 &\sim \mu_0 \in \mathcal{P}(\mathcal{X}) \\ X_{t+1} &= \Pi_{\mathcal{X}}\left[X_t - \eta \nabla f(X_t) + \sqrt{2\eta}\xi_t\right], \end{aligned}$$

where $\xi_t \sim \mathcal{N}\left(0, I_{d\times d}\right)$, has a stationary distribution π_η.

Proof Take $x \in \mathcal{X}$. By Lemma 43 and the completeness of the total variation norm, there exists a measure $\pi_\eta \in \mathcal{P}(\mathcal{X})$ such that $\lim_{n\to\infty} \|P^n(x,\cdot) - \pi_\eta\|_{\mathrm{TV}} = 0$.

By Cèsaro convergence, the sequence of measures defined by $\nu_n = \frac{1}{n}\sum_{k=1}^{n} P^k(x,\cdot)$ also converges to π_η in total variation. Notice that the sequence $(\nu_n)_{n\in\mathbb{N}}$ satisfies the identity

$$J\nu_n - \nu_n = \frac{1}{n}\left[P^{n+1}(x,\cdot) - P(x,\cdot)\right]. \tag{16}$$

We will now prove that $J\pi_\eta = \pi_\eta$. In order to do this, we will prove that for every continuous and bounded function g it holds that

$$\left|\int_{\mathcal{X}} g(x)(J\pi_\eta)(dx) - \int_{\mathcal{X}} g(x)\pi_\eta(dx)\right| < \varepsilon \tag{17}$$

for every $\varepsilon > 0$.

Let's fix a bounded and continuous function g. We can decompose the left side of (17) into three easier to bound parts via triangle inequality:

$$|\langle g, J\pi_\eta\rangle - \langle g, \pi_\eta\rangle| \le |\langle g, J\pi_\eta\rangle - \langle g, J\nu_n\rangle| + |\langle g, J\nu_n\rangle - \langle g, \nu_n\rangle| + |\langle g, \nu_n\rangle - \langle g, \pi_\eta\rangle|, \tag{18}$$

Let us define $n_1, n_2, n_3 \in \mathbb{N}$ as:

1. n_1 is such that $\forall m \ge n_1$:
$$|\langle J_* g, \pi_\eta\rangle - \langle J_* g, \nu_m\rangle| < \frac{\varepsilon}{3}.$$
This quantity exists because $J_* g$ is a bounded and measurable function and ν_m converges in total variation to π_η, so the integral converge.

2. n_2 is such that $\forall m \ge n_2$:
$$\frac{2\max_{x\in\mathcal{X}} |g(x)|}{m} < \frac{\varepsilon}{3}.$$
The existence of such quantity is obvious. The relevance of this inequality comes from the following:
$$\begin{aligned}|\langle g, J\nu_m\rangle - \langle g, \nu_m\rangle| &= |\langle g, J\nu_m - \nu_m\rangle| \\ &= \left|\left\langle g, \frac{1}{m}(P^{m+1}(x,\cdot) - P(x,\cdot))\right\rangle\right| \\ &\le \frac{2\max_{x\in\mathcal{X}} |g(x)|}{m} \\ &< \frac{\varepsilon}{3}\end{aligned}$$
for all $m \ge n_2$, where the second equality comes from (16).

3. n_3 is such that for all $m \ge n_3$
$$|\langle g, \nu_m\rangle - \langle g, \pi_\eta\rangle| < \frac{\varepsilon}{3}.$$
This quantity exists since ν_m converges in total variation to π_η, which implies that the integral converge.

Taking $n \ge \max\{n_1, n_2, n_3\}$ in (18), we conclude that $|\langle g, J\pi_\eta\rangle - \langle g, \pi_\eta\rangle| < \varepsilon$ and therefore $J\pi_\eta = \pi_\eta$. ■

References

Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 3788–3800, 2022.

Jason Altschuler and Kunal Talwar. Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling. In *Proceedings of the Thirty-Sixth Conference on Learning Theory (COLT)*, volume 201 of *Proceedings of Machine Learning Research (PMLR)*, pages 2509–2510. PMLR, 2023.

Shahab Asoodeh and Mario Diaz. Privacy loss of noisy stochastic gradient descent might converge even for non-convex losses. *arXiv preprint arXiv:2305.09903*, 2023.

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*, pages 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4381–4391, 2020.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. *Discrete & computational geometry*, 59(4): 757–783, 2018. ISSN 0179-5376.

Clément L. Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2023.

Niladri Chatterji, Jelena Diakonikolas, Michael Jordan, and Peter Bartlett. Langevin Monte Carlo without smoothness. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research (PMLR)*, pages 1716–1726. PMLR, 2020.

Eli Chien and Pan Li. Convergent privacy loss of Noisy-SGD without convexity and smoothness. *arXiv preprint arXiv:2410.01068*, 2024.

Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of Langevin diffusion and noisy gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 14771–14781, 2021.

Arnak Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(3):651–676, 2017a.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, Proceedings of Machine Learning Research (PMLR), pages 678–689. PMLR, 2017b.

Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

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

Martin Hairer. Ergodic properties of Markov processes. *Lecture notes*, 2006. URL `https://hairer.org/notes/Markov.pdf`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

Tim Johnston, Iosif Lytras, Nikolaos Makras, and Sotirios Sabanis. The performance of the unadjusted Langevin algorithm without smoothness assumptions. *arXiv preprint arXiv:2502.03458*, 2025.

Adam Tauman Kalai and Santosh Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

Weiwei Kong and Mónica Ribero. Privacy of the last iterate in cyclically-sampled DP-SGD on nonconvex composite losses. *arXiv preprint arXiv:2407.05237*, 2024.

Joseph Lehec. The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *The Annals of Applied Probability*, 33(6A):4858 – 4874, 2023.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research (PMLR)*, pages 5809–5819. PMLR, 2020.

David Levin and Yuval Peres. *Markov Chains and Mixing Times*, volume 107. American Mathematical Society, 2017.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103. IEEE, 2007.

Ilya Mironov. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Siddharth Mitra and Andre Wibisono. Fast convergence of Φ-divergence along the unadjusted Langevin algorithm and proximal sampler. *arXiv preprint arXiv:2410.10699*, 2024.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

Alexandre Tsybakov. *Introduction to Nonparametric Estimation.* Springer-Verlag New York, 2009.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Max Welling and Yee Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688. Citeseer, 2011.

Andre Wibisono. Proximal Langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*, 2019.

Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 703–715, 2022.