Journal of Machine Learning Research 25 (2024) 1-67 Submitted 12/22; Published 4/24

Penalized Overdamped and Underdamped Langevin Monte Carlo Algorithms for Constrained Sampling

Mert Gürbüzbalaban MG1366@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers Business School
Piscataway, NJ 08854, United States of America

Yuanhan Hu YH586@SCARLETMAIL.RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers Business School
Piscataway, NJ 08854, United States of America

Lingjiong Zhu ZHU@MATH.FSU.EDU
Department of Mathematics
Florida State University
Tallahassee, FL 32306, United States of America

Editor: Maxim Raginsky

Abstract

We consider the constrained sampling problem where the goal is to sample from a target distribution $\pi(x) \propto e^{-f(x)}$ when x is constrained to lie on a convex body $\mathcal{C} \subset \mathbb{R}^d$. Motivated by penalty methods from continuous optimization, we propose and study penalized Langevin Dynamics (PLD) and penalized underdamped Langevin Monte Carlo (PULMC) methods for constrained sampling that convert the constrained sampling problem into an unconstrained sampling problem by introducing a penalty function for constraint violations. When f is smooth and gradients of f are available, we show $\tilde{\mathcal{O}}(d/\varepsilon^{10})$ iteration complexity for PLD to sample the target up to an ε-error where the error is measured in terms of the total variation distance and $\tilde{\mathcal{O}}(\cdot)$ hides some logarithmic factors. For PULMC, we improve this result to $\tilde{\mathcal{O}}(\sqrt{d}/\varepsilon^{7})$ when the Hessian of f is Lipschitz and the boundary of $\mathcal{C}$ is sufficiently smooth. To our knowledge, these are the first convergence rate results for underdamped Langevin Monte Carlo methods in the constrained sampling setting that can handle non-convex choices of f and can provide guarantees with the best dimension dependency among existing methods for constrained sampling when the gradients are deterministically available. We then consider the setting where only unbiased stochastic estimates of the gradients of f are available, motivated by applications to large-scale Bayesian learning problems. We propose PSGLD and PSGULMC methods that are variants of PLD and PULMC that can handle stochastic gradients and that are scaleable to large datasets without requiring Metropolis-Hasting correction steps. For PSGLD and PSGULMC, when f is strongly convex and smooth, we obtain an iteration complexity of $\tilde{\mathcal{O}}(d/\varepsilon^{18})$ and $\tilde{\mathcal{O}}(d\sqrt{d}/\varepsilon^{39})$ respectively in the 2-Wasserstein distance. For the more general case, when f is smooth and f can be non-convex, we also provide finite-time performance bounds and iteration complexity results. Finally, we illustrate the performance of our algorithms on Bayesian LASSO regression and Bayesian constrained deep learning problems.

Keywords: Constrained sampling, Bayesian learning, Langevin Monte Carlo, penalty methods, stochastic gradient algorithms

1. Introduction

We consider the problem of sampling a distribution π on a convex constrained domain $\mathcal{C} \subsetneq \mathbb{R}^d$ with probability density function

$$\pi(x) \propto \exp(-f(x)),\ x \in \mathcal{C}, \tag{1}$$

for a function $f : \mathbb{R}^d \to \mathbb{R}$. This is a fundamental problem arising in many applications, including Bayesian statistical inference (Gelman et al., 1995), Bayesian formulations of inverse problems (Stuart, 2010), as well as Bayesian classification and regression tasks in machine learning (Andrieu et al., 2003; Teh et al., 2016; Gürbüzbalaban et al., 2021).

In the absence of constraints, i.e., when $\mathcal{C} = \mathbb{R}^d$ in (1), many algorithms in the literature are applicable (Geyer, 1992; Brooks et al., 2011) including the class of Langevin Monte Carlo algorithms. One popular algorithm for this setting is the *unadjusted Langevin algorithm*:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\xi_{k+1}, \tag{2}$$

where ξ_k are independent and identically distributed (i.i.d.) $\mathcal{N}(0, I_d)$ Gaussian vectors in $\mathbb{R}^d$. The classical Langevin algorithm (2) is the Euler discretization of the *overdamped (or first-order) Langevin diffusion*:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2}dW_t, \tag{3}$$

where W_t is a standard d-dimensional Brownian motion that starts at zero at time zero. Under some mild assumptions on f, the stochastic differential equation (SDE) (3) admits a unique stationary distribution with the density $\pi(x) \propto e^{-f(x)}$, known as the *Gibbs distribution* (Chiang et al., 1987; Holley et al., 1989). In computing practice, this diffusion is simulated by considering its discretization as in (2) whose stationary distribution may contain a bias that a Metropolis-Hasting step can correct. However, for many applications, including those in data science and machine learning, employing this correction step can be computationally expensive (Bardenet et al., 2017; Teh et al., 2016); therefore, our focus will be on unadjusted algorithms that avoid it.

Unadjusted Langevin algorithms have a long history and admit various asymptotic convergence guarantees (Talay and Tubaro, 1990; Mattingly et al., 2002; Gelfand and Mitter, 1991); however non-asymptotic performance bounds for them are relatively more recent (Dalalyan, 2017a; Durmus and Moulines, 2017, 2019; Durmus et al., 2018; Cheng and Bartlett, 2018). The unadjusted Langevin algorithm (2) assumes availability of the gradient ∇f. On the other hand, in many settings in machine learning, computing the full gradient ∇f is either infeasible or impractical. For example, in Bayesian regression or classification problems, f can have a finite-sum form as the sum of many component functions, i.e., $f(x) = \sum_{i=1}^n f_i(x)$ where $f_i(x)$ represents the loss of a predictive model with parameters x for the i-th data point and the number of data points n can be large (see, e.g., Gürbüzbalaban et al. (2021); Xu et al. (2018)). In such settings, algorithms that rely on *stochastic gradients*, i.e., unbiased stochastic estimates of the gradient obtained by a

randomized sampling of the data points, is often more efficient (Bottou, 2010). This fact motivated the development of Langevin algorithms that can support stochastic gradients. In particular, if one replaces the full gradient ∇f in (2) by a stochastic gradient, the resulting algorithm is known as the stochastic gradient Langevin dynamics (SGLD) (see, e.g., Welling and Teh (2011); Chen et al. (2015)).

Unadjusted underdamped Langevin Monte Carlo (ULMC) algorithms based on an alternative diffusion called underdamped (or second-order) Langevin diffusion have also been proposed; see e.g. Dalalyan and Riou-Durand (2020); Ma et al. (2021). Their versions that support stochastic gradients are also studied (see e.g. Chen et al. (2014); Zou and Gu (2021); Gao et al. (2022)). Although ULMC algorithms can often be faster than unadjusted (overdamped) Langevin algorithms on many practical problems (Chen et al., 2014), this is rigorously proven for particular choices of f (Chen et al., 2015; Gao et al., 2022; Mangoubi and Smith, 2021; Chen and Vempala, 2022) rather than general non-convex choices of f and the convergence of ULMC algorithms remains relatively less studied.

In this paper, we focus on the constrained setting when $\mathcal{C}$ is a convex body, i.e., when $\mathcal{C}$ is a compact convex set with a non-empty interior, and we consider both settings when f can be strongly convex or non-convex. We also consider both deterministic and stochastic gradients. Among the existing approaches that are the most closely related to our setting, Bubeck et al. (2015, 2018) studied the projected Langevin Monte Carlo algorithm that projects the iterates back to the constraint set after applying the Langevin step (2) where it is assumed that f is β-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$ for any $x, y \in \mathcal{C}$ and the norm of the gradient of f is bounded, i.e. $\|\nabla f(x)\| \leq L$. It is shown in Bubeck et al. (2018) that $\tilde{\mathcal{O}}(d^{12}/\varepsilon^{12})$ iterations are sufficient for having ε-error in the total variation (TV) metric with respect to the target distribution when the gradients are exact where the notation $\tilde{\mathcal{O}}(\cdot)$ hides some logarithmic factors. Lamperski (2021) considers the projected stochastic gradient Langevin dynamics (P-SGLD) in the setting of non-convex smooth Lipschitz f on a convex body where the gradient noise is assumed to have finite variance with a uniform sub-Gaussian structure. The author shows that $\tilde{\mathcal{O}}\left(d^4/\varepsilon^4\right)$ iterations suffice in the 1-Wasserstein metric. More recently, Zheng and Lamperski (2022) study P-SGLD for constrained sampling for a non-convex potential f that is strongly convex outside a radius of R where data variables are assumed to be L-mixing. They obtain an improved complexity of $\tilde{\mathcal{O}}\left(d^2/\varepsilon^2\right)$ for P-SGLD in the 1-Wasserstein metric with polyhedral constraints that are not necessarily bounded. Constrained sampling for convex f and strongly-convex f is also studied in Brosse et al. (2017), where a proximal Langevin Monte Carlo is proposed and a complexity of $\tilde{\mathcal{O}}\left(d^5/\varepsilon^6\right)$ is obtained. Salim and Richtárik (2020) further studies the proximal stochastic gradient Langevin algorithm from a primal-dual perspective. For constrained sampling when f is strongly convex, and the constraint set is convex, the proximal step corresponds to a projection step, and they obtain $\tilde{\mathcal{O}}(d/\varepsilon^2)$ complexity for the proximal stochastic gradient Langevin algorithm in terms of the 2-Wasserstein distance.

Mirror descent-based Langevin algorithms (see e.g. Hsieh et al. (2018); Chewi et al. (2020); Zhang et al. (2020); Li et al. (2022a); Ahn and Chewi (2021)) can also be used for constrained sampling. Mirrored Langevin dynamics was proposed in Hsieh et al. (2018), inspired by the classical mirror descent in optimization. For any target distribution with strongly-log-concave density (which corresponds to f being strongly convex), Hsieh et al. (2018) showed that their first-order algorithm requires $\tilde{\mathcal{O}}(\epsilon^{-2}d)$ iterations for ε error with

exact gradients and $\tilde{\mathcal{O}}(\epsilon^{-2}d^2)$ iterations for stochastic gradients. Zhang et al. (2020) establishes for the first time a non-asymptotic upper bound on the sampling error of the resulting Hessian Riemannian Langevin Monte Carlo algorithm that is closely related to the mirror-descent scheme. This bound is measured according to a Wasserstein distance induced by a Riemannian metric capturing the Hessian structure. In contrast to Hsieh et al. (2018), Zhang et al. (2020) studies a different scheme in which an appropriate diffusion term is used that entails a Gaussian noise in the discrete scheme with iteration-dependent covariances that account for the Hessian Riemannian structure instead of a standard Gaussian noise adopted in Hsieh et al. (2018). Moreover Zhang et al. (2020) relaxes the strong-convexity assumptions to relative versions. Motivated by Zhang et al. (2020), Chewi et al. (2020) propose a class of diffusions called Newton-Langevin diffusions and prove that they converge exponentially fast with a rate that has no dependence on the condition number of the target density in continuous time. They give an application of this result to the problem of sampling from the uniform distribution on a convex body using a strategy inspired by interior-point methods. In Jiang (2021), the author relaxes the strongly-log-concave density assumption in mirror-descent Langevin dynamics and assumes that the density function satisfies the mirror log-Sobolev inequality. Further improvements Zhang et al. (2020) have been achieved in Ahn and Chewi (2021); Li et al. (2022a). The analysis of Zhang et al. (2020) gives an error bound that contains a bias that does not vanish even if the stepsize goes to zero. The solution to this problem was first attempted by Ahn and Chewi (2021) who proposed an alternative discretization which achieves a vanishing bias, but requires an exact simulation of the Brownian motion with changing covariance. Finally, Li et al. (2022a) proved this bias is an artifact of analysis by building upon the mean-square analysis in Li et al. (2019, 2022b).

1.1 Our Approach and Contributions

Recent years have witnessed techniques and concepts from continuous optimization being used for analyzing and developing new Langevin algorithms (Dalalyan, 2017b; Balasubramanian et al., 2022; Chen et al., 2022; Gürbüzbalaban et al., 2021). In this paper, we develop Langevin algorithms for constrained sampling, leveraging penalty functions from continuous optimization. More specifically, penalty methods are frequently used in continuous optimization (Nocedal and Wright, 2006), where one converts the constrained optimization problem of minimizing an objective $f(x)$ subject to $x \in \mathcal{C}$ to an unconstrained optimization problem of minimizing $f_\delta(x) := f(x) + \frac{1}{\delta}S(x)$ on $\mathbb{R}^d$, where $\delta > 0$ is called the *penalty parameter*, and the function $S : \mathbb{R}^d \to [0, \infty)$ is called the *penalty function* with the property that $S(x) = 0$ for $x \in \mathcal{C}$ and $S(x)$ increases as x gets away from the constraint set $\mathcal{C}$. For $\delta > 0$ small enough, it can be seen that the global minimum of f_δ will approximate the global minimum of f on $\mathcal{C}$. Motivated by this technique, our main approach is to sample from a *penalized target distribution in an unconstrained fashion* with the modified target density:

$$\pi_\delta(x) \propto \exp\left(-\left(f(x) + \frac{1}{\delta}S(x)\right)\right), \qquad x \in \mathbb{R}^d,$$

for suitably chosen small enough $\delta > 0$. Here, a key challenge is to control the error between π_δ and π efficiently, leveraging the convex geometry of the constraint set and the proper-

ties of the penalty function. We then use the unconstrained SGLD or stochastic gradient underdamped Langevin Monte Carlo (SGULMC) algorithm to sample from the modified target distribution and call the resulting algorithms penalized SGLD (PSGLD) and penalized SGULMC (PSGULMC). If the gradients are deterministic, then we call the algorithms penalized Langevin dynamics (PLD) and penalized underdamped Langevin Monte Carlo (PULMC). Our detailed contributions are as follows:

- When f is smooth, meaning that its gradient is Lipschitz, we show $\tilde{\mathcal{O}}(d/\varepsilon^{10})$ iteration complexity in the TV distance for PLD. For PULMC, we improve this result to $\tilde{\mathcal{O}}(\sqrt{d}/\varepsilon^{7})$ when the Hessian of f is Lipschitz and the boundary of $\mathcal{C}$ is sufficiently smooth. To our knowledge, these are the first convergence rate results for underdamped MC methods in the constrained sampling setting that can handle non-convex choices of f and provide guarantees with the best dimension dependency among existing methods for constrained sampling when subject to deterministic gradients. To achieve these results, we develop a novel analysis and make a series of technical contributions. We first bound the Kullback-Leibler (KL) divergence between π_δ and π with a careful technical analysis and then apply weighted Csiszár-Kullback-Pinsker inequality to control the 2-Wasserstein distance between π_δ and π. To obtain the convergence rate to π_δ, we first regularize the convex domain $\mathcal{C}$ so that the regularized domain $\mathcal{C}^\alpha$ is α-strongly convex (a notation which will be defined rigorously in (13) and in the proof of Lemma D.1) and then show that $f+S^\alpha/\delta$ is strongly convex outside a compact domain, where S^α is the penalty function we construct for the regularized domain that has quadratic growth properties. Moreover, we quantify the differences between $\mathcal{C}^\alpha$ and $\mathcal{C}$, and between the regularized target π_δ^α (defined on the regularized domain $\mathcal{C}^\alpha$) and π_δ and show their differences are small for the choice of small values of α. Finally, we show that $f + S^\alpha/\delta$ is uniformly close to a function that is strongly convex everywhere and apply the convergence result for Langevin dynamics in the unconstrained setting to obtain our main result for PLD. The analysis for PULMC is similar but requires an additional technical result showing Hessian Lipschitzness.

- We then consider the setting of smooth f that can be non-convex subject to stochastic gradients. For the unconstrained sampling of π_δ, when the gradients of f are estimated from a randomly selected subset of data; the variance of the noise is not uniformly bounded over x but instead can grow linearly in $\|x\|^2$ (see e.g. Jain et al. (2018); Assran and Rabbat (2020)). Therefore, unlike the existing works for constrained sampling, we do not assume the variance of the stochastic gradient to be uniformly bounded but allow the gradient noise variance to grow linearly. For PSGLD and PSGULMC, we show an iteration complexity that scales as $\tilde{\mathcal{O}}(d^{17}/\lambda_*^9)$ and $\tilde{\mathcal{O}}(d^7/\mu_*^3)$ respectively in dimension d, where λ_* and μ_* are constants that relate to overdamped and underdamped Langevin SDEs and will be defined later in (29) and (36). These constants can scale exponentially in the dimension in the worst case (due to hardness of the non-convex setting) but can also be independent of the dimension (see Section 4 in Raginsky et al. (2017)). Our iteration complexity bounds for PSGLD and PSGULMC also scale polynomially in ε (see Table 1 for the details).[1] To our best knowledge, these

1. In Table 1, we used various metrics TV, $\mathcal{W}_1$, $\mathcal{W}_2$ and KL to measure the complexity and it is worth noting that they may scale differently. In general, it is always true that $\mathcal{W}_1 \leq \mathcal{W}_2$ and $\text{TV} \leq \mathcal{O}(\sqrt{\text{KL}})$

are the first results for ULMC algorithms in the constrained setting for general f that can be non-convex. Compared to Lamperski (2021), our dimension dependency is worse, but our noise assumption is more general, and we do not require sub-Gaussian noise. To achieve these results, in addition to bounding the difference between π_δ and π, we show that $f + S/\delta$ satisfies a dissipativity condition, which is the key technical result, that allows us to apply the convergence results in the literature for unconstrained Langevin algorithms with stochastic gradients where the target is non-convex and satisfies a dissipativity condition. Here, we also note that the standard penalty function we choose involves computing the distance of a point to the boundary of the constraint set. This is also the case for many algorithms in the literature, such as projected SGLD methods. However, often the set $\mathcal{C}$ is defined with convex constraints, i.e. $\mathcal{C} := \{x : h_i(x) \leq 0, i = 1, 2, \ldots, m\}$ where $h_i : \mathbb{R}^d \to \mathbb{R}$ are convex and m is the number of constraints. In this case, we discuss in Section 2.4 that the projections can be avoided when $h_i(x)$ satisfies some growth conditions.

- When f is strongly convex and smooth, we obtain iteration complexity of $\tilde{\mathcal{O}}(d/\varepsilon^{18})$ and $\tilde{\mathcal{O}}(d\sqrt{d}/\varepsilon^{39})$ for PSGLD and PSGULMC respectively. To achieve these results, in addition to bounding the difference between π_δ and π, we also extend the existing result in the unconstrained setting for ULMC with a deterministic gradient to allow stochastic gradient for strongly convex and smooth f, which is of independent interest.

The summary of our main results and their comparison with respect to most closely-related approaches are given in Table 1, where in our results it is assumed that the constraint set is compact and convex. We also note that when dealing with target densities where f is smooth but non-convex, the literature typically assumes growth conditions towards infinity such as dissipativity or isoperimetric inequalities (Raginsky et al., 2017; Gao et al., 2022; Jiang, 2021), but in our results we do not require such a condition. This is due to the fact that the constraint set is taken to be a convex body which is a compact set where the growth of f can be controlled.

1.2 Related Work

Mirror-descent Langevin algorithms can be viewed as a special case of Riemannian Langevin that can be used to sample from some subset $D \subseteq \mathbb{R}^d$ by endowing D with a Riemannian structure (Girolami and Calderhead, 2011; Patterson and Teh, 2013). Geodesic Langevin algorithm is proposed in Wang et al. (2020) that can sample a distribution supported on a manifold M. They showed that geodesic Langevin algorithm can sample a target distribution on a d-dimensional compact manifold M without boundary that satisfies a log-Sobolev inequality with parameter α with ε accuracy in KL divergence after $\mathcal{O}(\frac{d}{\alpha^2\epsilon}\log(1/\varepsilon))$ iterates. More recently, Gatmiry and Vempala (2022) showed that the Riemannian Langevin algorithm converges to the target that satisfies a log-Sobolev inequality with parameter α with accuracy ε in KL divergence after $\mathcal{O}(\frac{d^{5/2}}{\alpha^2\epsilon}\log(1/\varepsilon))$ iterates where d is the dimension for general Hessian manifolds that are second-order self-concordant where the log-density is gradient and Hessian Lipschitz. Very recently, Kook et al. (2022) used a Riemannian version

(Pinsker's inequality). On the other hand, $\mathcal{W}_2 \leq \mathcal{O}(\sqrt{\mathrm{KL}})$ (Otto and Villani Theorem) if a log-Sobolev inequality is satisfied and more generally $\mathcal{W}_2 \leq \mathcal{O}(\sqrt{\mathrm{KL}} + (\mathrm{KL})^{1/4})$ (Bolley and Villani (2005)).

Algorithms	Assump. on f	Assump. on $\mathcal{C}$	Stoc. grad.	Bdd. grad. noise var.[5]	Conv. meas.	Complexity
Projected LD (Bubeck et al., 2018)	Convex, Smooth, Lipschitz	Convex body	No	N/A	TV	$\tilde{\mathcal{O}}\left(\frac{d^{12}}{\varepsilon^{12}}\right)$
Projected SGLD (Lamperski, 2021)	Smooth, Lipschitz	Convex body	Yes	Yes[6]	$\mathcal{W}_1$	$\tilde{\mathcal{O}}\left(\frac{d^4}{\varepsilon^4}\right)$
Projected SGLD (Zheng and Lamperski, 2022)	Str. cvx. outside a ball,[1] Lipschitz	Polyhedral with 0 in interior	Yes	Yes	$\mathcal{W}_1$	$\tilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^2}\right)$
Proximal SGLD (Salim and Richtárik, 2020)	Str. cvx.	Convex[2]	Yes	Yes	$\mathcal{W}_2$	$\tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^2}\right)$
Mirrored LD (Hsieh et al., 2018)	Str. cvx.	Convex, Bounded	No	N/A	$\mathcal{W}_2$	$\tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^2}\right)$
Mirrored SGLD (Hsieh et al., 2018)	Str. cvx.	Convex, Bounded	Yes	Yes	KL	$\tilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^2}\right)$
MYULA (Brosse et al., 2017)	Convex, Smooth	Convex body	No	N/A	TV	$\tilde{\mathcal{O}}\left(\frac{d^5}{\varepsilon^6}\right)$
PLD Prop. 2.11 in our paper	Smooth	Convex body	No	N/A	TV	$\tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right)$
PULMC Prop. 2.15 in our paper	Smooth, Hessian Lipschitz	Convex body ‡	No	N/A	TV	$\tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^7}\right)$
PSGLD Prop. 2.21 in our paper	Str. cvx., Smooth	Convex body	Yes	No	$\mathcal{W}_2$	$\tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{18}}\right)$
PSGULMC Prop. 2.22 in our paper	Str. cvx., Smooth	Convex body	Yes	No	$\mathcal{W}_2$	$\tilde{\mathcal{O}}\left(\frac{d\sqrt{d}}{\varepsilon^{39}}\right)$
PSGLD Prop. 2.23 in our paper	Smooth	Convex body	Yes	No	$\mathcal{W}_2$	$\tilde{\mathcal{O}}\left(\frac{d^{17}}{\varepsilon^{392}\lambda_*^9}\right)$[3]
PSGULMC Prop. 2.24 in our paper	Smooth	Convex body	Yes	No	$\mathcal{W}_2$	$\tilde{\mathcal{O}}\left(\frac{d^7}{\varepsilon^{132}\mu_*^3}\right)$[4]

Table 1: Comparison of our methods and existing methods.

‡: $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex hypersurface of class C^3 and $\sup_{\xi\in\mathcal{C}} \|D^2 n(\xi)\|$ is bounded, where n is the unit normal vector of $\mathcal{C}$. [1] "Str. cvx." stands for "Strongly convex". Also, in Zheng and Lamperski (2022), it is assumed that f is μ-strongly convex outsize a Euclidean ball. [2] Salim and Richtárik (2020) consider the situation, where the target distribution is $\pi \propto e^{-V(x)}$ with $V(x) := f(x) + G(x)$. Function G is assumed to be nonsmooth and convex, and if G is the indicator function of $\mathcal{C}$, then proximal SGLD can sample from the constrained distribution.[3] λ_* is the spectral gap of penalized overdamped Langevin SDE (10) which is defined in (29). [4] μ_* is the convergence speed of penalized underdamped Langevin SDE (17)-(18) defined in (36). [5] This column specifies whether the methods assume that the gradient noise variance is uniformly bounded or not. [6] The gradient noise is assumed to have a uniform sub-Gaussian property.

of Hamiltonian Monte Carlo to sample ill-conditioned, non-smooth, constrained distributions that maintain sparsity where f is convex. Given a self-concordant barrier function for the constraint set, they empirically demonstrated that they could achieve a mixing rate independent of smoothness and condition numbers. Moreover, Chalkis et al. (2023) proposed reflective Hamiltonian Monte Carlo based on reflected underdamped Langevin diffusion to sample from a strongly-log-concave distribution restricted to a convex polytope. They showed that from a warm start, it mixes in $\tilde{\mathcal{O}}(\kappa d^2\ell^2 \log(1/\varepsilon))$ steps for a well-rounded polytope, where κ is the condition number of f, and ℓ is an upper bound on the number of reflections.

It is also worth mentioning that the idea of adding a penalty term to the Langevin diffusion (3) has appeared in the recent literature but in a very different context (Karagulyan and Dalalyan, 2020). By adding a penalty term to the Langevin diffusion with the log-concave target, the resulting target becomes strongly log-concave, and as the penalty term vanishes, Karagulyan and Dalalyan (2020) were able to obtain new convergence results for sampling a log-concave target.

SGLD algorithms have been studied in the unconstrained setting in a number of papers under various assumptions for f. Among these, we discuss closely related works. Dalalyan and Karagulyan (2019) study the convergence of SGLD for strongly convex smooth f. In a seminal work, Raginsky et al. (2017) show that when f is non-convex and smooth, under a dissipativity condition, SGLD iterates track the overdamped Langevin SDE closely and obtained finite-time performance bounds for SGLD. More recently, Xu et al. (2018) improve the ε dependency of the upper bounds of Raginsky et al. (2017) in the mini-batch setting and obtained several guarantees for the gradient Langevin dynamics and variance-reduced SGLD algorithms. Zou et al. (2021) improve the existing convergence guarantees of SGLD for unconstrained sampling, showing that $\mathcal{O}(d^4\varepsilon^{-2})$ stochastic gradient evaluations suffice for SGLD to achieve ε-sampling accuracy in terms of the TV distance for a class of distributions that can be non-log-concave. They further show that provided an additional Hessian Lipschitz condition on the log-density function, SGLD is guaranteed to achieve ε-sampling error within $\mathcal{O}(d^{15/4}\varepsilon^{-3/2})$ stochastic gradient evaluations. There have also been more recent works on SGLD algorithms that allow dependent data streams (Barkhagen et al., 2021; Chau et al., 2021) and require weaker assumptions on the target density (Zhang et al., 2023). Rolland et al. (2020) study a new annealing stepsize schedule for Unadjusted Langevin Algorithm (ULA) and they improve the convergence rate to $\mathcal{O}(d^3/T^{\frac{2}{3}})$ for unconstrained log-concave distribution, where d is the dimension and T is the number of iterates. They also apply the double-loop approach to the constrained sampling algorithm Moreau-Yoshida ULA (MYULA) from Brosse et al. (2017). They improve the convergence rate to $\mathcal{O}(d^{3.5}/\varepsilon^5)$ in the total variation distance for constrained log-concave distributions. Lan and Shahbaba (2016) propose a spherical augmentation method to sample constrained probability distributions by mapping the constrained domain to a sphere in the augmented space. Several other works have also studied SGULMC algorithms in the unconstrained setting. Zou and Gu (2021) propose a general framework for proving the convergence rate of Hamiltonian Monte Carlo with stochastic gradient estimators for sampling from strongly log-concave and log-smooth target distributions in the unconstrained setting. They show that the convergence to the target distribution in the 2-Wasserstein

distance can be guaranteed as long as the stochastic gradient estimator is unbiased and its variance is upper-bounded along the algorithm trajectory.

Lehec (2023) considers the projected Langevin algorithms and improves upon the work of Bubeck et al. (2018). The author considers the constrained sampling case when the potential f is a convex function that is Lipschitz on a convex constraint set $\mathcal{C} \subseteq \mathbb{R}^d$. In this setting, Lehec (2023) obtains an upper bound on the discretization error between the iterates x_k of the projected Langevin algorithm and its corresponding points in the Langevin diffusion based on the $\mathcal{W}_2$ distance (Lehec, 2023, Thm 1). Using this bound, under the additional assumptions that the target π satisfies a log-Sobolev inequality with constant C_{LS} and the initial iterate x_0 is a point in the support of π, a bound on the $\mathcal{W}_2$ distance between the law of the iterates and the target is proven (Lehec, 2023, Thm 2). Assuming further that the initial iterate x_0 is such that $\sigma_0 := f(x_0) - \min_x f(x) = \mathcal{O}(1)$, the latter result implies that $\mathcal{W}_2(\mathcal{L}(x_k), \pi) \leq \varepsilon$ after $k = \Theta^* \left(\frac{C_{LS}^3 d^2}{\varepsilon^4} \max \left(\frac{d}{r_0^2}, \frac{L_f^2}{d} \right) \right)$ iterations, where $\mathcal{L}(x_k)$ denotes the law of the k-th iterate x_k, where L_f is the Lipschitz constant of f on $\mathcal{C}$, r_0 is the distance of initial point x_0 to the boundary of $\mathcal{C}$, with the convention that Θ^* hides universal constants and possible polylog(d) dependencies. Here, when f is strongly convex and when the constraint set $\mathcal{C}$ is bounded, as discussed in Lehec (2023), we can take $C_{LS} = \frac{1}{\mu}$ where μ is the strong convexity constant of f. Also, when the constraint set is a ball of radius R and when the target measure $\pi(x) \propto e^{-f(x)}$ is isotropic in the sense that its covariance matrix is the identity matrix, then we can take C_{LS} to be R up to a universal constant where by the isotropy condition it holds that $R \geq \sqrt{d}$ (Lehec, 2023). Some convex choices of f may not necessarily satisfy the log-Sobolev inequality, but they do satisfy the Poincaré inequality for some finite constant C_P. For convex f (that does not necessarily satisfy the log-Sobolev inequality), Lehec (2023) also obtains Wasserstein bounds between the iterates and the target (that depends on the Poincaré constant C_P) when f is globally Lipschitz on the domain $\mathcal{C}$ under a warm-start strategy where the initialization x_0 is taken as a random point taking values in $\mathcal{C}$ whose chi-square divergence to π is finite (Lehec, 2023, Thm 2). This result is applicable to the case when the constraint set $\mathcal{C}$ is unbounded, and when $\sigma_0 = \mathcal{O}(1)$ and all the other parameters are at most polynomial in d, it implies in the unconstrained case that $\mathcal{W}_2(\mathcal{L}(x_k), \pi) \leq \varepsilon$ after $k = \Theta^* \left(\frac{C_P^3 L_f^2 d^4}{\varepsilon^4} \right)$ iterations. Compared to Lehec (2023), when f is strongly convex, we can get a better dimension dependency but our dependency on ε is worse. We can also allow f to be non-convex as long as it is smooth and our analysis can support stochastic gradients for both overdamped and underdamped dynamics; however, we require the constraint set $\mathcal{C}$ to be bounded.

In a recent work, Sato et al. (2022) considers the problem of constrained sampling when the potential f is C^4 with a Lipschitz gradient on the constraint set and when the constraint set $\mathcal{C}$ has a smooth (C^4) boundary, allowing it to be non-convex. The authors also assume that the projection to set $\mathcal{C}$ is unique and that the projections can be efficiently computed where they study a reflection-based overdamped Langevin algorithm that can be viewed as a discretization of a reflected Langevin diffusion, assuming access to (non-stochastic) exact gradients of f. To compute the reflections, their algorithm necessitates to compute projections at every step. The authors show that the optimization error converges to the target distribution and it suffices to have $\tilde{\mathcal{O}}(\frac{d^3}{\lambda_r \varepsilon^3})$ iterations for the suboptimality to be at

most ε in expectation where λ_r is the spectral gap of the reflected Langevin diffusion. In our paper, we require $\mathcal{C}$ to be convex but its boundary can be non-smooth. For the overdamped Langevin version of our algorithm which we call PLD, we require $\tilde{\mathcal{O}}(\frac{d}{\varepsilon^{10}})$ iterations which is better dependency to the dimension when $\mathcal{C}$ is convex; furthermore we can avoid projections and therefore we do not necessarily require the projections to be efficiently computable, in addition we do not necessarily require a smooth boundary. Moreover, our results can also handle underdamped dynamics and stochastic gradients which are key to handle machine learning applications, whereas the stochastic gradient setting is not considered in Sato et al. (2022).

Finally, we note that "hit-and-run walk" achieves a mixing time of $\tilde{\mathcal{O}}(d^4)$ iterations (Lovász and Vempala, 2007). However, they assume a "zeroth order oracle", i.e., assuming access to function values without access to its gradients. Thus, our setting is different where we work with gradients.

The notations to be used in the rest of the paper are summarized in Appendix A.

2. Main Results

Penalty methods in optimization convert a constrained optimization problem to an unconstrained one, where the idea is to add a term to the optimization objective that penalizes for being outside of the constraint set (Nocedal and Wright, 2006). Motivated by such methods, as discussed in the introduction, we propose to add a penalty term $\frac{1}{\delta}S(x)$ to the target distribution, and sample instead from the penalized target distribution in an unconstrained fashion with the modified target density:

$$\pi_\delta(x) \propto \exp\left(-f(x) - \frac{1}{\delta}S(x)\right), \qquad x \in \mathbb{R}^d,$$

where $S(x)$ is the penalty function that satisfies the following assumption and $\delta > 0$ is an adjustable parameter.

Assumption 2.1 *Assume that $S(x) = 0$ for any $x \in \mathcal{C}$ and $S(x) > 0$ for any $x \notin \mathcal{C}$.*

There are many simple choices of $S(x)$ for which Assumption 2.1 is satisfied. For instance, if we choose $S(x) = g(\delta_{\mathcal{C}}(x))$, where $\delta_{\mathcal{C}}(x) = \min_{c\in\mathcal{C}} \|x - c\|$ is the distance of the point x to a closed set $\mathcal{C}$ and $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a strictly increasing function with $g(0) = 0$, then Assumption 2.1 is satisfied. Throughout our paper, we will also discuss other choices of $S(x)$. In many of our results, we will also make the following assumption on the set $\mathcal{C}$.

Assumption 2.2 *Assume that $\mathcal{C}$ is a convex body, i.e., $\mathcal{C}$ is a compact convex set, contains an open ball centered at 0 with radius $r > 0$, and is contained in a Euclidean ball centered at 0 with radius $R > 0$.*

The fact that 0 is in the set $\mathcal{C}$ in Assumption 2.2 is made for simplifying the presentation but all our results will hold even if that is not the case. Assumption 2.2 has been commonly made in the literature (Bubeck et al., 2018, 2015; Lamperski, 2021; Brosse et al., 2017). In addition, for many applications including those arise in machine learning, this assumption naturally holds; for instance, when the constraints are polyhedral (Kook et al., 2022) or when the constraints are ℓ_p-norm constraints with $p \geq 1$ or for $p = \infty$ (Schmidt, 2005; Luo et al., 2016; Ma et al., 2019a; Gürbüzbalaban et al., 2022).

2.1 Bounding the Distance Between π_δ and π

In this section, we aim to bound the 2-Wasserstein distance between the modified target π_δ and the target π with an explicitly computable upper bound that goes to zero as δ tends to zero. We will first bound the KL divergence between π_δ and π and then apply weighted Csiszár-Kullback-Pinsker inequality (W-CKP) (see Lemma B.1) to bound the 2-Wasserstein distance between π_δ and π. To start with, we first bound the KL divergence between π_δ and π, which relies on a series of technical lemmas. The two main ideas are: (i) when the penalty value S is small, the Lebesgue measure of the set with small penalty values is also small so that its contribution is negligible; (ii) for small values of δ, the penalty $\frac{S}{\delta}$ is large and the integral with respect to that is also negligible. We start with the following lemma. The proofs of this lemma and our other results can be found in the appendix.

Lemma 2.3 *Suppose Assumption 2.1 holds and e^{-f} is integrable over $\mathcal{C}$. For any $\delta > 0$,*[2]

$$D(\pi \| \pi_\delta) \leq \frac{\int_{\mathbb{R}^d \backslash \mathcal{C}} e^{-\frac{1}{\delta} S(y) - f(y)} dy}{\int_{\mathcal{C}} e^{-f(y)} dy}. \tag{4}$$

Next, we provide a technical lemma that provides an upper bound on the Lebesgue measure of the set with small penalty values S. A special case of the following lemma can be found in Lemma 10.15 without a proof in Kallenberg (2002).[3]

Lemma 2.4 *Assume the constraint set $\mathcal{C}$ is a bounded closed set containing an open ball with radius $r > 0$. Let $S(x) = g\left(\delta_{\mathcal{C}}(x)\right)$, where $\delta_{\mathcal{C}}(x) = \min_{c \in \mathcal{C}} \|x - c\|$ is the distance of the point x to the set $\mathcal{C}$ and $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a strictly increasing function with $g(0) = 0$ with the property $g(x) \to \infty$ as $x \to \infty$. Then, for any $\epsilon > 0$,*

$$\left| x \in \mathbb{R}^d \backslash \mathcal{C} : S(x) \leq \epsilon \right| \leq \left(\left(1 + \frac{g^{-1}(\epsilon)}{r} \right)^d - 1 \right) |\mathcal{C}|,$$

where $|\cdot|$ denotes the Lebesgue measure and g^{-1} is the inverse function of g.

We are now ready to provide an upper bound for $D(\pi \| \pi_\delta)$, the KL divergence between the target distribution π and the penalized target distribution π_δ.

2. If $e^{-\frac{1}{\delta} S(y) - f(y)}$ is not integrable over $\mathbb{R}^d \backslash \mathcal{C}$, we take the term $\int_{\mathbb{R}^d \backslash \mathcal{C}} e^{-\frac{1}{\delta} S(y) - f(y)} dy$ to be ∞ as the convention and the upper bound in equation (4) becomes trivial.

3. Note that Lemma 10.15 in Kallenberg (2002) requires the set $\mathcal{C}$ to be convex since it estimates both the outer ϵ-collar of $\mathcal{C}$, defined as the set of all points that do not belong to $\mathcal{C}$ but lie within distance at most ϵ from it, as well as the inner ϵ-collar of $\mathcal{C}$, whereas we only need to consider the outer ϵ-collar of $\mathcal{C}$ so that we can remove the convexity assumption on the set $\mathcal{C}$.

Lemma 2.5 *In the setting of Lemma 2.4, assume e^{-f} is integrable over $\mathcal{C}$, then for any $\delta, \tilde{\alpha} > 0$, we have*[4]

$$D(\pi\|\pi_\delta) \leq \left(\left(1+\frac{g^{-1}(\tilde{\alpha}\delta\log(1/\delta))}{r}\right)^d - 1\right)\frac{\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}R^d e^{-\inf_{y\in\mathbb{R}^d\setminus\mathcal{C}:S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y)}}{\int_{\mathcal{C}} e^{-f(y)}dy} + \delta^{\tilde{\alpha}}\frac{\int_{\mathbb{R}^d\setminus\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}, \tag{5}$$

where Γ denotes the gamma function.

In Lemma 2.5, we obtained an upper bound of the KL divergence between π and π_δ. In the literature of Langevin Monte Carlo, it is common to use the 2-Wasserstein distance to measure the convergence to the target distribution (Cheng et al., 2018; Dalalyan and Karagulyan, 2019). The celebrated W-CKP inequality (see Lemma B.1) bounds the 2-Wasserstein distance by the KL divergence of any two probability distributions where some exponential integrability condition is satisfied (see Lemma B.1), which in our case can be applied to control the 2-Wasserstein distance between π_δ and π. From Lemma 2.4, recall the function $\delta_{\mathcal{C}}(x) = \text{distance}(x,\mathcal{C}) := \min_{c\in\mathcal{C}}\|x-c\|$, for $x\in\mathbb{R}^d$. The convexity of the set $\mathcal{C}$ implies that the function $S(x) = (\delta_{\mathcal{C}}(x))^2$ satisfies some differentiability and smoothness properties, which is provided in the following lemma.

Lemma 2.6 *If $\mathcal{C}$ is convex, then the function $S(x) = (\delta_{\mathcal{C}}(x))^2$ is convex, ℓ-smooth with $\ell = 4$ and continuously differentiable on $\mathbb{R}^d$ with a gradient $\nabla S(x) = 2(x - \mathcal{P}_{\mathcal{C}}(x))$, where $\mathcal{P}_{\mathcal{C}}(x)$ is the projection of x to the set $\mathcal{C}$, i.e. $P_{\mathcal{C}}(x) := \arg\min_{c\in\mathcal{C}}\|x-c\|$.*

In the rest of the paper (except in Section 2.4), we always take penalty function $S(x) = (\delta_{\mathcal{C}}(x))^2$ unless otherwise specified. Building on Lemma 2.6, we have the following result, which quantifies the 2-Wasserstein distance between the target π and the modified target π_δ corresponding to the penalized target distribution.

Theorem 2.7 *Suppose Assumptions 2.1 and 2.2 hold. Moreover, we assume that f is continuous and e^{-f} is integrable over $\mathcal{C}$ and there exist some $\hat{\alpha} > 0$ and $\hat{x}\in\mathbb{R}^d$ such that and $\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} e^{-\frac{S(x)}{\delta}-f(x)}dx < \infty$. Then, as $\delta\to 0$,*

$$\mathcal{W}_2(\pi_\delta,\pi) \leq \mathcal{O}\left(\delta^{1/8}\left(\log(1/\delta)\right)^{1/8}\right). \tag{6}$$

Theorem 2.7 shows that by choosing δ small enough, we can approximate the compactly supported target distribution π with the modified target π_δ which has full support on $\mathbb{R}^d$. This amounts to converting the problem of constrained sampling to the problem of unconstrained sampling with a modified target. In the next remark, we discuss that if we take $\mathcal{C}$ to be the closed ball and $g(x) = x^2$, and apply the W-CKP inequality, we obtain the same bound in (6) except the logarithmic factor. This shows that it is not possible to improve our bound with an approach that relies on W-CKP inequality except for logarithmic factors.

4. If $\inf_{y\in\mathbb{R}^d\setminus\mathcal{C}:S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y) = -\infty$, we take the right hand side of equation (5) to be ∞ as the convention and the upper bound in equation (5) becomes trivial.

Remark 2.8 *In the setting of Theorem 2.7, consider the special case $\mathcal{C} = \{x : \|x\| \leq R\}$ to be the closed ball of radius R. In this case $S(x) = s(r)$, with $r = \|x\|$ and $s(r) = (r-R)^2 1_{r \geq R}$, where $s(r) = 0$ for any $r \leq R$ and $s(r) > 0$ for any $r > R$ and moreover s is differentiable and $s(r)$ is strictly increasing in $r > R$. Moreover, we assume that $f \geq 0$. Then, by Lemma 2.3 and using the spherical symmetry, we can compute that*

$$D(\pi \| \pi_\delta) \leq \frac{\int_{\mathbb{R}^d \setminus \mathcal{C}} e^{-\frac{1}{\delta} S(y)} dy}{\int_{\mathcal{C}} e^{-f(y)} dy} = \frac{\int_{\|y\| \geq R} e^{-\frac{1}{\delta} s(\|y\|)} dy}{\int_{\|y\| < R} e^{-f(y)} dy} = \frac{\int_{r \geq R} e^{-\frac{1}{\delta} s(r)} r^{d-1} dr}{\int_{\|y\| < R} e^{-f(y)} dy}. \tag{7}$$

Since $s(r)$, for $r \geq R$, achieves the unique minimum at $r = R$ and $s'(R) = 0$, we can apply Laplace's method (see e.g. Bleistein and Handelsman (2010)), and obtain

$$\int_{r \geq R} e^{-\frac{1}{\delta} s(r)} r^{d-1} dr = \sqrt{\frac{\pi}{2s''(R)}} R^{d-1} \sqrt{\delta} \cdot (1 + o(1)), \qquad \text{as } \delta \to 0. \tag{8}$$

Therefore, it follows from (7) and (8) that for any sufficiently small $\delta > 0$,

$$D(\pi \| \pi_\delta) \leq \left(\frac{\sqrt{\frac{2\pi}{s''(R)}} R^{d-1}}{\int_{\|y\| < R} e^{-f(y)} dy} \right) \sqrt{\delta}. \tag{9}$$

By applying W-CKP inequality (see Lemma B.1) and (9), we conclude $\mathcal{W}_2(\pi_\delta, \pi) \leq \mathcal{O}\left(\delta^{1/8}\right)$.

2.2 Penalized Langevin Algorithms with Deterministic Gradient

In this section, we are interested in penalized Langevin algorithms with deterministic gradient when f is non-convex. Raginsky et al. (2017) and Gao et al. (2022) developed non-asymptotic convergence bounds for SGLD and SGULMC, respectively, when f belongs to the class of non-convex smooth functions that are dissipative. This is a relatively general class of non-convex functions that admit critical points on a compact set. In our case, since $\mathcal{C}$ is assumed to be a compact convex set, we will not need growth conditions such as the dissipativity of f. The only assumption we are going to make about f is that f is smooth, i.e. the gradient of f is Lipschitz. We will show that the penalty function S is dissipative and smooth, so that $f + \frac{1}{\delta} S$ is dissipative and smooth for $\delta > 0$ small enough.

Assumption 2.9 *Assume that f is L-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, for any $x, y \in \mathbb{R}^d$.*

If Assumption 2.9 and Assumption 2.2 hold, then the conditions in Theorem 2.7 are satisfied (see Lemma C.4 in the Appendix for details). Building on this result, next we derive iteration complexity corresponding to the penalized Langevin dynamics.

2.2.1 Penalized Langevin Dynamics

First, we introduce the penalized overdamped Langevin SDE:

$$dX(t) = -\nabla f(X(t))dt - \frac{1}{\delta} \nabla S(X(t))dt + \sqrt{2} dW_t, \tag{10}$$

where W_t is a standard d-dimensional Brownian motion, and under mild conditions, it admits a unique stationary distribution $\pi_\delta(x) \propto \exp\left(-f(x) - \frac{1}{\delta}S(x)\right)$; see e.g. Hérau and Nier (2004); Pavliotis (2014). Consider the penalized Langevin dynamics (PLD):

$$x_{k+1} = x_k - \eta\left(\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\eta}\xi_{k+1}, \tag{11}$$

where ξ_k are i.i.d. $\mathcal{N}(0, I_d)$ Gaussian noises in $\mathbb{R}^d$.

In many applications, the constrained set $\mathcal{C}$ is defined by functional constraints, i.e. $\mathcal{C} := \{x : h_i(x) \leq 0, i = 1, 2, \ldots, m\}$, where h_i is a (merely) convex function defined on an open set that contains $\mathcal{C}$ and m is the number of constraints. For example, when $\mathcal{C}$ is an ℓ_p ball with radius R with $p \geq 1$ or when $\mathcal{C}$ is an ellipsoid. In this case, we can write

$$\mathcal{C} := \{x : h(x) \leq 0\}, \tag{12}$$

where $h(x) := \max_i h_i(x)$ is convex and therefore locally Lipschitz continuous (see e.g. Roberts and Varberg (1974)). The choice of the $h(x)$ function here is clearly not unique. In fact, such an $h(x)$ can be constructed even if we do not possess an explicit formula for the functions $h_i(x)$. More specifically, Minkowski functional $\|\cdot\|_K$, also known as the gauge function, is defined as $\|x\|_K := \inf\{t \geq 0, x \in t\mathcal{C}\}$ such that given that 0 is in the interior of $\mathcal{C}$, we can write $\mathcal{C} := \{x : h(x) \leq 0\}$ where $h(x) = \|x\|_K - 1$ (Rockafellar, 1970; Thompson, 1996). It is also well-known that the gauge function is merely convex. Thus, we can conclude that any convex body $\mathcal{C}$ admits the representation (12) where $h(x)$ is convex and finite-valued and therefore Lipschitz continuous on $\mathcal{C}$ (Roberts and Varberg, 1974). Equipped with this representation given by (12), we now consider a regularized constraint set

$$\mathcal{C}^\alpha = \{x : h^\alpha(x) \leq 0\}, \qquad \text{where} \quad h^\alpha(x) := h(x) + \frac{\alpha}{2}\|x\|^2, \tag{13}$$

is α-strongly convex for $\alpha > 0$ as $h(x)$ is merely convex, and it holds that $\mathcal{C}^\alpha \subseteq \mathcal{C} \subseteq \mathbb{R}^d$. We define the regularized distribution π^α supported on $\mathcal{C}^\alpha$ with probability density function

$$\pi^\alpha(x) \propto \exp(-f(x)), \qquad x \in \mathcal{C}^\alpha.$$

We also consider adding a penalty term $\frac{1}{\delta}S^\alpha(x)$ to the regularized target distribution π^α, and sample instead from the "penalized target distribution" with the regularized target density:

$$\pi^\alpha_\delta(x) \propto \exp\left(-f(x) - \frac{1}{\delta}S^\alpha(x)\right), \qquad x \in \mathbb{R}^d,$$

where $S^\alpha(x) = \left(\delta_{\mathcal{C}^\alpha}(x)\right)^2$ is the penalty function that satisfies $S^\alpha(x) = 0$ for any $x \in \mathcal{C}^\alpha$ and $S^\alpha(x) > 0$ otherwise. Our motivation for considering the penalty function $S^\alpha(x) = \left(\delta_{\mathcal{C}^\alpha}(x)\right)^2$ is that as we show in the Appendix, under some conditions, S^α is strongly convex outside a compact set (Lemma D.1); it can be seen that the function $S(x) = \left(\delta_{\mathcal{C}}(x)\right)^2$ does not always have this property.[5] Consequently, as a corollary, the function $f + \frac{1}{\delta}S^\alpha$ becomes

5. For example, when $\mathcal{C}$ is the unit ℓ_∞ ball in dimension 2, the function S is not strongly convex at any point $(0, y)$ for $y \in \mathbb{R}$.

strongly convex outside a compact set for δ small enough (Corollary D.2). Our main result in this section builds on exploiting this structure to develop stronger iteration complexity results for sampling π_δ^α compared to sampling π directly, and then controlling the error between π_δ^α and π by choosing δ and α small enough appropriately. For this purpose, first we estimate the size of the set difference $\mathcal{C}\backslash\mathcal{C}^\alpha$.

Lemma 2.10 *For the constrained set $\mathcal{C}^\alpha$ defined in* (13)*, we have*

$$\frac{|\mathcal{C}\backslash\mathcal{C}^\alpha|}{|\mathcal{C}^\alpha|} \leq \mathcal{O}(\alpha), \qquad \text{as } \alpha \to 0. \tag{14}$$

Second, we show that there exists a function U that is strongly convex everywhere and the difference between U and $f + S^\alpha/\delta$ can be uniformly bounded (Lemma D.3). Then by combining all these technical results (Lemma D.1 and Corollary D.2, Lemma D.3, Lemma 2.10) discussed above, and estimating the distance of π_δ^α to π, we obtain the following result.

Proposition 2.11 *Suppose Assumptions 2.1, 2.2, and 2.9 hold. Given the constraint set $\mathcal{C}$, consider its representation as $\mathcal{C} = \{x : h(x) \leq 0\}$ given in* (12) *where $h(x) = \max_{1\leq i\leq m} h_i(x)$ for some $m \geq 1$ with h_i convex for $i = 1, 2, \ldots, m$. Let ν_K be the distribution of the K-th iterate x_K of penalized Langevin dynamics* (11) *with the constrained set $\mathcal{C}^\alpha$ that is defined in* (13) *and the initialization $\nu_0 = \mathcal{N}(0, \frac{1}{L_\delta} I_d)$, where we take $\alpha = 0$ if h is strongly convex and we take $\alpha = \varepsilon^2$ is h is merely convex. Then, we have $TV(\nu_K, \pi) \leq \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^4$ and*

$$K = \tilde{\mathcal{O}}\left(d/\varepsilon^{10}\right), \qquad \eta = \mathcal{O}\left(\varepsilon^{10}/d\right),$$

where $\tilde{\mathcal{O}}$ ignores the dependence on $\log d$ and $\log(1/\varepsilon)$.

Remark 2.12 *In Proposition 2.11, when h is β-strongly convex (with $\beta > 0$ and $\alpha = 0$) the leading-order complexity $K = \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right)$ does not depend on β. It can be seen from from the proof of Proposition 2.11 that the complexity K has a second-order dependence on β, such that $K = \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right) + \tilde{\mathcal{O}}\left(\frac{d}{\beta\varepsilon^6}\right)$, where we ignored the dependence on the other constants when we consider the second-order dependence on β. When h is merely convex (with $\beta = 0$ and $\alpha = \varepsilon^2$), we have $K = \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right) + \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^8}\right)$.*

2.2.2 Penalized Underdamped Langevin Monte Carlo

We can also design sampling algorithms based on the underdamped (also known as second-order, inertial, or kinetic) Langevin diffusion given by the following SDE:

$$dV(t) = -\gamma V(t)dt - \nabla f(X(t))dt + \sqrt{2\gamma}dW_t, \tag{15}$$
$$dX(t) = V(t)dt, \tag{16}$$

(see e.g. Cheng et al. (2018); Cheng et al. (2018); Dalalyan and Riou-Durand (2020); Gao et al. (2022, 2020); Ma et al. (2021); Cao et al. (2023)) where $\gamma > 0$ is the friction coefficient, $X(t), V(t) \in \mathbb{R}^d$ model the position and the momentum of a particle moving in a field of force (described by the gradient of f) plus a random (thermal) force described by the Brownian noise W_t, which is a standard d-dimensional Brownian motion that starts

at zero at time zero. It is known that under some mild assumptions on f, the Markov process $(X(t), V(t))_{t\geq 0}$ is ergodic and admits a unique stationary distribution π with density $\pi(x,v) \propto \exp\left(-\left(\frac{1}{2}\|v\|^2 + f(x)\right)\right)$ (see e.g. Hérau and Nier (2004); Pavliotis (2014)). Hence, the x-marginal distribution of the stationary distribution with the density $\pi(x,v)$ is exactly the invariant distribution of the overdamped Langevin diffusion. For approximate sampling, various discretization schemes of (15)-(16) have been used in the literature (see e.g. Cheng et al. (2018); Teh et al. (2016); Chen et al. (2016, 2015)).

To design a constrained sampling algorithm based on the underdamped Langevin diffusion, we propose the "penalized underdamped Langevin SDE":

$$dV(t) = -\gamma V(t)dt - \nabla f(X(t))dt - \frac{1}{\delta}\nabla S(X(t))dt + \sqrt{2\gamma}dW_t, \tag{17}$$

$$dX(t) = V(t)dt, \tag{18}$$

where W_t is a standard d-dimensional Brownian motion. Under mild conditions, it admits a unique stationary distribution $\pi_\delta(x,v) \propto \exp\left(-f(x) - \frac{1}{\delta}S(x) - \frac{1}{2}\|v\|^2\right)$, whose x-marginal distribution is $\pi_\delta(x) \propto \exp\left(-f(x) - \frac{1}{\delta}S(x)\right)$, which coincides with the stationary distribution of the penalized overdamped Langevin SDE (10).

A natural way to sample the penalized target distribution $\pi_\delta(x,v)$ is to consider the Euler discretization of (17)-(18). We adopt a more refined discretization, introduced by Cheng et al. (2018). We propose the penalized underdamped Langevin Monte Carlo (PULMC):

$$v_{k+1} = \psi_0(\eta)v_k - \psi_1(\eta)\left(\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi_{k+1}, \tag{19}$$

$$x_{k+1} = x_k + \psi_1(\eta)v_k - \psi_2(\eta)\left(\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi'_{k+1}, \tag{20}$$

(see e.g. Dalalyan and Riou-Durand (2020)) where (ξ_k, ξ'_k) are i.i.d. $2d$-dimensional Gaussian noises and independent of the initial condition v_0, x_0, and for any fixed k, the random vectors $((\xi_k)_2, (\xi'_k)_2), ((\xi_k)_2, (\xi'_k)_2), \dots ((\xi_k)_d, (\xi'_k)_d)$ are i.i.d. with the covariance matrix

$$C(\eta) := \int_0^\eta [\psi_0(t), \psi_1(t)]^\top [\psi_0(t), \psi_1(t)]dt, \tag{21}$$

where

$$\psi_0(t) := e^{-\gamma t} \quad \text{and} \quad \psi_{k+1}(t) := \int_0^t \psi_k(s)ds \quad \text{for every } k \geq 0. \tag{22}$$

Dalalyan and Riou-Durand (2020) studied the unconstrained kinetic (underdamped) Langevin Monte Carlo algorithms (subject to deterministic gradients) for strongly log-concave and smooth densities, and Ma et al. (2021) investigated the case when f is strongly convex outside a compact domain. When f is non-convex, Gao et al. (2022) studied the unconstrained underdamped Langevin Monte Carlo algorithms (which allows stochastic gradients) under a dissipativity assumption. Since the x-marginal distribution of the Gibbs distribution of the penalized underdamped Langevin SDE (17)-(18) coincides that with the penalized overdamped Langevin SDE (10), we can bound $\mathcal{W}_2(\pi, \pi_\delta)$ using Theorem 2.7 with the same bounds as in the overdamped case.

Under some additional assumptions, we showed in Lemma D.1 that $f + S/\delta$ is strongly convex outside a compact domain, and thus one can leverage the non-asymptotic guarantees in Ma et al. (2021) for unconstrained underdamped Monte Carlo to obtain better performance guarantees for the penalized underdamped Langevin Monte Carlo. Before we proceed, we first provide a technical lemma that shows that under some additional assumptions on $\mathcal{C}$, S is Hessian Lipschitz.

Lemma 2.13 *Suppose $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex hypersurface of class C^3 and $\sup_{\xi\in\mathcal{C}} \|D^2 n(\xi)\|$ is bounded, where n is unit normal vector of $\mathcal{C}$. Then S is M_S-Hessian Lipschitz for some $M_S > 0$.*

As a corollary, if f is Hessian Lipschitz, then $f + S/\delta$ is Hessian Lipschitz and we immediately have the following result.

Corollary 2.14 *Under assumptions of Lemma 2.13 and assume that f is M_f-Hessian Lipschitz for some $M_f > 0$. Then $f + S/\delta$ is M_δ-Hessian Lipschitz, where $M_\delta := M_f + \frac{M_S}{\delta}$.*

Now, we are ready to state the following proposition that provides performance guarantees for the penalized underdamped Langevin Monte Carlo.

Proposition 2.15 *Suppose Assumptions 2.1, 2.2, and 2.9 hold, and also assume the conditions in Corollary 2.14 are satisfied. Given the constraint set $\mathcal{C}$, consider its representation as $\mathcal{C} = \{x : h(x) \le 0\}$ given in (12) where $h(x) = \max_{1\le i\le m} h_i(x)$ for some $m \ge 1$ with h_i convex for $i = 1, 2, \dots, m$. Let ν_K be the distribution of the K-th iterate x_K of penalized underdamped Langevin Monte Carlo (19)-(20) for the constrained set $\mathcal{C}^\alpha$ defined in (13) and the distribution of (v_0, x_0) follows $\mathcal{N}(0, \frac{1}{L_\delta} I_d) \otimes \mathcal{N}(0, \frac{1}{L_\delta} I_d)$, where we take $\alpha = 0$ if h is strongly convex and we take $\alpha = \varepsilon^2$ if h is merely convex. Then, we have $TV(\nu_K, \pi) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^4$, $\alpha = \varepsilon^2$ and*

$$K = \tilde{\mathcal{O}}\left(\sqrt{d}/\varepsilon^7\right),$$

where $\tilde{\mathcal{O}}$ ignores the dependence on $\log d$ and $\log(1/\varepsilon)$.

Remark 2.16 *When we compare the algorithmic complexity in Proposition 2.15 with Proposition 2.11, we see that for the underdamped-Langevin-based penalized underdamped Langevin Monte Carlo has complexity $K = \tilde{\mathcal{O}}(\sqrt{d}/\varepsilon^7)$, which improves the dependence on both the dimension d and the accuracy level ε compared to the overdamped-Langevin-based penalized Langevin dynamics where the complexity is $K = \tilde{\mathcal{O}}(d/\varepsilon^{10})$. This is obtained under additional assumptions on the smoothness of the boundary of $\mathcal{C}$ and Hessian Lipschitzness of f. To the best of our knowledge, $\tilde{\mathcal{O}}(\sqrt{d})$ is the best dependency on dimension for the constrained sampling.*

Remark 2.17 *In Proposition 2.15, when h is β-strongly convex (with $\beta > 0$ and $\alpha = 0$) the leading-order complexity $K = \tilde{\mathcal{O}}\left(\sqrt{d}/\varepsilon^7\right)$ does not depend on β. It can be seen from the proof of Proposition 2.15 that the complexity K has a second-order dependence on β, such that $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^7}\right) + \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\beta\varepsilon^3}\right)$, where we ignored the dependence on the other constants when we consider the second-order dependence on β. When h is merely convex (with $\beta = 0$ and $\alpha = \varepsilon^2$), we have $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^7}\right) + \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^5}\right)$.*

2.3 Penalized Langevin Algorithms with Stochastic Gradient

In the previous sections, we studied penalized Langevin algorithms with deterministic gradient when the objective f is non-convex. In this section, we study the extension to allow stochastic estimates of the gradients in our algorithms. Supporting stochastic gradients becomes especially key in machine learning and data science applications where the exact gradients can be computationally expensive but stochastic estimates can be obtained efficiently from data. We start with the case when f is assumed to be strongly convex and smooth.

2.3.1 Strongly Convex Case

In this section, we assume that the target f is strongly convex and its gradient is Lipschitz. More precisely, we make the following assumption.

Assumption 2.18 *Assume that f is μ-strongly convex and L-smooth.*

Assumption 2.18 is equivalent to assuming that the target density $\pi(x) \propto e^{-f(x)}$ is strongly log-concave and smooth. This assumption has also been made frequently in the literature (see, e.g., Bubeck et al. (2018, 2015); Lamperski (2021)). Such densities arise in several applications including but not limited to Bayesian linear regression and Bayesian logistic regression (see, e.g., Castillo et al. (2015); O'Brien and Dunson (2004)). Under Assumption 2.18, we have the following property for the target function f.

Lemma 2.19 *Under Assumptions 2.18 and Assumption 2.2, the minimizers of $f + \frac{S}{\delta}$ are uniformly bounded in δ such that there exists some $c \geq 0$ and the norm of any minimizer of $f + \frac{S}{\delta}$ is bounded by $(1 + c)R$.*

When δ is large, the minimizers of $f + \frac{S}{\delta}$ are close to the minimizers of f, which are uniformly bounded, and when δ is small, by the definition of the penalty function S, the minimizers of $f + \frac{S}{\delta}$ will concentrate on the set $\mathcal{C}$. Moreover, if the minimizers of f are inside the constrained set $\mathcal{C}$, then, the minimizers of $f + \frac{S}{\delta}$ must also lie in the set because $S(x) = (\delta_{\mathcal{C}}(x))^2 = 0$ for $x \in \mathcal{C}$. Hence, the above lemma naturally holds.

Moreover, under Assumptions 2.18 and 2.2, the conditions in Theorem 2.7 are satisfied (see Lemma C.5 in the Appendix for details). Building on this result, in the following subsections, we study penalized Langevin algorithms and the number of iterations needed to sample from a distribution within ε distance to the target.

Penalized Stochastic Gradient Langevin Dynamics. We now consider the extension to allow stochastic gradients, known as the stochastic gradient Langevin dynamics in the literature (see, e.g., Welling and Teh (2011); Chen et al. (2015); Raginsky et al. (2017)). In particular, we propose the penalized stochastic gradient Langevin dynamics (PSGLD):

$$x_{k+1} = x_k - \eta \left(\nabla \tilde{f}(x_k) + \frac{1}{\delta} \nabla S(x_k) \right) + \sqrt{2\eta} \xi_{k+1}, \tag{23}$$

where ξ_k are i.i.d. $\mathcal{N}(0, I_d)$ Gaussian noises in $\mathbb{R}^d$ and we assume that we have access to noisy estimates $\tilde{\nabla} f(x_k)$ of the actual gradients satisfying the following assumption:

Assumption 2.20 *We assume at iteration k, we have access to $\tilde{\nabla} f(x_k, w_k)$ which is a random estimate of $\nabla f(x_k)$ where w_k is a random variable independent from $\{w_j\}_{j=0}^{k-1}$ and satisfies $\mathbb{E}\left[\nabla \tilde{f}(x_k, w_k) - \nabla f(x_k) | x_k\right] = \nabla f(x_k)$ and*

$$\mathbb{E}\left\|\nabla \tilde{f}(x_k, w_k) - \nabla f(x_k)\Big| x_k\right\|^2 \leq 2\sigma^2\left(L^2\|x_k\|^2 + \|\nabla f(0)\|^2\right). \tag{24}$$

To simplify the notation, we suppress the w_k dependence and denote $\nabla \tilde{f}(x_k, w_k)$ by $\tilde{\nabla} f(x_k)$.

We note that the assumption (24) has been commonly made in data science and machine learning applications (see, e.g., Raginsky et al. (2017)) and arises when gradients are estimated from randomly sampled subsets of data points in the context of stochastic gradient methods. It is more general than the assumption $\mathbb{E}\left\|\nabla \tilde{f}(x_k) - \nabla f(x_k)\Big| x\right\|^2 \leq \sigma^2 d$ that has also been used in the literature (Dalalyan and Karagulyan, 2019) and allows handling gradient noise arising in many machine learning applications where the variance is not uniformly bounded (Raginsky et al., 2017; Aybat et al., 2019; Gürbüzbalaban et al., 2021). In (24), if $f(x)$ takes the form $f(x) = \sum_{i=1}^n f_i(x)$, and $\nabla \tilde{f}(x) = \frac{1}{b}\sum_{j\in\Omega} \nabla f_j(x)$, where Ω is a random subset of $\{1, 2, \ldots, n\}$ with batch-size b, due to the central limit theorem, we can assume that $\sigma^2 = \mathcal{O}(1/b)$, where b is the batch-size of the mini-batch. We have the following proposition, which characterizes the number of iterations necessary to sample from the target up to an ε error using the penalized stochastic gradient Langevin dynamics.

Proposition 2.21 *Suppose Assumptions 2.2, 2.18 and 2.20 hold. Let ν_K denote the distribution of the K-th iterate x_K of penalized stochastic gradient Langevin dynamics* (23)*. We have $\mathcal{W}_2(\nu_K, \pi) \leq \tilde{\mathcal{O}}(\varepsilon)$, where $\tilde{\mathcal{O}}$ ignores the dependence on $\log(1/\varepsilon)$, provided that $\delta = \varepsilon^8$, the batch-size b is of the constant order, and the stochastic gradient computations $\hat{K} := Kb$ and the stepsize η satisfy:*

$$\hat{K} = \tilde{\mathcal{O}}\left(\frac{d(L\varepsilon^8 + 4)^2}{\varepsilon^{18}\mu^3}\right), \qquad \eta = \frac{\varepsilon^{18}\mu^2}{d(L\varepsilon^8 + 4)^2}.$$

In terms of the dependence on the condition number $\kappa := L/\mu$, Proposition 2.21 implies that the batch-size b is of constant order, the stochastic gradient computations $\hat{K} = \tilde{\mathcal{O}}(\kappa^2/\mu)$ and the stepsize $\eta = \Theta(1/\kappa^2)$.

Penalized Stochastic Gradient Underdamped Langevin Monte Carlo. Next, we consider the extension to allow stochastic gradient, which we refer to as the stochastic gradient underdamped Langevin Monte Carlo (SGULMC). Such algorithms have been studied previously in the unconstrained setting in the literature (Chen et al., 2014, 2015; Gao et al., 2022). We propose the penalized stochastic gradient underdamped Langevin Monte Carlo (PSGULMC):

$$v_{k+1} = \psi_0(\eta)v_k - \psi_1(\eta)\left(\nabla \tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi_{k+1}, \tag{25}$$

$$x_{k+1} = x_k + \psi_1(\eta)v_k - \psi_2(\eta)\left(\nabla \tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi'_{k+1}, \tag{26}$$

where (ξ_k, ξ_k') are i.i.d. $2d$-dimensional Gaussian noises independent of the initial condition v_0, x_0, centered with covariance matrix given in (21) and $\psi_k(t)$ are defined in (22), where we recall that the gradient noise satisfies Assumption 2.20. Then we can provide the following proposition for the number of iterations we need to sample from the target distribution within ε error using PSGULMC with a stochastic gradient that satisfies Assumption 2.20.

Proposition 2.22 *Suppose Assumptions 2.2, 2.18 and 2.20 hold. Let ν_K denote the distribution of the K-th iterate x_K of penalized stochastic gradient underdamped Langevin Monte Carlo* (25)-(26) *and (v_0, x_0) follows the product distribution $\mathcal{N}(0, I_d) \otimes \nu_0$. We have $\mathcal{W}_2(\nu_K, \pi) \leq \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^8$, and the batch-size b satisfies:*

$$b = \Omega\left(\sigma^{-2}\right) = \Omega\left(\frac{L^2(L\varepsilon^8+4)(\varepsilon^{16}d\mu + (L\varepsilon^8+4)^2)}{\varepsilon^{26}\mu^4}\right),$$

and the stochastic gradient computations $\hat{K} := Kb$ and the stepsize η satisfy:

$$\hat{K} = \tilde{\mathcal{O}}\left(\frac{L^2(L\varepsilon^8+4)^2(\varepsilon^{16}d\mu + (L\varepsilon^8+4)^2)\sqrt{(\mu+L)\varepsilon^8+4}}{\varepsilon^{39}\mu^6} \max\left(\sqrt{d}, \frac{\sqrt{(L+\mu)\varepsilon^8+4}}{\varepsilon^3}\right)\right),$$

and

$$\eta = \min\left(\frac{1}{\sqrt{d}}\frac{\varepsilon^9\mu}{(L\varepsilon^8+4)}, \frac{1}{\sqrt{(\mu+L)\varepsilon^8+4}}\frac{\varepsilon^{12}\mu}{(L\varepsilon^8+4)}\right).$$

In terms of the dependence on the condition number $\kappa = L/\mu$, Proposition 2.22 implies that the batch-size $b = \Omega(L^5/\mu^4) = \Omega(L\kappa^4)$, the stochastic gradient computations $\hat{K} = \tilde{\mathcal{O}}(\kappa^6)$ and the stepsize $\eta = \Theta(1/(\sqrt{L}\kappa))$.

2.3.2 Non-Convex Case

This section discusses the case when f is non-convex and smooth.

Penalized Stochastic Gradient Langevin Dynamics. First, we consider the penalized stochastic gradient Langevin dynamics (PSGLD):

$$x_{k+1} = x_k - \eta\left(\nabla\tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\eta}\xi_{k+1}, \tag{27}$$

whereby following Raginsky et al. (2017) we assume that the initial distribution x_0 satisfies the exponential integrability condition

$$\kappa_0 := \log \mathbb{E}\left[e^{\|x_0\|^2}\right] < \infty, \tag{28}$$

and we recall that the gradient noise satisfies Assumption 2.20. For instance, we could take x_0 to be a Dirac measure or any distribution that is compactly supported. Similar to Proposition 2.21, we have the following proposition about the complexity analysis of the PSGLD for the non-convex case.

Proposition 2.23 *Suppose Assumptions 2.1, 2.2, 2.20 and 2.9 hold. Let ν_K be the distribution of the K-th iterate x_K of penalized stochastic gradient Langevin dynamics* (27)*. We have $\mathcal{W}_2(\nu_K, \pi) \leq \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^8$, the batch-size $b = \Omega(\eta^{-1})$ and the stochastic gradient computations $\hat{K} := Kb$ and the stepsize η satisfy:*

$$\hat{K} = \tilde{\mathcal{O}}\left(\frac{d^{17}\lambda_*^{-9}(\log(\lambda_*^{-1}))^8}{\varepsilon^{392}}\right), \qquad \eta = \tilde{\Theta}\left(\frac{\varepsilon^{196}}{d^8\lambda_*^{-4}(\log(\lambda_*^{-1}))^4}\right),$$

where $\tilde{\mathcal{O}}$ and $\tilde{\Theta}$ ignore the dependence on $\log d$ and $\log(1/\varepsilon)$, and λ_* is the spectral gap of the penalized overdamped Langevin SDE (10)[6]*:*

$$\lambda_* := \inf\left\{\frac{\int_{\mathbb{R}^d}\|\nabla g\|^2 d\pi_\delta}{\int_{\mathbb{R}^d} g^2 d\pi_\delta} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_\delta), g \neq 0, \int_{\mathbb{R}^d} g d\pi_\delta = 0\right\}. \tag{29}$$

Moreover, if we further assume that the assumptions of Corollary D.2 hold, then $\frac{1}{\lambda_*} \leq \mathcal{O}(1)$, and we have $\hat{K} = \tilde{\mathcal{O}}\left(\frac{d^{17}}{\varepsilon^{392}}\right)$ and $\eta = \tilde{\Theta}\left(\frac{\varepsilon^{196}}{d^8}\right)$.

Penalized Stochastic Gradient Underdamped Langevin Monte Carlo. Next, we consider the extension of underdamped Langevin Monte Carlo to allow stochastic gradient, which we refer to as the stochastic gradient underdamped Langevin Monte Carlo (SGULMC). Such algorithms have been studied in the unconstrained setting in the literature (Chen et al., 2014, 2015; Gao et al., 2022). We now consider the penalized stochastic gradient underdamped Langevin Monte Carlo (PSGULMC):

$$v_{k+1} = \psi_0(\eta)v_k - \psi_1(\eta)\left(\nabla\tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi_{k+1}, \tag{30}$$

$$x_{k+1} = x_k + \psi_1(\eta)v_k - \psi_2(\eta)\left(\nabla\tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi'_{k+1}, \tag{31}$$

where (ξ_k, ξ'_k) are i.i.d. $2d$-dimensional Gaussian noises independent of the initial condition v_0, x_0, centered with covariance matrix given in (21) and $\psi_k(t)$ are defined in (22), and finally, we recall that the gradient noise satisfies Assumption 2.20. We follow Gao et al. (2022) by assuming that the probability law μ_0 of the initial state (x_0, v_0) satisfies the following exponential integrability condition:

$$\int_{\mathbb{R}^{2d}} e^{\alpha\mathcal{V}(x,v)}\mu_0(dx, dv) < \infty, \tag{32}$$

where $\mathcal{V}$ is a Lyapunov function:

$$\mathcal{V}(x, v) := f(x) + \frac{S(x)}{\delta} + \frac{1}{4}\gamma^2\left(\left\|x + \gamma^{-1}v\right\|^2 + \left\|\gamma^{-1}v\right\|^2 - \lambda\|x\|^2\right), \tag{33}$$

and λ is a positive constant less than $\min(1/4, m_\delta/(L_\delta + \gamma^2/2))$, and $\alpha = \lambda(1 - 2\lambda)/12$, where we recall from Lemma C.2 that $f + \frac{S}{\delta}$ is (m_δ, b_δ)-dissipative where m_δ, b_δ are defined in (39). Notice that there exists a constant $A \in (0, \infty)$ so that

$$\left\langle x, \nabla f(x) + \frac{\nabla S(x)}{\delta}\right\rangle \geq m_\delta\|x\|^2 - b_\delta \geq 2\lambda\left(f(x) + \gamma^2\|x\|^2/4\right) - 2A.$$

6. This definition of the spectral gap can be found in Raginsky et al. (2017).

Indeed, Gao et al. (2020) showed that one can take

$$\lambda := \frac{1}{2}\min(1/4, m_\delta/(L_\delta + \gamma^2/2)), \tag{34}$$

$$A := \frac{m_\delta}{2L_\delta + \gamma^2}\left(\frac{\|\nabla f(0)\|^2}{2L_\delta + \gamma^2} + \frac{b_\delta}{m_\delta}\left(L_\delta + \frac{1}{2}\gamma^2\right) + f(0)\right), \tag{35}$$

where we recall from Lemma C.2 that $f + \frac{S}{\delta}$ is L_δ-smooth with $L_\delta = L + \frac{\ell}{\delta}$.

The Lyapunov function (33) is constructed in Eberle et al. (2019) as a key ingredient to show the convergence speed of the penalized underdamped Langevin SDE (17)-(18) to the Gibbs distribution $\pi_\delta(x, v) \propto \exp(-f(x) - \frac{1}{\delta}S(x) - \frac{1}{2}\|v\|^2)$. Eberle et al. (2019) shows that the convergence speed of (17)-(18) to the Gibbs distribution π_δ is governed by

$$\mu_* := \frac{\gamma}{768}\min\left\{\lambda L_\delta \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda} L_\delta \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda}\right\}, \tag{36}$$

where

$$\Lambda := \frac{12}{5}\left(1 + 2\alpha_1 + 2\alpha_1^2\right)(d + A)L_\delta \gamma^{-2}\lambda^{-1}(1 - 2\lambda)^{-1}, \qquad \alpha_1 := \left(1 + \Lambda^{-1}\right) L_\delta \gamma^{-2}.$$

The Lyapunov function (33) also plays a key role in Gao et al. (2022) that obtains the non-asymptotic convergence guarantees for (unconstrained) stochastic gradient underdamped Langevin Monte Carlo. We have the following proposition about the complexity of PS-GULMC with stochastic gradient that satisfies Assumption 2.20 for the non-convex case.

Proposition 2.24 *Suppose Assumptions 2.1, 2.2, 2.20 and 2.9 hold. Let ν_K be the distribution of the K-th iterate x_K of penalized stochastic gradient underdamped Langevin Monte Carlo* (30)-(31). *We have $\mathcal{W}_2(\nu_K, \pi) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^8$, the batch-size $b = \Omega(\eta^{-1})$ and the stochastic gradient computations $\hat{K} := Kb$ and the stepsize η satisfy:*

$$\hat{K} = \tilde{\mathcal{O}}\left(\frac{d^7\left(\log(1/\mu_*)\right)^5}{\varepsilon^{132}\mu_*^3}\right), \qquad \eta = \tilde{\Theta}\left(\frac{\varepsilon^{50}\mu_*}{d^3\left(\log(1/\mu_*)\right)^2}\right),$$

where $\tilde{\Theta}$ ignores the dependence on $\log d$ and $\log(1/\varepsilon)$.

In Proposition 2.24 (resp. Proposition 2.23), μ_* (resp. λ_*) governs the speed of convergence of the continuous-time penalized underdamped (resp. overdamped) Langevin SDEs to the Gibbs distribution. It is shown in Proposition 1 in Gao et al. (2022) that when the surface of the target is relatively flat, μ_* can be better than λ_* by a square root factor, i.e. $1/\mu_* = \mathcal{O}\left(\sqrt{1/\lambda_*}\right)$.

2.4 Avoiding Projections

We recall our discussion from Section 2.2.1 that the constraint set is often defined by functional inequalities of the form

$$\mathcal{C} := \{x : h_i(x) \le 0, \text{ for } i = 1, 2, \ldots, m\},$$

where $h_i(x) : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable for every i. This would, for instance, be the case if $\mathcal{C}$ is the ℓ_p-ball in $\mathbb{R}^d$ for $p \geq 1$. So far, our main complexity results involve the choice of $S(x) = (\delta_{\mathcal{C}}(x))^2$ as a penalty function where computing $S(x)$ requires calculating projection of x to the set $\mathcal{C}$. Computing such projections can be carried out in polynomial time, but it can be costly in some cases, for instance, when the number of constraints m is large or if the constraints are not simple. A natural question to ask is whether our results will hold if we use

$$S(x) = \sum\nolimits_{i=1}^{m} \max\left(0, h_i(x)\right)^2,$$

as a penalty function and sample from the modified target

$$\pi_\delta(x) \propto \exp\left(-f(x) - \frac{1}{\delta}\sum\nolimits_{i=1}^{m} \max\left(0, h_i(x)\right)^2\right), \qquad x \in \mathbb{R}^d, \tag{37}$$

instead. After all, this (alternative) choice of $S(x)$ would still satisfy our Assumption 2.1. In this section, we will show that this is indeed possible, provided that the functions $h_i(x)$ satisfy some growth conditions. The advantage of the formulation (37) is that the modified target does not require computing the projection and the distance function $\delta_{\mathcal{C}}(x)$ to the constraint set as before, and it allows directly working with the functions that define the constraint set. This is computationally more efficient when computing the projections to the constraint set is not straightforward. For example, if $h_i(x)$'s are affine (in which case the constraint set $\mathcal{C}$ is a polyhedral set as an intersection of half-planes) and the number of constraints m is large, computing the projection will be typically slower than evaluating the derivative of the penalized objective in (37).

We first show that when $h_i(x)$'s are differentiable and convex for every i, then the function S and therefore the density (37) is differentiable despite the presence of the non-smooth $\max(0, \cdot)$ part in (37). Under some further assumptions, we also show in the next result that S is ℓ-smooth with appropriate constants.

Lemma 2.25 *If $h_i(x)$ is differentiable and convex on $\mathbb{R}^d$ for every $i = 1, 2, \ldots, m$, then $\sum_{i=1}^m \max(0, h_i(x))^2$ is differentiable and convex on $\mathbb{R}^d$. Furthermore, assume that on the set $\mathcal{B}_i := \{x \in \mathbb{R}^d : \ h_i(x) \geq 0\}$, $h_i(x)$ satisfies the following three properties for every $i = 1, 2, \ldots, m$: (i) $h_i(x)$ is continuously twice differentiable, (ii) the gradient of $h_i(x)$ is bounded $\|\nabla h_i(x)\| \leq N_i$, (iii) the Hessian of $h_i(x)$ satisfies $|h_i(x)|\nabla^2 h_i(x) \preceq P_i I$, i.e., the large eigenvalue of the matrix $|h_i(x)|\nabla^2 h_i(x)$ is smaller than or equal to a non-negative scalar P_i. Then, $\sum_{i=1}^m \max(0, h_i(x))^2$ is ℓ-smooth, where $\ell := 2\sum_{i=1}^m \left(N_i^2 + P_i\right)$.*

The ℓ_p ball constraint arises in several applications that we will also discuss in the numerical experiments section (Section 3). Next, we show that the conditions in Lemma 2.25 can be satisfied for ℓ_p ball constraints.

Corollary 2.26 *If we choose $\mathcal{C} = \{x : h(x) \leq 0\}$ with $h(x) = \|x\|_p - R$ for a given $R > 0$ with $p \geq 2$, then $\max(0, h(x))^2$ is ℓ-smooth on $\mathbb{R}^d$, where $\ell := \left(\frac{2}{R} + (d-1)\right)(p-1)$.*

In the rest of this section, we will argue that our results can be extended to the penalty function $S(x) = \sum_{i=1}^m \max(0, h_i(x))^2$, when h_i satisfies certain growth conditions so that

projections required by the distance-based penalty functions can be avoided. First of all, by applying the same arguments as in Lemma 2.3, we can show that for any $\delta > 0$,

$$D(\pi\|\pi_\delta) \leq \frac{\int_{\mathbb{R}^d\backslash\mathcal{C}_1} e^{-\frac{1}{\delta}\sum_{i=1}^m \max(0,h_i(x))^2 - f(x)} dx}{\int_{\mathcal{C}_1} e^{-f(x)} dx},$$

where

$$\mathcal{C}_1 := \left\{x \in \mathbb{R}^d : \sum\nolimits_{i=1}^m \max(0, h_i(x))^2 \leq 0\right\} = \left\{x \in \mathbb{R}^d : \max\nolimits_{1\leq i\leq m} h_i(x) \leq 0\right\}.$$

Next, we provide an analog of Lemma 2.4 that upper bounds of the Lebesgue measure of the set of all points that are outside $\mathcal{C}_1$ yet in a small neighborhood of $\mathcal{C}_1$. Consider the constraint map $H : \mathbb{R}^d \to \mathbb{R}^m$ defined as $H(x) := [h_1(x), h_2(x), \ldots, h_m(x)]^\top$. We assume that H is *metrically subregular* everywhere on the boundary of the constrained set, i.e. we assume there exists some constant $\bar{K} > 0$ such that for any sufficiently small $\epsilon > 0$,

$$\left|x \in \mathbb{R}^d\backslash\mathcal{C}_1 : \sum\nolimits_{i=1}^m \max(0, h_i(x))^2 \leq \epsilon\right| \leq \left|x \in \mathbb{R}^d\backslash\mathcal{C}_1 : \delta_{\mathcal{C}_1}(x) \leq \sqrt{\bar{K}\epsilon}\right|,$$

(see, e.g., Ioffe (2016a,b) for more about metric subregularity and its consequences). For instance, the last inequality is satisfied when the constraint set is the ℓ_1 ball with radius R which is a a polyhedral set that can be expressed in the form (62) with affine choices of $h_i(x)$. Another example, would be the ℓ_p ball of radius R; i.e when $\mathcal{C} = \{x : h(x) \leq 0\}$ with $h(x) = \max_i h_i(x) = \|x\|_p - R$ and $p > 1$. By applying the same arguments as in Lemma 2.4, we conclude that there exists some constant $\bar{K} > 0$ such that for any sufficiently small $\epsilon > 0$,

$$\left|x \in \mathbb{R}^d\backslash\mathcal{C}_1 : \sum\nolimits_{i=1}^m \max(0, h_i(x))^2 \leq \epsilon\right| \leq \left(\left(1 + \sqrt{\bar{K}\epsilon}/r_1\right)^d - 1\right)|\mathcal{C}_1|,$$

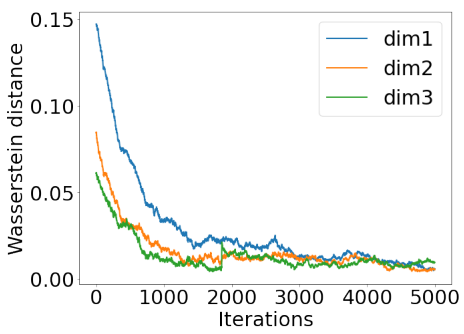
where we assumed that $\mathcal{C}_1$ contains an open ball of radius r_1 centered at 0. Furthermore, Lemma 2.5, Lemma 2.6 and Theorem 2.7 still apply with minor modifications and it follows that as $\delta \to 0$, $\mathcal{W}_2(\pi_\delta, \pi) \leq \mathcal{O}\left((\delta \log(1/\delta))^{1/8}\right)$, which is an analogue of Theorem 2.7. We can then obtain analogous results for PSGLD, and PSGULMC in Section 2.3. We can then utilize the conclusions of previous sections to get the convergence rate and complexity by using penalized Langevin and underdamped Langevin Monte Carlo algorithms in this setting.

3. Numerical Experiments

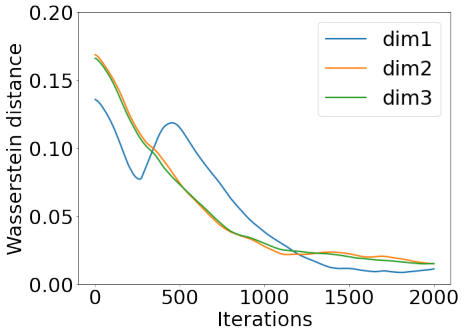
3.1 Synthetic Experiment for Dirichlet Posterior

As a toy experiment, we consider our proposed PLD and PULMC algorithms for sampling from a 3-dimensional Dirichlet posterior distribution. The Dirichlet distribution is commonly used in machine learning, especially in Latent Dirichlet allocation (LDA) problems; see, e.g., Blei et al. (2003). The Dirichlet distribution of dimension $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$ has a probability density function with respect to Lebesgue measure on $\mathbb{R}^{K-1}$ given by:

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1}, \tag{38}$$

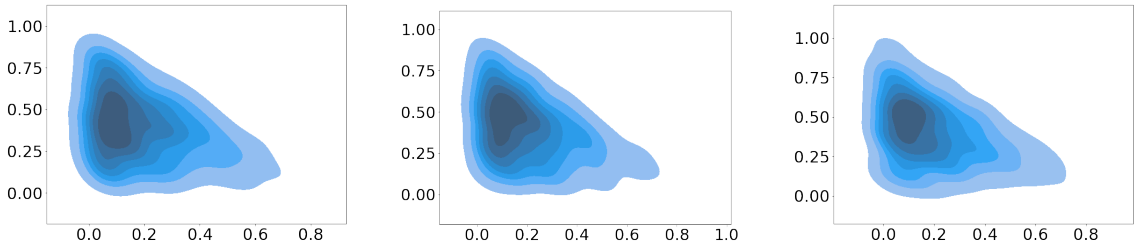


(a) Penalized LD (PLD)



(b) Penalized Underdamped Langevin Monte Carlo (PULMC)

Figure 1: Wasserstein distance between the target distribution and our proposed methods.

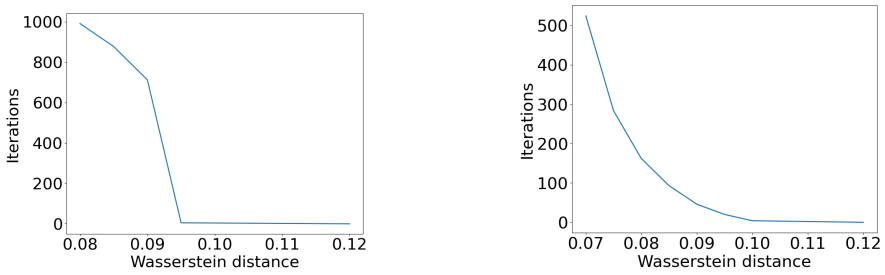


(a) True distribution (b) Penalized LD (PLD) (c) Penalized ULMC (PULMC)

Figure 2: Density plots of the target distribution and samples obtained by PLD and PULMC.

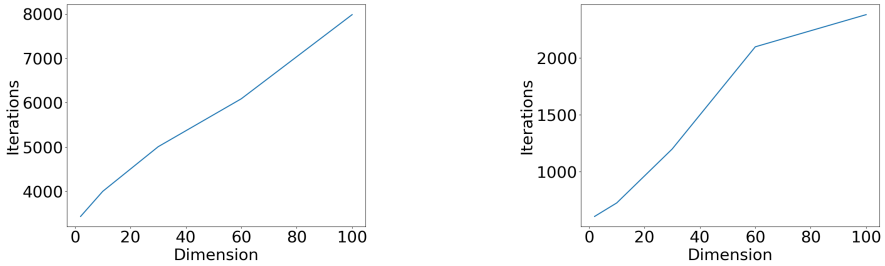
where $\{x_k\}_{k=1}^K$ belongs to the standard $K-1$ simplex, i.e. $\sum_{i=1}^K x_i = 1$ and $x_i \geq 0$ for all $i \in \{1, \ldots, K\}$, and the normalizing constant $B(\alpha)$ in equation (38) is the multivariate alpha function, which can be expressed as $B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$, for any $\alpha := (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_{\geq 0}^K$, where $\Gamma(\cdot)$ denotes the gamma function.

In our experiment, we set $\alpha = (1, 2, 2)$ and use uniform distribution on the simplex as the prior distribution. For PLD, we set $\delta = 0.005$ and learning rate $\eta = 0.0001$, and η is decreased by 25% every 1000 iterations. For PULMC, we set $\delta = 0.01$, $\gamma = 0.6$, and we take the learning rate $\eta = 0.0012$, where η is decreased by 10% every 200 iterations. We obtain 1000 samples from the posterior distribution using our methods and calculate the 2-Wasserstein distance for each of the three (coordinates) dimensions with respect to the true distribution based on 1000 runs. The results in Figure 1 illustrates the convergence of our methods where we observe that the 2-Wasserstein distance decays to zero in each dimension for both PLD and PULMC methods. In Figure 2, on the left panel we illustrate the target distribution whereas in the middle and right panels, we illustrate the density of the samples obtained by PLD and PULMC methods, based on 1000 samples. These



(a) PLD for Dirichlet sampling (b) PULMC for Dirichlet sampling

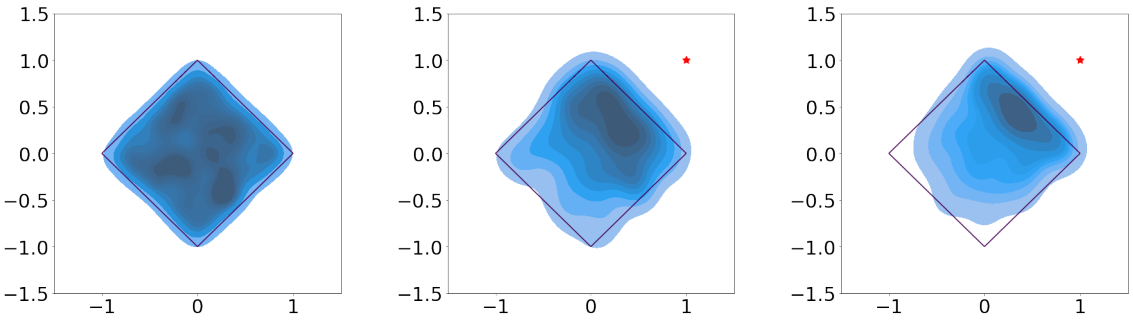
Figure 3: Average number of iterations required for achieving a target accuracy ε (measured in terms of the Wasserstein distance) for the Dirichlet sampling problem as ε is varied for PLD (left panel) and PULMC (right panel).



(a) PLD for Dirichlet sampling (b) PULMC for Dirichlet sampling

Figure 4: Dimension (d) dependency of PLD and PULMC on the Dirichlet distribution sampling problem.

figures illustrate that PLD and PULMC can sample successfully from the true Dirichlet distribution for this problem. In Figure 3, we also plot the (expected) average number of iterations k required for achieving an accuracy ε, i.e. for achieving $\mathcal{W}_2(\mathcal{L}(x_k), \pi) \leq \varepsilon$ where x_k are the iterates and π is the target Dirichlet distribution. In Figure 3, the x-axis is the accuracy ε whereas the y-axis is the number of iterations required. PULMC and PLD performs similarly, especially when the accuracy required is not small. It may be that both algorithms admit a better scaling in practice on this example with respect to ε than the worst-case theoretical bounds we provide in Table 1. In Figure 4, we also vary the dimension d while keeping the target accuracy ε fixed. More specifically, we report the (estimated) expected number of iterations needed to achieve the Wasserstein distance at most $\varepsilon = 0.25$. The parameter $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ of the Dirichlet distribution in dimension d is generated randomly, where α_i is set to a uniformly random integer from 1 to 5 independently for every $i = 1, 2, \ldots, d$. We tuned the parameters for both algorithms. In the PLD case, we use $\delta = 0.0004, \eta = 0.0003/d$. In the PULMC case, we use $\delta = 0.001, \eta = 0.0012/d, \gamma = 0.7$. We observe that the number of iterations required for PLD grows (approximately) linearly in the dimension d, whereas for PULMC we have roughly a sublinear growth in the dimension.



(a) Prior (b) Penalized SGLD (c) Penalized SGULMC

Figure 5: Prior and posterior distribution with 1-norm constraint in dimension 2.

The experimental results are more or less inline with our theoretical findings, where we prove that for the TV distance, PULMC admits better $(\mathcal{O}(\sqrt{d}))$ guarantees compared to $\mathcal{O}(d)$ guarantees of PLD.

3.2 Bayesian Constrained Linear Regression

We consider Bayesian constrained linear regression models in our next set of experiments. Such models have many applications in data science and machine learning (Brosse et al., 2017; Bubeck et al., 2018). For example, if the constraint set is an ℓ_p-ball around the origin, for $p = 1$, we obtain the Bayesian Lasso regression, and for $p = 2$, we get the Bayesian Ridge regression. We will consider both synthetic data and real-world data settings.

3.2.1 Synthetic 2-Dimensional Problem

In our first set of experiments, we will consider the case when $p = 1$, which corresponds to the Bayesian Lasso regression (Hans, 2009). For better visualization, we start with a synthetic 2-dimensional problem. We generate 10,000 data points (a_j, y_j) according to the model:

$$\delta_j \sim \mathcal{N}(0, 0.25), \quad a_j \sim \mathcal{N}(0, I), \quad y_j = {x_\star}^\top a_j + \delta_j, \quad x_\star = [1, 1]^\top.$$

We take the constraint set to be

$$\mathcal{C} = \{x : \|x\|_1 \leq 1\}.$$

The prior distribution is the uniform distribution, where the constraints are satisfied. This is illustrated in Figure 5(a). The posterior distribution of this model is given by

$$\pi(x) \propto e^{\sum_{j=1}^{n} -\frac{1}{2}(y_j - x^\top a_j)^2} \cdot \mathbb{1}_{\mathcal{C}},$$

where $\mathbb{1}_{\mathcal{C}}$ is the indicator function for the constraint set $\mathcal{C}$ and $n = 10,000$ is the number of data points. For this set of experiments, we take the batch size $b = 50$ and run PSGLD with $\delta = 0.001$, the learning rate $\eta = 10^{-5}$ where we reduce η by 15% every 5000 iterations. The total number of iterations is set to 50,000. For PSGULMC, we have a similar setting

with $\delta = 0.001$, $\gamma = 0.1$, and learning rate $\eta = 0.0001$, where we reduce η by 15% every 5000 iterations. The results are shown in Figure 5 where the point $x_\star$ is marked with a red asterisk. In Figure 5, we estimate the density of the samples obtained by both PSGLD and PSGULMC methods based on 500 runs. We see that the densities obtained by PSGLD and PSGULMC algorithms are compatible with the constraints, and they sample from a target distribution that puts higher weights into regions closer to $x_\star$ as expected (where without any constraints $x_\star$ would be the peak of the target).

We also consider an ellipsoidal constraint set

$$\mathcal{C} := \left\{ x \ : \ (x - \bar{a}_1)^\top Q_1 (x - \bar{a}_1) \leq \bar{b}_1 \right\},$$

for the same posterior distribution

$$\pi(x) \propto e^{\sum_{j=1}^n -\frac{1}{2}(y_j - x^\top a_j)^2} \cdot \mathbb{1}_{\mathcal{C}},$$

where $Q_1 \in \mathbb{R}^{2\times 2}$ is positive definite, $\bar{a}_1 \in \mathbb{R}^2$ is a real vector and $\bar{b}_1 > 0$ is a real scalar. We take $x_\star = [2, 2]^\top$ and

$$\bar{a}_1 = [1, 0]^\top, \quad \bar{b}_1 = 1, \quad Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

If we use the squared distances $S(x) = (\min_{y\in\mathcal{C}} \|x - y\|)^2$ as a penalty, then this will necessitate calculating projections to the ellipsoid constraint. However, we can avoid projections by following the methodology described in Section 2.4. Namely, we take

$$S(x) = \max\left(0, (x - \bar{a}_1)^\top Q_1 (x - \bar{a}_1) - \bar{b}_1\right).$$

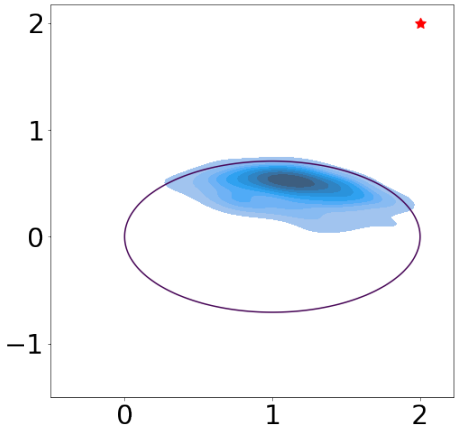
The ellipsoid constraint set and a contour plot of the densities obtained by PSGLD and PSGULMC algorithms are reported in Figure 6, where the lighter colors in the contour plots (including the white and light blue colors) correspond to regions with a smaller estimated density compared to darker blue regions. In these experiments, we tuned the parameters for each algorithm: For PSGLD, we set $\delta = 0.0001, \eta = 0.00005$, the number of iterations $k = 20,000$ and we reduce the stepsize η by 15% every $10,000$ iterations. For PSGULMC, we set $\delta = 0.001, \gamma = 0.1, \eta = 0.00001$, the number of iterations $k = 17,000$ where the stepsize η is reduced by 5% every $5,000$ iterations. We see that the densities lie within the constraints, and PSGLD and PSGULMC sample from a target distribution that puts higher weights into regions closer to $x_\star$ as expected.

We also considered another example where the aim is to sample from a Gaussian mixture

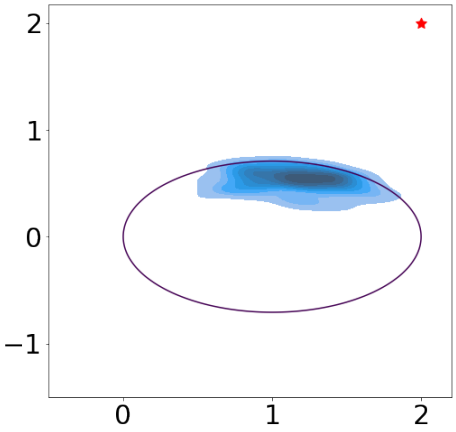
$$\pi(x) \propto \frac{2}{3}\exp(-\|x - z_1\|^2/2) + \frac{1}{3}\exp(-\|x - z_2\|^2/2),$$

where $z_1 = [2, 2]^\top, z_2 = [-2, -2]^\top$. We consider the non-convex constraint set $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$ obtained by intersecting the ellipsoids

$$\mathcal{C}_i := \left\{ x \ : \ (x - \bar{a}_i)^\top Q_i (x - \bar{a}_i) - \bar{b}_1 \leq 0 \right\}, \qquad \text{for } i = 1, 2.$$



(a) PSGLD for ellipsoid constraints



(b) PSGULMC for ellipsoid constraints

Figure 6: The density plot of the posterior distribution with ellipsoid constraints.

We take

$$\bar{a}_1 = [1,0]^\top, \quad \bar{a}_2 = [1,0]^\top, \quad \bar{b}_2 = 60, \quad \bar{b}_1 = 40, \quad Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix},$$

and consider the penalty

$$S(x) = \max\left(0, (x-\bar{a}_1)^\top Q_1 (x-\bar{a}_1) - \bar{b}_1\right) + \max\left(0, (x-\bar{a}_2)^\top Q_2 (x-\bar{a}_2) - \bar{b}_2\right).$$

For both PLD and PULMC, we take 10,000 iterations. The results are given in Figure 7 where we densities of the distributions that are outputs of PLD and PULMC algorithms are given as a contour plot and where the constraint set $\mathcal{C}$ is the intersection of the two ellipsoids displayed in the figure. We can see that the output distributions obtained PLD and PULMC are within both of the constraints, and the peaks of the two Gaussian distributions that are part of the mixture can be clearly observed in the figures.

3.2.2 Diabetes Dataset Experiment

Besides the synthetic dataset, we consider the Bayesian constrained linear regression on the Diabetes dataset.[7] Similar to Brosse et al. (2017), we take the constraint set to be

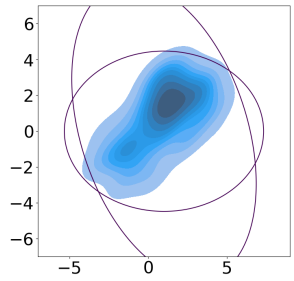
$$\mathcal{C} = \{x : \|x\|_1 \leq s\|x_{\text{OLS}}\|_1\},$$

where s is the shrinkage factor and x_{OLS} is the solution to the ordinary least squares problem without any constraints. The posterior distribution of this model is given by

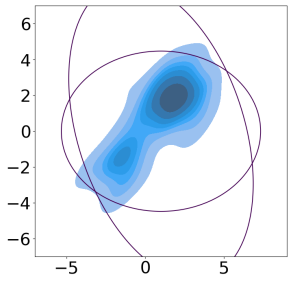
$$\pi(x) \propto e^{\sum_{j=1}^n -\frac{1}{2}(y_j - x^\top a_j)^2} \cdot \mathbb{1}_{\mathcal{C}},$$

where $\mathbb{1}_{\mathcal{C}}$ is the indicator function for the constraint set $\mathcal{C}$, and $(a_j, y_j), j = 1, 2, \ldots n$ are the data points in the Diabetes dataset. We experiment with different choices of s ranging from

7. This dataset is available online at `https://archive.ics.uci.edu/ml/datasets/diabetes`.



(a) PLD for sampling a Gaussian mixture subject to ellipsoid constraints



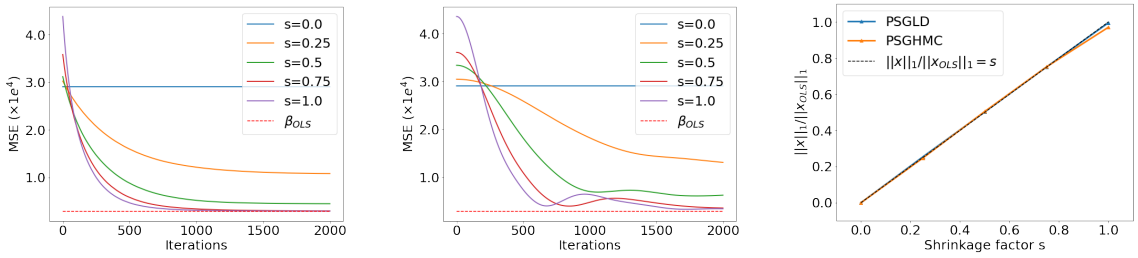
(b) PULMC for sampling a Gaussian mixture subject to ellipsoid constraints

Figure 7: The contour plot of the density for a Gaussian mixture with ellipsoid constraints.

0 to 1. For penalized SGLD, we set $\eta = s\|x_{\rm OLS}\| \times 10^{-5}, b = 50$, and $\delta = 0.05$, for penalized SGULMC, we set $\eta = s\|x_{\rm OLS}\| \times 10^{-5}, b = 50, \delta = 0.05$, and $\gamma = 0.6$. We take the prior distribution to be the uniform distribution on $\mathcal{C}$. We run our algorithms 100 times, and for the ℓ-th run, we let $x_k^{(\ell)}$ denote the k-th iterate of the ℓ-th run of our algorithms. First, we compute the mean squared error $\text{MSE}_k^{(\ell)} := \frac{1}{n}\sum_{j=1}^n (y_j - (x_k^{(\ell)})^\top a_j)^2$ corresponding to the k-th iterate of the ℓ-th run. In Figure 8(a) and 8(b), we report the average of the mean squared error values of each iteration, averaged over 100 runs, i.e. we plot $\text{MSE}_k := \frac{1}{100}\sum_{\ell=1}^{100} \text{MSE}_k^{(\ell)}$ over the iterations k. The results of averaged MSE over 100 samples are shown in Figure 8(a) and 8(b). We can observe from these figures that with s increasing from 0 to 1, the average mean squared error will decrease to the mean squared error of $x_{\rm OLS}$ for $p = 1$ as expected. As the number of iterations increases, the error of iterates decreases to a steady state. To illustrate that the final iterates of both algorithms are still lying in the constraint set $\mathcal{C}$, we plot the maximum values of $\|x_{\rm last}\|_1/\|x_{\rm OLS}\|_1$ calculated among 100 samples in Figure 8(c), where $x_{\rm last}$ is the last iterates of each sample from both algorithms, against the shrinkage factor s. The results from PSGLD and PSGULMC are shown as the blue and orange lines in the figure, where we plot the equation $\|x_{\rm last}\|_1/\|x_{\rm OLS}\|_1 = s$ as a dashed black line. We can observe that $\|x_{\rm last}\|_1/\|x_{\rm OLS}\|_1$ is always smaller than s for various s values. We illustrates that the final iterates for both PSGLD and PSGULMC stay in the constrained set $\mathcal{C}$ as expected. In Figure 9, we also plotted the (expected) average number of iterations required for achieving a target MSE value, as the target value is varied for both PSGLD and PSGULMC algorithms. Here, the dimension d is fixed and is determined by the Diabetes dataset. Although we do not provide theoretical guarantees for the MSE, comparing both algorithms in practice, PSGLD admits slightly better accuracy (measured in terms of MSE) on this example for the same number of iterations.

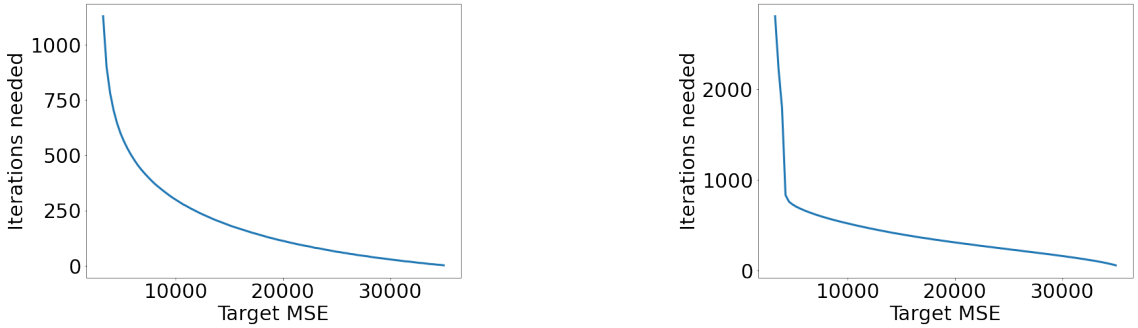
3.3 Bayesian Constrained Deep Learning

Non-Bayesian formulation of deep learning is based on minimizing the so-called *empirical risk* $f := \frac{1}{n}\sum_{i=1}^n f_i(x, z_i)$ where f_i is the loss function corresponding to the i-the data point based on the dataset $z = (z_1, z_2, \ldots, z_n)$ and has a particular structure as a composition



(a) Penalized SGLD (b) Penalized SGULMC (c) Norm of parameters

Figure 8: Penalized SGLD and Penalized SGULMC results for Diabetes dataset with 1-norm ball constraints in dimension 2.



(a) PSGLD for the Diabetes dataset (b) PSGULMC for the Diabetes dataset

Figure 9: Target MSE vs. average number of iterations needed to reach the target for the Diabetes dataset.

of non-linear but smooth functions when smooth activation functions (such as the sigmoid function or the ELU function) are used (Clevert et al., 2016). Furthermore, here x denotes the weights of the neural network and is a concatenation of vectors $x = \left[x^{(1)}, x^{(2)}, \ldots, x^{(I)}\right]$ (Hu et al., 2020) where $x^{(i)}$ are the (vectorized) weights of the i-th layer for $i = 1, 2, \ldots, I$ and I is the number of layers. We refer the reader to Deisenroth et al. (2020) for the details.

Constraining the weights x to lie on a compact set has been proposed in the deep learning practice for regularization purposes (Goodfellow et al., 2016). From the Bayesian sampling perspective, instead of minimizing the empirical risk function, we are interested in sampling from the posterior distribution $\pi(x) \propto e^{-f}$ (see, e.g., Gürbüzbalaban et al. (2021) for a similar approach) subject to constraints. We first consider the unconstrained setting where we run SGLD for 400 epochs and draw 20 samples from the posterior. We let $x_{\text{optimal}} = \left[x^{(1)}_{\text{optimal}}, x^{(2)}_{\text{optimal}}, \ldots, x^{(I)}_{\text{optimal}}\right]$ denote the average of the samples, which is an approximation to the solution of the unconstrained minimization problem. We consider the constraints $\|x^{(i)}\|_p \leq s\|x^{(i)}_{\text{optimal}}\|_p$ for the i-th layer in the network with $p = 1$. Since ℓ_1-norm promotes sparsity (Hastie et al., 2009), by adding these layer-wise constraints, we expect

to get a sparser model compared to the original model. Sparser models can be preferable as they would be more memory efficient, if they have similar prediction power (Srivastava et al., 2014).

Note that f will be smooth on the constraint set if smooth activation functions are used in which case our theory will apply. In our experiments, we use a four-layer fully connected network with hidden layer width $d = 400$ on the MNIST dataset.[8] The results are shown in Table 2 and Table 3, where the results are based on the average of 20 independent samples. We set the stepsize $\eta = 10^{-7}$ for PSGLD and SGLD methods, and we decay η by 10% every 100 epochs. We set penalty term $\delta = 0.1$ and report results after 350 epochs. For PSGULMC and SGULMC methods, we set $\gamma = 0.1$ and the stepsize $\eta = 5 \times 10^{-8}$, which decreases 10% every 100 epochs. We set penalty term $\delta = 100$ and report results after 400 epochs. For both PSGLD and PSGULMC, we use the batch size $b = 128$. In Table 2 we report the accuracy of the prediction in the training and test datasets, where we compared SGLD (without any constraints) to PSGLD algorithms with constraints defined by $s = 0.9$ and $s = 0.8$. We also report the maximum values of $\hat{s}$ among 20 samples, where $\hat{s}$ is defined as $\hat{s} := \frac{\|x\|_1}{\|x_{\text{optimal}}\|_1}$ and x are the parameters from the last iteration, after running the algorithms for 400 epochs. We can see from the results that for different s values, the value of $\hat{s}$ is always smaller than s, which indicates that the parameters of our algorithms satisfy the constraints. Table 3 reports similar results for PSGULMC. Basically, by enforcing ℓ_1 constraints, we can make the models sparser with a smaller ℓ_1 constraint at the cost of a relatively small decrease in training and test accuracy.

	training accuracy	**testing accuracy**	$\hat{s} := \frac{\|x\|_1}{\|x_{\text{optimal}}\|_1}$
SGLD	90.60%	89.95%	1
PSGLD (s=0.9)	89.37%	88.89%	0.8954
PSGLD (s=0.8)	87.35%	87.80%	0.7999

Table 2: Training and testing accuracy of fully-connected network with different constraints using PSGLD based on 20 samples.

	training accuracy	**testing accuracy**	$\hat{s} := \frac{\|x\|_1}{\|x_{\text{optimal}}\|_1}$
SGULMC	89.88%	90.22%	1
PSGULMC (s=0.9)	89.72%	89.49%	0.8918
PSGULMC (s=0.8)	87.28%	87.80%	0.7931

Table 3: Training and testing accuracy of fully-connected network with different constraints using PSGULMC based on 20 samples.

8. This dataset is available online at `http://yann.lecun.com/exdb/mnist/`.

4. Conclusion

In this paper, we considered the problem of constrained sampling where the goal is to sample from a target distribution $\pi(x) \propto e^{-f(x)}$ when x is constrained to lie on a convex body $\mathcal{C}$. We proposed and studied penalty-based overdamped Langevin and underdamped Langevin Monte Carlo (ULMC) methods. We considered targets where f is smooth and strongly convex as well as the more general case where f can be non-convex. In both cases, under some assumptions, we characterized the number of iterations and samples required to sample the target up to an ε-error while the error is measured in terms of the 2-Wasserstein or the total variation distance. Our methods improve upon the dimension dependency of the existing approaches in a number of settings and to our knowledge provides the first convergence results for ULMC-based methods for non-convex f in the context of constrained sampling. Our methods can also handle unbiased stochastic noise on the gradients that arise in machine learning applications. Finally, we illustrated the efficiency of our methods on the Bayesian Lasso linear regression and Bayesian deep learning problems.

Acknowledgements

The authors thank the acting editor and two anonymous referees for helpful comments and suggestions. The authors also thank Sam Ballas and Andrzej Ruszczyński for helpful discussions. Mert Gürbüzbalaban and Yuanhan Hu's research is partly supported by the grants Office of Naval Research Award Numbers N00014-21-1-2244 and N00014-24-1-2628, National Science Foundation (NSF) CCF-1814888, NSF DMS-1723085, NSF DMS-2053485. Lingjiong Zhu is partially supported by the grants NSF DMS-2053454, NSF DMS-2208303, and a Simons Foundation Collaboration Grant.

References

K. Ahn and S. Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

M. Assran and M. Rabbat. On the convergence of Nesterov's accelerated gradient method in stochastic settings. In *Proceedings of the 37th International Conference on Machine Learning*, pages 410–420. PMLR, 2020.

N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, Cham, 2014.

M. V. Balashov and M. O. Golubev. About the Lipschitz property of the metric projection in the Hilbert space. *Journal of Mathematical Analysis and Applications*, 394(2):545–551, 2012.

K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and S. Zhang. Towards a theory of non-log-concave sampling: First-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

R. Bardenet, A. Doucet, and C. C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

M. Barkhagen, N. Chau, E. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

N. Bleistein and R. A. Handelsman. *Asymptotic Expansions of Integrals*. Dover, New York, 2010.

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales-Faculté des sciences Toulouse Mathematiques*, 14 (3):331–352, 2005.

J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer, New York, 2nd edition, 2005.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press, 2011.

N. Brosse, A. Durmus, E. Mouliness, and M. Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 319–342. PMLR, 2017.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

S. Bubeck, R. Eldan, and J. Lehec. Finite-time analysis of projected Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Y. Cao, J. Lu, and L. Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(90):1–34, 2023.

I. Castillo, J. Schmidt-Hieber, and A. Van der Vaart. Bayesian linear regression with sparse priors. *Annals of Statistics*, 43(5):1986–2018, 2015.

A. Chalkis, V. Fisikpoulos, M. Papachristou, and E. Tsigaridas. Truncated log-concave sampling for convex bodies with Reflective Hamiltonian Monte Carlo. *ACM Transactions on Mathematical Software*, 49(2):1–25, 2023.

N. H. Chau, E. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2278–2286, 2015.

C. Chen, D. Carlson, Z. Gan, C. Li, and L. Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1051–1060, 2016.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

Y. Chen, S. Chewi, A. Salim, and A. Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, volume 178, pages 2984–3014. PMLR, 2022.

Z. Chen and S. S. Vempala. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *Theory of Computing*, 18(1):1–18, 2022.

X. Cheng and P. L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, pages 186–211, 2018.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv:1805.01648*, 2018.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory*, pages 300–323. PMLR, 2018.

S. Chewi, T. Le Gouic, C. Lu, T. Maunu, P. Rigollet, and A. Stromme. Exponential ergodicity of mirror-Langevin diffusions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations*, 2016.

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

A. S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, volume 65, pages 678–689. PMLR, 2017b.

A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov Chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, 47(4):1982–2010, 2019.

K. Fan. Note on circular disks containing the eigenvalues of a matrix. *Duke Mathematical Journal*, 25(3):441–445, 1958.

H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93 (3):418–491, 1959.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Global convergence of Stochastic Gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022.

K. Gatmiry and S. S. Vempala. Convergence of the Riemannian Langevin algorithm. *arXiv:2204.10818*, 2022.

S. B. Gelfand and S. K. Mitter. Simulated annealing type algorithms for multivariate optimization. *Algorithmica*, 6(1):419–436, 1991.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, 1995.

C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (2):123–214, 2011.

I. Goodfellow, Y. Bengio, and A. Courville. Regularization for deep learning. In *Deep Learning*. MIT Press, 2016.

M. Gürbüzbalaban, X. Gao, Y. Hu, and L. Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22:1–69, 2021.

M. Gürbüzbalaban, A. Ruszczyński, and L. Zhu. A stochastic subgradient method for distributionally robust non-convex and non-smooth learning. *Journal of Optimization Theory and Applications*, 194(3):1014–1041, 2022.

C. Hans. Bayesian Lasso regression. *Biometrika*, 96(4):835–845, 2009.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

F. Hérau and F. Nier. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Archive for Rational Mechanics and Analysis*, 171(2):151–218, 2004.

R. A. Holley, S. Kusuoka, and D. W. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

R. A. Holley, S. Kusuoka, and D. W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *Journal of Functional Analysis*, 83(2): 333–347, 1989.

Y.-P. Hsieh, A. Kavis, P. Rolland, and V. Cevher. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Y. Hu, X. Wang, X. Gao, M. Gürbüzbalaban, and L. Zhu. Non-convex stochastic optimization via nonreversible stochastic gradient Langevin dynamics. *arXiv:2004.02823*, 2020.

A. D. Ioffe. Metric regularity–a survey part 1. theory. *Journal of the Australian Mathematical Society*, 101:188–243, 2016a.

A. D. Ioffe. Metric regularity—a survey part ii. applications. *Journal of the Australian Mathematical Society*, 101(3):376–417, 2016b.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory*, pages 545–604. PMLR, 2018.

Q. Jiang. Mirror Langevin Monte Carlo: the case under isoperimetry. In *Advances in Neural Information Processing Systems*, volume 34, pages 715–725, 2021.

O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, 2nd edition, 2002.

A. Karagulyan and A. Dalalyan. Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Y. Kook, Y. T. Lee, R. Shen, and S. S. Vempala. Sampling with Riemannian Hamiltonian Monte Carlo in a constrained space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

A. Lamperski. Projected stochastic gradient Langevin algorithms for constrained sampling and non-convex learning. In *Proceedings of The 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1–47. PMLR, 2021.

S. Lan and B. Shahbaba. Sampling constrained probability distributions using spherical augmentation. In H. Q. Minh and V. Murino, editors, *Algorithmic Advances in Riemannian Geometry and Applications: For Machine Learning, Computer Vision, Statistics, and Optimization*, pages 25–71. Springer International Publishing, Cham, 2016.

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

G. Leobacher and A. Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of Global Analysis and Geometry*, 60(3):559–587, 2021.

R. Li, M. Tao, S. S. Vempala, and A. Wibisono. The mirror Langevin algorithm converges with vanishing bias. In S. Dasgupta and N. Haghtalab, editors, *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167, pages 718–742. PMLR, 2022a.

R. L. Li, H. Zha, and M. Tao. Sqrt(d) dimension dependence of Langevin Monte Carlo. In *Internatonal Conference on Learning Representations*, 2022b.

X. Li, D. Wu, L. Mackey, and M. A. Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

X. Luo, X. Chang, and X. Ban. Regression and classification using extreme learning machine based on L_1-norm and L_2-norm. *Neurocomputing*, 174(Part A):179–186, 2016.

R. Ma, J. Miao, L. Niu, and P. Zhang. Transformed ℓ_1 regularization for learning sparse deep neural networks. *Neural Networks*, 119:286–298, 2019a.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(24):20881–20885, 2019b.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942–1992, 2021.

O. Mangoubi and A. Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions: Continuous dynamics. *Annals of Applied Probability*, 31(5):2019 – 2045, 2021.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.

S. M. O'Brien and D. B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60 (3):739–746, 2004.

S. Patterson and Y. W. Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 3102–3110, 2013.

G. A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, volume 65, pages 1674–1703. PMLR, 2017.

A. W. Roberts and D. E. Varberg. Another proof that convex functions are locally Lipschitz. *The American Mathematical Monthly*, 81(9):1014–1016, 1974.

R. T. Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1970.

P. Rolland, A. Eftekhari, K. Ali, and V. Cevher. Double-loop Unadjusted Langevin Algorithm. In *International Conference on Machine Learning*, volume 119, pages 8169–8177. PMLR, 2020.

A. Salim and P. Richtárik. Primal dual interpretation of the proximal stochastic gradient Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

K. Sato, A. Takeda, R. Kawai, and T. Suzuki. Convergence error analysis of reflected gradient Langevin dynamics for globally optimizing non-convex constrained problems. *arXiv preprint arXiv:2203.10215*, 2022.

M. Schmidt. Least squares optimization with L1-norm regularization. *CS542B Project Report*, 504:195–221, 2005.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(1):193–225, 2016.

A. C. Thompson. *Minkowski Geometry*. Cambridge University Press, 1996.

J.-P. Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9 (1-2):187–205, 1982.

C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

X. Wang, Q. Lei, and I. Panageas. Fast convergence of Langevin dynamics on manifold: Geodesics meet log-Sobolev. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688. PMLR, 2011.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 3122–3133, 2018.

K. S. Zhang, G. Peyré, J. Fadili, and M. Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Conference on Learning Theory*, volume 125, pages 3814–3841. PMLR, 2020.

Y. Zhang, O. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *Applied Mathematics and Optimization*, 87(25):1–41, 2023.

Y. Zheng and A. Lamperski. Constrained Langevin algorithms with L-mixing external random variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

D. Zou and Q. Gu. On the convergence of Hamiltonian Monte Carlo with stochastic gradients. In *International Conference on Machine Learning*, volume 139, pages 13012–13022. PMLR, 2021.

D. Zou, P. Xu, and Q. Gu. Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, volume 161, pages 1152–1162. PMLR, 2021.

Appendix A. Notations

A function $f:\mathbb{R}^d\to\mathbb{R}$ is said to be μ-strongly convex if there exists $\mu>0$ such that for any $x,y\in\mathbb{R}^d$,

$$f(x)-f(y)-g^\top(x-y)\geq\frac{\mu}{2}\|x-y\|^2,\qquad\text{for all } g\in\partial f(y),$$

where ∂f denotes the subdifferential. If f is differentiable at y, then $\partial f(y)=\{\nabla f(y)\}$ is a singleton set. If the latter inequality holds for $\mu=0$, we say f is merely convex (see e.g. Nesterov (2013)).

The function $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth if for any $x,y\in\mathbb{R}^d$, the gradients $\nabla f(x),\nabla f(y)$ exist and satisfy $\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|$. If f is both μ-strongly convex and L-smooth, it holds that (see e.g. Bubeck (2015)):

$$\frac{\mu}{2}\|x-y\|^2\leq f(x)-f(y)-\nabla f(y)^\top(x-y)\leq\frac{L}{2}\|x-y\|^2,\qquad\text{for any } x,y\in\mathbb{R}^d.$$

We say that a function $f:\mathbb{R}^d\to\mathbb{R}$ is (m,b)-dissipative if for some $m,b>0$, $\langle\nabla f(x),x\rangle\geq m\|x\|^2-b$, for any $x\in\mathbb{R}^d$.

For any $x,y\in\mathbb{R}$, $x\vee y$ denotes $\max(x,y)$ and $x\wedge y$ denotes $\min(x,y)$. For any $x=(x_1,\ldots,x_d)\in\mathbb{R}^d$, its ℓ_p-norm (also referred to as p-norm) is denoted by $\|x\|_p:=\left(\sum_{i=1}^d|x_i|^p\right)^{1/p}$. For any measurable set $\mathcal{A}\subset\mathbb{R}^d$, we use $|\mathcal{A}|$ to denote the Lebesgue measure of $\mathcal{A}$. For a set $\mathcal{A}$, the indicator function $1_{\mathcal{A}}(y)=1$ for $y\in\mathcal{A}$ and $1_{\mathcal{A}}(y)=0$ otherwise. We denote $\mathbb{R}_{\geq 0}$ the set of non-negative real scalars.

A subset $\mathcal{C}$ of $\mathbb{R}^d$ is called a hypersurface of class C^k, if for every $x_0\in\mathcal{C}$ there is an open set $V\subset\mathbb{R}^d$ containing x_0 and a real-valued function $\phi\in C^k(V)$ such that $\nabla\phi$ is non-vanishing on $S\cap V=\{x\in V:\phi(x)=0\}$, where $C^k(V)$ is the set of functions defined on V with k continuous derivatives. We denote $Dn(\xi)$ and $D^2n(\xi)$ as the first and second-order derivatives of unit normal vector n in the sense of Leobacher and Steinicke (2021).

Next, we introduce three standard notions often used to quantify the distances between two probability measures. For a survey on distances between two probability measures, we refer to Gibbs and Su (2002).

Wasserstein metric. For any $p\geq 1$, define $\mathcal{P}_p(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^d$ with the finite p-th moment (based on the Euclidean norm). For any two Borel probability measures $\nu_1,\nu_2\in\mathcal{P}_p(\mathbb{R}^d)$, we define the standard p-Wasserstein metric (Villani, 2009): $\mathcal{W}_p(\nu_1,\nu_2):=\left(\inf\mathbb{E}\left[\|Z_1-Z_2\|^p\right]\right)^{1/p}$, where the infimum is taken over all joint distributions of the random variables Z_1,Z_2 with marginal distributions ν_1,ν_2.

Kullback-Leibler (KL) divergence. KL divergence, also known as relative entropy, between two probability measures μ and ν on $\mathbb{R}^d$, where μ is absolutely continuous with respect to ν, is defined as: $D(\mu\|\nu):=\int_{\mathbb{R}^d}\frac{d\mu}{d\nu}\log\left(\frac{d\mu}{d\nu}\right)d\nu$.

Total variation distance. The total variation (TV) distance between two probability measures P and Q on a sigma-algebra $\mathcal{F}$ is defined as $\sup_{A\in\mathcal{F}}|P(A)-Q(A)|$.

Appendix B. Weighted Csiszár-Kullback-Pinsker Inequality

The KL divergence can bound the Wasserstein distances on $\mathbb{R}^d$ under some technical conditions, known as the weighted Csiszár-Kullback-Pinsker (W-CKP) inequality.

Lemma B.1 (page 337 in Bolley and Villani (2005)) *For any two probability measures μ and ν on $\mathbb{R}^d$, we have*

$$\mathcal{W}_2(\mu,\nu) \leq \hat{C}\left(D(\mu\|\nu)^{\frac{1}{2}} + \left(\frac{D(\mu\|\nu)}{2}\right)^{\frac{1}{4}}\right),$$

where $\hat{C} := 2\inf_{\hat{x}\in\mathbb{R}^d,\hat{\alpha}>0}\left(\frac{1}{\hat{\alpha}}\left(\frac{3}{2}+\log\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2}d\nu(x)\right)\right)^{\frac{1}{2}}$, provided that there exists some $\hat{\alpha}>0$ and $\hat{x}\in\mathbb{R}^d$ such that $\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2}d\nu(x)<\infty$.

Appendix C. Technical Lemmas

In this section, we provide some technical lemmas that are used in the proofs of the main results. The proofs of these technical lemmas will be provided in Appendix D.

Lemma C.1 *If Assumption 2.2 holds, then the penalty function $S(x) = (\delta_{\mathcal{C}}(x))^2$ is continuously differentiable, ℓ-smooth with $\ell = 4$ and (m_S, b_S)-dissipative with $m_S = 1, b_S = R^2/4$, i.e. $\langle x, \nabla S(x)\rangle \geq m_S\|x\|^2 - b_S$, for any $x\in\mathbb{R}^d$.*

Lemma C.2 *If Assumption 2.9 and Assumption 2.2 hold, then $f+\frac{1}{\delta}S$ is L_δ-smooth, with $L_\delta := L + \frac{\ell}{\delta}$ and moreover $f+\frac{1}{\delta}S$ is (m_δ, b_δ)-dissipative with*

$$m_\delta := -L - \frac{1}{2} + \frac{m_S}{\delta} > 0, \qquad b_\delta := \frac{1}{2}\|\nabla f(0)\|^2 + \frac{b_S}{\delta}, \tag{39}$$

provided that $\delta < m_S/(L+\frac{1}{2})$, where m_S, b_S are defined in Lemma C.1.

Lemma C.3 *Under Assumption 2.9 and Assumption 2.2, then $f+\frac{S}{\delta}$ is lower bounded for $S(x) = (\delta_{\mathcal{C}}(x))^2$, i.e. there exists a real non-negative scalar M such that $f(x)+\frac{S(x)}{\delta} \geq -M$ for any $x\in\mathbb{R}^d$, where we can take*

$$M := -f(0) + \frac{1}{2}\|\nabla f(0)\|^2 + \frac{b_S}{2\delta}\log 3, \tag{40}$$

provided that $\delta \leq \frac{2m_S}{3(1+L)}$ where m_S, b_S are defined in Lemma C.1.

Lemma C.4 *If Assumption 2.9 and Assumption 2.2 hold, then the conditions in Theorem 2.7 are satisfied with $\hat{\alpha} = \frac{m_\delta}{6}$ and $\hat{x} = 0$, where m_δ is defined in (39).*

Lemma C.5 *If Assumptions 2.18 and 2.2 hold, then the assumptions in Theorem 2.7 are satisfied with $\hat{\alpha} = \frac{\mu}{4}$ and $\hat{x} = x_*$, where x_* is the unique minimizer of f.*

Appendix D. Technical Proofs

In this section, we provide technical proofs of the main results in our paper.

Proof of Lemma 2.3

Note that π is supported on $\mathcal{C}$ whereas π_δ is supported on $\mathbb{R}^d$, and π is absolutely continuous with respect to π_δ. We can compute that the KL divergence between π and π_δ is given by

$$\begin{aligned}
&D(\pi\|\pi_\delta)\\
&= \int_{\mathbb{R}^d} \log\left(\frac{\pi(x)}{\pi_\delta(x)}\right)\pi(x)dx = \int_{\mathcal{C}} \log\left(e^{\frac{1}{\delta}S(x)}\frac{\int_{\mathbb{R}^d} e^{-f(y)-\frac{1}{\delta}S(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}\right)\frac{e^{-f(x)}}{\int_{\mathcal{C}} e^{-f(y)}dy}dx \quad (41)\\
&= \int_{\mathcal{C}} \log\left(\frac{\int_{\mathbb{R}^d} e^{-f(y)-\frac{1}{\delta}S(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}\right)\frac{e^{-f(x)}}{\int_{\mathcal{C}} e^{-f(y)}dy}dx, \quad (42)\\
&= \log\left(\frac{\int_{\mathbb{R}^d} e^{-f(y)-\frac{1}{\delta}S(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}\right), \quad (43)
\end{aligned}$$

where we used the definition of π and π_δ to obtain (41) and the fact that $S(x) = 0$ for any $x \in \mathcal{C}$ to obtain (42). We can further compute from (43) that

$$\begin{aligned}
D(\pi\|\pi_\delta) &= \log\left(\frac{\int_{\mathcal{C}} e^{-f(y)-\frac{1}{\delta}S(y)}dy + \int_{\mathbb{R}^d\setminus\mathcal{C}} e^{-f(y)-\frac{1}{\delta}S(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}\right)\\
&= \log\left(1 + \frac{\int_{\mathbb{R}^d\setminus\mathcal{C}} e^{-f(y)-\frac{1}{\delta}S(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}\right) \le \frac{\int_{\mathbb{R}^d\setminus\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}, \quad (44)
\end{aligned}$$

where we used the fact that $S(y) = 0$ for any $y \in \mathcal{C}$ to obtain the equality in (44) and $\log(1+x) \le x$ for any $x \ge 0$ to obtain the inequality in (44). This completes the proof. □

Proof of Lemma 2.4

By the definitions of $S(y)$ and g, we have

$$\left|y \in \mathbb{R}^d\setminus\mathcal{C} : S(y) \le \epsilon\right| = \left|y \in \mathbb{R}^d\setminus\mathcal{C} : \delta_{\mathcal{C}}(y) \le \delta\right|,$$

with $\delta := g^{-1}(\epsilon)$, where g^{-1} denotes the inverse function of g which exists due to the assumptions on g. Translate $\mathcal{C}$ so that the largest ball it contains is centered at 0. The set $\left(1+\frac{\delta}{r}\right)\mathcal{C} = \mathcal{C} + \frac{\delta}{r}\mathcal{C}$ contains the δ-neighborhood of $\mathcal{C}$ since $\frac{\delta}{r}\mathcal{C}$ contains a ball of radius δ. The volume of $\left(1+\frac{\delta}{r}\right)\mathcal{C}$ is $(1+\delta/r)^d|\mathcal{C}|$, where we used the fact that for any Lebesgue measurable set A in $\mathbb{R}^d$ the dilation of A by $\lambda > 0$ defined as λA is also Lebesgue measurable with the Lebesgue measure $\lambda^d|A|$. Therefore, the volume of the set of all points that do not belong to $\mathcal{C}$ but lie within distance at most δ from it has volume at most $\left(\left(1+\frac{\delta}{r}\right)^d - 1\right)|\mathcal{C}|$. The proof is complete. □

Proof of Lemma 2.5

First, we recall from Lemma 2.3 that the KL divergence between π and π_δ is bounded by: $D(\pi\|\pi_\delta) \leq \frac{\int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy}{\int_{\mathcal{C}} e^{-f(y)}dy}$. It is easy to compute that for any $\theta > 0$,

$$\begin{aligned}
&\int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy \\
&= \int_{y\in\mathbb{R}^d\backslash\mathcal{C}:S(y)\leq\theta} e^{-\frac{1}{\delta}S(y)-f(y)}dy + \int_{y\in\mathbb{R}^d\backslash\mathcal{C}:S(y)>\theta} e^{-\frac{1}{\delta}S(y)-f(y)}dy \\
&\leq \left|y\in\mathbb{R}^d\backslash\mathcal{C} : S(y)\leq\theta\right| e^{-\inf_{y\in\mathbb{R}^d\backslash\mathcal{C}:S(y)\leq\theta} f(y)} + e^{-\frac{\theta}{\delta}} \int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy, \qquad (45)
\end{aligned}$$

where we used $S(y) \geq 0$ for any $y \in \mathbb{R}^d$ to obtain the inequality (45).

By taking $\theta = \tilde{\alpha}\delta\log(1/\delta)$ with $\tilde{\alpha} > 0$ in (45), and by applying Lemma 2.4, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy \\
&\leq \left|y\in\mathbb{R}^d\backslash\mathcal{C} : S(y)\leq\tilde{\alpha}\delta\log(1/\delta)\right| e^{-\inf_{y\in\mathbb{R}^d\backslash\mathcal{C}:S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y)} + \delta^{\tilde{\alpha}} \int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy \\
&\leq \left(\left(1+\frac{g^{-1}(\tilde{\alpha}\delta\log(1/\delta))}{r}\right)^d - 1\right)|\mathcal{C}| e^{-\inf_{y\in\mathbb{R}^d\backslash\mathcal{C}:S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y)} \\
&\qquad + \delta^{\tilde{\alpha}} \int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy \\
&\leq \left(\left(1+\frac{g^{-1}(\tilde{\alpha}\delta\log(1/\delta))}{r}\right)^d - 1\right)\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}R^d e^{-\inf_{y\in\mathbb{R}^d\backslash\mathcal{C}:S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y)} \\
&\qquad + \delta^{\tilde{\alpha}} \int_{\mathbb{R}^d\backslash\mathcal{C}} e^{-\frac{1}{\delta}S(y)-f(y)}dy,
\end{aligned}$$

where we used the fact that $\mathcal{C}$ is contained in an Euclidean ball with radius R (Assumption 2.2) so that $|\mathcal{C}|$ is less than or equal to the volume of a Euclidean ball with radius R which is $\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}R^d$, where Γ denotes the gamma function. The proof is complete. □

Proof of Lemma 2.6

Since $\mathcal{C}$ is convex, for every $x \in \mathbb{R}^d$ there exists a unique point of $\mathcal{C}$ nearest to x. Then the fact that $S(x) = (\delta_{\mathcal{C}}(x))^2$ is convex, ℓ-smooth and continuously differentiable with a gradient $\nabla S(x) = 2(x - \mathcal{P}_{\mathcal{C}}(x))$ is a direct consequence of Federer (1959, Theorem 4.8). To show that $S(x)$ is convex, consider two points x_1 and $x_2 \in \mathbb{R}^d$, and their projections c_1 and c_2 to the set $\mathcal{C}$. By the convexity of the set $\mathcal{C}$, we have $\bar{c} := \frac{(c_1+c_2)}{2} \in \mathcal{C}$ and by the definition of S, we obtain

$$\begin{aligned}
S\left(\frac{x_1+x_2}{2}\right) &\leq \left\|\frac{x_1+x_2}{2} - \bar{c}\right\|^2 = \frac{\|(x_1-c_1)+(x_2-c_2)\|^2}{4} \\
&\leq \frac{2\|x_1-c_1\|^2 + 2\|x_2-c_2\|^2}{4} = \frac{S(x_1)+S(x_2)}{2},
\end{aligned}$$

where we used the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any two vectors a, b in the last inequality. Finally, note that by the triangle inequality,

$$\|\nabla S(y) - \nabla S(x)\| \leq 2\|y-x\| + 2\left\|P_{\mathcal{C}}(y) - P_{\mathcal{C}}(x)\right\| \leq 4\|y-x\|,$$

where we used the non-expansiveness of the projection step. Therefore, we can take the smoothness constant of $S(x)$ to be $\ell = 4$. This completes the proof. □

Proof of Theorem 2.7

By weighted Csiszár-Kullback-Pinsker (W-CKP) inequality (see Lemma B.1), we have

$$\mathcal{W}_2(\pi, \pi_\delta) \leq \hat{C}\left(D(\pi\|\pi_\delta)^{\frac{1}{2}} + \left(\frac{D(\pi\|\pi_\delta)}{2}\right)^{\frac{1}{4}}\right), \tag{46}$$

where $\hat{C} := 2\inf_{\hat{x}\in\mathbb{R}^d, \hat{\alpha}>0}\left(\frac{1}{\hat{\alpha}}\left(\frac{3}{2} + \log\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} d\pi_\delta(x)\right)\right)^{\frac{1}{2}} < \infty$, provided that there exists some $\hat{\alpha} > 0$ and $\hat{x} \in \mathbb{R}^d$ so that $\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} d\pi_\delta(x) < \infty$. Furthermore, we can compute the following:

$$\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} d\pi_\delta(x) = \frac{\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} e^{-f(x)-\frac{S(x)}{\delta}} dx}{\int_{\mathbb{R}^d} e^{-f(x)-\frac{S(x)}{\delta}} dx} \leq \frac{\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} e^{-f(x)-\frac{S(x)}{\delta}} dx}{\int_{\mathcal{C}} e^{-f(x)} dx} < \infty,$$

provided that $\int_{\mathbb{R}^d} e^{\hat{\alpha}\|x-\hat{x}\|^2} e^{-f(x)-\frac{S(x)}{\delta}} dx < \infty$ which is increasing in δ and hence uniformly bounded as $\delta \to 0$.

We now take $g(x) = x^2$ in Lemma 2.5 with $S(x) = (\delta_{\mathcal{C}}(x))^2$ so that by Lemma 2.6, we have that S is convex, ℓ-smooth and continuously differentiable. Moreover, since $S(x) = (\delta_{\mathcal{C}}(x))^2$ and f is continuous and the set $\{y \in \mathbb{R}^d : S(y) \leq \tilde{\alpha}\delta\log(1/\delta)\}$ is compact and we have that in equation (5) in Lemma 2.5,

$$\inf_{y\in\mathbb{R}^d\setminus\mathcal{C}: S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y) \geq \inf_{y\in\mathbb{R}^d: S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y) = \min_{y\in\mathbb{R}^d: S(y)\leq\tilde{\alpha}\delta\log(1/\delta)} f(y) > -\infty,$$

and it is uniform in δ as $\delta \to 0$ and hence by applying Lemma 2.5 we obtain

$$D(\pi\|\pi_\delta) \leq \mathcal{O}\left((\delta\log(1/\delta))^{1/2}\right). \tag{47}$$

Finally, we get the desired result by plugging (47) into W-CKP inequality (46). The proof is complete. □

Proof of Lemma 2.10

Recall that we have the representation $\mathcal{C} = \{x : h(x) \leq 0\}$ given in (12) where $h(x) = \max_{1\leq i\leq m} h_i(x)$ for some $m \geq 1$ with $h_i : \mathbb{R}^d \to \mathbb{R}$ being convex for $i = 1, 2, \ldots, m$. Furthermore, $h(0) \leq 0$ as we assumed in Assumption 2.2 that $0 \in \mathcal{C}$. We first define the function $p_m : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$,

$$p_m(x) := \inf\{t \geq 0 : h_i(x/t) \leq 0 \text{ for every } i = 1, 2, \ldots, m\}.$$

By the convexity of h_i, it is easy to check that p_m is subadditive satisfying $p_m(x+y) \leq p_m(x) + p_m(y)$ for every x and y, and it is homogeneous with $p_m(sx) = sp_m(x)$ for any x and scalar $s \geq 0$. Therefore, p_m is convex and consequently locally Lipschitz continuous (Roberts and Varberg, 1974) and Lipschitz continuous on compact sets. The function $h(x)$ is also convex; hence, there exists a positive constant B such that $\|y\| \leq B$ for any $y \in \partial h(x)$ and $x \in \mathcal{C}$. We note that there exist some constants $c_K, C_K > 0$ such that

$$c_K p_m(x) \leq \|x\| \leq C_K p_m(x), \qquad \text{for any } x \in \mathbb{R}^d,$$

where $\|x\|$ is the Euclidean norm of $x \in \mathbb{R}^d$. To show this, let $\mathrm{bd}(\mathcal{C})$ denote the boundary of $\mathcal{C}$ and let

$$c_K := \min\{\|x\| : x \in \mathrm{bd}(\mathcal{C})\} \qquad \text{and} \qquad C_K := \max\{\|x\| : x \in \mathrm{bd}(\mathcal{C})\}.$$

Note that $p_m(x) = 1$ for $x \in \mathrm{bd}(\mathcal{C})$ and furthermore p_m is homogeneous. For any x, there exists $t > 0$ such that $tx \in \mathrm{bd}(\mathcal{C})$. Moreover, $c_K \leq \|tx\| \leq C_K$ and $p_m(tx) = 1$. Therefore, $p_m(x) = 1/t$ and $c_K/t \leq \|x\| \leq C_K/t$. Hence, we showed that

$$c_K p_m(x) \leq p_m(tx) \leq C_K p_m(x).$$

Next, we can compute that

$$\left|x : -\frac{\alpha}{2}\|x\|^2 \leq h(x) \leq 0\right| = \left|x : 1 - \frac{\alpha}{2}\|x\|^2 \leq p_m(x) \leq 1\right|.$$

For any x such that $p_m(x) \leq 1$, we have $\|x\| \leq C_K$. Thus, for any x such that $p_m(x) \geq 1 - \frac{\alpha}{2}\|x\|^2$ and $p_m(x) \leq 1$, we have $p_m(x) \geq 1 - \frac{\alpha}{2}C_K^2$, which implies that

$$\left|x : 1 - \frac{\alpha}{2}\|x\|^2 \leq p_m(x) \leq 1\right| \leq \left|x : 1 - \frac{\alpha}{2}C_K^2 \leq p_m(x) \leq 1\right|,$$

provided that $\alpha < 2/C_K^2$. Furthermore, by the definition of the functional $p_m(x)$, we have $p_m(x) \leq 1$ if and only if $x \in \mathcal{C}$. Therefore,

$$\begin{aligned}\left|x : 1 - \frac{\alpha}{2}C_K^2 \leq p_m(x) \leq 1\right| &= |x : p_m(x) \leq 1| - \left|x : p_m(x) \leq 1 - \frac{\alpha}{2}C_K^2\right| \\ &= |\mathcal{C}| - \left(1 - \frac{\alpha}{2}C_K^2\right)^d |\mathcal{C}|.\end{aligned}$$

On the other hand,

$$\begin{aligned}\left|x : h(x) + \frac{\alpha}{2}\|x\|^2 \leq 0\right| &= \left|x : p_m(x) + \frac{\alpha}{2}\|x\|^2 \leq 1\right| \\ &\geq \left|x : p_m(x) + \frac{\alpha}{2}C_K^2\|x\|_K^2 \leq 1\right| \\ &\geq \left|x : p_m(x) \leq 1 - \frac{\alpha}{2}C_K^2\right| = \left(1 - \frac{\alpha}{2}C_K^2\right)^d |\mathcal{C}|,\end{aligned}$$

provided that $\alpha < 2/C_K^2$. Hence, we conclude that

$$\frac{|\mathcal{C} \backslash \mathcal{C}^\alpha|}{|\mathcal{C}^\alpha|} \leq \frac{1 - \left(1 - \frac{\alpha}{2}C_K^2\right)^d}{\left(1 - \frac{\alpha}{2}C_K^2\right)^d} \leq \mathcal{O}(\alpha),$$

as $\alpha \to 0$. Therefore, the inequality (14) is satisfied, and the proof is complete. □

Proof of Proposition 2.11

Before we proceed to the proof of Proposition 2.11, we first state a few technical lemmas whose proofs will be provided at the end of the Appendix. The next technical lemma states that the penalty function $S^\alpha(x)$ is strongly-convex outside a compact domain for $\alpha \geq 0$ if the boundary function $h(x)$ is strongly convex, or for $\alpha > 0$ when h is merely convex. Before we proceed, let us recall that since the function $h(x)$ is also convex, there exists a positive constant B such that $\|y\| \leq B$ for any $y \in \partial h(x)$ and $x \in \mathcal{C}$.

Lemma D.1 *Consider the constrained set $\mathcal{C}^\alpha$ that is defined in (13) for $\alpha \geq 0$. Let β be the strong convexity constant of h with the convention that $\beta = 0$ if h is merely convex. If $\alpha + \beta > 0$, then the penalty function $S^\alpha(x)$ is strongly convex with constant $\frac{2(\alpha+\beta)\rho}{B+(\alpha+\beta)\rho}$ on the set $\mathbb{R}^d \backslash U(\mathcal{C}^\alpha, \rho)$, where $U(\mathcal{C}^\alpha, \rho)$ is the open ρ-neighborhood of $\mathcal{C}^\alpha$ i.e. $U(\mathcal{C}^\alpha, \rho) := \{x : dist(x, \mathcal{C}^\alpha) < \rho\}$.*

We have the following corollary as an immediate consequence of Lemma D.1.

Corollary D.2 *Under Assumption 2.9 and the assumptions of Lemma D.1, $f + \frac{S^\alpha}{\delta}$ is μ_δ-strongly convex outside an Euclidean ball with radius $R + \rho$, where $\mu_\delta := \frac{2(\alpha+\beta)\rho}{\delta(B+(\alpha+\beta)\rho)} - L$ provided that $\delta < \frac{2(\alpha+\beta)\rho}{L(B+(\alpha+\beta)\rho)}$.*

When $f + S^\alpha/\delta$ is strongly convex outside a compact domain, one can leverage the non-asymptotic guarantees in Ma et al. (2019b) for Langevin dynamics to obtain the following performance guarantees for the penalized Langevin dynamics. Before we proceed, we introduce the following technical lemma, which states that $f + \frac{S^\alpha}{\delta}$ is close to a strongly-convex function.

Lemma D.3 *Under the assumptions in Corollary D.2, for any given $m > 0$, there exists a C^1 function U such that U is s_0-strongly convex on $\mathbb{R}^d$ with*

$$\sup_{x\in\mathbb{R}^d}\left(U(x) - \left(f(x) + \frac{S^\alpha(x)}{\delta}\right)\right) - \inf_{x\in\mathbb{R}^d}\left(U(x) - \left(f(x) + \frac{S^\alpha(x)}{\delta}\right)\right)$$
$$\leq R_0 := 2(R+\rho)^2\left(\frac{m+L}{2} + \frac{(m+L)^2}{\mu_\delta}\right),$$

where $s_0 := \min(m, \mu_\delta/2)$.

Finally, we can proceed to the proof of Proposition 2.11. We first consider the case that $\mathcal{C} = \{x : h(x) \leq 0\}$, where h is β-strongly convex.

First of all, by running the penalized Langevin dynamics (11), we have

$$\mathrm{TV}(\nu_K, \pi) \leq \mathrm{TV}(\nu_k, \pi_\delta) + \mathrm{TV}(\pi_\delta, \pi),$$

where TV standards for the total variation distance. We recall from (47) that in KL divergence: $D(\pi\|\pi_\delta) \leq \mathcal{O}\left((\delta\log(1/\delta))^{1/2}\right)$. By Pinsker's inequality, we have

$$\mathrm{TV}(\pi_\delta, \pi) \leq \sqrt{\frac{1}{2}D(\pi\|\pi_\delta)} \leq \mathcal{O}\left((\delta\log(1/\delta))^{1/4}\right).$$

Therefore, $\mathrm{TV}(\pi_\delta, \pi) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^4$.

By Lemma C.2, $f + S/\delta$ is L_δ-smooth with $L_\delta := L + \frac{\ell}{\delta}$. Note that in Lemma D.3, we showed that there exists a C^1 function U that is s_0-strongly convex and satisfies

$$\sup_{x\in\mathbb{R}^d}\left(U(x) - f(x) - \frac{S(x)}{\delta}\right) - \inf_{x\in\mathbb{R}^d}\left(U(x) - f(x) - \frac{S(x)}{\delta}\right) \le R_0.$$

By Proposition 2 in Ma et al. (2019b), π_δ satisfies a log-Sobolev inequality with constant $\rho_* \ge s_0 e^{-R_0}$. Moreover, with $\delta = \varepsilon^4$, we recall from Lemma C.2 that $s_0 = \min(m, \mu_\delta/2)$ and $R_0 := 2(R+\rho)^2\left(\frac{m+L}{2} + \frac{(m+L)^2}{\mu_\delta}\right)$ so that $\mu_\delta = \frac{2\beta\rho}{\delta(B+\beta\rho)} - L = \Theta\left(\frac{1}{\varepsilon^4}\right)$ and thus $s_0 = \Theta(1)$, $R_0 = \Theta(1)$. By the proof of Theorem 1 in Ma et al. (2019b), $\mathrm{TV}(\nu_K, \pi_\delta) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that

$$\eta = \mathcal{O}\left(\frac{\rho_*}{L_\delta^2}\frac{\varepsilon^2}{d}\right) = \mathcal{O}\left(\frac{\varepsilon^{10}}{d}\right), \quad \text{and} \quad K = \tilde{O}\left(\frac{1}{\rho_*\eta}\right) = \tilde{O}\left(\frac{L_\delta^2 d}{\rho_*^2\varepsilon^2}\right) = \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right).$$

This completes the proof when h is β-strongly convex.

Indeed, we can see that the leading-order term for K we derived above does not depend on β. However, we can also spell out the dependence on β through the second-order term as follows. Notice that, by taking into account β, we have $\mu_\delta = \Theta\left(\frac{\beta}{\varepsilon^4}\right)$ and thus $s_0 = \Theta(1)$ and $R_0 = \Theta(1) + \Theta\left(\frac{\varepsilon^4}{\beta}\right)$. Then, we have $\mathrm{TV}(\nu_K, \pi_\delta) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that

$$\eta = \mathcal{O}\left(\frac{\rho_*}{L_\delta^2}\frac{\varepsilon^2}{d}\right), \quad \text{and} \quad K = \tilde{O}\left(\frac{1}{\rho_*\eta}\right) = \tilde{O}\left(\frac{L_\delta^2 d}{\rho_*^2\varepsilon^2}\right) = \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right) + \tilde{\mathcal{O}}\left(\frac{d}{\beta\varepsilon^6}\right),$$

where we ignored the dependence on the other constants B, m, L, ρ when we consider the second-order dependence on β.

Next, we consider the case when h is merely convex so that

$$h^\alpha(x) := h(x) + \frac{\alpha}{2}\|x\|^2$$

is α-strongly convex. By the previous discussions, $\mathrm{TV}(\nu_K, \pi_\delta^\alpha) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that

$$\eta = \mathcal{O}\left(\frac{\rho_*}{L_\delta^2}\frac{\varepsilon^2}{d}\right), \quad \text{and} \quad K = \tilde{O}\left(\frac{1}{\rho_*\eta}\right) = \tilde{O}\left(\frac{L_\delta^2 d}{\rho_*^2\varepsilon^2}\right),$$

where we can take $\rho_* = s_0 e^{-R_0}$. Next, we can compute that

$$\begin{aligned}
D(\pi^\alpha\|\pi) &= \int_{\mathbb{R}^d}\log\left(\frac{\pi^\alpha(x)}{\pi(x)}\right)\pi^\alpha(x)dx = \int_{\mathcal{C}^\alpha}\log\left(\frac{\int_{\mathcal{C}} e^{-f(x)}dx}{\int_{\mathcal{C}^\alpha} e^{-f(x)}dx}\right)\frac{e^{-f(x)}}{\int_{\mathcal{C}^\alpha} e^{-f(y)}dy}dx \\
&= \log\left(\frac{\int_{\mathcal{C}} e^{-f(x)}dx}{\int_{\mathcal{C}^\alpha} e^{-f(x)}dx}\right) = \log\left(1 + \frac{\int_{\mathcal{C}\setminus\mathcal{C}^\alpha} e^{-f(x)}dx}{\int_{\mathcal{C}^\alpha} e^{-f(x)}dx}\right) \\
&\le \frac{\int_{\mathcal{C}\setminus\mathcal{C}^\alpha} e^{-f(x)}dx}{\int_{\mathcal{C}^\alpha} e^{-f(x)}dx} \le e^{\sup_{x\in\mathcal{C}} f(x) - \inf_{x\in\mathcal{C}} f(x)}\frac{|\mathcal{C}\setminus\mathcal{C}^\alpha|}{|\mathcal{C}^\alpha|},
\end{aligned}$$

where $\sup_{x\in\mathcal{C}} f(x) - \inf_{x\in\mathcal{C}} f(x)$ is finite since $\mathcal{C}$ is compact. We recall from Lemma 2.10 that $\frac{|\mathcal{C}\setminus\mathcal{C}^\alpha|}{|\mathcal{C}^\alpha|} \leq \mathcal{O}(\alpha)$, as $\alpha \to 0$. Finally, by Pinsker's inequality,

$$\mathrm{TV}(\pi^\alpha, \pi) \leq \sqrt{\frac{1}{2} D(\pi^\alpha \| \pi)} \leq \mathcal{O}(\sqrt{\alpha}),$$

as $\alpha \to 0$. Therefore $\mathrm{TV}(\pi^\alpha, \pi) \leq \mathcal{O}(\varepsilon)$ provided that $\alpha = \varepsilon^2$. We recall from Lemma C.2 that $s_0 = \min(m, \mu_\delta/2)$, and $R_0 = 2(R+\rho)^2 \left(\frac{m+L}{2} + \frac{(m+L)^2}{\mu_\delta}\right)$, so that $\mu_\delta = \frac{2\alpha\rho}{\delta(B+\alpha\rho)} - L = \Theta\left(\frac{1}{\varepsilon^2}\right)$ with the choice of $\alpha = \varepsilon^2$ and $\delta = \varepsilon^4$. Hence, we conclude that $\mathrm{TV}(\nu_K, \pi) \leq \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^4$, $\alpha = \varepsilon^2$ and

$$\eta = \mathcal{O}\left(\frac{\rho_*}{L_\delta^2}\frac{\varepsilon^2}{d}\right) = \mathcal{O}\left(\frac{\varepsilon^{10}}{d}\right), \quad \text{and} \quad K = \tilde{O}\left(\frac{1}{\rho_*\eta}\right) = \tilde{O}\left(\frac{L_\delta^2 d}{\rho_*^2\varepsilon^2}\right) = \tilde{\mathcal{O}}\left(\frac{d}{\varepsilon^{10}}\right).$$

This completes the proof. □

Proof of Lemma 2.13

Since $\mathcal{C}$ is a convex set, every point in $\mathbb{R}^d$ has an unique projection on $\mathcal{C}$, which leads to $reach(\mathcal{C}) = \infty$, where

$$reach(\mathcal{C}) := \sup\left\{\zeta \in [0,\infty] : \text{every point in } \mathcal{C}^\zeta \text{ has unique projection on } \mathcal{C}\right\},$$

with $\mathcal{C}^\zeta := \{x \in \mathbb{R}^d : \inf\{\|x-\xi\| : \xi \in \mathcal{C}\} < \zeta\}$. According to Corollary 4 in Leobacher and Steinicke (2021), we can get that $D^2\mathcal{P}_\mathcal{C}$ is bounded on $\mathbb{R}^d$, where $\mathcal{P}_\mathcal{C}$ is the projection operator on $\mathcal{C}$. Then there exists some constant $M_\mathcal{P} > 0$ such that $\|D^2\mathcal{P}_\mathcal{C}\|_F \leq M_\mathcal{P}$, where $\|\cdot\|_F$ is the Frobenius norm. Moreover, we can compute that $S(x) = (x - \mathcal{P}_\mathcal{C}(x))^2$, $\nabla S(x) = 2(x - \mathcal{P}_\mathcal{C}(x))$ and $\nabla^2 S(x) = 2(I - D\mathcal{P}_\mathcal{C}(x))$. Note that for $x, y \in \mathbb{R}^d$,

$$\left\|\nabla^2 S(x) - \nabla^2 S(y)\right\|_F = 2\left\|D\mathcal{P}_\mathcal{C}(x) - D\mathcal{P}_\mathcal{C}(y)\right\|_F \leq 2M_\mathcal{P}\|x-y\|.$$

The proof is complete. □

Proof of Corollary 2.14

The result follows from Lemma 2.13 immediately. □

Proof of Proposition 2.15

We first consider the case that $\mathcal{C} = \{x : h(x) \leq 0\}$, where $h(x)$ is β-strongly convex. First of all, by running the penalized underdamped Langevin Monte Carlo (19)-(20), in total variation distance (TV), we have

$$\mathrm{TV}(\nu_K, \pi) \leq \mathrm{TV}(\nu_k, \pi_\delta) + \mathrm{TV}(\pi_\delta, \pi).$$

We recall from (47) that the KL divergence between π and π_δ can be bounded as: $D(\pi\|\pi_\delta) \leq \mathcal{O}\left((\delta\log(1/\delta))^{1/2}\right)$. By Pinsker's inequality, we have

$$\mathrm{TV}(\pi_\delta, \pi) \leq \sqrt{\frac{1}{2} D(\pi\|\pi_\delta)} \leq \mathcal{O}\left((\delta\log(1/\delta))^{1/4}\right).$$

Therefore, $\mathrm{TV}(\pi_\delta, \pi) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^4$. Moreover, by Corollary D.2, $f + S/\delta$ is μ_δ strongly convex outside an Euclidean ball with radius $R + \rho$ with $\mu_\delta := \frac{2\beta\rho}{\delta(B+\beta\rho)} - L$ and by Lemma C.2, $f + S/\delta$ is L_δ-smooth with $L_\delta := L + \frac{\ell}{\delta}$. By Theorem 1 in Ma et al. (2021) and Pinsker's inequality, we have

$$\mathrm{TV}(\nu_K, \pi_\delta) \le \sqrt{\frac{1}{2} D(\nu_K \| \pi_\delta)} \le \tilde{\mathcal{O}}(\varepsilon),$$

provided that

$$K = \tilde{\mathcal{O}}\left(\max\left\{ \frac{L_\delta^{3/2}}{\hat{\mu}_*^2}, \frac{M_\delta}{\hat{\mu}_*^2} \right\} \frac{\sqrt{d}}{\varepsilon} \right),$$

where $\hat{\mu}_* = \min\{\rho_*, 1\}$, where ρ_* is the log-Sobolev constant for π_δ. Note that in Lemma D.3, we showed that there exists a C^1 function U that is s_0-strongly convex and satisfies:

$$\sup_{x \in \mathbb{R}^d} \left(U(x) - f(x) - \frac{S(x)}{\delta} \right) - \inf_{x \in \mathbb{R}^d} \left(U(x) - f(x) - \frac{S(x)}{\delta} \right) \le R_0.$$

Therefore, by Holley-Stroock perturbation principle (see Holley et al. (1987)), the log-Sobolev constant for π_δ can be lower bounded as $\rho_* \ge s_0 e^{-R_0}$ where we recall from Lemma C.2 that $s_0 = \min(m, \mu_\delta/2)$ and $R_0 := 2(R+\rho)^2 \left(\frac{m+L}{2} + \frac{(m+L)^2}{\mu_\delta} \right)$, so that we can take $\hat{\mu}_* = \min\left\{ s_0 e^{-R_0}, 1 \right\}$. Finally, we notice that $L_\delta = L + \frac{\ell}{\delta} = \mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$ and $M_\delta = M_f + \frac{M_S}{\delta} = \mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$ with the choice $\delta = \varepsilon^4$. Moreover, $\mu_\delta = \frac{2\beta\rho}{\delta(B+\beta\rho)} - L = \mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$ so that $s_0 = \Theta(1)$ and $R_0 = \Theta(1)$ and thus $\hat{\mu}_* = \Theta(1)$. Hence, we conclude that $\mathrm{TV}(\nu_K, \pi) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^7} \right)$.

Indeed, we can see that the leading-order term for K we derived above does not depend on β. However, we can also spell out the dependence on β through the second-order term as follows. Notice that, by taking into account β, we have $\mu_\delta = \Theta\left(\frac{\beta}{\varepsilon^4}\right)$ and thus $s_0 = \Theta(1)$ and $R_0 = \Theta(1) + \Theta\left(\frac{\varepsilon^4}{\beta}\right)$. Then, we have $\mathrm{TV}(\nu_K, \pi_\delta) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^7} \right) + \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\beta\varepsilon^3} \right)$, where we ignored the dependence on the other constants B, m, L, ρ when we consider the second-order dependence on β.

Next, we consider the case when h is merely convex so that

$$h^\alpha(x) := h(x) + \frac{\alpha}{2}\|x\|^2$$

is α-strongly convex and consider the constraint set

$$\mathcal{C}^\alpha := \left\{ x : h(x) + \frac{\alpha}{2}\|x\|^2 \le 0 \right\}.$$

In the previous discussions, we showed that $\mathrm{TV}(\nu_K, \pi_\delta^\alpha) \le \tilde{\mathcal{O}}(\varepsilon)$ provided that $K = \tilde{\mathcal{O}}\left(\max\left\{ \frac{L_\delta^{3/2}}{\hat{\mu}_*^2}, \frac{M_\delta}{\hat{\mu}_*^2} \right\} \frac{\sqrt{d}}{\varepsilon} \right)$, where we can take $\hat{\mu}_* = \min\left\{ s_0 e^{-R_0}, 1 \right\}$. By following the proof of Proposition 2.11, we have $\mathrm{TV}(\pi^\alpha, \pi) \le \mathcal{O}(\varepsilon)$ provided that $\alpha = \varepsilon^2$. We recall from Lemma C.2 that $s_0 = \min(m, \mu_\delta/2)$ and $R_0 = 2(R+\rho)^2 \left(\frac{m+L}{2} + \frac{(m+L)^2}{\mu_\delta} \right)$, so that

$\mu_\delta = \frac{2\alpha\rho}{\delta(B+\alpha\rho)} - L = \Theta\left(\frac{1}{\varepsilon^2}\right)$ with the choice of $\alpha = \varepsilon^2$ and $\delta = \varepsilon^4$ so that $s_0 = \Theta(1)$ and $R_0 = \Theta(1)$ and $\hat{\mu}_* = \Theta(1)$. Finally, we notice that $L_\delta = L + \frac{\ell}{\delta} = \mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$ and $M_\delta = M_f + \frac{M_S}{\delta} = \mathcal{O}\left(\frac{1}{\varepsilon^4}\right)$ with the choice $\delta = \varepsilon^4$. Hence, we conclude that $\mathrm{TV}(\nu_K, \pi) \leq \tilde{\mathcal{O}}(\varepsilon)$ provided that $\delta = \varepsilon^4$ and $\alpha = \varepsilon^2$ and $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon^r}\right)$. This completes the proof. □

Proof of Lemma 2.19

Since f is strongly convex, it admits a unique minimizer, say $x_{*,f}$. If $x_{*,f} \in \mathcal{C}$, then for any $x \notin \mathcal{C}$, $S(x) > 0$ and

$$f(x) + \frac{S(x)}{\delta} > f(x_{*,f}) + \frac{S(x_{*,f})}{\delta} = f(x_{*,f}),$$

which implies the minimizer of $f + \frac{S}{\delta}$ must lie within $\mathcal{C}$ and hence the conclusion follows. If $x_{*,f} \notin \mathcal{C}$, then $S(x_{*,f}) = (\delta_{\mathcal{C}}(x_{*,f}))^2 > 0$. Then, for any x such that $S(x) > S(x_{*,f})$, we have

$$f(x) + \frac{S(x)}{\delta} > f(x_{*,f}) + \frac{S(x_{*,f})}{\delta} > f(x_{*,f}),$$

which implies that any minimizer x_* of $f + \frac{S}{\delta}$ must satisfy $S(x_*) \leq S(x_{*,f})$ so that $\delta_{\mathcal{C}}(x_*) \leq \delta_{\mathcal{C}}(x_{*,f})$. Since $\mathcal{C}$ is contained in a Euclidean ball centered at 0 with radius $R > 0$, we conclude that $\|x_*\| \leq R + \delta_{\mathcal{C}}(x_{*,f})$, which is independent of δ. This completes the proof. □

Proof of Proposition 2.21

First of all, we notice that with $S(x) = (\delta_{\mathcal{C}}(x))^2$, by Lemma 2.6, S is convex, ℓ-smooth (with $\ell = 4$) and continuously differentiable.

We will first show that we can uniformly bound the variance of the gradient noise. Let x_* be the unique minimizer of $f(x) + \frac{1}{\delta}S(x)$ (the minimizer is unique since $f(x) + \frac{1}{\delta}S(x)$ is strongly convex by Assumption 2.18 and Lemma 2.6). By Lemma 2.19, $\|x_*\| \leq (1+c)R$ for some $c, R \geq 0$. This implies that for any $\frac{\eta L_\delta}{2} < 1$,

$$\begin{aligned}
&\mathbb{E}\|x_{k+1} - x_*\|^2 \\
&= \mathbb{E}\left\|x_k - x_* - \eta\left(\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right)\right\|^2 + \eta^2\mathbb{E}\left\|\nabla\tilde{f}(x_k) - \nabla f(x_k)\right\|^2 + \mathbb{E}\left\|\sqrt{2\eta}\xi_{k+1}\right\|^2 \\
&\leq \mathbb{E}\left\|x_k - x_*\right\|^2 - 2\eta\mathbb{E}\left\langle x_k - x_*, \nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right\rangle + \eta^2\mathbb{E}\left\|\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right\|^2 \\
&\qquad + 2\eta^2\sigma^2\left(L^2\mathbb{E}\|x_k\|^2 + \|\nabla f(0)\|^2\right) + 2\eta d \\
&\leq \mathbb{E}\left\|x_k - x_*\right\|^2 - 2\eta\left(1 - \frac{\eta L_\delta}{2}\right)\mathbb{E}\left\langle x_k - x_*, \nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right\rangle \\
&\qquad + 2\eta^2\sigma^2\left(L^2\mathbb{E}\|x_k\|^2 + \|\nabla f(0)\|^2\right) + 2\eta d \\
&\leq (1 - 2\eta\mu + \eta^2\mu L_\delta)\mathbb{E}\left\|x_k - x_*\right\|^2 + 2\eta^2\sigma^2\left(L^2\mathbb{E}\|x_k\|^2 + \|\nabla f(0)\|^2\right) + 2\eta d \\
&\leq (1 - 2\eta\mu + \eta^2\mu L_\delta)\mathbb{E}\left\|x_k - x_*\right\|^2 \\
&\qquad + 2\eta^2\sigma^2\left(2L^2\mathbb{E}\|x_k - x_*\|^2 + 2L^2(1+c)^2R^2 + \|\nabla f(0)\|^2\right) + 2\eta d,
\end{aligned}$$

where we used $\frac{\eta L_\delta}{2} < 1$, and the fact that $f + \frac{1}{\delta}S$ is μ-strongly convex and L_δ-smooth. Hence, for any $\eta \leq \frac{\mu}{\mu L_\delta + 4\sigma^2 L^2}$ and $\frac{\eta L_\delta}{2} < 1$, we get

$$\mathbb{E}\|x_{k+1} - x_*\|^2 \leq (1 - \eta\mu)\mathbb{E}\left\|x_k - x_*\right\|^2 + 2\eta^2\sigma^2\left(2L^2(1+c)^2R^2 + \|\nabla f(0)\|^2\right) + 2\eta d,$$

which implies that

$$\begin{aligned}\mathbb{E}\|x_k\|^2 &\leq 2\mathbb{E}\|x_k - x_*\|^2 + 2(1+c)^2R^2 \\ &\leq \frac{4\eta\sigma^2}{\mu}\left(2L^2(1+c)^2R^2 + \|\nabla f(0)\|^2\right) + \frac{4d}{\mu} + 2(1+c)^2R^2.\end{aligned}$$

Hence, we conclude that

$$\mathbb{E}\left\|\nabla\tilde{f}(x_k) - \nabla f(x_k)\right\|^2 \leq 2\sigma^2\left(L^2\mathbb{E}\|x_k\|^2 + \|\nabla f(0)\|^2\right) \leq \sigma_V^2 d, \tag{48}$$

where

$$\sigma_V^2 := \sigma^2\left(\frac{8\eta\sigma^2L^2}{\mu d}\left(2L^2(1+c)^2R^2 + \|\nabla f(0)\|^2\right) + \frac{8L^2}{\mu} + \frac{4L^2(1+c)^2R^2}{d} + \frac{2\|\nabla f(0)\|^2}{d}\right). \tag{49}$$

Let ν_K be the distribution of the K-th iterate of the penalized stochastic gradient Langevin dynamics given by (23). By applying Theorem 4 in Dalalyan and Karagulyan (2019), under the assumption that $f(x) + \frac{1}{\delta}S(x)$ is μ-strongly convex and L_δ-smooth and the variance of the gradient noise is uniformly bounded (i.e., (48)) and the stepsize satisfies $\eta \leq \frac{\mu}{\mu L_\delta + 4\sigma^2L^2}$ and $\eta < \min\left(\frac{\mu}{\mu L_\delta + 4\sigma^2L^2}, \frac{2}{L_\delta}\right)$ (so that (48) holds), we have

$$\mathcal{W}_2(\nu_K, \pi_\delta) \leq (1-\mu\eta)^K\mathcal{W}_2(\nu_0, \pi_\delta) + \frac{1.65L_\delta}{\mu}\sqrt{\eta d} + \frac{\sigma_V^2\sqrt{\eta d}}{1.65L_\delta + \sigma_V\sqrt{\mu}},$$

where σ_V is defined in (49), so that together with Theorem 2.7 we have

$$\mathcal{W}_2(\nu_K, \pi) \leq (1-\mu\eta)^K\mathcal{W}_2(\nu_0, \pi_\delta) + \frac{1.65L_\delta}{\mu}\sqrt{\eta d} + \frac{\sigma_V^2\sqrt{\eta d}}{1.65L_\delta + \sigma_V\sqrt{\mu}} + \mathcal{O}\left((\delta\log(1/\delta))^{1/8}\right).$$

Moreover, we can compute that

$$\mathcal{W}_2(\nu_0, \pi_\delta) \leq \left(\mathbb{E}_{X\sim\nu_0}\|X\|^2\right)^{1/2} + \left(\mathbb{E}_{X\sim\pi_\delta}\|X\|^2\right)^{1/2},$$

and by the definition of π_δ,

$$\mathbb{E}_{X\sim\pi_\delta}\|X\|^2 = \frac{\int_{\mathbb{R}^d}\|x\|^2 e^{-f(x) - \frac{S(x)}{\delta}}dx}{\int_{\mathbb{R}^d} e^{-f(x)-\frac{S(x)}{\delta}}dx} \leq \frac{\int_{\mathbb{R}^d}\|x\|^2 e^{-f(x)}dx}{\int_{\mathcal{C}} e^{-f(x)}dx}, \tag{50}$$

where the upper bound in (50) is finite and independent of δ since f is μ-strongly convex.

By taking $\delta = \varepsilon^8$, $\eta = \frac{\varepsilon^{18}\mu^2}{d(L\varepsilon^8+\ell)^2}$, and $K = \tilde{\mathcal{O}}\left(\frac{d(L\varepsilon^8+\ell)^2}{\varepsilon^{18}\mu^3}\right)$, we get

$$\begin{aligned}\mathcal{W}_2(\nu_K, \pi) &\leq \tilde{\mathcal{O}}(\varepsilon) + \frac{\sigma_V^2\sqrt{\eta d}}{1.65L_\delta + \sigma_V\sqrt{\mu}} \\ &\leq \tilde{\mathcal{O}}(\varepsilon) + \tilde{\mathcal{O}}\left(\frac{\sigma_V^2\sqrt{\eta d}\varepsilon^8}{L\varepsilon^8 + \ell}\right) \leq \tilde{\mathcal{O}}(\varepsilon) + \tilde{\mathcal{O}}\left(\frac{\sigma_V^2\varepsilon^{17}\mu}{(L\varepsilon^8+\ell)^2}\right).\end{aligned}$$

Therefore $\mathcal{W}_2(\nu_K,\pi)\leq\tilde{\mathcal{O}}(\varepsilon)$ provided that $\sigma_V^2=\tilde{\mathcal{O}}\left(\frac{(L\varepsilon^8+\ell)^2}{\varepsilon^{16}\mu}\right)$. This implies that σ_V^2 and hence σ^2 and the batch-size b can simply be taken as the constant order, and therefore the stochastic gradient computations satisfy: $\hat{K}:=Kb=\tilde{\mathcal{O}}\left(\frac{d(L\varepsilon^8+\ell)^2}{\varepsilon^{18}\mu^3}\right)$. Finally, by Lemma 2.6, we can take $\ell=4$. The proof is complete. □

Proof of Proposition 2.22

Before we proceed to the technical proof of Proposition 2.22, we make the following remark regarding Lemma C.3.

Remark D.4 *Note that in Lemma C.3 without loss of generality, we can always assume $M=0$ so that $f+\frac{S}{\delta}\geq 0$. This is because, if $M>0$, we can always consider the "shifted" function $\hat{f}:=f+M$ which will satisfy $\hat{f}\geq 0$ and then apply the proof arguments to $e^{-\hat{f}(x)-\frac{S(x)}{\delta}}/\int_{x\in\mathcal{C}}e^{-\hat{f}(x)-\frac{S(x)}{\delta}}dx$ which will be proportional to $e^{-f(x)-\frac{S(x)}{\delta}}$. Therefore, in the rest of the paper and the proofs, we will assume $M=0$ in Lemma C.3.*

Now, we are ready to present the technical proof of Proposition 2.22.

First of all, we notice that with $S(x)=(\delta_{\mathcal{C}}(x))^2$, by Lemma 2.6, S is convex, ℓ-smooth and continuously differentiable. One technical challenge is that we cannot apply the results directly from Dalalyan and Riou-Durand (2020) because the results in Dalalyan and Riou-Durand (2020) are for the underdamped Langevin Monte Carlo without the gradient noise. Therefore, we need to adapt their approach to allow the additional gradient noise. First, we will obtain uniform L^2 bounds on penalized SGULMC v_k and x_k in (25)–(26).

Under Assumption 2.18 and by Lemma 2.6, $f+\frac{S}{\delta}$ is μ-strongly convex so that we have

$$\left\langle\nabla f(x)+\frac{1}{\delta}\nabla S(x),x-x_*\right\rangle\geq\mu\|x-x_*\|^2,\tag{51}$$

where x_* is the unique minimizer of $f+\frac{1}{\delta}S$. By Lemma 2.19, $\|x_*\|\leq(1+c)R$ for some $c,R\geq 0$. On the other hand,

$$\left|\left\langle\nabla f(x)+\frac{1}{\delta}\nabla S(x),x_*\right\rangle\right|\leq L_\delta\|x_*\|\cdot\|x-x_*\|\leq L_\delta(1+c)R\|x-x_*\|,$$

which together with (51) implies that

$$\begin{aligned}\left\langle\nabla f(x)+\frac{1}{\delta}\nabla S(x),x\right\rangle&\geq\mu\|x-x_*\|^2-L_\delta(1+c)R\|x-x_*\|\\&\geq\frac{\mu}{2}\|x-x_*\|^2-\frac{L_\delta^2(1+c)^2R^2}{2\mu}\\&\geq\frac{\mu}{4}\|x\|^2-\frac{\mu}{2}\|x_*\|^2-\frac{L_\delta^2(1+c)^2R^2}{2\mu}\\&\geq\frac{\mu}{4}\|x\|^2-\frac{\mu}{2}(1+c)^2R^2-\frac{L_\delta^2(1+c)^2R^2}{2\mu},\end{aligned}$$

and therefore $f + \frac{1}{\delta}S$ is (m_0, b_0)-dissipative with

$$m_0 := \frac{\mu}{4}, \qquad b_0 := \frac{\mu}{2}(1+c)^2R^2 + \frac{L_\delta^2(1+c)^2R^2}{2\mu}, \tag{52}$$

and moreover by Lemma C.3 and Remark D.4, $f+\frac{1}{\delta}S \geq 0$, and it follows from Lemma EC.5 in Gao et al. (2022) that uniformly in k, we have

$$\begin{aligned}
\mathbb{E}\|x_k\|^2 &\leq C_x^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\gamma^2}, \\
\mathbb{E}\|v_k\|^2 &\leq C_v^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{4}(1-2\lambda)},
\end{aligned} \tag{53}$$

where

$$\lambda := \frac{1}{2}\min(1/4, m_0/(L_\delta + \gamma^2/2)), \tag{54}$$

$$A := \frac{m_0}{2L_\delta + \gamma^2}\left(\frac{\|\nabla f(0)\|^2}{2L_\delta+\gamma^2} + \frac{b_0}{m_0}\left(L_\delta + \frac{1}{2}\gamma^2\right) + f(0)\right), \tag{55}$$

and μ_0 is the distribution of (x_0, v_0) and

$$\mathcal{V}(x,v) := f(x) + \frac{S(x)}{\delta} + \frac{1}{4}\gamma^2\left(\left\|x + \gamma^{-1}v\right\|^2 + \left\|\gamma^{-1}v\right\|^2 - \lambda\|x\|^2\right).$$

Next, we will bound the difference between v_{k+1}, x_{k+1} of the penalized SGULMC and $\tilde{v}_{k+1}, \tilde{x}_{k+1}$, which are the penalized ULMC without gradient noise that also start from v_k, x_k of the penalized SGULMC at k-th iterate. We recall from (25)-(26) that

$$\begin{aligned}
v_{k+1} &= \psi_0(\eta)v_k - \psi_1(\eta)\left(\nabla\tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi_{k+1}, \\
x_{k+1} &= x_k + \psi_1(\eta)v_k - \psi_2(\eta)\left(\nabla\tilde{f}(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi'_{k+1},
\end{aligned}$$

and next, we define

$$\begin{aligned}
\tilde{v}_{k+1} &:= \psi_0(\eta)v_k - \psi_1(\eta)\left(\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi_{k+1}, \\
\tilde{x}_{k+1} &:= x_k + \psi_1(\eta)v_k - \psi_2(\eta)\left(\nabla f(x_k) + \frac{1}{\delta}\nabla S(x_k)\right) + \sqrt{2\gamma}\xi'_{k+1},
\end{aligned}$$

so that one can easily check that

$$\mathbb{E}\|v_{k+1} - \tilde{v}_{k+1}\|^2 \leq (\psi_1(\eta))^2 2\sigma^2\left(L^2\mathbb{E}\|x_k\|^2 + \|\nabla f(0)\|^2\right) \leq 2\eta^2\sigma^2\left(L^2C_x^d + \|\nabla f(0)\|^2\right), \tag{56}$$

and moreover,

$$\mathbb{E}\|x_{k+1} - \tilde{x}_{k+1}\|^2 \leq (\psi_2(\eta))^2 2\sigma^2\left(L^2\mathbb{E}\|x_k\|^2 + \|\nabla f(0)\|^2\right) \leq 2\eta^4\sigma^2\left(L^2C_x^d + \|\nabla f(0)\|^2\right). \tag{57}$$

Since $\tilde{v}_{k+1}, \tilde{x}_{k+1}$ are the updates without the gradient noise, by using the synchronous coupling and following the same argument as in the proof of Theorem 2 in Dalalyan and Riou-Durand (2020), one can show that

$$\left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}\tilde{v}_{k+1}-V((k+1)\eta)\\ \tilde{x}_{k+1}-X((k+1)\eta)\end{array}\right]\right\|^2\right]\right)^{1/2}$$
$$\leq \left(1-\frac{0.75\mu\eta}{\gamma}\right)\left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}v_k-V(k\eta)\\ x_k-X(k\eta)\end{array}\right]\right\|^2\right]\right)^{1/2}+0.75L_\delta\eta^2\sqrt{d},$$

where $(X(t), V(t))$ is the continuous-time penalized underdamped Langevin diffusion (17)-(18) starting from the Gibbs distribution π_δ and

$$P := \frac{1}{\gamma}\left[\begin{array}{cc}0_{d\times d} & -\gamma I_d\\ I_d & I_d\end{array}\right].$$

This implies that

$$\left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}v_{k+1}-V((k+1)\eta)\\ x_{k+1}-X((k+1)\eta)\end{array}\right]\right\|^2\right]\right)^{1/2}$$
$$\leq \left(1-\frac{0.75\mu\eta}{\gamma}\right)\left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}v_k-V(k\eta)\\ x_k-X(k\eta)\end{array}\right]\right\|^2\right]\right)^{1/2}+0.75L_\delta\eta^2\sqrt{d}$$
$$+\left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}\tilde{v}_{k+1}-v_{k+1}\\ \tilde{x}_{k+1}-x_{k+1}\end{array}\right]\right\|^2\right]\right)^{1/2},$$

where we can compute from (56) and (57) that

$$\left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}\tilde{v}_{k+1}-v_{k+1}\\ \tilde{x}_{k+1}-x_{k+1}\end{array}\right]\right\|^2\right]\right)^{1/2} \leq \left\|P^{-1}\right\|\left(\mathbb{E}\|\tilde{v}_{k+1}-v_{k+1}\|^2+\mathbb{E}\|\tilde{x}_{k+1}-x_{k+1}\|^2\right)^{1/2}$$
$$\leq 2\eta\sigma\left\|P^{-1}\right\|\left(L^2C_x^d+\|\nabla f(0)\|^2\right)^{1/2},$$

provided that $\eta \leq 1$, which implies that

$$A_{k+1} \leq \left(1-\frac{0.75\mu\eta}{\gamma}\right)A_k+0.75L_\delta\eta^2\sqrt{d}+2\eta\sigma\left\|P^{-1}\right\|\left(L^2C_x^d+\|\nabla f(0)\|^2\right)^{1/2},$$

where

$$A_k := \left(\mathbb{E}\left[\left\|P^{-1}\left[\begin{array}{c}v_k-V(k\eta)\\ x_k-X(k\eta)\end{array}\right]\right\|^2\right]\right)^{1/2}.$$

This implies that

$$\begin{aligned}\mathcal{W}_2(\nu_k,\pi) &\le \gamma^{-1}\sqrt{2}A_k \\ &\le \frac{L_\delta\eta\sqrt{2d}}{\mu} + \frac{8\sqrt{2}}{3\mu}\sigma\left\|P^{-1}\right\|\left(L^2C_x^d + \|\nabla f(0)\|^2\right)^{1/2} + \sqrt{2}\left(1-\frac{0.75\mu\eta}{\gamma}\right)^k\frac{A_0}{\gamma} \\ &= \frac{L_\delta\eta\sqrt{2d}}{\mu} + \frac{8\sqrt{2}}{3\mu}\sigma\left\|P^{-1}\right\|\left(L^2C_x^d + \|\nabla f(0)\|^2\right)^{1/2} \\ &\qquad + \sqrt{2}\left(1-\frac{0.75\mu\eta}{\gamma}\right)^k\mathcal{W}_2(\nu_0,\pi),\end{aligned}$$

where ν_K denotes the distribution of the K-th iterate x_K of penalized stochastic gradient underdamped Langevin Monte Carlo (25)-(26). By the same argument as in the proof of Proposition 2.21, we can show that $\mathcal{W}_2(\nu_0,\pi_\delta)$ can be bounded uniformly in δ.

Hence, by taking $\delta=\varepsilon^8$, we get

$$\mathcal{W}_2(\nu_K,\pi) \le \tilde{\mathcal{O}}(\varepsilon) + \frac{8\sqrt{2}}{3\mu}\sigma\left\|P^{-1}\right\|\left(L^2C_x^d + \|\nabla f(0)\|^2\right)^{1/2},$$

where $\tilde{\mathcal{O}}$ ignores the dependence on $\log(1/\varepsilon)$, provided that

$$\eta = \min\left(\frac{1}{\sqrt{d}}\frac{\varepsilon^9\mu}{(L\varepsilon^8+\ell)}, \frac{1}{\sqrt{(\mu+L)\varepsilon^8+\ell}}\frac{\varepsilon^{12}\mu}{(L\varepsilon^8+\ell)}\right),$$

and

$$K = \tilde{\mathcal{O}}\left(\frac{\sqrt{(\mu+L)\varepsilon^8+\ell}(L\varepsilon^8+\ell)}{\varepsilon^{13}\mu^2}\max\left(\sqrt{d}, \frac{\sqrt{(L+\mu)\varepsilon^8+\ell}}{\varepsilon^3}\right)\right),$$

where $\tilde{\mathcal{O}}$ ignores the dependence on $\log(1/\varepsilon)$. Next, we recall from (53) that

$$C_x^d = \frac{\int_{\mathbb{R}^{2d}}\mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\gamma^2} \tag{58}$$

where we recall from (54)-(55) that

$$\lambda = \frac{1}{2}\min(1/4, m_0/(L_\delta+\gamma^2/2)), \tag{59}$$

$$A = \frac{m_0}{2L_\delta+\gamma^2}\left(\frac{\|\nabla f(0)\|^2}{2L_\delta+\gamma^2} + \frac{b_0}{m_0}\left(L_\delta+\frac{1}{2}\gamma^2\right)+f(0)\right), \tag{60}$$

where we recall from (52) that $m_0=\frac{\mu}{4}$ and $b_0=\frac{\mu}{2}(1+c)^2R^2+\frac{L_\delta^2(1+c)^2R^2}{2\mu}$. Since $L_\delta = L+\frac{\ell}{\delta} = L+\frac{\ell}{\varepsilon^8}$, we conclude from (59) and (60) that $\lambda=\Omega\left(\frac{\mu\varepsilon^8}{\varepsilon^8L+\ell}\right)$ and $A=\mathcal{O}\left(\left(L+\frac{\ell}{\varepsilon^8}\right)^2\frac{1}{\mu}\right)$, and it follows from (58) that $C_x^d=\mathcal{O}\left(\frac{\varepsilon^{16}d\mu(L\varepsilon^8+\ell)+(L\varepsilon^8+\ell)^3}{\mu^2\varepsilon^{24}}\right)$, which implies that $\mathcal{W}_2(\nu_K,\pi)\le\tilde{\mathcal{O}}(\varepsilon)$ provided that $\sigma=\mathcal{O}\left(\frac{\varepsilon^{13}\mu^2}{L\sqrt{L\varepsilon^8+\ell}\sqrt{\varepsilon^{16}d\mu+(L\varepsilon^8+\ell)^2}}\right)$, so that we can take

$$b=\Omega\left(\sigma^{-2}\right)=\Omega\left(\frac{L^2(L\varepsilon^8+\ell)(\varepsilon^{16}d\mu+(L\varepsilon^8+\ell)^2)}{\varepsilon^{26}\mu^4}\right).$$

Hence, $\mathcal{W}_2(\nu_K,\pi)\le\tilde{\mathcal{O}}(\varepsilon)$ with stochastic gradient computations $\hat{K}:=Kb$:

$$\hat{K}=\tilde{\mathcal{O}}\left(\frac{L^2(L\varepsilon^8+\ell)^2(\varepsilon^{16}d\mu+(L\varepsilon^8+\ell)^2)\sqrt{(\mu+L)\varepsilon^8+\ell}}{\varepsilon^{39}\mu^6}\max\left(\sqrt{d},\frac{\sqrt{(L+\mu)\varepsilon^8+\ell}}{\varepsilon^3}\right)\right).$$

Finally, by Lemma 2.6, we can take $\ell=4$. The proof is complete. □

Proof of Proposition 2.23

Let ν_K be the distribution of the K-th iterate x_K of penalized stochastic gradient Langevin dynamics (27). We recall from Lemma C.2 that $f+\frac{1}{\delta}S$ is (m_δ,b_δ)-dissipative with $m_\delta:=-L-\frac{1}{2}+\frac{m_S}{\delta}$ and $b_\delta:=\frac{1}{2}\|\nabla f(0)\|^2+\frac{b_S}{\delta}$ from (67) and $f+\frac{1}{\delta}S$ is L_δ-smooth with $L_\delta:=L+\frac{\ell}{\delta}$ and we also recall from Lemma C.3 and Remark D.4 that $f+\frac{1}{\delta}S\ge 0$. Under Assumption 2.9 and the assumption that $\eta\in(0,1\wedge\frac{m_\delta}{4L_\delta^2})$ and $k\eta\ge 1$, by Proposition 10 in Raginsky et al. (2017), we have

$$\mathcal{W}_2(\nu_K,\pi)\le\left(\tilde{C}_0\sigma^{1/2}+\tilde{C}_1\eta^{1/4}\right)(K\eta)+\tilde{C}_2e^{-K\eta/c_{LS}}+\mathcal{O}\left((\delta\log(1/\delta))^{1/8}\right),\tag{61}$$

where $\tilde{C}_1$, $\tilde{C}_2$ are defined as,

$$\begin{aligned}
\tilde{C}_0&:=(12+8(\kappa_0+2b_\delta+2d))\left(C_0+\sqrt{C_0}\right),\\
\tilde{C}_1&:=(12+8(\kappa_0+2b_\delta+2d))\left(6L_\delta^2(C_0+d)+\sqrt{6L_\delta^2(C_0+d)}\right),\\
\tilde{C}_2&:=\sqrt{2c_{LS}}\left(\log\|p_0\|_\infty+\frac{d}{2}\log\frac{3\pi}{m_\delta}+\frac{L_\delta\kappa_0}{3}+\|\nabla f(0)\|\sqrt{\kappa_0}+f(0)+\frac{b_\delta}{2}\log 3\right)^{1/2},
\end{aligned}$$

where κ_0 is given in (28) and

$$C_0:=L_\delta^2\left(\kappa_0+2\left(1\vee\frac{1}{m_\delta}\right)\left(b_\delta+2\|\nabla f(0)\|^2+d\right)\right)+\|\nabla f(0)\|^2,$$

where p_0 is the density of x_0, κ_0 is defined in (28) and c_{LS} is the constant for the logarithmic Sobolev inequality that π_δ satisfies which can be bounded as

$$c_{LS}\le\frac{2m_\delta^2+8L_\delta^2}{m_\delta^2L_\delta}+\frac{1}{\lambda_*}\left(\frac{6L_\delta(d+1)}{m_\delta}+2\right),$$

where λ_* is the spectral gap of the penalized overdamped Langevin SDE (10) that is defined in (29). Moreover, we observe that $\tilde{C}_0=\mathcal{O}(\tilde{C}_1)$. By (61), we have $\mathcal{W}_2(\nu_K,\pi)\le\tilde{\mathcal{O}}(\varepsilon)$ with

$$\begin{aligned}
\eta&=\Theta\left(\frac{\varepsilon^4}{\tilde{C}_1^4c_{LS}^4(\log\tilde{C}_2)^4}\right)=\tilde{\Theta}\left(\frac{\varepsilon^{196}}{d^8\lambda_*^{-4}(\log(\lambda_*^{-1}))^4}\right),\\
K&=\tilde{\mathcal{O}}\left(\frac{d^9\lambda_*^{-5}(\log(\lambda_*^{-1}))^4}{\varepsilon^{196}}\right),
\end{aligned}$$

and

$$\sigma^2 = \mathcal{O}(\eta) = \tilde{\Theta}\left(\frac{\varepsilon^{196}}{d^8\lambda_*^{-4}(\log(\lambda_*^{-1}))^4}\right),$$

where λ_* is defined in (29) so that

$$b = \Omega\left(\sigma^{-2}\right) = \tilde{\Omega}\left(\frac{d^8\lambda_*^{-4}(\log(\lambda_*^{-1}))^4}{\varepsilon^{196}}\right).$$

Hence, the stochastic gradient computations require

$$\hat{K} = Kb = \tilde{\mathcal{O}}\left(\frac{d^{17}\lambda_*^{-9}(\log(\lambda_*^{-1}))^8}{\varepsilon^{392}}\right).$$

Finally, under the further assumptions of Corollary D.2, $f + \frac{S}{\delta}$ is μ_δ-strongly convex (with $\mu_\delta := \frac{2\alpha\rho}{\delta(B+\alpha\rho)} - L$) outside of an Euclidean ball with radius $R+\rho$ and by Lemma C.2, $f + S/\delta$ is L_δ-smooth with $L_\delta := L + \frac{\ell}{\delta}$. By applying Lemma D.3, there exists a C^1 function U such that U is s_0-strongly convex on $\mathbb{R}^d$ with

$$\sup_{x\in\mathbb{R}^d}\left(U(x) - \left(f(x) + \frac{S(x)}{\delta}\right)\right) - \inf_{x\in\mathbb{R}^d}\left(U(x) - \left(f(x) + \frac{S(x)}{\delta}\right)\right) \leq R_0,$$

where s_0, R_0 are defined in Lemma D.3. We define π_U as the Gibbs measure such that $\pi_U \propto e^{-U(x)}$ and we also define:

$$\lambda_U := \inf\left\{\frac{\int_{\mathbb{R}^d}\|\nabla g\|^2 d\pi_U}{\int_{\mathbb{R}^d} g^2 d\pi_U} : g \in C^1(\mathbb{R}^d)\cap L^2(\pi_U), g \neq 0, \int_{\mathbb{R}^d} g d\pi_U = 0\right\}.$$

Since U is s_0-strongly convex, by Bakry-Émery criterion (see Corollary 4.8.2 in Bakry et al. (2014)), we have $\frac{1}{\lambda_U} \leq \frac{1}{s_0}$. Finally, by the Holley-Stroock perturbation principle (see Holley et al. (1987) and Proposition 5.1.6 and the discussion thereafter in Bakry et al. (2014)), we have $\frac{1}{\lambda_*} \leq \frac{1}{s_0}e^{R_0} \leq \mathcal{O}(1)$, which is a dimension-free bound, where we chose $\delta = \varepsilon^8$. Hence, we have $\hat{K} = \tilde{\mathcal{O}}\left(\frac{d^{17}}{\varepsilon^{392}}\right)$ and $\eta = \tilde{\Theta}\left(\frac{\varepsilon^{196}}{d^8}\right)$. The proof is complete. □

Proof of Proposition 2.24

Let ν_k be the distribution of the k-th iterate x_k of penalized stochastic gradient underdamped Langevin Monte Carlo (30)-(31). We recall from Lemma C.2 that $f + \frac{1}{\delta}S$ is (m_δ, b_δ)-dissipative with $m_\delta := -L - \frac{1}{2} + \frac{m_S}{\delta}$ and $b_\delta := \frac{1}{2}\|\nabla f(0)\|^2 + \frac{b_S}{\delta}$ from (67) and $f + \frac{1}{\delta}S$ is L_δ-smooth with $L_\delta := L + \frac{\ell}{\delta}$ and we also recall from Lemma C.3 and Remark D.4 that $f + \frac{1}{\delta}S \geq 0$. Then, under Assumption 2.9, it follows from Theorem EC.1 and Lemma EC.6 in Gao et al. (2022) that when the stepsize $\eta \leq \min\{1, \frac{\gamma}{\hat{K}_2}(d+A), \frac{\gamma\lambda}{2\hat{K}_1}, \frac{2}{\gamma\lambda}\}$ where λ, A are defined in (34)-(35) where $\hat{K}_1 := K_1 + Q_1\frac{4}{1-2\lambda} + Q_2\frac{8}{(1-2\lambda)\gamma^2}$ and $\hat{K}_2 := K_2 + Q_3$, where $Q_1 = \Theta(L_\delta)$, $Q_2 = \Theta\left((L_\delta)^3\right)$, $Q_3 = \Theta\left(L_\delta d\right)$, $K_1 = \Theta\left((L_\delta)^2\right)$, $K_2 = \Theta(1)$, (see Lemma EC.6 in Gao et al. (2022) for the precise definitions of K_1, K_2 and Q_1, Q_2, Q_3) and $k\eta \geq e$, we have

$$\mathcal{W}_2(\nu_k, \pi_\delta) \leq \left(C_0\sigma^{1/2} + C_1\eta^{1/2}\right)\cdot(k\eta)^{1/2}\cdot\sqrt{\log(k\eta)} + C\sqrt{\overline{\mathcal{H}}_\rho(\mu_0)}e^{-\mu_* k\eta},$$

where C_1 is given by

$$C_1 := \hat{\gamma} \cdot \left(\frac{3L_\delta^2}{2\gamma} \left(C_v^d + \left(2L_\delta^2 C_x^d + 2\|\nabla f(0)\|^2 \right) + \frac{2d\gamma}{3} \right) + \sqrt{\frac{3L_\delta^2}{2\gamma} \left(C_v^d + \left(2L_\delta^2 C_x^d + 2\|\nabla f(0)\|^2 \right) + \frac{2d\gamma}{3} \right)} \right)^{1/2}, \tag{62}$$

where $\hat{\gamma}$ is given by:

$$\hat{\gamma} := \frac{2\sqrt{2}}{\sqrt{\alpha}} \left(\frac{5}{2} + \log \left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha \mathcal{V}(x,v)} \mu_0(dx,dv) + \frac{1}{4} e^{\frac{\alpha(d+A)}{3\lambda}} \alpha\gamma(d+A) \right) \right)^{1/2}, \tag{63}$$

where μ_0 is the initial distribution for (x_0, v_0) and λ, A are defined in (34)-(35) and $\alpha := \lambda(1-2\lambda)/12$ and $\mathcal{V}(x,v)$ is the Lyapunov function defined in (33) and moreover

$$C_x^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\gamma^2}, \quad C_v^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{4}(1-2\lambda)}, \tag{64}$$

where $\hat{\gamma}, C_x^d, C_v^d$ are finite due to (32) and furthermore,

$$\mu_* := \frac{\gamma}{768} \min \left\{ \lambda L_\delta \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda} L_\delta \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda} \right\}, \tag{65}$$

$$C := \sqrt{2} e^{1+\frac{\Lambda}{2}} \frac{1+\gamma}{\min\{1,\alpha_1\}} \sqrt{\max\{1, 4(1+2\alpha_1+2\alpha_1^2)(d+A)\gamma^{-1}\mu_*^{-1} / \min\{1, R_1\}\}},$$

$$\Lambda := \frac{12}{5}(1+2\alpha_1+2\alpha_1^2)(d+A)L_\delta\gamma^{-2}\lambda^{-1}(1-2\lambda)^{-1}, \quad \alpha_1 := (1+\Lambda^{-1})L_\delta\gamma^{-2},$$

$$\varepsilon_1 := 4\gamma^{-1}\mu_*/(d+A), \quad R_1 := 4 \cdot (6/5)^{1/2}(1+2\alpha_1+2\alpha_1^2)^{1/2}(d+A)^{1/2}\gamma^{-1}(\lambda - 2\lambda^2)^{-1/2},$$

and moreover,

$$\overline{\mathcal{H}}_\rho(\mu_0) := R_1 + R_1\varepsilon_1 \max\left\{ L_\delta + \frac{1}{2}\gamma^2, \frac{3}{4} \right\} \|(x,v)\|^2_{L^2(\mu_0)} + R_1\varepsilon_1 \left(L_\delta + \frac{1}{2}\gamma^2 \right) \frac{b_\delta + d}{m_\delta} + R_1\varepsilon_1 \frac{3}{4} d + 2R_1\varepsilon_1 \left(f(0) + \frac{\|\nabla f(0)\|^2}{2L_\delta} \right),$$

where $\|(x,v)\|^2_{L^2(\mu_0)} := \int_{\mathbb{R}^{2d}} \|(x,v)\|^2 \mu_0(dx,dv)$, and finally, C_0 is defined as:

$$C_0 := \hat{\gamma} \cdot \left(\left(L_\delta^2 C_x^d + \|\nabla f(0)\|^2 \right) \frac{1}{\gamma} + \sqrt{\left(L_\delta^2 C_x^d + \|\nabla f(0)\|^2 \right) \frac{1}{\gamma}} \right)^{1/2},$$

where $\hat{\gamma}$ is defined in (63) and C_x^d is defined in (64). Thus, it is easy to see that $C_0 = \mathcal{O}(C_1)$, where C_1 is given in (62) and we can choose $\sigma^2 = \mathcal{O}(\eta)$ with

$$\eta = \tilde{\Theta} \left(\frac{\varepsilon^{50}\mu_*}{d^3 \left(\log(1/\mu_*) \right)^2} \right),$$

and the batch-size $b = \Omega(\sigma^{-2})$ such that

$$b = \tilde{\Theta}\left(\frac{d^3 \left(\log(1/\mu_*)\right)^2}{\varepsilon^{50}\mu_*}\right),$$

where μ_* is defined in (65). Hence, the stochastic gradient computations require

$$\hat{K} = Kb = \tilde{\mathcal{O}}\left(\frac{d^7 \left(\log(1/\mu_*)\right)^5}{\varepsilon^{132}\mu_*^3}\right).$$

The proof is complete. □

Proof of Lemma 2.25

We first show that $H_i(x) := \max(0, h_i(x))^2$ is continuously differentiable and convex. Note that the functions $h_i(x)$ and $\max(0, x)$ are both convex in x. Since the composition of convex functions is convex, $H_i(x)$ is convex. By the chain rule for convex functions in Section 3.3 of Borwein and Lewis (2005), the subdifferential of $H_i(x)$ is given by

$$\partial H_i(x) = \begin{cases} 0 & \text{if } h_i(x) \leq 0, \\ 2h_i(x)\nabla h_i(x) & \text{if } h_i(x) > 0, \end{cases}$$

where in the case $h_i(x) = 0$, we used the fact that the subdifferential of the convex function $\max(0, x)$ at $x = 0$ is given by the interval $[0, 1]$; so that by the chain rule, the subdifferential of $\partial H_i(x) = 2\max(0, h_i(x))[0, 1] = 0$ is single-valued for $h_i(x) = 0$. Since the subdifferential of $H_i(x)$ is single-valued for any x and is also continuous, we conclude that $H_i(x)$ is continuously differentiable.

Let $\mathcal{C}_i$ be the convex set on which $h_i(x) \leq 0$, i.e. $\mathcal{C}_i := \{x \in \mathbb{R}^d \ : \ h_i(x) \leq 0\}$. Since $h_i(x)$ is continuous, $x \in \mathrm{bd}(\mathcal{C}_i)$ if and only if $h_i(x) = 0$ where $\mathrm{bd}(\cdot)$ denotes the boundary of a set. Note that the Hessian of H_i, denoted by Hess_i, is continuous except at the boundary of $\mathcal{C}_i$ and can be computed as

$$\mathrm{Hess}_i(x) = 2\left[\nabla h_i(x) \cdot (\nabla h_i(x))^\top + h_i(x)\nabla^2 h_i(x)\right],$$

if $x \notin \mathcal{C}_i$. This is the case when $h_i(x) > 0$. On the other hand, for $x \in \mathrm{int}(\mathcal{C}_i)$, we have $H_i(x) = 0$ and $\mathrm{Hess}_i(x) = 0$ where $\mathrm{int}(\cdot)$ denotes the interior of a set. Therefore, for any $x \in \mathbb{R}^d \setminus \mathrm{bd}(\mathcal{C}_i)$,

$$\|\mathrm{Hess}_i(x)\| \leq 2\left(\|\nabla h_i(x)\|^2 + \max_{x\in\mathbb{R}^d} |h_i(x)| \left\|\nabla^2 h_i(x)\right\|\right) \leq \ell_i := 2\left(N_i^2 + P_i\right), \qquad (66)$$

where $\|\cdot\|$ denotes the matrix 2-norm (largest singular value), and we used the triangular inequality and sub-multiplicativity of the matrix 2-norm in (66). So far, we have shown that $H_i(x)$ is ℓ_i-smooth on the open set that excludes the boundary points of $\mathcal{C}_i$. For establishing smoothness at the boundary points $x \in \mathrm{bd}(\mathcal{C}_i)$, our proof relies on a more technical argument as the Hessian of H_i may not even exist for $x \in \mathrm{bd}(\mathcal{C}_i)$.[9] Our argument will roughly use the

9. For example, in dimension one; the unit ball around origin is defined by $m = 2$ constraints with $h_1(x) = x - 1 \leq 0$ and $h_2(x) = -x - 1 \leq 0$ where $\mathrm{bd}(\mathcal{C}_1) = \{1\}$ and $\mathrm{bd}(\mathcal{C}_2) = \{-1\}$. In this case, $\mathrm{Hess}_1(x) = 0$ for $-1 < x < 1$ and $\mathrm{Hess}_1(x) = 2$ for $x > 1$ and the Hessian does not exist at $x = 1$.

fact that boundary points constitute a measure zero set and the gradient of H_i is continuous at the boundary. For this purpose, next, we consider the line $\ell(t) := x + t(y-x)$ that passes through the points x and y, parameterized by the scalar $t \in \mathbb{R}$. Let

$$T := \{t \in [0,1] \ : \ \ell(t) \in \mathrm{bd}(\mathcal{C}_i)\}$$

correspond to the set of times t when the line segment between x and y crosses the boundary of the set $\mathcal{C}_i$. If we introduce $z(t) := \nabla H_i(\ell(t))$, then $z(t)$ is continuous, and it is continuously differentiable except when $t \in T$. Since $\mathcal{C}_i$ is closed, T is closed. Recalling that $\mathcal{C}_i$ is convex, roughly speaking, the line segment cannot go strictly out of the set $\mathcal{C}_i$ and then re-enter. We have three different cases:

I. T is the empty set: This case can arise when the line segment of x and y (including the endpoints) never intersects the set $\mathcal{C}_i$. In this case, H_i is twice continuously differentiable along the line segment. Thus, by Taylor's theorem with a remainder, we have

$$\begin{aligned}\|\nabla H_i(x) - \nabla H_i(y)\| = \|z(1) - z(0)\| &= \left\| \int_{t=0}^{1} z'(t) dt \right\| \\ &= \left\| \int_{t\in[0,1]} \mathrm{Hess}_i(\ell(t)) dt \right\| \leq L_i \|x-y\|,\end{aligned}$$

where we used (66).

II. $T = [t_1, t_2]$ for some $t_1 \leq t_2$ with the convention that T is a singleton when $t_1 = t_2$. In this case, $z(t)$ may not be differentiable for some points in $[0,1]$; however, we can approximate the interval [0,1] with unions of intervals where $z(t)$ is differentiable. More specifically, for any given $\varepsilon > 0$, we consider the closed intervals $I_1 = [\varepsilon, t_1 - \varepsilon], I_2 = [t_1 + \varepsilon, t_2 - \varepsilon], I_3 = [t_2 + \varepsilon, 1 - \varepsilon]$ with the convention that $[a,b]$ denotes the empty set when $a > b$. The union $\bigcup_{i=1,2,3} I_i$ approximates the interval $[0,1]$ when ε is sufficiently small. The function $z(t)$ is continuously differentiable for every $t \in I_i$ for $i = 1,2,3$ if $\varepsilon > 0$ is small enough except when $t \in T$. Furthermore, by the continuity of $z(t)$ and the fact that $z(t) = 0$ for $t \in T$, we have

$$\begin{aligned}\nabla H_i(x) - \nabla H_i(y) &= z(0) - z(1) \\ &= \int_{t\in I_1 \cap T^c} z'(t) dt + \int_{t\in I_2 \cap T^c} z'(t) dt + \int_{t\in I_3 \cap T^c} z'(t) dt + o(\varepsilon) \\ &= \int_{t\in (I_1 \cup I_2 \cup I_c) \cap T^c} \mathrm{Hess}_i(x + t(y-x)) dt + o(\varepsilon),\end{aligned}$$

where T^c denotes the complement of the set T and we used the fact that $z(t)$ is continuously differentiable on the set $t \in I_i \cap T^c$ for any $i \in \{1,2,3\}$. Taking the limit as $\varepsilon \to 0$, by a similar argument to Case I, we obtain $\|\nabla H_i(x) - \nabla H_i(y)\| \leq \ell_i \|x-y\|$, where $\ell_i := 2(N_i^2 + M_i P_i)$.

III. $T = \{t_1, t_2\}$ for some $t_1 \neq t_2$. This case can be treated similarly to Case II by considering the intervals I_1, I_2, and I_3.

Combining these cases, we can conclude that $H_i(x)$ is ℓ_i-smooth on $\mathbb{R}^d$, where $\ell_i := 2(N_i^2 + M_i P_i)$. Hence $\sum_{i=1}^m \max(0, h_i(x))^2 = \sum_{i=1}^m H_i(x)$ is ℓ-smooth, where $\ell := \sum_{i=1}^m \ell_i = 2\sum_{i=1}^m \left(N_i^2 + M_i P_i\right)$. This completes the proof. □

Proof of Corollary 2.26

To use Lemma 2.25, we need to prove that $h(x)$ satisfies its assumptions. Note that the function $t_i(x) := |x_i|^p$ is twice continuously differentiable in $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ for $p \geq 2$, $i = 1, 2, \ldots, d$ and the function $t_0 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined as $t_0(z) := z^{1/p}$ is twice continuously differentiable unless $z = 0$. Since the sum and composition of twice continuously differentiable functions remain twice continuously differentiable, we conclude that the p-norm $\|x\|_p := t_0\left(\sum_{i=1}^d |x_i|^p\right) = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ is twice continuously differentiable unless $x = 0$. Therefore, we conclude that $h(x)$ is twice continuously differentiable on the set

$$\mathcal{B} := \left\{x \in \mathbb{R}^d : \ h(x) \geq 0\right\},$$

which does not include $x = 0$. Since the p-norm is convex, $h(x)$ is also convex. For the rest, it suffices to prove that on the set $\mathcal{B}$, $h(x)$ has bounded gradients, and the product of $|h(x)|$ and the Hessian is bounded. For any $x \neq 0$, the gradient of $h(x)$ is given by:

$$\nabla h(x) = \left((|x_i|/\|x\|_p)^{p-1} \operatorname{sgn}(x_i), \ 1 \leq i \leq d\right),$$

with $\operatorname{sgn}(x) := -1$ if $x < 0$, 1 if $x > 0$, and 0 if $x = 0$. According to the definition of p-norm $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$, we have $|x_i| \leq \|x\|_p$ for any i, such that $|(\nabla h(x))_i| \leq 1$, which implies that $\|\nabla h(x)\| \leq \sqrt{d}$. Next, we consider the Hessian matrix of $h(x)$. After some computations, the entries (i, j) of the Hessian matrix of h are given by

$$\left[\nabla^2 h(x)\right]_{i,j} = \begin{cases} (p-1)\frac{1}{\|x\|_p^p}\left(|x_i|^{p-2}\|x\|_p - \frac{|x_i|^{2p-2}}{\|x\|_p^{p-1}}\right) & \text{if } i = j, \\ -\operatorname{sgn}(x_i x_j)(p-1)\frac{|x_i|^{p-1}|x_j|^{p-1}}{\|x\|_p^{2p-1}} & \text{if } i \neq j, \end{cases}$$

provided that $x \neq 0$. Note that the Hessian matrix $\nabla^2 h(x)$ is continuous unless $x = 0$.

Since $|x_i| \leq \|x\|_p$ for any i, we obtain the following bounds for the elements of Hessian matrix on the set $\mathcal{B} = \{x \in \mathbb{R}^d : \ h(x) \geq 0\} = \{x \in \mathbb{R}^d : \ \|x\|_p \geq R\}$:

$$0 \leq |h(x)| \left[\nabla^2 h(x)\right]_{i,i} \leq (p-1)\frac{\|x\|_p - R}{\|x\|_p^p}\left(2\|x\|^{p-1}\right) \leq 2(p-1)\frac{\|x\|_p - R}{\|x\|_p} \leq \frac{2(p-1)}{R},$$

and for $i \neq j$, we have

$$|h(x)| \cdot \left|\left[\nabla^2 h(x)\right]_{i,j}\right| \leq (p-1)\frac{\|x\|_p - R}{\|x\|_p} \leq p - 1.$$

Therefore, by applying the Gershgorin circle theorem (see, e.g., Fan (1958)), we obtain

$$|h(x)|\nabla^2 h(x) \preceq \left(\frac{2}{R} + (d-1)\right)(p-1)I.$$

Hence, $\max(0, h(x))^2$ is ℓ-smooth with $\ell = \left(\frac{2}{R} + (d-1)\right)(p-1)$ and the proof is complete.

Proof of Lemma C.1

The proof is similar to the proof of Lemma 2.6 with some minor differences to the potential non-convexity of the set $\mathcal{C}$. By the assumption for every $x \in \mathbb{R}^d$ there exists a unique point of $\mathcal{C}$ nearest to x. Then the fact that $S(x) = (\delta_{\mathcal{C}}(x))^2$ is ℓ-smooth and continuously differentiable with the gradient $\nabla S(x) = 2(x - \mathcal{P}_{\mathcal{C}}(x))$ is a direct consequence of Federer (1959, Theorem 4.8). Note that for $x_1, x_2 \in \mathbb{R}^d$,

$$\|\nabla S(x_1) - \nabla S(x_2)\| \leq 2\|x_1 - x_2\| + 2\|\mathcal{P}_{\mathcal{C}}(x_1) - \mathcal{P}_{\mathcal{C}}(x_2)\| \leq 4\|x_1 - x_2\|,$$

where in the last step we applied Federer (1959, Theorem 4.8, part (8)). Therefore, S is ℓ-smooth with $\ell = 4$. Also,

$$\langle x, \nabla S(x)\rangle = \langle x, 2(x - \mathcal{P}_{\mathcal{C}}(x))\rangle \geq 2\|x\|^2 - R\|x\| \geq m_S\|x\|^2 - b_S,$$

for $m_S = 1$, $b_S = R^2/4$. This completes the proof. □

Proof of Lemma C.2

Lemma C.1 shows that $S(x)$ is (m_S, b_S)-dissipative and ℓ-smooth. Then it follows that $f + \frac{1}{\delta}S$ is also L_δ-smooth, where $L_\delta := L + \frac{\ell}{\delta}$. By (m_S, b_S)-dissipativity of S, we have

$$\begin{aligned}
\left\langle x, \nabla f(x) + \frac{1}{\delta}\nabla S(x)\right\rangle &\geq \langle x, \nabla f(x)\rangle + \frac{m_S}{\delta}\|x\|^2 - \frac{b_S}{\delta} \\
&\geq \langle x, \nabla f(x) - \nabla f(0)\rangle - \|x\| \cdot \|\nabla f(0)\| + \frac{m_S}{\delta}\|x\|^2 - \frac{b_S}{\delta} \\
&\geq -L\|x\|^2 - \|x\| \cdot \|\nabla f(0)\| + \frac{m_S}{\delta}\|x\|^2 - \frac{b_S}{\delta} \\
&\geq -L\|x\|^2 - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|\nabla f(0)\|^2 + \frac{m_S}{\delta}\|x\|^2 - \frac{b_S}{\delta}, \qquad (67)
\end{aligned}$$

where we used L-smoothness of f. Therefore, $f + \frac{1}{\delta}S$ is also (m_δ, b_δ)-dissipative with $m_\delta := -L - \frac{1}{2} + \frac{m_S}{\delta} > 0$ and $b_\delta := \frac{1}{2}\|\nabla f(0)\|^2 + \frac{b_S}{\delta}$, provided that $\delta < m_S/(L + \frac{1}{2})$. This completes the proof. □

Proof of Lemma C.3

Since f is L-smooth, we have

$$f(x) \geq f(0) - \|\nabla f(0)\| \cdot \|x\| - \frac{L}{2}\|x\|^2,$$

and since S is (m_S, b_S)-dissipative and bounded below by 0, by Lemma 2 in Raginsky et al. (2017), we have

$$S(x) \geq \frac{m_S}{3}\|x\|^2 - \frac{b_S}{2}\log 3,$$

for any $x \in \mathbb{R}^d$ and thus

$$\begin{aligned}
f(x) + \frac{S(x)}{\delta} &\geq f(0) - \|\nabla f(0)\| \cdot \|x\| - \frac{L}{2}\|x\|^2 + \frac{m_S}{3\delta}\|x\|^2 - \frac{b_S}{2\delta}\log 3 \\
&\geq f(0) - \frac{1}{2}\|\nabla f(0)\|^2 - \frac{1}{2}\|x\|^2 - \frac{L}{2}\|x\|^2 + \frac{m_S}{3\delta}\|x\|^2 - \frac{b_S}{2\delta}\log 3 \geq -M,
\end{aligned}$$

where $M := -f(0) + \frac{1}{2}\|\nabla f(0)\|^2 + \frac{b_S}{2\delta}\log 3$, provided that $\delta \leq \frac{2m_S}{3(1+L)}$. This completes the proof. □

Proof of Lemma C.4

If Assumption 2.9 and Assumption 2.2 hold, according to Lemma C.3, function $f + \frac{1}{\delta}S$ is uniformly lower bounded, i.e. $f + \frac{1}{\delta}S \geq -M$ for an explicit non-negative scalar M defined in (40), which leads to the result that $f + \frac{1}{\delta}S + M$ is non-negative. Then according to Lemma C.2, function $f + \frac{1}{\delta}S$ is L_δ-smooth and (m_δ, b_δ)-dissipative, where $L_\delta, m_\delta, b_\delta$ is defined in (39). By Lemma 2 in Raginsky et al. (2017), we have

$$f(x) + \frac{S(x)}{\delta} + M \geq \frac{m_\delta}{3}\|x\|^2 - \frac{b_\delta}{2}\log 3,$$

for any $x \in \mathbb{R}^d$. Hence e^{-f} is integrable over $\mathcal{C}$, and moreover,

$$\int_{\mathbb{R}^d} e^{\frac{m_\delta}{6}\|x\|^2} e^{-f(x) - \frac{S(x)}{\delta}} dx \leq e^{\frac{b_\delta}{2}\log 3 + M} \int_{\mathbb{R}^d} e^{-\frac{m_\delta}{6}\|x\|^2} dx < \infty.$$

So that the assumptions in Theorem 2.7 are satisfied with $\hat{\alpha} = \frac{m_\delta}{6}$ and $\hat{x} = 0$. This completes the proof. □

Proof of Lemma C.5

Since it follows from Lemma 2.6 that $S(x)$ is convex and ℓ-smooth, it follows that under Assumption 2.18, $f + \frac{1}{\delta}S$ is also μ-strongly convex and L_δ-smooth, where $L_\delta := L + \frac{\ell}{\delta}$. Moreover, we notice that since f is μ-strongly convex, $f(x) \geq f(x_*) + \frac{\mu}{2}\|x - x_*\|^2$, where x_* is the unique minimizer of f. Hence, e^{-f} is integrable over $\mathcal{C}$ and moreover

$$\int_{\mathbb{R}^d} e^{\frac{\mu}{4}\|x-x_*\|^2} e^{-\frac{S(x)}{\delta} - f(x)} dx \leq \int_{\mathbb{R}^d} e^{\frac{\mu}{4}\|x-x_*\|^2} e^{-f(x)} dx \leq e^{-f(x_*)} \int_{\mathbb{R}^d} e^{-\frac{\mu}{4}\|x-x_*\|^2} dx < \infty,$$

so that the assumptions in Theorem 2.7 are satisfied with $\hat{\alpha} = \frac{\mu}{4}$ and $\hat{x} = x_*$. This completes the proof. □

Proof of Lemma D.1

We denote $x_{\mathcal{C}^\alpha}$ and $y_{\mathcal{C}^\alpha}$ as the projections of x and y onto $\mathcal{C}^\alpha$. Since we can compute that $S^\alpha(x) = \|x - x_{\mathcal{C}^\alpha}\|^2$ and $\nabla S(x) = 2(x - x_{\mathcal{C}^\alpha})$, we have:

$$\nabla S^\alpha(x) - \nabla S^\alpha(y) = 2\left(x - x_{\mathcal{C}^\alpha}\right) - 2\left(y - y_{\mathcal{C}^\alpha}\right) = 2(x-y) - 2\left(x_{\mathcal{C}^\alpha} - y_{\mathcal{C}^\alpha}\right).$$

It follows that:

$$\begin{aligned}\left(\nabla S^\alpha(x) - \nabla S^\alpha(y)\right)^\top (x-y) &= 2\|x-y\|^2 - 2\left(x_{\mathcal{C}^\alpha} - y_{\mathcal{C}^\alpha}\right)^\top (x-y) \\ &\geq 2\|x-y\|^2 - 2\|x_{\mathcal{C}^\alpha} - y_{\mathcal{C}^\alpha}\|\,\|x-y\|.\end{aligned}$$

By the assumptions, $\mathcal{C}^\alpha := \{x : h^\alpha(x) \leq 0\}$, where $h^\alpha(x)$ is a continuous $(\alpha+\beta)$-strongly convex function. By the convexity of h^α, it is Lipschitz on compact sets (Roberts and Varberg, 1974) and therefore there exists a positive constant B such that $\|y\| \leq B$ for any

$y \in \partial h^\alpha(x)$ and $x \in \mathcal{C}^\alpha$. According to Corollary 2 in Vial (1982), the set $\mathcal{C}^\alpha$ is strongly convex with radius $B/(\alpha+\beta)$ in the sense of Definition 1.1 in Balashov and Golubev (2012).[10] Then by applying Corollary 2.1 in Balashov and Golubev (2012), for any $x, y \in \mathbb{R}^d \backslash U(\mathcal{C}^\alpha, \rho)$, we have:

$$\|x_{\mathcal{C}^\alpha} - y_{\mathcal{C}^\alpha}\| \leq \frac{B}{B + (\alpha+\beta)\rho}\|x - y\|.$$

By combining these two inequalities, we have:

$$(\nabla S^\alpha(x) - \nabla S^\alpha(y))^\top (x - y) \geq \frac{2(\alpha+\beta)\rho}{B + (\alpha+\beta)\rho}\|x - y\|^2.$$

By Theorem 2.1.10 in Nesterov (2013), we conclude that the penalty function $S^\alpha(x)$ is strongly convex with constant $\frac{2(\alpha+\beta)\rho}{B+(\alpha+\beta)\rho}$ outside the ρ-neighborhood of the set $\mathcal{C}^\alpha$. The proof is complete. □

Proof of Corollary D.2

By Lemma D.1, the penalty function $S^\alpha(x)$ is strongly convex with constant $\frac{2(\alpha+\beta)\rho}{B+(\alpha+\beta)\rho}$ on the set $\mathbb{R}^d \backslash U(\mathcal{C}^\alpha, \rho)$, where $U(\mathcal{C}^\alpha, \rho)$ is the open ρ-neighborhood of $\mathcal{C}^\alpha$ i.e.

$$U(\mathcal{C}^\alpha, \rho) := \{x : \mathrm{dist}(x, \mathcal{C}^\alpha) < \rho\}.$$

Since $\mathcal{C}^\alpha$ is contained in an Euclidean ball centered at 0 of radius R, it follows that S^α is strongly convex with constant $\frac{2(\alpha+\beta)\rho}{B+(\alpha+\beta)\rho}$ outside a Euclidean ball with radius $R + \rho$ and moreover,

$$S^\alpha(x) \geq S^\alpha(y) + \langle \nabla S^\alpha(x), y - x\rangle + \frac{1}{2}\frac{2(\alpha+\beta)\rho}{B + (\alpha+\beta)\rho}\|x - y\|^2,$$

for any x, y outside an Euclidean ball with radius $R + \rho$. On the other hand, by Assumption 2.9, it follows that for any x, y: $f(x) \geq f(y) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|x - y\|^2$, which implies:

$$f(x) + \frac{S^\alpha(x)}{\delta} \geq f(y) + \frac{S^\alpha(y)}{\delta} + \left\langle \nabla f(x) + \frac{S^\alpha(x)}{\delta}, y - x \right\rangle + \frac{1}{2}\left(\frac{2(\alpha+\beta)\rho}{\delta(B + (\alpha+\beta)\rho)} - L\right)\|x - y\|^2,$$

for any x, y outside an Euclidean ball with radius $R + \rho$. This completes the proof. □

Proof of Lemma D.3

We start with defining

$$U(x) := f(x) + \frac{S^\alpha(x)}{\delta} + u(x),$$

10. A nonempty subset $\mathcal{C} \subset \mathbb{R}^d$ is called strongly convex of radius $R > 0$ if it can be represented as the intersection of closed balls of radius $R > 0$, i.e., there exists a subset $X \subset \mathbb{R}^d$ such that $\mathcal{C} = \bigcap_{x \in X} B_R(x)$, where $B_R(x)$ is a closed ball with radius R centered with x, see Def. 1.1 in Balashov and Golubev (2012).

where

$$u(x) := \begin{cases} \frac{m+L}{2}\|x\|^2 & \text{for} \quad \|x\| < R+\rho, \\ -\frac{\mu_\delta}{4}\|x\|^2 + a_\delta\|x\| + b_\delta & \text{for} \quad R+\rho \le \|x\| \le (R+\rho)\left(1+\frac{2(m+L)}{\mu_\delta}\right), \\ c_\delta & \text{for} \quad \|x\| > (R+\rho)\left(1+\frac{2(m+L)}{\mu_\delta}\right), \end{cases}$$

with

$$\begin{aligned} a_\delta &:= (m+L+\mu_\delta/2)(R+\rho), \\ b_\delta &:= -\frac{1}{2}(R+\rho)^2(m+L+\mu_\delta/2), \\ c_\delta &:= (R+\rho)^2\left(m+L+\frac{2(m+L)^2}{\mu_\delta}\right). \end{aligned}$$

In the first region, when $\|x\| < R+\rho$, we observe that the function $u(x)$ is a piecewise-defined quadratic that is clearly $(m+L)$-strongly convex. Since $S^\alpha(x)$ is convex and f is L-smooth, this implies that U is m-strongly convex in the first region when $\|x\| < R+\rho$.

In the second region, when $R+\rho \le \|x\| \le (R+\rho)\left(1+\frac{2(m+L)}{\mu_\delta}\right)$, u is a quadratic that is $\mu_\delta/2$-strongly concave (or equivalently $-u(x)$ is $\mu_\delta/2$-strongly convex) and $f + S/\delta$ is strongly convex with constant μ_δ, consequently U is strongly convex with constant $\mu_\delta/2$.

In the third region, outside the Euclidean ball with radius $(R+\rho)\left(1+\frac{2(m+L)}{\mu_\delta}\right)$, we observe that $u(x) \equiv c_\delta$ is a constant. Therefore $U = f + S^\alpha/\delta + u$ is μ_δ-strongly convex.

Moreover, it is straightforward to check that the piecewise function u has continuous derivatives and is of class C^1 and therefore $U = f + \frac{S^\alpha}{\delta} + u$ is a C^1 function. Finally, it is easy to check that $\sup_{x\in\mathbb{R}^d}\|u(x)\| = c_\delta$. Therefore,

$$\sup_{x\in\mathbb{R}^d}\left(U(x) - \left(f(x) + \frac{S^\alpha(x)}{\delta}\right)\right) - \inf_{x\in\mathbb{R}^d}\left(U(x) - \left(f(x) + \frac{S^\alpha(x)}{\delta}\right)\right) \le 2c_\delta,$$

and the result follows. The proof is complete. □