Fast sampling from constrained spaces using the Metropolis-adjusted Mirror Langevin algorithm

Vishwak Srinivasan VISHWAKS@MIT.EDU
Department of Electrical Engineering and Computer Science, MIT

Andre Wibisono ANDRE.WIBISONO@YALE.EDU
Department of Computer Science, Yale University

Ashia Wilson ASHIA07@MIT.EDU
Department of Electrical Engineering and Computer Science, MIT

Editors: Shipra Agrawal and Aaron Roth

Abstract

We propose a new method called the Metropolis-adjusted Mirror Langevin algorithm for approximate sampling from distributions whose support is a compact and convex set. This algorithm adds an accept-reject filter to the Markov chain induced by a single step of the Mirror Langevin algorithm (Zhang et al., 2020), which is a basic discretisation of the Mirror Langevin dynamics. Due to the inclusion of this filter, our method is unbiased relative to the target, while known discretisations of the Mirror Langevin dynamics including the Mirror Langevin algorithm have an asymptotic bias. For this algorithm, we also give upper bounds for the number of iterations taken to mix to a constrained distribution whose potential is relatively smooth, convex, and Lipschitz continuous with respect to a self-concordant mirror function. As a consequence of the reversibility of the Markov chain induced by the inclusion of the Metropolis-Hastings filter, we obtain an exponentially better dependence on the error tolerance for approximate constrained sampling.

Keywords: Sampling, constrained spaces, Langevin algorithm, mirror methods

1. Introduction

Continuous distributions supported on high-dimensional spaces are prevalent in various areas of science, more commonly so in machine learning and statistics. Samples drawn from such distributions can be used to generate confidence intervals for a point estimate, or provide Monte Carlo estimates for functionals of distributions. This motivates development of efficient algorithms to sample from such distributions. However, not all distributions are supported on the entire space (for e.g., $\mathbb{R}^d$). In one dimension, some common examples are the Gamma and Beta distributions (supported on $(0, \infty)$ and $(0, 1)$, respectively), and the latter's generalisation to multiple dimensions is the Dirichlet distribution (supported on a simplex Δ_{d+1}, which is a compact and convex subset of $\mathbb{R}^d$). Such constrained distributions not only occur in theory, but also in practice: for instance, in several Bayesian models (Pakman and Paninski, 2014), latent Dirichlet allocation (Blei et al., 2003) for topic modelling, regularised regression (Celeux et al., 2012), more recently in language models (Kumar et al., 2022), and for modelling metabolic networks (Heirendt et al., 2019; Kook et al., 2022). Analogous to sampling, optimisation over constrained domains are also of interest, and constrained optimisation is generally harder than unconstrained optimisation. Two popular families of approaches for solving constrained optimisation problems include interior-point methods (Nesterov and Nemirovski, 1994), and mirror descent (Nemirovski and Yudin, 1983).

The constrained sampling problem More formally, in this paper, we are interested in generating (approximate) samples from a target distribution Π with support $\mathcal{K}$ that is a compact, convex subset of $\mathbb{R}^d$, and whose density π is of the form

$$\pi(x) \propto e^{-f(x)} .$$

Here, $f : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ is termed the *potential* of Π.

The constrained sampling problem has been long-studied and a variety of prior approaches exist. The earliest inventions initially focused on obtaining uniform samples over $\mathcal{K}$ (i.e., when $f = 0$). This task had applications in estimating the volume of $\mathcal{K}$, and these methods were subsequently generalised to sampling from log-concave distributions supported on $\mathcal{K}$ (equivalently when f is convex in its domain). These inventions include the Hit-And-Run (Smith, 1984; Lovász, 1999; Lovász and Vempala, 2003), and BallWalk algorithms (Lovász and Simonovits, 1993; Kannan et al., 1997). Amongst other later modifications and analyses of these algorithms, the DikinWalk (Kannan and Narayanan, 2009; Narayanan and Rakhlin, 2017) is particularly notable due to the use of Dikin ellipsoids which are extensively used in the design and analysis of interior-point methods in optimisation. A commonality of all the aforementioned methods is that they only require calls to the (unnormalised) density of the target distribution.

To see how the gradients of f could be useful, consider the unconstrained setting where $\mathcal{K} = \mathbb{R}^d$. A popular algorithm for sampling in this setting is the unadjusted Langevin algorithm (ULA), which is the Euler-Maruyama discretisation of the continuous-time Langevin dynamics (LD).

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}\, dB_t \tag{LD}$$

$$X_{k+1} - X_k = -h \cdot \nabla f(X_k) + \sqrt{2h} \cdot \xi_k; \quad \xi_k \sim \mathcal{N}(0, I_d) \tag{ULA}$$

In the seminal work of Jordan et al. (1998), the Langevin dynamics was shown to be the gradient flow of the KL divergence in the space of probability measures equipped with the Wasserstein metric via the Fokker-Planck equation – thus making the connection between sampling and optimisation more substantive; see also the paper by Wibisono (2018). However, in the constrained sampling problem described earlier, the Langevin dynamics (or algorithm) is not applicable since the solutions (or iterates) are no longer guaranteed to remain in $\mathcal{K}$. The Mirror Langevin dynamics (MLD) (Zhang et al., 2020; Chewi et al., 2020)[1] was proposed to address this inapplicability, which draws from the idea from mirror descent in optimisation. Specifically, an invertible *mirror map* is used, which is often defined as the gradient of a barrier function (also called the *mirror function*) ϕ over $\mathcal{K}$.

$$Y_t = \nabla\phi(X_t); \quad dY_t = -\nabla f(X_t)dt + \sqrt{2}\, \nabla^2\phi(X_t)^{1/2}\, dB_t \tag{MLD}$$

Fundamentally, the mirror function serves two key purposes: (1) it changes the geometry of the primal space ($\mathcal{K}$) suitably; the underlying metric is given by the Hessian of the mirror function, and (2) it allows us to perform unconstrained LD-style diffusion in the dual space (the range of the mirror map) which is unconstrained when the domain is bounded. Due to the occurrence of the

1. A closely related dynamics was proposed in Hsieh et al. (2018) which is also termed Mirror Langevin dynamics – for a detailed comparison between the two dynamics, see Zhang et al. (2020, §1.2).

time-varying diffusion matrix in MLD, the dynamics can be discretised in a variety of ways. The most basic discretisation (or also called the Euler-Maruyama discretisation) of MLD is commonly referred to as the Mirror Langevin algorithm (MLA) (Zhang et al., 2020; Li et al., 2022a).

$$\begin{cases} Y_k = \nabla\phi(X_k)\ ; \\ Y_{k+1} - Y_k = -h \cdot \nabla f(X_k) + \sqrt{2h} \cdot \nabla^2\phi(X_k)^{1/2}\, \xi_k\ , \qquad \xi_k \sim \mathcal{N}(0, I_d)\ ; \\ X_{k+1} = (\nabla\phi)^{-1}(Y_{k+1}) \end{cases} \tag{MLA}$$

Other discretisations have since been proposed and analysed (Ahn and Chewi, 2021; Jiang, 2021); however, the advantage of MLA over these other discretisations is that MLA can be implemented exactly, whereas the latter methods involve simulating a stochastic differential equation (SDE) which cannot be performed exactly, and has to be approximated for practical purposes.

Theoretically, for MLA, the distribution of X_k as $k \to \infty$ is not guaranteed to coincide with Π, and such an algorithm is said to be *biased*, where the distance between this limit and Π is termed the *bias*. This phenomenon is not specific to MLA; for instance, the more popular ULA is also biased, and this bias is shown to scale with the step size h under relatively weak assumptions made on the target distribution – see Durmus and Moulines (2017); Dalalyan (2017b,a); Dalalyan and Karagulyan (2019); Cheng and Bartlett (2018); Vempala and Wibisono (2019); Li et al. (2022b) for a variety of analyses of ULA. Interestingly, due to the dependence on h, the bias of ULA tends to 0 as $h \to 0$ (thus termed a "vanishing bias"). In the case of MLA however, the first analysis by Zhang et al. (2020) suggested that MLA had a non-vanishing bias, and conjectured that this was unavoidable. Notably, their analysis assumed a *modified self-concordance condition* on ϕ, which is sufficient to ensure the existence of strong solutions for the continuous-time dynamics (MLD). In contrast, a more natural assumption placed on ϕ is *self-concordance*, especially considered in optimisation. A subsequent but different analysis of MLA by Jiang (2021) assumed the more standard self-concordance condition over ϕ, but their analysis was also unable to improve the non-vanishing bias of MLA. They also presented an analysis of an alternate yet impractical discretisation of MLD developed and analysed by Ahn and Chewi (2021) that we refer to as $\mathsf{MLA}_{\mathsf{FD}}$, and independent analyses in both of these works showed that $\mathsf{MLA}_{\mathsf{FD}}$ has a vanishing bias when the more natural self-concordance condition over ϕ is considered. More recently, Li et al. (2022a) improved the result due to Zhang et al. (2020), and showed that MLA has a vanishing bias when ϕ satisfies the modified self-concordance condition. This newer result is based on a mean-square analysis technique developed in Li et al. (2019), for which the modified self-concordance condition is more amenable.

A SUMMARY OF OUR WORK Despite these recent results, a question that has been left unanswered is if MLA, or any other exactly implementable algorithm based on discretising MLD has a vanishing bias when ϕ is simply self-concordant. In this paper, we give an algorithm that is *unbiased*, and with provable guarantees while assuming self-concordance of ϕ. Our proposed algorithm applies a Metropolis-Hastings filter to the proposal induced by a single step of MLA – we call this the *Metropolis-adjusted Mirror Langevin algorithm* (or MAMLA in short). Each iteration of the algorithm is composed of three key steps, and these are formalised in Algorithm 1 in the sequel.

1. Generate a proposal Z_k from iterate X_k using a single step of MLA.
2. Compute the Metropolis-Hastings acceptance probability p_{accept}.
3. With probability p_{accept}, set $X_{k+1} = Z_k$ (accept); otherwise set $X_{k+1} = X_k$ (reject).

In principle, this is similar to the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996), which applies a Metropolis-Hastings filter to the proposal induced by a single step of ULA. Due to the form of MLA, MAMLA is exactly implementable much like MALA, which crucially relies on the ability to compute the proposal densities. Moreover by construction, the inclusion of a Metropolis-Hastings filter results in a Markov chain that is reversible with respect to the target distribution. Consequently, the algorithm converges to the target distribution (and is therefore unbiased) unlike MLA and other discretisations of MLD, and does so exponentially quickly. In particular, let f be μ-strongly convex, λ-smooth, and β-Lipschitz relative to a self-concordant mirror function ϕ (see Section 2.1 for formal definitions). We show in Theorem 6 that

1. when $\mu > 0$, the δ-mixing time of MAMLA scales as $\mathcal{O}\left(\frac{1}{\mu} \cdot \max\left\{d^3, d\lambda, \beta^2\right\} \cdot \log\left(\frac{1}{\delta}\right)\right)$, and
2. when $\mu = 0$, the δ-mixing time of MAMLA scales as $\mathcal{O}\left(\nu \cdot \max\{d^3, d\lambda, \beta^2\} \cdot \log\left(\frac{1}{\delta}\right)\right)$, where ν is a constant that depends on the structure of $\mathcal{K}$ as induced by ϕ.

We obtain these guarantees through the classical one-step overlap technique due to Lovász and Simonovits (1993), which was also used to give mixing time guarantees for MALA in Dwivedi et al. (2018); Chewi et al. (2021). Our analysis is however not a direct consequence of these aforementioned results specifically due to the occurrence of the mirror map $\nabla\phi$ and metric $\nabla^2\phi$ in MLA, which poses certain difficulties. We handle these newer quantities by strongly leveraging the self-concordant nature of the mirror function ϕ and use recent isoperimetry results for distributions supported on a Hessian manifold induced by a self-concordant function (Gopi et al., 2023) to provide guarantees for MAMLA. A notable special case is when the mirror function ϕ is chosen to be the potential f, and the Mirror Langevin dynamics (MLD) specialises to the *Newton Langevin dynamics* (NLD), which is the sampling analogue of the continuous time version of Newton's method in optimisation. NLD particularly has parameter-free convergence rates when f is strictly convex, and we defer an extended discussion pertaining to this to Appendix C due to space constraints.

1.1. Related work

The original version of DikinWalk (Kannan and Narayanan, 2009) was developed with the goal of uniform sampling over polytopes, and this algorithm was based on generating either (a) uniform samples from Dikin ellipsoids, or (b) samples from a Gaussian distribution with covariance given by the Dikin ellipsoid (Sachdeva and Vishnoi, 2016; Narayanan, 2016). Other related algorithms are the JohnWalk (Gustafson and Narayanan, 2018), VaidyaWalk (Chen et al., 2018), and WeightedDikinWalk (Laddha et al., 2020). Narayanan and Rakhlin (2017) modify the filter in DikinWalk to make it amenable for the more general constrained sampling problem. This was recently modified by Mangoubi and Vishnoi (2023) who propose a modified Soft-Threshold DikinWalk, which primarily focuses on the case where $\mathcal{K}$ is a polytope. Kook and Vempala (2023) develop a broader theory of interior point sampling methods while generalising the ideas from the analysis of DikinWalk in Laddha et al. (2020). As remarked in Narayanan and Rakhlin (2017), while the existence of such a self-concordant barrier is a stronger assumption than a separation oracle for $\mathcal{K}$, the self-concordant barrier enables leveraging the geometry of $\mathcal{K}$ better.

The aforementioned methods related to DikinWalk do not require access to gradients of f. Drawing inspiration from projected gradient descent for optimising functions over constrained feasibility sets, Bubeck et al. (2018) propose a projected Langevin algorithm, where each step of ULA is projected onto the domain. Another class of approaches (Brosse et al., 2017; Gürbüzbalaban et al., 2022) propose obtaining a good approximation $\widetilde{\Pi}$ of the target Π such that the support of $\widetilde{\Pi}$ is $\mathbb{R}^d$, but this does not eliminate the likelihood of iterates lying outside $\mathcal{K}$. This enables running the simpler unadjusted Langevin algorithm over $\widetilde{\Pi}$ to obtain approximate samples from Π. MALA (Roberts and Tweedie, 1996; Dwivedi et al., 2018; Chewi et al., 2021), or Hamiltonian Monte Carlo (HMC) (Neal, 2011; Durmus et al., 2017; Bou-Rabee et al., 2020) is also a viable option for sampling from log-concave distributions over $\mathcal{K}$; however, the isoperimetric inequalities used in deriving these guarantees are based on the Euclidean metric / distance (i.e., Euclidean geometry) over $\mathcal{K}$ as opposed a non-Euclidean metric / distance. Riemannian Hamiltonian Monte Carlo (RHMC) (Girolami and Calderhead, 2011; Lee and Vempala, 2018) extends HMC to allow for sampling in such settings and with provable guarantees, and this has recently been modified to handle constrained domains better in Kook et al. (2022); Noble et al. (2023), while assuming that there exists a computable self-concordant barrier for $\mathcal{K}$ Girolami and Calderhead (2011) also proposed the Riemannian Langevin algorithm as a generalisation of ULA, and this was subsequently analysed in Gatmiry and Vempala (2022); Li and Erdogdu (2023). Relatedly, drawing from proximal methods for sampling, Gopi et al. (2023) propose a novel proximal sampler for sampling over non-Euclidean spaces.

Condition number independence Here, the *condition number* is a property of the domain $\mathcal{K}$ and is defined as $\mathcal{C}_{\mathcal{K}} = {}^{R}/_{r}$, where R and r are the radii of the smallest ball containing $\mathcal{K}$ and the largest ball contained entirely in $\mathcal{K}$ respectively (Kannan and Narayanan, 2009). A Euclidean ball in $\mathbb{R}^d$ of arbitrary radius has condition number 1, and non-unitary affine transformations of this ball results in sets with different condition numbers. Constrained sampling methods over $\mathcal{K}$ that have mixing time guarantees that are independent of $\mathcal{C}_{\mathcal{K}}$ are valuable, and this is because the complexity of approximate sampling from distributions with affine invariant properties over a possibly ill-conditioned domain $\mathcal{K}$ is the same as approximate sampling from similar distributions over a well-conditioned (or *isotropic*) domain like a ball. It is important to note that $\mathcal{C}_{\mathcal{K}}$ is not the same as the value $\kappa = {}^{\lambda}/_{\mu}$; the latter corresponds to the conditioning of the potential relative to a mirror function defined over $\mathcal{K}$. The dependence on κ is expected; intuitively, it is harder to sample from sharper / peakier distributions (equivalent to a higher value of κ), even when supported over a well-conditioned domain.

DikinWalk is a popular example of a sampling algorithm whose mixing time is independent of $\mathcal{C}_{\mathcal{K}}$, and this is primarily due to the involvement of Dikin ellipsoids. As far as discretisations of MLD are concerned, the analysis of $\mathsf{MLA}_{\mathsf{FD}}$ from Ahn and Chewi (2021) gives mixing time guarantees that depend on constants in the assumptions made. Notably, these assumptions are affine invariant, which implies these constants are independent of $\mathcal{C}_{\mathcal{K}}$, and results in mixing time guarantees that are *independent* of $\mathcal{C}_{\mathcal{K}}$. On the other hand, the recent analysis of MLA due to Li et al. (2022a) assumes a *modified self-concordance* condition. This is not affine invariant in the sense that the parameter in this condition can change when the domain is transformed by an affine map (Li et al., 2022a, §D). Put simply, this parameter could depend on $\mathcal{C}_{\mathcal{K}}$. As this parameter appears in their mixing time guarantee, it suggests that the mixing time of MLA is also possibly dependent on $\mathcal{C}_{\mathcal{K}}$. Our mixing time guarantees for MAMLA crucially does not depend on $\mathcal{C}_{\mathcal{K}}$ much like the guarantees for $\mathsf{MLA}_{\mathsf{FD}}$, and this is due to the affine invariance of the assumptions we make for the analysis.

Organisation The rest of this paper is organised as follows: in Section 2, we review notation and definitions that are key for this work, in Section 3, we define the algorithm and provide mixing time guarantees. We discuss corollaries of our guarantees for applying MAMLA to a variety of sampling tasks in Section 4, and conclude with a discussion of our results in Section 5. Other details including the proofs of our mixing time guarantees and corollaries are deferred to the Appendix.

2. Preliminaries

We begin by introducing some general notation that will be used throughout this work. We remind the reader that $\mathcal{K}$ is a compact, convex subset of $\mathbb{R}^d$.

Notation The set $\{1, \ldots, m\}$ is denoted by $[m]$. We denote the set of positive reals by $\mathbb{R}_+$. Let $A \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix. For $x, y \in \mathbb{R}^d$, we define $\langle x, y\rangle_A = \langle x, Ay\rangle$, and $\|x\|_A = \sqrt{\langle x, x\rangle_A}$. When the subscript is omited as in $\|x\|$, then this corresponds to $A = I_{d\times d}$, or the Euclidean norm of x. For a set $\mathcal{A}$, $\mathrm{int}(\mathcal{A})$ denotes its interior, and collection of all measurable subsets of $\mathcal{A}$ is denoted by $\mathcal{F}(\mathcal{A})$. For a measurable map $T : \mathcal{A} \to \mathcal{B}$ and a distribution P supported on $\mathcal{A}$, we denote $T_{\#}P$ to be the pushforward measure of P using T; in other words, $T_{\#}P$ is the law of $T(x)$ where x is distributed according to P, and $T_{\#}P$ has support $\mathcal{B}$. Unless specified explicitly, the density of a distribution P (if it exists) at x is denoted by $dP(x)$.

2.1. Function classes

Throughout this work, we make the standard assumption that the mirror function is of *Legendre type*. A function with domain $\mathcal{K}$ is of Legendre type if it is differentiable and strictly convex in $\mathrm{int}(\mathcal{K})$, and has gradients that become unbounded as we approach the boundary of $\mathcal{K}$; we give the precise definition of Legendre type functions as stated in Rockafellar (1970, Chap. 26) in the appendix (Definition 9). The convex conjugate of ϕ is denoted by ϕ^*, and is defined as

$$\phi^*(y) = \max_{x\in\mathcal{K}} \langle x, y\rangle - \phi(x) \ .$$

For both the mirror descent algorithm in optimisation and MLA, a method to compute the inverse of $\nabla\phi$ is required, as the respective updates are made in the dual space (range of $\nabla\phi$). Crucially, when ϕ is of Legendre type, then $\nabla\phi$ is an invertible map between $\mathcal{K}$ and the domain of ϕ^*, and the inverse of this map is $\nabla\phi^*$. Additionally, when $\mathcal{K}$ is a bounded subset of $\mathbb{R}^d$, the domain of ϕ^* is $\mathbb{R}^d$ (Rockafellar, 1970, Corr. 13.3.1). This implies that for any $y \in \mathbb{R}^d$, there exists a unique $x \in \mathrm{int}(\mathcal{K})$ such that $\nabla\phi^*(y) = (\nabla\phi)^{-1}(y) = x$. This fact is essential for MLA as the random vector ξ_k in MLA is unrestricted, and X_{k+1} could be undefined unless the domain of ϕ^* is $\mathbb{R}^d$.

Self-concordance Self-concordant functions are ubiquitous in the study and design of interior-point methods in optimisation, and are defined as follows.

Definition 1 (Nesterov (2018, §5.1.3)) *A thrice differentiable strictly convex function $\psi : \mathcal{K} \to \mathbb{R}\cup\{\infty\}$ is said to be self-concordant with parameter $\alpha \geq 0$ if for all $x \in \mathrm{int}(\mathcal{K})$ and $u \in \mathbb{R}^d$,*

$$|\nabla^3\psi(x)[u, u, u]| \leq 2\alpha\|u\|^3_{\nabla^2\psi(x)} \ .$$

Relative convexity and smoothness These classes of functions can be viewed as generalisations of convex and smooth functions with different geometrical properties, and were independently studied in Bauschke et al. (2017) and Lu et al. (2018).

Definition 2 *Let $g,\ \psi : \mathcal{K} \to \mathbb{R}\cup\{\infty\}$ be differentiable convex functions. We say that g is μ-relatively convex with respect to ψ if $g - \mu\cdot\psi$ is convex, where $\mu \geq 0$. Equivalently, when $g,\ \psi$ are twice differentiable, then $\mu\cdot\nabla^2\psi(x) \preceq \nabla^2 g(x)$ for all $x \in \mathrm{int}(\mathcal{K})$.*

Definition 3 *Let $g,\ \psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be differentiable convex functions. We say that g is λ-relatively smooth with respect to ψ if $\lambda \cdot \psi - g$ is convex, where $\lambda \geq 0$. Equivalently, when $g,\ \psi$ are twice differentiable, then $\lambda \cdot \nabla^2\psi(x) \succeq \nabla^2 g(x)$ for all $x \in \mathrm{int}(\mathcal{K})$.*

Relative Lipschitz continuity This class of functions is a generalisation of Lipschitz continuity of a differentiable function, and has been useful in the analysis of $\mathsf{MLA}_{\mathsf{FD}}$ (Ahn and Chewi, 2021).

Definition 4 (Jiang (2021); Ahn and Chewi (2021)) *Let $g : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a differentiable function. We say that g is β-relatively Lipschitz continuous with respect to a twice differentiable strictly convex function $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ with parameter β if for all $x \in \mathrm{int}(\mathcal{K})$, it holds that $\|\nabla g(x)\|_{\nabla^2\psi(x)^{-1}} \leq \beta$, or equivalently, $\beta^2 \cdot \nabla^2\psi(x) \succeq \nabla g(x) \nabla g(x)^\top$.*

Self-concordance, relative convexity, relative smoothness and relative Lipschitz continuity are invariant to affine transformations of the domain. More precisely, let T_{Aff} be an affine transformation of suitable dimensions. If ψ is a self-concordant function with parameter α, then $\psi \circ T_{\text{Aff}}$ is also a self-concordant function with parameter α (Nesterov, 2018, Thm. 5.1.2). If g is a μ-relatively convex (or λ-relatively smooth, or β-relatively Lipschitz continuous) function with respect to ψ, then $g \circ T_{\text{Aff}}$ is also a μ-relatively convex (or λ-relatively smooth, or β-relatively Lipschitz continuous, resp.) function with respect to $\psi \circ T_{\text{Aff}}$ (Lu et al. (2018, Prop. 1.2); Nesterov (2018, Thm. 5.3.3)).

Symmetric Barrier Symmetric barriers were introduced in Laddha et al. (2020) where it was used to obtain mixing time bounds for the Dikin walk and weighted variant. This property was originally introduced in Gustafson and Narayanan (2018) in the development of the John walk.

Definition 5 (Laddha et al. (2020, Def. 2)) *Let $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a twice differentiable function, and let $\mathcal{E}_x^{\psi}(r) = \{y : \|y - x\|_{\nabla^2\psi(x)} \leq r\}$ be the Dikin ellipsoid of radius r centered at x. We say that ψ is a symmetric barrier with parameter $\nu > 0$ if for all $x \in \mathrm{int}(\mathcal{K})$,*

$$\mathcal{E}_x^{\psi}(1) \subseteq \mathcal{K} \cap (2x - \mathcal{K}) \subseteq \mathcal{E}_x^{\psi}(\sqrt{\nu}) \ .$$

Examples of such barriers are the log-barrier of a polytope formed by m constraints is a symmetric barrier with parameter $\nu = m$ (Lemma 10), and the log-barrier of an ellipsoid with unit radius is also a symmetric barrier with parameter $\nu = 2$ (Lemma 11). Notably, the symmetric barrier parameters of these log-barriers is independent of the conditioning of these sets.

2.2. Markov chains, conductance and mixing time

Markov chains Let the domain of interest be $\mathcal{K}$. A (time-homogeneous) Markov chain over $\mathcal{K}$ is characterised by a set of transition kernels $\mathbf{P} = \{\mathcal{P}_x : x \in \mathcal{K}\}$, where $\mathcal{P}_x$ is the one-step distribution that maps measurable subsets of $\mathcal{K}$ to non-negative values. With this setup, we have a transition operator $\mathbb{T}_{\mathbf{P}}$ on the space of probability measures defined by

$$(\mathbb{T}_{\mathbf{P}}\mu)(S) = \int_{\mathcal{K}} \mathcal{P}_y(S) \cdot d\mu(y) \qquad \forall\, S \in \mathcal{F}(\mathcal{K}) \ .$$

The distribution after k applications of the transition operator to μ is denoted by $\mathbb{T}_{\mathbf{P}}^k\mu$. A probability measure π is called a *stationary* measure of a Markov chain $\mathbf{P}$ if $\pi = \mathbb{T}_{\mathbf{P}}\pi$. A Markov chain $\mathbf{P}$ is said to be *reversible* with respect to a measure π if for any $A, B \in \mathcal{F}(\mathcal{K})$,

$$\int_A \mathcal{P}_x(B) \cdot d\pi(x) = \int_B \mathcal{P}_y(A) \cdot d\pi(y) .$$

If $\mathbf{P}$ is reversible with respect to π, then π is a stationary measure of $\mathbf{P}$.

Conductance, total variation distance and mixing time The *conductance* of a Markov chain $\mathbf{P} = \{\mathcal{P}_x : x \in \mathcal{K}\}$ with stationary measure π supported on $\mathcal{K}$ is defined as

$$\Phi_{\mathbf{P}} = \inf_{A \in \mathcal{F}(\mathcal{K})} \frac{1}{\min\{\pi(A), 1 - \pi(A)\}} \int_{x \in A} \mathcal{P}_x(\mathcal{K} \setminus A) \cdot d\pi(x) .$$

To measure how quickly a Markov chain mixes to its stationary distribution, we use the *total variation (TV) distance*. The TV distance between two distributions μ and ν with support $\mathcal{K}$ is

$$\mathrm{d_{TV}}(\mu, \nu) = \sup_{A \in \mathcal{F}(\mathcal{K})} \mu(A) - \nu(A) .$$

Let $\mathbf{P}$ be a Markov chain with reversible distribution π. For $\delta \in (0, 1)$, the δ-*mixing time* from an initial distribution μ_0 of $\mathbf{P}$, denoted by $\tau_{\mathrm{mix}}(\delta; \mathbf{P}, \mu_0)$, is defined as

$$\tau_{\mathrm{mix}}(\delta; \mathbf{P}, \mu_0) = \inf\{k \geq 0 : \mathrm{d_{TV}}(\mathbb{T}_{\mathbf{P}}^k \mu_0, \pi) \leq \delta\} .$$

Finally, we introduce the notion of a *warm distribution*, which is useful for obtaining mixing time guarantees. A distribution μ supported on $\mathcal{K}$ is said to be M-warm with respect to another distribution Π also supported on $\mathcal{K}$ if for all $A \in \mathcal{F}(\mathcal{K})$, $\mu(A) \leq M \cdot \Pi(A)$.

3. Metropolis-adjusted Mirror Langevin algorithm

In this section, we introduce the Metropolis-adjusted Mirror Langevin algorithm (MAMLA). Let $\mathbf{P}$ be a Markov chain which defines a collection of proposal distributions at each $x \in \mathcal{K}$ with densities $\{p_x : x \in \mathcal{K}\}$. The *acceptance ratio* of z with respect to x (given a target density π) is defined as

$$p_{\mathrm{accept}}^{\mathbf{P}}(z; x) = \min\left\{1, \frac{\pi(z)p_z(x)}{\pi(x)p_x(z)}\right\} . \tag{1}$$

We recap the general outline of the Metropolis-adjustment of a Markov chain from Section 1. From a point x, the Markov chain $\mathbf{P}$ generates a proposal $z \sim \mathcal{P}_x$. This proposal z is accepted to be the next iterate with probability $p_{\mathrm{accept}}^{\mathbf{P}}(z; x)$, and if not accepted, x is retained. Let $\mathbf{T} \stackrel{\mathrm{def}}{=} \{\mathcal{T}_x : x \in \mathcal{K}\}$ denote the Metropolis-adjusted Markov chain, where $\mathcal{T}_x$ is the one-step distribution after this adjustment. As noted earlier, by construction, $\mathbf{T}$ is reversible with respect to Π with density π.

In our setting, the Markov chain $\mathbf{P}$ is induced by one step of MLA. Let ξ be an independent standard normal vector in d dimensions. From any $x \in \mathrm{int}(\mathcal{K})$, one step of MLA returns the point

$$x' = \nabla\phi^*(\nabla\phi(x) - h \cdot \nabla f(x) + \sqrt{2h \cdot \nabla^2\phi(x)}\, \xi) .$$

We consider $\mathcal{P}_x$ to be the law of x' for a given x, and $\mathbf{P} = \{\mathcal{P}_x : x \in \mathcal{K}\}$[2]. To compute the acceptance ratio (eq. (1)), we require the density of the $\mathcal{P}_x$ for each x, and this can be obtained by the change of the variable formula. Let $\mathcal{N}(x; \mu, \Sigma)$ denote the density of a multivariate normal distribution with mean μ and covariance Σ at x, and the density of $\mathcal{P}_x$ at any $z \in \mathrm{int}(\mathcal{K})$ is given by

$$p_x(z) = \mathcal{N}(\nabla\phi(z); \nabla\phi(x) - h \cdot \nabla f(x), 2h \cdot \nabla^2\phi(x)) \cdot |\det \nabla^2\phi(z)| \,. \tag{2}$$

With these definitions, the algorithm is composed of three key steps described previously. For completeness, we derive eq. (2), and state the algorithm (MAMLA) formally in Appendix B.

3.1. Mixing time analysis

Here, we state our main theorem concerning the mixing time of MAMLA, under assumptions made on both the potential f, and the mirror function ϕ. Recall that ϕ is assumed to be of Legendre type. The other key assumptions are

($\mathbf{A}_1$) ϕ is a self-concordant function with parameter α (Definition 1),

($\mathbf{A}_2$) f is μ-relatively convex and λ-relatively smooth with respect to ϕ (Definitions 2 and 3),

($\mathbf{A}_3$) f is β-relatively Lipschitz continuous with respect to ϕ (Definition 4), and

($\mathbf{A}_4$) ϕ is a symmetric barrier with parameter ν (Definition 5).

Define the constants $\bar{\alpha} = \max\{1, \alpha\}$ and $\gamma = \frac{\lambda}{2} + \alpha \cdot \beta$, which appear in the theorem below.

Theorem 6 *Consider a distribution Π with density $\pi(x) \propto e^{-f(x)}$ that is supported on a compact and convex set $\mathcal{K} \subset \mathbb{R}^d$, and mirror map $\phi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$. If f and ϕ satisfy assumptions ($\mathbf{A}_1$)-($\mathbf{A}_3$), then there exists a maximum step size $h_{\max} > 0$ given by*

$$h_{\max} = \min\left\{1, \frac{C^{(1)}}{\alpha^2 d^3}, \frac{C^{(2)}}{d\gamma}, \frac{C^{(3)}}{\alpha^{4/3}\beta^{2/3}}, \frac{C^{(4)}}{\beta^{2/3}\gamma^{2/3}}, \frac{C^{(5)}}{\beta^2}\right\} \tag{3}$$

for universal constants $C^{(1)}, \ldots, C^{(5)}$ such that for any $0 < h \leq h_{\max}$, $\delta \in (0, 1)$, and M-warm initial distribution Π_0 with respect to Π, MAMLA has the following mixing time guarantees.

When $\mu > 0$ (strongly convex),

$$\tau_{\mathrm{mix}}(\delta; \mathbf{T}, \Pi_0) = \mathcal{O}\left(\max\left\{1, \frac{\bar{\alpha}^4}{\mu \cdot h}\right\} \cdot \log\left(\frac{\sqrt{M}}{\delta}\right)\right) \,. \tag{4}$$

When $\mu = 0$ (weakly convex), *and additionally assuming that ϕ satisfies ($\mathbf{A}_4$),*

$$\tau_{\mathrm{mix}}(\delta; \mathbf{T}, \Pi_0) = \mathcal{O}\left(\max\left\{1, \frac{\nu \cdot \bar{\alpha}^2}{h}\right\} \cdot \log\left(\frac{\sqrt{M}}{\delta}\right)\right) \,. \tag{5}$$

The proof of Theorem 6 is given in Section D.2 of the Appendix.

Setting of μ in $(\mathbf{A}_2)$	MAMLA (this work)	MLA*	$\mathsf{MLA}_{\mathsf{FD}}$	$\mathsf{MLA}_{\mathsf{BD}}$
$\mu > 0$	$\widetilde{\mathcal{O}}\left(\frac{1}{\mu}\max\left\{d^3, d\lambda\right\}\log\left(\frac{1}{\delta}\right)\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\lambda^2}{\mu^3\delta^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\lambda}{\mu\delta^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{\lambda^2+d^3}}{\mu^{3/2}\delta}\right)$
$\mu = 0$	$\widetilde{\mathcal{O}}\left(\nu\max\left\{d^3, d\lambda\right\}\log\left(\frac{1}{\delta}\right)\right)$	N/A	$\widetilde{\mathcal{O}}\left(\frac{d^2\lambda}{\delta^4}\right)$	N/A

Table 1: Comparison of mixing time guarantees for algorithms based on discretisations of MLD. We use the $\widetilde{\mathcal{O}}$ notation to only showcase the dependence on λ, μ, d, and δ, and omit the dependence on other parameters. For $\mathsf{MLA}_{\mathsf{BD}}$ in Jiang (2021), μ is the constant in the mirror log-Sobolev condition, as relative strong convexity is not considered, and λ is the constant in the alternative relative smoothness condition assumed on f.

3.1.1. A DISCUSSION OF THE RESULT IN THEOREM 6

We begin by discussing the assumptions $(\mathbf{A}_2)$-$(\mathbf{A}_4)$. $(\mathbf{A}_2)$ states that f is a relatively convex and smooth function with respect to ϕ. Prior works that analyse MLA and other discretisations of MLD (Zhang et al., 2020; Ahn and Chewi, 2021; Jiang, 2021; Li et al., 2022a) consider this assumption. We note that Jiang (2021) assumes that Π satisfies a mirror log-Sobolev inequality in lieu of relative convexity of f with respect to ϕ to analyse MLA, $\mathsf{MLA}_{\mathsf{FD}}$, and another discretisation proposed in their work that we refer to as $\mathsf{MLA}_{\mathsf{BD}}$. While the usual strong convexity of f implies the log-Sobolev inequality (Bakry and Émery, 1985), it is not known if relative strong convexity of f with respect to ϕ yields a mirror log-Sobolev inequality. This makes it hard to assess whether this substitution in Jiang (2021) is a weaker assumption than relative (strong) convexity in $(\mathbf{A}_2)$. Additionally, the mixing time guarantees for MLA in Li et al. (2022a) (who work with a subset of assumptions in Zhang et al. (2020)) is only meaningful when $\mu > 0$ in $(\mathbf{A}_2)$, and a guarantee for MLA in the case where f is (weakly) convex ($\mu = 0$) is unknown still. On the other hand, Ahn and Chewi (2021) give an analysis of $\mathsf{MLA}_{\mathsf{FD}}$ for both cases i.e., $\mu > 0$ and $\mu = 0$. Next, $(\mathbf{A}_3)$ states that the gradients of f are bounded in the local norm $\|.\|_{\nabla^2\phi(.)^{-1}}$. This is used in the analysis of $\mathsf{MLA}_{\mathsf{FD}}$ and $\mathsf{MLA}_{\mathsf{BD}}$ in Ahn and Chewi (2021) and Jiang (2021) respectively, but is not used in any existing analysis of MLA. Finally, $(\mathbf{A}_4)$ is a geometric property of ϕ, and is useful to obtain guarantees in the case where $\mu = 0$ in $(\mathbf{A}_2)$. In this case, we rely on isoperimetric inequalities for sampling from log-concave densities over convex bodies (Vempala, 2005), which we get through $(\mathbf{A}_4)$. This assumption has been employed in prior work to analyse the Dikin walk (Laddha et al., 2020; Kook and Vempala, 2023), constrained RHMC (Kook et al., 2023), hybrid RHMC (Gatmiry et al., 2023).

Under these assumptions, we are interested in how the mixing time guarantees scale with the error tolerance δ, and the dimension d of $\mathcal{K}$. Table 1 summarises the comparison between MAMLA and other algorithms based on discretisations of MLD that have been discussed above. The analyses of MLA, $\mathsf{MLA}_{\mathsf{FD}}$, and $\mathsf{MLA}_{\mathsf{BD}}$ in the aforementioned works consider varying definitions of mixing time to the δ-mixing time in TV distance we define in Section 2.2, and additional assumptions which we highlight as follows. The most recent mixing time guarantees for MLA in Li et al. (2022a) is given in terms of the mirrored 2-Wasserstein distance, as was previously done in Zhang et al. (2020); this is indicated by an asterisk in Table 1. Both of these works also assume that ϕ satisfies a *modified*

2. The transition kernel for any $x \in \partial\mathcal{K}$ is analytically undefined since ϕ is of Legendre type, but this will not have any influence since the boundary $\partial\mathcal{K}$ is a Lebesgue null set as $\mathcal{K}$ is convex.

self-concordance condition instead of self-concordance as we do in $(\mathbf{A}_1)$. Furthermore, the relation between the mirrored 2-Wasserstein distance and a more canonical functional like the KL divergence ($\mathrm{d_{KL}}$) / TV distance ($\mathrm{d_{TV}}$) cannot be easily established without assuming that ϕ satisfies additional properties. For $\mathsf{MLA_{FD}}$, Ahn and Chewi (2021) establish bounds on K such that $\mathrm{d_{KL}}(\overline{\Pi}_K, \Pi) \leq \delta$. Here, $\overline{\Pi}_K$ is a uniform mixture distribution composed of the sequence of iterates $\{\Pi_k\}_{k=1}^{K}$, where Π_k is the distribution at iteration k. For $\mathsf{MLA_{BD}}$, Jiang (2021) gives upper bounds on K such that $\mathrm{d_{KL}}(\Pi_K, \Pi) \leq \delta$. To do so, they place a different relative smoothness assumption over f, and additionally assume that ϕ is strongly convex. For $\mathsf{MLA_{FD}}$ and $\mathsf{MLA_{BD}}$, we infer mixing time guarantees in TV distance from guarantees in KL divergence using Pinsker's inequality which states that $\mathrm{d_{TV}}(\mu, \nu) \leq \sqrt{\frac{1}{2}\mathrm{d_{KL}}(\mu, \nu)}$. We call MAMLA *fast* due to the dependence on δ in the mixing time guarantees ($\log(1/\delta)$), which is exponentially better than the dependence of δ in the mixing time guarantees for MLA, $\mathsf{MLA_{FD}}$, and $\mathsf{MLA_{BD}}$ ($\mathrm{poly}(1/\delta)$). This echoes the improvement observed in the mixing time guarantees for MALA relative to ULA (Dwivedi et al., 2018; Chewi et al., 2021). In contrast, the dependence on d is better in the mixing time guarantees for the unadjusted Mirror Langevin discretisations; fixing β, λ, μ, the mixing time bound for MAMLA scales as d^3 compared to d^γ with $\gamma \in \{1, 1.5, 2\}$ in the mixing time bounds for the other methods that we compare to.

4. Applications of MAMLA with provable guarantees

In this subsection, we discuss some applications of MAMLA for which we can infer mixing time guarantees from Theorem 6. These are (1) uniform sampling from polytopes and regions defined by the intersection of ellipsoids, and (2) sampling from Dirichlet distributions. The proofs of the statements given in this subsection are stated in Section D.4. We use C to denote a universal positive constant in the corollaries, which can change between corollaries.

4.1. Uniform sampling over polytopes and intersection of ellipsoids

For uniform sampling, the target density π is a constant function, and consequently $\nabla f(x) = 0$ for any $x \in \mathrm{int}(\mathcal{K})$. In this setting, MAMLA can be viewed as a Gaussian DikinWalk in the dual space $(\mathbb{R}^d, \nabla^2\phi^*)$. We are interested in approximate uniform sampling from the following sets.

- $\mathsf{Polytope}(A, b)$: a bounded polytope with non-zero volume defined by $\{x \in \mathbb{R}^d : Ax \leq b\}$ for matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, and
- $\mathsf{Ellipsoids}(\{(c_i, M_i)\}_{i=1}^m)$: a non-empty region defined by the intersection of ellipsoids $\{x \in \mathbb{R}^d : \|x - c_i\|_{M_i}^2 \leq 1 \; \forall\, i \in [m]\}$ for a sequence of $d \times d$ symmetric positive definite matrices $\{M_i\}_{i=1}^m$ and centres $\{c_i\}_{i=1}^m$. The radius of 1 does not affect the generality of this region.

The following corollary establishes mixing time guarantees for uniform sampling over these sets.

Corollary 7 *Let Π_0 be a M-warm distribution with respect to the uniform distribution over either $\mathcal{K} = \mathsf{Polytope}(A, b)$, or $\mathcal{K} = \mathsf{Ellipsoids}(\{(c_i, M_i)\}_{i=1}^m)$, and let ϕ be the log-barrier of $\mathcal{K}$. Then, for any $\delta \in (0, 1)$, the mixing time of MAMLA is*

$$C \cdot m \cdot d^3 \cdot \log\left(\frac{\sqrt{M}}{\delta}\right) .$$

Remark. For polytopes, Laddha et al. (2020) show that DikinWalk satisfies a mixing time of $m \cdot d$ owing to a *strong self-concordance* condition that holds in this setting. They also propose WeightedDikinWalk, which has a mixing time that scales as d^2 (independent of the number of constraints). In a similar vein, Gatmiry et al. (2023) propose a modification to RHMC, and prove a mixing time guarantee that scales as $m^{1/3} \cdot d^{4/3}$ for their method. Kook et al. (2023) study a constrained RHMC algorithm applicable to this setting i.e., uniform sampling from both polytopes and intersection of ellipsoids, and show a mixing time guarantee for this algorithm that scales as $m \cdot d^3$. Our analysis echoes the mixing time guarantee of the latter method and scales as $m \cdot d^3$.

4.2. Sampling from Dirichlet distributions

The Dirichlet distribution is the multi-dimensional generalisation of the Beta distribution. A sample from the Dirichlet distribution $x' \in \mathbb{R}^{d+1}$ satisfies $x'_i \geq 0$ for all $i \in [d+1]$, and $\mathbf{1}^\top x' = 1$, and thus an element of Δ_d. Equivalently, we can also express this sample with the first d elements $x \in \mathbb{R}^d_+$ satisfying an inequality constraint $\mathbf{1}^\top x \leq 1$, and write $x'_{d+1} = 1 - \mathbf{1}^\top x$. We work with the latter definition, and $\mathcal{K} = \{x \in \mathbb{R}^d_+ : \mathbf{1}^\top x \leq 1\}$. The Dirichlet distribution is parameterised by a vector of positive reals[3] $\boldsymbol{a} \in \mathbb{R}^{d+1}_+$ called the concentration parameter. The density is

$$\pi(x) \propto \exp(-f(x))\,; \qquad f(x) = -\sum_{i=1}^{d} a_i \cdot \log x_i - a_{d+1} \cdot \log\left(1 - \sum_{i=1}^{d} x_i\right)\,. \tag{6}$$

We use $\boldsymbol{a}_{\min}$ to denote the minimum of $\boldsymbol{a}$. Since the sample space is a special polytope, the log-barrier of $\mathcal{K}$ is a natural consideration for the mirror function ϕ, and with this choice of ϕ, we generate approximate samples from a Dirichlet distribution using MAMLA. The following corollary states a mixing time upper bound for this task.

Corollary 8 *Let Π_0 be a M-warm distribution with respect to a Dirichlet distribution parameterised by $\boldsymbol{a} \in \mathbb{R}^{d+1}_+$. Let ϕ be the log-barrier of $\mathcal{K}$, and f be as defined in eq. (6). If $\|\boldsymbol{a}\| \geq 1$, then for any $\delta \in (0,1)$, the mixing time of MAMLA is*

$$C \cdot \frac{\max\{d^3,\ d \cdot \|\boldsymbol{a}\|,\ \|\boldsymbol{a}\|^2\}}{\boldsymbol{a}_{\min}} \cdot \log\left(\frac{\sqrt{M}}{\delta}\right)\,.$$

Remark. In comparison, a corollary of the guarantee for MLA as shown in Li et al. (2022a) gives a mixing time guarantee that scales as $\frac{d\boldsymbol{a}^2_{\max}}{\boldsymbol{a}^3_{\min}} \cdot \frac{1}{\delta^2}$ for this task. It is worth noting that this holds only when the *modified self-concordance* parameter used in its analysis is at most $\boldsymbol{a}_{\min}$, and that this is in the mirrored 2-Wasserstein distance. $\mathsf{MLA}_{\mathsf{FD}}$ (Ahn and Chewi, 2021) on the other hand satisfies a mixing time guarantee that scales as $\frac{d\|\boldsymbol{a}\|}{\boldsymbol{a}_{\min}} \cdot \frac{1}{\delta^2}$. We cannot establish mixing time guarantees for $\mathsf{MLA}_{\mathsf{BD}}$ in this case due to the use of a mirror log-Sobolev inequality instead of relative strong convexity, and this former condition is hard to verify in general.

5. Conclusion

To summarise, we introduce the Metropolis-adjusted Mirror Langevin algorithm (MAMLA), and provide non-asymptotic mixing time guarantees for it. This algorithm adds a Metropolis-Hastings

3. More generally, $a_i > -1$, but we focus on when $a_i > 0$. The case where $a_i \in (-1, 0)$ results in antimodes.

accept-reject filter to the proposal Markov chain defined by a single step of the Mirror Langevin algorithm (MLA) which is the Euler-Maruyama discretisation of the Mirror Langevin dynamics (MLD), at each point in $\mathcal{K}$. The resulting Markov chain is reversible with respect to the target distribution that we seek to sample from, unlike the Markov chain induced by MLA.

Our mixing time guarantees are strongly related to the maximum permissible stepsize that results in non-vanishing acceptance rates, and our current analysis shows that this scales as $1/d^3$ as d grows. We believe that for structured domains like polytopes where ϕ is chosen to be the log-barrier of the domain, a step size that scales as $1/d^\gamma$ for $\gamma < 3$ can lead to non-vanishing acceptance rates for MAMLA, and consequently yield better mixing time guarantees. Another question that we leave for the future is to check if the relative Lipschitz continuity assumption over f can be relaxed or removed altogether for the mixing time analysis of MAMLA, especially when used in conjuction with relative convexity and smoothness of f. This is motivated by the analysis in Li et al. (2022a), which does not employ such a condition to derive mixing time guarantees for MLA.

Acknowledgments

We would like to thank the reviewers for their feedback, and Sinho Chewi for helpful remarks.

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. In *Advances in Neural Information Processing Systems*, volume 34, pages 28405–28418. Curran Associates, Inc., 2021.

D. Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, Berlin, 1985.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.*, 30(3):1209–1250, 2020.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo 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.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59:757–783, 2018.

Gilles Celeux, Mohammed El Anbari, Jean-Michel Marin, and Christian P. Robert. Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation. *Bayesian Analysis*, 7(2):477 – 502, 2012.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast MCMC Sampling Algorithms on Polytopes. *Journal of Machine Learning Research*, 19(55):1–86, 2018.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 2018.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-Langevin diffusions. In *Advances in Neural Information Processing Systems*, volume 33, pages 19573–19585. Curran Associates, Inc., 2020.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1260–1300. PMLR, 2021.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR, 2017a.

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

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

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the Unadjusted Langevin Algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1705.00166*, 2017.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797. PMLR, 2018.

Khashayar Gatmiry and Santosh S Vempala. Convergence of the Riemannian Langevin Algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

Khashayar Gatmiry, Jonathan Kelner, and Santosh S Vempala. Sampling with Barriers: Faster Mixing via Lewis Weights. *arXiv preprint arXiv:2303.00480*, 2023.

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

Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Algorithmic Aspects of the Log-laplace Transform and a Non-Euclidean Proximal Sampler. In *Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2399–2439. PMLR, 2023.

Mert Gürbüzbalaban, Yuanhan Hu, and Lingjiong Zhu. Penalized Langevin and Hamiltonian Monte Carlo Algorithms for Constrained Sampling. *arXiv preprint arXiv:2212.00570*, 2022.

Adam Gustafson and Hariharan Narayanan. John's walk. *arXiv preprint arXiv:1803.02032*, 2018.

Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. *Nature Protocols*, 14(3):639–702, 2019.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin Dynamics. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Qijia Jiang. Mirror Langevin Monte Carlo: the Case Under Isoperimetry. In *Advances in Neural Information Processing Systems*, volume 34, pages 715–725. Curran Associates, Inc., 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

Ravi Kannan and Hariharan Narayanan. Random Walks on Polytopes and an Affine Interior Point Method for Linear Programming. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, pages 561–570. Association for Computing Machinery, 2009. ISBN 9781605585062.

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

Yunbum Kook and Santosh S Vempala. Gaussian Cooling and Dikin Walks: The interior-point method for Logconcave Sampling. *arXiv preprint arXiv:2307.12943*, 2023.

Yunbum Kook, Yin-Tat Lee, Ruoqi Shen, and Santosh Vempala. Sampling with Riemannian Hamiltonian Monte Carlo in a Constrained Space. In *Advances in Neural Information Processing Systems*, volume 35, pages 31684–31696. Curran Associates, Inc., 2022.

Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh Vempala. Condition-number-independent Convergence Rate of Riemannian Hamiltonian Monte Carlo with Numerical Integrators. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4504–4569. PMLR, 2023.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. Gradient-based Constrained Sampling from Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, 2022.

Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Strong Self-Concordance and Sampling. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pages 1212–1222. Association for Computing Machinery, 2020.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

Yin Tat Lee and Santosh S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 1115–1121. Association for Computing Machinery, 2018.

Mufan Li and Murat A. Erdogdu. Riemannian Langevin Algorithm for Solving Semidefinite Programs. *Bernoulli*, 29(4):3093–3113, 2023.

Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. The Mirror Langevin Algorithm Converges with Vanishing Bias. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 718–742. PMLR, 2022a.

Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt (d) dimension dependence of Langevin Monte Carlo. In *The International Conference on Learning Representations*, 2022b.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

László Lovász. Hit-and-run mixes fast. *Mathematical programming*, 86:443–461, 1999.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

László Lovász and Santosh Vempala. Hit-and-run is fast and fun. *preprint, Microsoft Research*, 2003.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively-Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Oren Mangoubi and Nisheeth K. Vishnoi. Sampling from Structured Log-Concave Distributions via a Soft-Threshold Dikin Walk. In *Advances in Neural Information Processing Systems*, volume 36, pages 31908–31942. Curran Associates, Inc., 2023.

Hariharan Narayanan. Randomized Interior Point methods for Sampling and Optimization. *Ann. Appl. Probab.*, 26(1):597–641, 2016.

Hariharan Narayanan and Alexander Rakhlin. Efficient Sampling from Time-Varying Log-Concave Distributions. *The Journal of Machine Learning Research*, 18(1):4017–4045, 2017.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, A Wiley-Interscience Publication.

Yu. E. Nesterov and M. J. Todd. On the Riemannian Geometry Defined by Self-Concordant Barriers and Interior-Point Methods. *Found. Comput. Math.*, 2(4):333–361, 2002.

Yuri Nesterov and Arkadi Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, second edition, 2018.

Maxence Noble, Valentin De Bortoli, and Alain Durmus. Unbiased constrained sampling with Self-Concordant Barrier Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 36, pages 32672–32719. Curran Associates, Inc., 2023.

Ari Pakman and Liam Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *J. Comput. Graph. Statist.*, 23(2):518–542, 2014.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

R. Tyrrell Rockafellar. *Convex Analysis*, volume No. 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.

Sushant Sachdeva and Nisheeth K Vishnoi. The Mixing Time of the Dikin Walk in a Polytope — A Simple Proof. *Operations Research Letters*, 44(5):630–634, 2016.

Robert L Smith. Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research*, 32(6):1296–1308, 1984.

Santosh Vempala. Geometric Random Walks: A Survey. *Combinatorial and computational geometry*, 52(573-612):2, 2005.

Santosh Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 2018.

Andre Wibisono, Molei Tao, and Georgios Piliouras. Alternating Mirror Descent for Constrained Min-Max Games. In *Advances in Neural Information Processing Systems*, volume 35, pages 35201–35212. Curran Associates, Inc., 2022.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein Control of Mirror Langevin Monte Carlo. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3814–3841. PMLR, 2020.

Appendices

Organisation In Appendix A, we provide an extended discussion about the function classes that were previously discussed in Section 2. In Appendix B, we give the complete pseudocode for MAMLA, and include a derivation of the density of the proposal distribution (eq. (2)) defined by a single step of MLA. We provide a short discussion of the Newton Langevin dynamics and its Euler-Maruyama discretisation, the Newton Langevin algorithm, and provide mixing time guarantees for the Metropolis-adjusted Newton Langevin algorithm in Appendix C. Finally, we give the complete proofs of Theorem 6, and its corollaries stated in Section 4 in Appendix D, and some miscellaneous mathematical facts used in these proofs in Appendix E.

Appendix A. Extended discussion about the function classes

Legendre Type Formally, a function of Legendre type is defined as follows.

Definition 9 (Rockafellar (1970, Chap. 26)) *Let $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a proper, lower semicontinuous, convex function. ψ is of Legendre type if*

1. *The set $\mathrm{int}(\mathcal{K})$ is non-empty,*
2. *ψ is differentiable and strictly convex in $\mathrm{int}(\mathcal{K})$, and*
3. *For any sequence $\{x_k\}_{k\in\mathbb{N}}$ in $\mathrm{int}(\mathcal{K})$ whose limit lies on $\partial\mathcal{K}$, $\lim_{k\to\infty} \|\nabla\phi(x_k)\| \to \infty$.*

A naive approach to compute the inverse of the mirror map is by inverting the mirror map exactly. However, when the mirror function is of Legendre type, there are two other approaches by which this can be computed. The first is by computing the convex conjugate in closed form, and taking its gradient at the query point, which yields the inverse of the mirror map by definition. The second approach is an approximate method, and involves solving the optimisation problem in the definition of the convex conjugate. At optimality, the solution $\hat{x}$ satisfies $y = \nabla\phi(\hat{x}) \Leftrightarrow \hat{x} = (\nabla\phi)^{-1}(y)$.

Relative convexity and smoothness Let $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a differentiable convex function. The Bregman divergence of ψ at y with respect to x is defined as $D_\psi(y; x) = \psi(y) - \psi(x) - \langle\nabla\psi(x), y - x\rangle$. The μ-relative convexity and λ-relative smoothness of a function g with respect to ψ imply respectively that for any $x, y \in \mathrm{int}(\mathcal{K})$,

$$g(y) \geq g(x) + \langle\nabla g(x), y - x\rangle + \mu \cdot D_\psi(y; x) \ ,$$
$$g(y) \leq g(x) + \langle\nabla g(x), y - x\rangle + \lambda \cdot D_\psi(y; x) \ .$$

When $\psi(x) = \frac{\|x\|^2}{2}$, and $\mathcal{K} = \mathbb{R}^d$, $D_\psi(y; x) = \frac{\|y-x\|^2}{2}$ for any $x, y \in \mathbb{R}^d$, and substituting the equation for D_ψ in the inequalities above recovers the standard first order definitions of convexity and smoothness (Nesterov, 2018, §2.1).

Relative Lipschitz continuity Let $g : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a differentiable function, and $\psi : \mathcal{K} \to \mathbb{R}\cup\{\infty\}$ be a twice differentiable, strictly convex function. If g is β-relatively Lipschitz continuous with respect to ψ, then for any $x \in \mathrm{int}(\mathcal{K})$,

$$\|\nabla g(x)\|_{\nabla^2\psi(x)^{-1}} \leq \beta \ .$$

When $\psi(x) = \frac{\|x\|^2}{2}$, and $\mathcal{K} = \mathbb{R}^d$, then $\nabla^2\psi(x) = I$ for all $x \in \mathbb{R}^d$, and consequently this relative Lipschitz continuity condition reads

$$\|\nabla g(x)\| \leq \beta \qquad \forall\, x \in \mathbb{R}^d\,,$$

which is an equivalent characterisation of standard Lipschitz continuity of g.

A special case to make note of is when $\psi = g$, and a function g that satisfies $\|\nabla g(x)\|_{\nabla^2 g(x)^{-1}} \leq \beta$ for all $x \in \mathrm{int}(\mathcal{K})$ is termed a *barrier function* (Nesterov, 2018, §5.3.2). This property is useful in the analysis of Newton's method in optimisation. In this work, we use this property to give guarantees for the Metropolis-adjusted Newton Langevin algorithm (Corollary 12).

Symmetric barriers The logarithmic barriers of $\mathsf{Polytope}(A, b)$ and $\mathsf{Ellipsoids}(\{(c_i, M_i)\}_{i=1}^m)$ are symmetric barriers, which is useful to providing guarantees for approximate uniform sampling over these domains. The following two lemmas formally state this.

Lemma 10 *Let $\mathcal{K} = \mathsf{Polytope}(A, b)$ where $A \in \mathbb{R}^{m\times d}, b \in \mathbb{R}^m$ be a non-empty, bounded polytope, and ϕ be the log-barrier of $\mathcal{K}$. Then, ϕ is a symmetric barrier with parameter m.*

Lemma 11 *Let $\mathcal{K} = \mathsf{Ellipsoids}(\{M_i\}_{i=1}^m)$ for a sequence of positive definite matrices $\{M_i\}_{i=1}^m$, and ϕ be the log-barrier of $\mathcal{K}$. Then, ϕ is a symmetric barrier with parameter $2m$.*

Lemma 10 was stated as Kook and Vempala (2023, Lem. 5.5). To the best of our knowledge, Lemma 11 has not been discussed in any prior work, and we give a proof of this lemma in Section D.5.1.

Appendix B. Additional details about the Metropolis-adjusted Mirror Langevin algorithm

We begin by giving the formal definition of the Metropolis-adjusted Mirror Langevin algorithm (MAMLA), as introduced in Section 3.

In eq. (2), we gave the formula for the density of the proposal distribution. This can be derived via the change of variable formula as shown below. Recall that

$$\mathcal{P}_x = (\nabla\phi^*)_\#\mathcal{N}(.; \nabla\phi(x) - h\cdot\nabla f(x), 2h\cdot\nabla^2\phi(x))\,,$$

where $\mathcal{N}(.; \mu, \Sigma)$ is the multivariate normal distribution with mean μ and Σ. The change of variable formula states that for a probability distribution μ with a density, and a differentiable invertible map T,

$$dT_\#\mu(x) = d\mu(T^{-1}(x))|\det JT^{-1}(x)|\,,$$

where $JT^{-1}(x)$ is the Jacobian of T^{-1} evaluated at x. Since ϕ is assumed to be of Legendre type, $\nabla\phi$ (and equivalently, $\nabla\phi^*$) is an invertible map. Hence,

$$\begin{aligned} p_x(z) &= \mathcal{N}((\nabla\phi^*)^{-1}(z); \nabla\phi(x) - h\cdot\nabla f(x), 2h\cdot\nabla^2\phi(x))\cdot|\det\nabla(\nabla\phi^*)^{-1}(z)| \\ &= \frac{\det\nabla^2\phi(z)}{(4h\pi)^{d/2}\cdot\sqrt{\det\nabla^2\phi(x)}}\exp\left(-\frac{\|\nabla\phi(z) - \nabla\phi(x) + h\cdot\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}}}{4h}\right)\,. \end{aligned}$$

Algorithm 1: Metropolis-adjusted Mirror Langevin algorithm (MAMLA)

Input : Potential f of Π, mirror function ϕ, iterations K, initial distribution Π_0, step size $h > 0$

Sample $x_0 \sim \Pi_0$.

for $k \leftarrow 0$ **to** $K - 1$ **do**

  Sample a random vector $\xi_k \sim \mathcal{N}(0, I)$.

  Generate proposal $z = \nabla\phi^*(\nabla\phi(x_k) - h \cdot \nabla f(x_k) + \sqrt{2h \cdot \nabla^2\phi(x_k)}\ \xi_k)$.

  Compute acceptance ratio $p_{\text{accept}}(z; x_k) = \min\left\{1, \frac{\pi(z)p_z(x_k)}{\pi(x_k)p_{x_k}(z)}\right\}$ using eq. (2).

  Obtain $U \sim \text{Unif}([0, 1])$.

  if $U \leq p_{\text{accept}}(z; x_k)$ **then**

    Set $x_{k+1} = z$.

  else

    Set $x_{k+1} = x_k$.

  end

end

Output: x_K

Appendix C. Discussion on the Newton Langevin dynamics and algorithm

As remarked in the introduction, a special case of the Mirror Langevin dynamics and algorithm is when $\phi = f$, which results in the Newton Langevin dynamics (NLD) and the Newton Langevin algorithm (NLA) respectively.

$$Y_t = \nabla f(X_t); \quad dY_t = -\nabla f(X_t)dt + \sqrt{2}\ \nabla^2 f(X_t)^{1/2}\ dB_t\ . \tag{NLD}$$

$$\begin{cases} Y_k = \nabla f(X_k)\ ; \\ Y_{k+1} - Y_k = -h \cdot \nabla f(X_k) + \sqrt{2h} \cdot \nabla^2 f(X_k)^{1/2}\ \xi_k\ , \qquad \xi_k \sim \mathcal{N}(0, I_d)\ ; \\ X_{k+1} = (\nabla f)^{-1}(Y_{k+1})\ . \end{cases} \tag{NLA}$$

Chewi et al. (2020) showed that when f is strictly convex over $\mathcal{K}$, NLD converges exponentially quickly to the target in χ^2-divergence at a rate that is invariant to affine transformations of $\mathcal{K}$. This result is due to the Brascamp-Lieb inequality. Hence, it is reasonable to expect that the Newton Langevin algorithm (NLA), which is the specialisation of the Mirror Langevin algorithm (MLA) in this case, also has mixing time guarantees that are invariant to affine transformations. But unfortunately, the most recent analysis of this method in Li et al. (2022a) cannot show this, owing to the modified self-concordance assumption as elaborated in section 1. Despite this, a special case of our analysis yields a guarantee for the Metropolis-adjusted Newton Langevin algorithm that is invariant to affine transformations of $\mathcal{K}$, which we discuss next.

C.1. Mixing time for the Metropolis-adjusted Newton Langevin algorithm

This setting satisfies assumption ($\mathbf{A}_2$) with constants $\mu = \lambda = 1$, and when f is self-concordant with parameter α, and a barrier function with parameter β, assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_3$) are satisfied as well. This leads to a corollary of Theorem 6 for the Metropolis-adjusted Newton Langevin algorithm, which is also unbiased with respect to the target, as stated in the following corollary.

Corollary 12 *Consider a distribution Π with density $\pi(x) \propto e^{-f(x)}$ that is supported on compact and convex set $\mathcal{K} \subset \mathbb{R}^d$. When f is a self-concordant function with parameter α and a barrier function with parameter β, there exists a maximum step size $h_{\max} > 0$ of the form in eq. (3) where $\gamma = \frac{1}{2} + \alpha \cdot \beta$, such that for any $0 < h \leq h_{\max}$, $\delta \in (0, 1)$, and M-warm initial distribution Π_0 with respect to Π, MAMLA with $\phi = f$ satisfies*

$$\tau_{\mathrm{mix}}(\delta; \mathbf{T}, \Pi_0) = \mathcal{O}\left(\max\left\{1, \frac{\bar{\alpha}^4}{h}\right\} \cdot \log\left(\frac{\sqrt{M}}{\delta}\right)\right) .$$

Appendix D. Proofs

This section is devoted providing the proofs of the main theorem and its corollaries in Section 3. In Section D.1, we give the main lemmas that form the proof of the Theorem 6, whose proof follows in Section D.2. The proofs of these main lemmas are given in Section D.3, and finally in Section D.4, we give the proofs of the corollaries from Section 4.

D.1. Results for conductance using one-step overlap

The following classical result by Lovász and Simonovits (1993) is the basis of the mixing time guarantee of MAMLA, which is a common tool for studying reversible Markov chains.

Proposition 13 *Let $\mathbf{Q}$ be a lazy, reversible Markov chain over $\mathcal{K}$ with stationary distribution Π and conductance $\Phi_{\mathbf{Q}}$. For any Π_0 that is M-warm with respect to Π, we have for all $k \geq 1$ that*

$$\mathrm{d}_{\mathrm{TV}}(\mathbb{T}_{\boldsymbol{Q}}^k \Pi_0, \Pi) \leq \sqrt{M}\left(1 - \frac{\Phi_{\mathbf{Q}}^2}{2}\right)^k .$$

The above result indicates that it suffices to show a non-vacuous lower bound on the conductance of a reversible Markov chain to derive mixing time upper bounds for it. To do so, we use the *one-step overlap* technique developed in Lovász (1999). The gist of this technique is to show that when given a reversible Markov chain $\mathbf{Q} = \{\mathcal{Q}_x : x \in \mathcal{K}\}$, the distance $\mathrm{d}_{\mathrm{TV}}(\mathbb{T}_{\boldsymbol{Q}}(\delta_x), \mathbb{T}_{\boldsymbol{Q}}(\delta_y))$ (also equal to $\mathrm{d}_{\mathrm{TV}}(\mathcal{Q}_x, \mathcal{Q}_y)$) is uniformly bounded away from 1 for any $x, y \in \mathrm{int}(\mathcal{K})$ that are close. For MAMLA, recall that $\mathbf{T} = \{\mathcal{T}_x : x \in \mathcal{K}\}$ denotes the Markov chain induced by a single iteration of the algorithm, and $\mathcal{P}_x$ is proposal distribution defined whose density is given in eq. (2).

Lemmas 14 and 15 precisely state how a one-step overlap can yield lower bounds on the conductance of $\mathbf{T}$, whose stationary distribution has density $\pi \propto e^{-f}$. These are two separate lemmas, one for when f is μ-strongly convex relative to a self-concordant mirror function ϕ with $\mu > 0$, and the other for when the potential f is simply convex i.e., when $\mu = 0$.

Lemma 14 *Let $\phi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a self-concordant function with parameter α, and let f be μ-relatively convex with respect to ϕ. Assume that for any $x, y \in \mathrm{int}(\mathcal{K})$, if $\|x - y\|_{\nabla^2\phi(y)} \leq \Delta$ where $\Delta \leq \frac{1}{2\alpha}$, then $\mathrm{d}_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{T}_y) \leq \frac{1}{4}$. Then, for any measurable partition $\{A_1, A_2\}$ of $\mathcal{K}$,*

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \geq \frac{3}{16} \cdot \min\left\{1, \frac{\sqrt{\mu} \cdot \Delta}{2 \cdot (8\alpha + 4)}\right\} \cdot \min\{\Pi(A_1), \Pi(A_2)\} .$$

Lemma 15 *Let $\phi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a symmetric barrier with parameter ν, and let $f : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a convex function. Assume that for any $x, y \in \mathrm{int}(\mathcal{K})$, if $\|x - y\|_{\nabla^2\phi(y)} \leq \Delta$, then $\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{T}_y) \leq \frac{1}{4}$. Then, for any measurable partition $\{A_1, A_2\}$ of $\mathcal{K}$,*

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \geq \frac{3}{16} \cdot \min\left\{1, \frac{1}{8} \cdot \frac{\Delta}{\sqrt{\nu}}\right\} \cdot \min\{\Pi(A_1), \Pi(A_2)\} .$$

These lemmas are based on two different isoperimetric inequalities. Specifically, Lemma 14 relies on an isoperimetric inequality for distributions whose potentials f are relatively convex with respect to a self-concordant function, and Lemma 15 uses another isoperimetric inequality for log-concave distributions supported on bounded sets from Lovász and Vempala (2003) in conjunction with the notion of symmetric barriers to arrive at the result.

Next, we establish the one-step overlap that we have assumed in the aforementioned lemmas. The triangle inequality for the TV distance gives for any $x, y \in \mathrm{int}(\mathcal{K})$

$$\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{T}_y) \leq \mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{P}_x) + \mathrm{d_{TV}}(\mathcal{P}_x, \mathcal{P}_y) + \mathrm{d_{TV}}(\mathcal{P}_y, \mathcal{T}_y) .$$

The goal is to show that $\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{T}_y)$ is sufficiently bounded away from 1, and one method to achieve this is by showing that each of the three quantities above are also sufficiently bounded away from 1. In the following lemma, we give bounds for each of these quantities. First, we define the function $b : [0, 1) \to (0, 1)$ as

$$b(\varepsilon; d, \alpha, \beta, \lambda) = \min\left\{1, \frac{\mathcal{C}_1(\varepsilon)}{\alpha^2 \cdot d^3}, \frac{\mathcal{C}_2(\varepsilon)}{d \cdot \gamma}, \frac{\mathcal{C}_3(\varepsilon)}{\beta^{2/3} \cdot \alpha^{4/3}}, \frac{\mathcal{C}_4(\varepsilon)}{\beta^{2/3} \cdot \gamma^{2/3}}, \frac{\varepsilon}{4 \cdot \beta^2}\right\} , \tag{7}$$

where $\mathcal{C}_1(\varepsilon), \dots, \mathcal{C}_4(\varepsilon)$ are functions of ε defined in eq. (12) later.

Lemma 16 *Let ϕ and f satisfy (A_1), (A_2), and (A_3). Then, the following statements hold.*

1. *For any two points $x, y \in \mathrm{int}(\mathcal{K})$ such that $\|x - y\|_{\nabla^2\phi(y)} \leq \frac{\sqrt{h}}{10 \cdot \bar{\alpha}}$ and any step size $h \in (0, 1)$,*

$$\mathrm{d_{TV}}(\mathcal{P}_x, \mathcal{P}_y) \leq \frac{1}{2}\sqrt{\frac{h \cdot d}{20} + \frac{1}{80} + \frac{9}{2} \cdot h \cdot \beta^2} .$$

2. *For any $x \in \mathrm{int}(\mathcal{K})$, $\varepsilon < 1$, and step size $h \in (0, b(\varepsilon; d, \alpha, \beta, \lambda)]$.*

$$\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{P}_x) \leq \frac{\varepsilon}{2} .$$

where $\bar{\alpha} = \max\{1, \alpha\}$, $\gamma = \frac{\lambda}{2} + \alpha \cdot \beta$.

The proofs of Lemmas 14 to 16 are given in Section D.3. With these results, we proceed to the proof of Theorem 6.

D.2. Complete proof of Theorem 6

Proof Let $h \in (0, b(\varepsilon; d, \alpha, \beta, \lambda)] \subseteq (0, 1]$. From Lemma 16, we have that for any $x, y \in \mathrm{int}(\mathcal{K})$ such that $\|x - y\|_{\nabla^2\phi(y)} \le \frac{\sqrt{h}}{10\cdot\bar{\alpha}}$,

$$\mathrm{d_{TV}}(\mathcal{P}_x, \mathcal{P}_y) \le \frac{1}{2}\sqrt{\frac{h\cdot d}{20} + \frac{1}{80} + \frac{9}{2}\cdot h\cdot\beta^2} \le \frac{1}{2}\sqrt{\frac{1}{20} + \frac{1}{80} + \frac{9}{8}\cdot\varepsilon}\,.$$

For $\varepsilon = 1/24$,

$$\mathrm{d_{TV}}(\mathcal{P}_x, \mathcal{P}_y) \le \frac{1}{2}\sqrt{\frac{1}{20} + \frac{1}{80} + \frac{3}{64}} \le \frac{1}{6}$$

$$\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{P}_x)\,, \mathrm{d_{TV}}(\mathcal{T}_y, \mathcal{P}_y) \le \frac{1}{48}$$

Consequently, $\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{T}_y) \le \frac{1}{24} + \frac{1}{6} \le \frac{1}{4}$.
Let $\{A_1, A_2\}$ be an arbitrary measurable partition of $\mathcal{K}$. When f is μ-strongly convex with respect to ϕ such that $\mu > 0$, we invoke Lemma 14 with $\Delta = \frac{\sqrt{h}}{10\cdot\bar{\alpha}}$ to obtain

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \ge \frac{3}{16}\cdot\min\left\{1, \frac{C_\alpha\cdot\sqrt{\mu}\cdot\sqrt{h}}{20\cdot\bar{\alpha}}\right\}\cdot\min\{\Pi(A_1), \Pi(A_2)\}\,; \quad C_\alpha = \frac{1}{8\alpha + 4}\,.$$

Consequently,

$$\Phi_{\mathbf{T}} \ge \frac{3}{16}\cdot\min\left\{1, \frac{C_\alpha\cdot\sqrt{\mu}\cdot\sqrt{h}}{20\cdot\bar{\alpha}}\right\}\,.$$

Instead, when $\mu = 0$ i.e., when f is convex, the additional assumption about ϕ being a symmetric barrier with parameter ν allows us to use Lemma 15 with $\Delta = \frac{\sqrt{h}}{10\cdot\bar{\alpha}}$ to get

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \ge \frac{3}{16}\cdot\min\left\{1, \frac{\sqrt{h}}{80\cdot\sqrt{\nu}\cdot\bar{\alpha}}\right\}\cdot\min\{\Pi(A_1), \Pi(A_2)\}\,,$$

and as a result,

$$\Phi_{\mathbf{T}} \ge \frac{3}{16}\cdot\min\left\{1, \frac{\sqrt{h}}{80\cdot\sqrt{\nu}\cdot\bar{\alpha}}\right\}\,.$$

Since Π_0 is M-warm with respect to Π, we instantiate Proposition 13, which provides bounds on the mixing time for any $\delta \in (0, 1)$ in the two cases discussed above. To be precise,

$$\mathrm{d_{TV}}(\mathbb{T}^k_{\mathbf{T}}\Pi_0, \Pi) \le \sqrt{M}\cdot\left(1 - \frac{\Phi^2_{\mathbf{T}}}{2}\right)^k \le \sqrt{M}\cdot\exp\left(-k\cdot\frac{\Phi^2_{\mathbf{T}}}{2}\right)\,.$$

When $k = \frac{2}{\Phi^2_{\mathbf{T}}}\log\left(\frac{\sqrt{M}}{\delta}\right)$, $\mathrm{d_{TV}}(\mathbb{T}^k_{\mathbf{T}}\Pi_0, \Pi) \le \delta$. Thus,

when $\mu > 0$

$$\tau_{\mathrm{mix}}(\delta; \mathbf{T}, \Pi_0) = \mathcal{O}\left(\max\left\{1, \frac{\bar{\alpha}^4}{\mu\cdot h}\right\}\cdot\log\left(\frac{\sqrt{M}}{\delta}\right)\right)\,, \text{ and}$$

when $\mu = 0$

$$\tau_{\mathrm{mix}}(\delta; \mathbf{T}, \Pi_0) = \mathcal{O}\left(\max\left\{1, \frac{\bar{\alpha}^2 \cdot \nu}{h}\right\} \cdot \log\left(\frac{\sqrt{M}}{\delta}\right)\right) .$$

■

D.3. Proofs of conductance results and one-step overlap in Section D.1

Prior to presenting the the proofs of Lemmas 14 and 15, we first introduce and elaborate on two specific isoperimetric lemmas which we use in the proofs, which follow after this introduction.

The proof of Lemma 14 uses an isoperimetric lemma for a distribution supported on a compact, convex subset $\mathcal{K} \subset \mathbb{R}^d$ with density $\pi \propto e^{-f}$ where f is relatively strongly convex with respect to a self-concordant function ψ with domain $\mathcal{K}$. This lemma is due to Gopi et al. (2023); their original lemma assumes the self-concordance parameter of ψ to be 1 while we generalise it for ψ with an arbitrary self-concordance parameter α. Additionally, we have the following metric $d_\psi : \mathcal{K} \times \mathcal{K} \to \mathbb{R}_+ \cup \{0\}$. For any two points $x, y \in \mathcal{K}$, $d_\psi(x, y)$ is the Riemannian distance between x, y as measured with respect to the Hessian metric $\nabla^2\psi$ over $\mathcal{K}$. We overload this notation to take sets as arguments; for two disjoint sets A, B of $\mathcal{K}$, $d_\psi(A, B)$ is defined as $d_\psi(A, B) = \inf_{x \in A, y \in B} d_\psi(x, y)$. We state the lemma due to Gopi et al. (2023) as stated in Kook and Vempala (2023), which addresses the subtlety of f only taking finite values in $\mathrm{int}(\mathcal{K})$, and also highlight the dependence on the self-concordance parameter.

Proposition 17 (Kook and Vempala (2023, Lemma 2.7)) *Let $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a self-concordant function with parameter α, and let f be a μ-relatively convex function with respect to ψ. Then, for any given partition $\{A_1, A_2, A_3\}$ of $\mathcal{K}$,*

$$\Pi(A_3) \geq C_\alpha \cdot \sqrt{\mu} \cdot d_\psi(A_1, A_2) \cdot \min\{\Pi(A_1), \Pi(A_2)\}$$

for positive constant $C_\alpha = \frac{1}{8\alpha+4}$.

To operationalise this in order to obtain lower bounds on the conductance using the one-step overlap which uses a closeness criterion in terms of $\|x - y\|_{\nabla^2\phi(y)}$, we require a relation between $d_\phi(x, y)$ and $\|x - y\|_{\nabla^2\phi(y)}$. The self-concordance property of ϕ ensures that the metric $\nabla^2\phi$ is relatively stable between two points that are close, and thus implying that the Riemannian distance between x, y is approximately the local norm $\|x - y\|_{\nabla^2\phi(y)}$. The following lemma from Nesterov and Todd (2002) formalises this intuition. The lemma below is a slight modification of their original lemma that makes the dependence on the self-concordance parameter explicit.

Proposition 18 (Nesterov and Todd (2002, Lemma 3.1)) *Let $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a convex and self-concordant with parameter α. For any $\Delta \in [0, 1/\alpha)$, if for any $x, y \in \mathrm{int}(\mathcal{K})$ it holds that $d_\psi(x, y) \leq \Delta - \alpha \cdot \frac{\Delta^2}{2}$, then $\|x - y\|_{\nabla^2\psi(x)} \leq \Delta$.*

Next, for the proof of Lemma 15, we use an isoperimetric inequality for a log-concave distribution Π from Lovász and Vempala (2003). Before we state this, we introduce the notion of the cross ratio between two points in $\mathcal{K}$, which is compact. Let $x, y \in \mathrm{int}(\mathcal{K})$, and consider a line segment between x and y. Let the endpoints of this extension of the line segment to the boundary of $\mathcal{K}$ be p and q respectively (thus forming a chord), with points in the order p, x, y, q. The cross ratio between x and y (with respect to $\mathcal{K}$) is

$$\mathrm{CR}(x, y; \mathcal{K}) = \frac{\|y - x\| \cdot \|q - p\|}{\|y - q\| \cdot \|p - x\|} .$$

We overload this notation to define the cross ratio between sets as

$$\mathrm{CR}(A_2, A_1; \mathcal{K}) = \inf_{y \in A_2,\, x \in A_1} \mathrm{CR}(y, x; \mathcal{K}) .$$

Proposition 19 (Lovász and Vempala (2003, Thm. 2.2)) *Let Π be a log-concave distribution on $\mathbb{R}^d$ with support $\mathcal{K}$. For any measurable partition $\{A_1, A_2, A_3\}$ of $\mathcal{K}$,*

$$\Pi(A_3) \geq \mathrm{CR}(A_2, A_1; \mathcal{K}) \cdot \Pi(A_2) \cdot \Pi(A_1) .$$

As seen previously, to operationalise this isoperimetric inequality, we have to relate the cross ratio between two points to the local norm $\|x - y\|_{\nabla^2\phi(y)}$. When ϕ is a symmetric barrier, this is possible due to the following lemma.

Proposition 20 (Laddha et al. (2020, Lem. 2.3)) *Let $\psi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a symmetric barrier with parameter ν. Then, for any $x, y \in \mathrm{int}(\mathcal{K})$, $\mathrm{CR}(y, x; \mathcal{K}) \geq \frac{\|y-x\|_{\nabla^2\psi(y)}}{\sqrt{\nu}}$.*

With all of these ingredients, we give the proofs of Lemmas 14 and 15. The proofs of both lemmas only differ in the isoperimetric inequalities used, and hence are presented together.

D.3.1. PROOFS OF LEMMAS 14 AND 15

Proof By the reversibility of the Markov chain $\mathbf{T}$, we have for any partition $\{A_1, A_2\}$ of $\mathcal{K}$ that

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx = \int_{A_2} \mathcal{T}_x(A_1)\pi(x)dx . \tag{8}$$

Define the subsets A_1' and A_2' of A_1 and A_2 respectively as

$$A_1' = \left\{x \in A_1 \;:\; \mathcal{T}_x(A_2) < \frac{3}{8}\right\}; \qquad A_2' = \left\{x \in A_2 \;:\; \mathcal{T}_x(A_1) < \frac{3}{8}\right\} .$$

We consider two cases, which cover all possibilities

Case 1 : $\Pi(A_1') \leq \frac{\Pi(A_1)}{2}$ or $\Pi(A_2') \leq \frac{\Pi(A_2)}{2}$.

Case 2 : $\Pi(A_i') > \frac{\Pi(A_i)}{2}$ for $i \in \{1, 2\}$

Lower bound from case 1: Let $\Pi(A_1') \le \frac{\Pi(A_1)}{2}$. Then,

$$\Pi(A_1 \cap (\mathcal{K} \setminus A_1')) = \Pi(A_1) - \Pi(A_1') \ge \frac{\Pi(A_1)}{2}.$$

Consequently,

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \ge \int_{A_1\cap(\mathcal{K}\setminus A_1')} \mathcal{T}_x(A_2)\pi(x)dx \ge \int_{A_1\cap(\mathcal{K}\setminus A_1')} \frac{3}{8}\cdot\pi(x)dx \ge \frac{3}{16}\cdot\Pi(A_1)\ .$$

The first inequality is due to $A_1 \cap (\mathcal{K} \setminus A_1') \subset A_1$. The second inequality uses the fact that since $x \in A_1 \cap (\mathcal{K} \setminus A_1')$, $x \in A_1 \setminus A_1'$, and for such x, $\mathcal{T}_x(A_2) \ge \frac{3}{8}$. The final step uses the fact shown right above. We can use this same technique to show that when $\Pi(A_2') \le \frac{\Pi(A_2)}{2}$,

$$\int_{A_2} \mathcal{T}_x(A_1)\pi(x)dx \ge \frac{3}{16}\cdot\Pi(A_2)\ .$$

Combining these gives,

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \ge \frac{3}{16}\cdot\min\{\Pi(A_1),\Pi(A_2)\}\ .$$

Lower bound from case 2: We start by considering arbitrary $x' \in A_1'$ and $y' \in A_2'$. From eq. (8),

$$\begin{aligned}
\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx &= \frac{1}{2}\left(\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx + \int_{A_2} \mathcal{T}_y(A_1)\pi(y)dy\right)\\
&\ge \frac{1}{2}\left(\int_{A_1\cap(\mathcal{K}\setminus A_1')} \mathcal{T}_x(A_2)\pi(x)dx + \int_{A_2\cap(\mathcal{K}\setminus A_2')} \mathcal{T}_y(A_1)\pi(y)dy\right)\\
&\ge \frac{3}{16}\left(\int_{A_1\cap(\mathcal{K}\setminus A_1')} \pi(x)dx + \int_{A_2\cap(\mathcal{K}\setminus A_2')} \pi(y)dy\right)\\
&= \frac{3}{16}\cdot\Pi(\mathcal{K}\setminus(A_1'\cup A_2')) = \frac{3}{16}\cdot\Pi(\mathcal{K}\setminus A_1'\setminus A_2')\ .
\end{aligned} \tag{9}$$

Also, by the definition of $\mathrm{d_{TV}}(\mathcal{T}_{x'}, \mathcal{T}_{y'})$, and using the fact that A_1 and A_2 form a partition,

$$\mathrm{d_{TV}}(\mathcal{T}_{x'}, \mathcal{T}_{y'}) \ge \mathcal{T}_{x'}(A_1) - \mathcal{T}_{y'}(A_1) = 1 - \mathcal{T}_{x'}(A_2) - \mathcal{T}_{y'}(A_1) > 1 - \frac{3}{8} - \frac{3}{8} = \frac{1}{4}.$$

At this juncture, the proofs of Lemmas 14 and 15 will differ due to the different isoperimetric inequalities discussed previously.

When $\mu > 0$ We first discuss **Case 2** in the context of Lemma 14. From the assumption of the lemma, this implies by contraposition that $\|x' - y'\|_{\nabla^2\phi(y')} > \Delta$ for $\Delta \le \frac{1}{2\alpha}$. Let $A_3' = \mathcal{K}\setminus A_1'\setminus A_2'$. Note that $\{A_1', A_2', A_3'\}$ forms a partition of $\mathcal{K}$. Since f is μ-relatively convex with respect to ϕ, which is self-concordant with parameter α, Proposition 17 then gives

$$\Pi(A_3') \ge C_\alpha\cdot\sqrt{\mu}\cdot d_\phi(A_2', A_1')\cdot\min\left\{\Pi(A_1'),\Pi(A_2')\right\}\ .$$

From Proposition 18, if $\|x' - y'\|_{\nabla^2 \phi(y')} > \Delta$ for $\Delta \le \frac{1}{2\alpha}$, $d_\phi(y', x') > \Delta - \alpha \cdot \frac{\Delta^2}{2}$. Since $x' \in A_1', y' \in A_2'$ are arbitrary, this holds for all pairs of $(x', y') \in A_1' \times A_2'$, and hence

$$d_\phi(A_2', A_1') = \inf_{x' \in A_1', y' \in A_2'} d_\phi(y', x') > \Delta - \alpha \cdot \frac{\Delta^2}{2} \ge \frac{\Delta}{2}$$

where the final inequality uses the fact that $t - \frac{t^2}{2} \ge \frac{t}{2}$ for $t \in [0, 0.5]$ with $t = \alpha \cdot \Delta$. Substituting the above two inequalities in eq. (9), we get the lower bound

$$\begin{aligned}\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx &\ge C_\alpha \cdot \sqrt{\mu} \cdot \Delta \cdot \frac{3}{32} \cdot \min\{\Pi(A_1'), \Pi(A_2')\} \\ &\ge C_\alpha \cdot \sqrt{\mu} \cdot \Delta \cdot \frac{3}{64} \cdot \min\{\Pi(A_1), \Pi(A_2)\} .\end{aligned}$$

where the final inequality uses the assumption of the case that $\Pi(A_i') > \frac{\Pi(A_i)}{2}$ for $i \in \{1, 2\}$.

When $\mu = 0$ Next, we discuss **Case 2** in the context of Lemma 15. From Proposition 19,

$$\Pi(\mathcal{K} \setminus A_1' \setminus A_2') \ge \mathrm{CR}(A_2', A_1'; \mathcal{K}) \cdot \Pi(A_1') \cdot \Pi(A_2') .$$

From Proposition 20, $\mathrm{CR}(y', x'; \mathcal{K}) \ge \frac{\|y' - x'\|_{\nabla^2 \phi(y')}}{\sqrt{\nu}}$. Since $x' \in A_1', y' \in A_2'$ are arbitrary, this holds for all pairs of $(x', y') \in A_1' \times A_2'$, and hence

$$\mathrm{CR}(A_2', A_1'; \mathcal{K}) = \inf_{y' \in A_2', x' \in A_1'} \mathrm{CR}(y', x'; \mathcal{K}) \ge \inf_{y' \in A_2', x' \in A_1'} \frac{\|y' - x'\|_{\nabla^2 \phi(y')}}{\sqrt{\nu}} \ge \frac{\Delta}{\sqrt{\nu}} .$$

Substituting the above two inequalities in eq. (9), we get the lower bound

$$\begin{aligned}\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx &\ge \frac{3}{16} \cdot \frac{\Delta}{\sqrt{\nu}} \cdot \Pi(A_2') \cdot \Pi(A_1') \\ &\ge \frac{3}{64} \cdot \frac{\Delta}{\sqrt{\nu}} \cdot \Pi(A_2) \cdot \Pi(A_1) \\ &\ge \frac{3}{128} \cdot \frac{\Delta}{\sqrt{\nu}} \cdot \min\{\Pi(A_1), \Pi(A_2)\}\end{aligned}$$

where the second inequality uses the assumption of the case that $\Pi(A_i') > \frac{\Pi(A_i)}{2}$ for $i \in \{1, 2\}$, and the final inequality uses the simple fact that $t(1-t) \ge 0.5 \cdot \min\{t, (1-t)\}$ when $t \in [0, 1]$.

We collate the inequalities from **Case 1** and **2** to get the following inequalities

when $\mu > 0$ as in Lemma 14

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \ge \min\left\{\frac{3}{16}, \frac{3}{64} \cdot C_\alpha \cdot \sqrt{\mu} \cdot \Delta\right\} \cdot \min\{\Pi(A_1), \Pi(A_2)\} ,$$

when $\mu = 0$ as in Lemma 15

$$\int_{A_1} \mathcal{T}_x(A_2)\pi(x)dx \ge \min\left\{\frac{3}{16}, \frac{3}{128} \cdot \frac{\Delta}{\sqrt{\nu}}\right\} \cdot \min\{\Pi(A_1), \Pi(A_2)\} ,$$

which concludes the proofs. ∎

D.3.2. PROOF OF LEMMA 16

In this subsection, we give the proof of both statements of the lemma. Specifically,

Part 1 provides the proof for the first statement about $\mathrm{d}_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y)$, and

Part 2 provides the proof for the second statement about $\mathrm{d}_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{P}_x)$.

Before giving the proofs, we state some key properties of ϕ due to its self-concordance.

Proposition 21 (Nesterov (2018, §5.1.4)) *Let $\phi : \mathcal{K} \to \mathbb{R} \cup \{\infty\}$ be a self-concordant function with parameter α.*

1. *For any $x, y \in \mathrm{int}(\mathcal{K})$ such that $\|x - y\|_{\nabla^2\phi(y)} < \frac{1}{\alpha}$,*
$$(1 - \alpha \cdot \|x - y\|_{\nabla^2\phi(y)})^2 \cdot \nabla^2\phi(y) \preceq \nabla^2\phi(x) \preceq \frac{1}{(1 - \alpha \cdot \|x - y\|_{\nabla^2\phi(y)})^2} \cdot \nabla^2\phi(y) \, .$$

2. *For any $x, y \in \mathrm{int}(\mathcal{K})$ such that $\|x - y\|_{\nabla^2\phi(y)} < 1/\alpha$, the matrix*
$$G_y = \int_0^1 \nabla^2\phi(y + \tau(x - y))d\tau$$
satisfies
$$G_y \preceq \frac{1}{1 - \alpha \cdot \|x - y\|_{\nabla^2\phi(y)}} \cdot \nabla^2\phi(y) \, .$$

3. *ϕ^* is also a self-concordant function with parameter α.*

Proof [Part 1.] By definition of $\mathcal{P}_x$, $\mathcal{P}_x = (\nabla\phi^*)_{\#}\widetilde{\mathcal{P}}_x$, where $\widetilde{\mathcal{P}}_x = \mathcal{N}(\nabla\phi(x) - h \cdot \nabla f(x), 2h \cdot \nabla^2\phi(x))$. Starting with Pinsker's inequality, we have

$$\begin{aligned}
\mathrm{d}_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) &\leq \sqrt{\frac{1}{2}\mathrm{d}_{\mathrm{KL}}(\mathcal{P}_x, \mathcal{P}_y)} \\
&= \sqrt{\frac{1}{2}\mathrm{d}_{\mathrm{KL}}((\nabla\phi^*)_{\#}\widetilde{\mathcal{P}}_x, (\nabla\phi^*)_{\#}\widetilde{\mathcal{P}}_y)} \\
&= \sqrt{\frac{1}{2}\mathrm{d}_{\mathrm{KL}}(\widetilde{\mathcal{P}}_x, \widetilde{\mathcal{P}}_y)} \, .
\end{aligned}$$

Since ϕ is of Legendre type, $\nabla\phi^*$ is a bijective and differentiable map. This allows us to instantiate Vempala and Wibisono (2019, Lemma 15) to get the final equality.

To further work with $\mathrm{d}_{\mathrm{KL}}(\widetilde{\mathcal{P}}_x, \widetilde{\mathcal{P}}_y)$, we use the identity

$$\mathrm{d}_{\mathrm{KL}}(\mathcal{N}(m_1, \Sigma_1), \mathcal{N}(m_2, \Sigma_2)) = \frac{1}{2}\left(\mathrm{trace}(\Sigma_2^{-1}\Sigma_1 - I) + \log\frac{\det \Sigma_2}{\det \Sigma_1} + \|m_2 - m_1\|^2_{\Sigma_2^{-1}}\right) \, .$$

In our case,

$$\Sigma_1 = 2h\nabla^2\phi(x) \, , \; \Sigma_2 = 2h\nabla^2\phi(y) \, , \; m_1 = \nabla\phi(x) - h\nabla f(x) \, , \; m_2 = \nabla\phi(y) - h\nabla f(y) \, .$$

This yields a closed form expression $\mathrm{d}_{\mathrm{KL}}(\widetilde{\mathcal{P}}_x, \widetilde{\mathcal{P}}_y) = \frac{1}{2}\left(T_1^P + T_2^P\right)$ where

$$T_1^P = \mathrm{trace}(\nabla^2\phi(y)^{-1}\nabla^2\phi(x) - I) - \log\det\nabla^2\phi(y)^{-1}\nabla^2\phi(x)\ ,$$
$$T_2^P = \frac{1}{2h}\|(\nabla\phi(y) - \nabla\phi(x)) - h(\nabla f(y) - \nabla f(x))\|^2_{\nabla^2\phi(y)^{-1}}\ .$$

The rest of the proof is dedicated to showing that

$$T_1^P \leq \frac{h \cdot d}{20}\ ; \qquad T_2^P \leq \frac{1}{80} + \frac{9}{2}\cdot h \cdot \beta\ ,$$

under assumptions made in the statement of the lemma.

For convenience, we denote $\|x - y\|_{\nabla^2\phi(y)}$ by r_y. From the statement of the first part of the lemma, $r_y \leq \frac{\sqrt{h}}{10\cdot\bar{\alpha}} \leq \frac{\sqrt{h}}{10\cdot\alpha}$, since $\bar{\alpha} = \max\{1, \alpha\}$.

Bounding T_1^P Owing to the cyclic property of trace and the product property of determinants, we have

$$\mathrm{trace}(\nabla^2\phi(y)^{-1}\nabla^2\phi(x)) = \mathrm{trace}(\nabla^2\phi(y)^{-1/2}\nabla^2\phi(x)\nabla^2\phi(y)^{-1/2})$$
$$\det\nabla^2\phi(y)^{-1}\nabla^2\phi(x) = \det\nabla^2\phi(y)^{-1/2}\nabla^2\phi(x)\nabla^2\phi(y)^{-1/2}.$$

Let $M = \nabla^2\phi(y)^{-1/2}\nabla^2\phi(x)\nabla^2\phi(y)^{-1/2}$ for convenience, and let $\{\lambda_i(M)\}_{i=1}^d$ be its eigenvalues.

$$T_1^P = \mathrm{trace}(M - I) - \log\det M = \sum_{i=1}^{d}\{\lambda_i(M) - 1 - \log\lambda_i(M)\}\ .$$

Lemma 22 states that $\lambda_i(M) - 1 - \log\lambda_i(M) \leq \frac{(\lambda_i(M)-1)^2}{\lambda_i(M)}$. Therefore, we have the upper bound

$$T_1^P \leq \sum_{i=1}^{d}\frac{(\lambda_i(M) - 1)^2}{\lambda_i(M)}\ .$$

From Proposition 21(1), each $\lambda_i(M)$ is bounded as

$$\lambda_i(M) \in \left[\frac{1}{(1 - \alpha\cdot r_y)^2}\ ,\ (1 - \alpha\cdot r_y)^2\right]\ .$$

The function $t \mapsto \frac{(t-1)^2}{t}$ is strictly convex for $t > 0$. This is because its second derivative is $t \mapsto \frac{1}{t^3}$. This implies that when t is restricted to a closed interval $[a, b]$, the maximum of $\frac{(t-1)^2}{t}$ is attained at the end points.

$$\max_{t\in[a,b]}\frac{(t-1)^2}{t} = \max\left\{\frac{(a-1)^2}{a}, \frac{(b-1)^2}{b}\right\}\ .$$

When $b = \frac{1}{a}$, the maximum on the right hand side is $\frac{(a-1)^2}{a}$. We use this fact with bounds on $t = \lambda_i(M)$, and thus have for all $i \in [d]$ that

$$\frac{(\lambda_i(M) - 1)^2}{\lambda_i(M)} \leq \frac{((1 - \alpha\cdot r_y)^2 - 1)^2}{(1 - \alpha\cdot r_y)^2} = \frac{\alpha^2\cdot r_y^2\cdot(2 - \alpha\cdot r_y)^2}{(1 - \alpha\cdot r_y)^2} \leq \frac{4\cdot\alpha^2\cdot r_y^2}{(1 - \alpha\cdot r_y)^2}\ .$$

and consequently,

$$T_1^P \leq d \cdot \frac{4 \cdot \alpha^2 \cdot r_y^2}{(1 - \alpha \cdot r_y)^2} \,.$$

The function $t \mapsto \frac{4t^2}{(1-t)^2}$ is an increasing function since it is a product of two increasing functions $t \mapsto 4t^2$ and $t \mapsto \frac{1}{(1-t)^2}$. As stated earlier, $r_y \leq \frac{\sqrt{h}}{10 \cdot \alpha}$, which implies

$$T_1^P \leq d \cdot \frac{4 \cdot \frac{h}{100}}{(1 - \frac{\sqrt{h}}{10})^2} \leq \frac{h \cdot d}{20} \,.$$

The final inequality uses the fact that $h \leq 1$ which implies $\left(1 - \frac{\sqrt{h}}{10}\right)^{-2} \leq \frac{100}{81} \leq \frac{5}{4}$.

Bounding T_2^P Using the fact that $\|a + b\|_M^2 \leq 2\|a\|_M^2 + 2\|b\|_M^2$ successively,

$$\begin{aligned}
T_2^P &= \frac{1}{2h} \|(\nabla\phi(y) - \nabla\phi(x)) - h(\nabla f(y) - \nabla f(x))\|^2_{\nabla^2\phi(y)^{-1}} \\
&\leq \frac{1}{2h} \left(2 \cdot \|\nabla\phi(y) - \nabla\phi(x)\|^2_{\nabla^2\phi(y)^{-1}} + 2h^2 \cdot \|\nabla f(y) - \nabla f(x)\|^2_{\nabla^2\phi(y)^{-1}}\right) \\
&\leq \frac{1}{h} \cdot \|\nabla\phi(y) - \nabla\phi(x)\|^2_{\nabla^2\phi(y)^{-1}} + 2h \cdot \|\nabla f(y)\|^2_{\nabla^2\phi(y)^{-1}} + 2h \cdot \|\nabla f(x)\|^2_{\nabla^2\phi(y)^{-1}} \,.
\end{aligned}$$

Note that

$$\nabla\phi(x) - \nabla\phi(y) = \int_0^1 \frac{d}{d\tau} \nabla\phi(y + \tau(x-y)) d\tau = \int_0^1 \nabla^2\phi(y + \tau(x-y))(x-y) d\tau = G_y(x-y).$$

From Proposition 21(2), and using the fact that $r_y < 1/\alpha$, we have $G_y \preceq \frac{1}{(1-\alpha \cdot r_y)} \cdot \nabla^2\phi(y)$. By virtue of ϕ being of Legendre type, $G_y \succ 0$. As a result,

$$G_y^{-1} \succeq (1 - \alpha \cdot r_y) \cdot \nabla^2\phi(y)^{-1} \Leftrightarrow G_y \preceq \frac{1}{1 - \alpha \cdot r_y} \cdot \nabla^2\phi(y) \,, \tag{10}$$

where the equivalence is due to the Loewner ordering. Let $v = G_y^{1/2}(x - y)$.

$$\begin{aligned}
\|\nabla\phi(x) - \nabla\phi(y)\|^2_{\nabla^2\phi(y)^{-1}} &= \|G_y(x-y)\|^2_{\nabla^2\phi(y)^{-1}} \\
&= \langle v, G_y^{1/2} \nabla^2\phi(y)^{-1} G_y^{1/2} v\rangle \\
&\leq \frac{\langle v, v\rangle}{1 - \alpha \cdot r_y} \\
&= \frac{\langle x - y, G_y(x-y)\rangle}{1 - \alpha \cdot r_y} \\
&\leq \frac{r_y^2}{(1 - \alpha \cdot r_y)^2} \,.
\end{aligned}$$

The inequalities above are due to eq. (10). Next, from Proposition 21(1), we have $\nabla^2\phi(x) \succeq \frac{1}{(1-\alpha \cdot r_y)^2} \cdot \nabla^2\phi(y)$, and using the fact that ϕ is strictly convex yields

$$\|\nabla f(x)\|^2_{\nabla^2\phi(y)^{-1}} \leq \frac{1}{(1 - \alpha \cdot r_y)^2} \cdot \|\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}} \,.$$

This gives a bound for T_2^P in terms of r_y as

$$\begin{aligned} T_2^P &\leq \frac{r_y^2}{h \cdot (1 - \alpha \cdot r_y)^2} + 2h \cdot \|\nabla f(y)\|^2_{\nabla^2 \phi(y)^{-1}} + \frac{2h \cdot \|\nabla f(x)\|^2_{\nabla^2 \phi(x)^{-1}}}{(1 - \alpha \cdot r_y)^2} \\ &\leq \frac{r_y^2}{h \cdot (1 - \alpha \cdot r_y)^2} + 2h \cdot \beta^2 + \frac{2h \cdot \beta^2}{(1 - \alpha \cdot r_y)^2} , \end{aligned}$$

where in the final inequality, we have used the fact that f is a β-relatively Lipschitz with respect to ϕ. Finally, since $\alpha \cdot r_y \leq \frac{\sqrt{h}}{10} \leq \frac{1}{10}$, $(1 - \alpha \cdot r_y)^{-2} \leq \frac{5}{4}$ as noted earlier. This gives the upper bound

$$T_2^P \leq \frac{5}{4} \cdot \frac{r_y^2}{h} + \frac{9}{2} \cdot h \cdot \beta^2 .$$

From the statement of the lemma, $r_y \leq \frac{\sqrt{h}}{10 \cdot \max\{1, \alpha\}}$ which also implies that $r_y \leq \frac{\sqrt{h}}{10}$. Finally, substituting this bound over r_y, we get

$$T_2^P \leq \frac{5}{4} \cdot \frac{1}{h} \cdot \frac{h}{100} + \frac{9}{2} \cdot h \cdot \beta \leq \frac{1}{80} + \frac{9}{2} \cdot h \cdot \beta^2 .$$

Using the bounds derived for T_1^P and T_2^P, we can complete the proof.

$$\mathrm{d}_{\mathrm{KL}}(\widetilde{\mathcal{P}}_x, \widetilde{\mathcal{P}}_y) \leq \frac{1}{2}\left(\frac{h \cdot d}{20} + \frac{1}{80} + \frac{9}{2} \cdot h \cdot \beta^2\right) .$$

Finally, applying Pinsker's inequality as stated in the beginning of the proof yields

$$\mathrm{d}_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) \leq \frac{1}{2}\sqrt{\frac{h \cdot d}{20} + \frac{1}{80} + \frac{9}{2} \cdot h \cdot \beta^2} .$$

■

Proof [Part 2.]

For $\varepsilon \in (0, 1)$, define the following quantities.

$$\mathsf{N}_\varepsilon = 1 + \sqrt{2 \log\left(\frac{8}{\varepsilon}\right)} + 2 \log\left(\frac{8}{\varepsilon}\right); \qquad \mathsf{I}_\varepsilon = \sqrt{2 \log\left(\frac{8}{\varepsilon}\right)} . \tag{11}$$

The explicit forms of the $\mathcal{C}_1(\varepsilon), \dots, \mathcal{C}_4(\varepsilon)$ in eq. (7) are given below.

$$\left.\begin{aligned} \mathcal{C}_1(\varepsilon) &= \left(\left(\frac{\varepsilon}{24}\right)^{2/3} \frac{1}{6 \cdot \mathsf{N}_\varepsilon}\right)^3 , \ \mathcal{C}_2(\varepsilon) = \frac{\varepsilon}{16} \cdot \frac{1}{6 \cdot \mathsf{N}_\varepsilon} \\ \mathcal{C}_3(\varepsilon) &= \left(\left(\frac{\varepsilon}{72}\right)^2 \cdot \frac{1}{6\sqrt{2} \cdot \mathsf{I}_\varepsilon}\right)^{2/3} , \ \mathcal{C}_4(\varepsilon) = \frac{\varepsilon}{16} \cdot \frac{1}{6\sqrt{2} \cdot \mathsf{I}_\varepsilon} \end{aligned}\right\} \tag{12}$$

Recall that the transition kernel has an atom i.e., $\mathcal{T}_x(\{x\}) \neq 0$. The explicit form of this is given by

$$\mathcal{T}_x(\{x\}) = 1 - \int_{z \in \mathcal{K}} \alpha_{\mathbf{P}}(z; x) p_x(z) dz = 1 - \int_{z \in \mathcal{K}} \min\left\{1, \frac{\pi(z) p_z(x)}{\pi(x) p_x(z)}\right\} p_x(z) dz .$$

Using this, we have the expression for $\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{P}_x)$ as (c.f. Dwivedi et al. (2018, Eq. 46))

$$\mathrm{d_{TV}}(\mathcal{T}_x, \mathcal{P}_x) = 1 - \mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z)p_z(x)}{\pi(x)p_x(z)}\right\}\right] .$$

Our goal is to bound this quantity from above, which equivalently implies bounding the expectation on the right hand side from below. This can be achieved using Markov's inequality; for any $t > 0$,

$$\mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z)p_z(x)}{\pi(x)p_x(z)}\right\}\right] \geq t\,\mathbb{P}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z)p_z(x)}{\pi(x)p_x(z)}\right\} \geq t\right] .$$

If $t \leq 1$, then

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z)p_z(x)}{\pi(x)p_x(z)}\right\} \geq t\right] = \mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)p_z(x)}{\pi(x)p_x(z)} \geq t\right].$$

Due to eq. (2), we can write the explicit expression for this ratio

$$\begin{aligned}\frac{\pi(z)\cdot p_z(x)}{\pi(x)\cdot p_x(z)} = \exp\Bigg(f(x) - f(z) + \frac{3}{2}\left\{\log\det\nabla^2\phi(x) - \log\det\nabla^2\phi(z)\right\} \\ + \frac{1}{4h}\|\nabla\phi(z) - \nabla\phi(x) + h\cdot\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}} \\ - \frac{1}{4h}\|\nabla\phi(x) - \nabla\phi(z) + h\cdot\nabla f(z)\|^2_{\nabla^2\phi(z)^{-1}}\Bigg).\end{aligned}$$

Concisely, $\frac{\pi(z)\cdot p_z(x)}{\pi(x)\cdot p_x(z)} = \exp(\mathcal{A}(x,z))$, where $\mathcal{A}(x,z) = T_1^A + T_2^A + T_3^A + T_4^A + T_5^A$ and each of these terms are

$$T_1^A := \frac{\|\nabla\phi(z) - \nabla\phi(x)\|^2_{\nabla^2\phi(x)^{-1}}}{4h} - \frac{\|\nabla\phi(x) - \nabla\phi(z)\|^2_{\nabla^2\phi(z)^{-1}}}{4h} \tag{13a}$$

$$T_2^A := \frac{3}{2}\log\det\nabla^2\phi(z)^{-1}\nabla^2\phi(x) \tag{13b}$$

$$T_3^A := \frac{1}{2}\left(f(x) - f(z) - \langle\nabla f(z), \nabla\phi(x) - \nabla\phi(z)\rangle_{\nabla^2\phi(z)^{-1}}\right) \tag{13c}$$

$$T_4^A := \frac{1}{2}\left(f(x) - f(z) - \langle\nabla f(x), \nabla\phi(x) - \nabla\phi(z)\rangle_{\nabla^2\phi(x)^{-1}}\right) \tag{13d}$$

$$T_5^A := \frac{h^2}{4h}\left(\|\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}} - \|\nabla f(z)\|^2_{\nabla^2\phi(z)^{-1}}\right). \tag{13e}$$

The self-concordant nature of ϕ enables using Proposition 21 to control some of these quantities above, but only when $\nabla\phi(x)$ and $\nabla\phi(z)$ are close in the local norm. Therefore, we condition on an event $\mathfrak{E}$ which implies that $\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi(x)^{-1}} \leq \frac{3}{10\cdot\alpha}$. To be specific,

$$\begin{aligned}\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)p_z(x)}{\pi(x)p_x(z)} \geq t\right] &\geq \mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)p_z(x)}{\pi(x)p_x(z)} \geq t \wedge \mathfrak{E}\right] \\ &= \mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)p_z(x)}{\pi(x)p_x(z)} \geq t \;\middle|\; \mathfrak{E}\right]\cdot\mathbb{P}_{z\sim\mathcal{P}_x}(\mathfrak{E}) .\end{aligned}$$

The remainder of the proof consists of three parts.

- In Part D.3.2.1, we identify an event $\mathfrak{E}$ satisfying our requirements, and show that it occurs with high probability. We show that for the choice of step size h assumed, $\mathfrak{E}$ also implies that $\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi(x)^{-1}} \leq \frac{3}{10\cdot\alpha}$.
- Next, in Part D.3.2.2, we condition on event $\mathfrak{E}$ and give lower bounds for each of the T_i^A quantities by leveraging the implication in the previous part.
- Finally, in Part D.3.2.3, we show that for the choice of h in the statement of the lemma there exists $t \leq 1$ such that the conditional probability is 1, thus reducing the lower bound to $t \cdot \mathbb{E}_{z\sim\mathcal{P}_x}[\mathfrak{E}]$.

D.3.2.1. Identifying an event $\mathfrak{E}$ Since $\mathcal{P}_x = (\nabla\phi^*)_\#\widetilde{\mathcal{P}}_x$,

$$z \sim \mathcal{P}_x \;\Leftrightarrow\; \nabla\phi(z) \sim \widetilde{\mathcal{P}}_x = \mathcal{N}(\nabla\phi(x) - h\cdot\nabla f(x), 2h\nabla^2\phi(x)) \;.$$

We use this to show that $\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi(x)^{-1}}$ concentrates well, since given $\xi \sim \mathcal{N}(0, I)$, $\nabla\phi(z)$ is distributionally equivalent to $\nabla\phi(x) - h\cdot\nabla f(x) + \sqrt{2h}\nabla^2\phi(x)^{1/2}\xi$ when $z \sim \mathcal{P}_x$.
We begin by expanding the squared norm as

$$\begin{aligned}&\|\nabla\phi(z) - \nabla\phi(x)\|^2_{\nabla^2\phi(x)^{-1}} \\ &\qquad = h^2\cdot\langle\nabla f(x), \nabla^2\phi(x)^{-1}\nabla f(x)\rangle + 2h\cdot\langle\xi,\xi\rangle + 2\sqrt{2}h\sqrt{h}\cdot\langle\xi, \nabla^2\phi(x)^{-1/2}\nabla f(x)\rangle.\end{aligned}$$

Consider the following probability

$$\begin{aligned}&\mathbb{P}_{z\sim\mathcal{P}_x}(\|\nabla\phi(x) - \nabla\phi(z)\|^2_{\nabla^2\phi(x)^{-1}} > h^2\cdot\|\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}} + 2h\cdot d\cdot\mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h}\cdot\mathsf{I}_\varepsilon) \\ &\quad = \mathbb{P}_{\xi\sim\mathcal{N}(0,I_d)}(2h\cdot\|\xi\|^2 + 2\sqrt{2}h\sqrt{h}\cdot\langle\xi, \nabla^2\phi(x)^{-1/2}\nabla f(x)\rangle > 2h\cdot d\cdot\mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h}\cdot\beta\cdot\mathsf{I}_\varepsilon) \\ &\quad \leq \mathbb{P}_{\xi\sim\mathcal{N}(0,I_d)}(\|\xi\|^2 > d\cdot\mathsf{N}_\varepsilon) + \mathbb{P}_{\xi\sim\mathcal{N}(0,I_d)}(\langle\xi, \nabla^2\phi(x)^{-1/2}\nabla f(x)\rangle > \beta\cdot\mathsf{I}_\varepsilon)\end{aligned}$$

where the final inequality uses the fact that for random variables a, b,

$$\begin{gathered}(\mathsf{a} \leq a)\wedge(\mathsf{b} \leq b) \Rightarrow \mathsf{a}+\mathsf{b} \leq a+b \\ \mathbb{P}(\mathsf{a}+\mathsf{b} > a+b) \leq \mathbb{P}(\mathsf{a} > a \vee \mathsf{b} > b) \leq \mathbb{P}(\mathsf{a} > a) + \mathbb{P}(\mathsf{b} > b)\;.\end{gathered}$$

Through χ^2 concentration guarantees (Laurent and Massart, 2000, Lem. 1), we get

$$\mathbb{P}(\|\xi\|^2 > d\cdot\mathsf{N}_\varepsilon) \leq \frac{\varepsilon}{8}\;,$$

where N_ε was previously defined in eq. (11). Also, since ξ is Gaussian, $\langle\xi, \nabla^2\phi(x)^{-1/2}\nabla f(x)\rangle$ is a mean zero, sub-Gaussian random variable with variance parameter $\|\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}}$. This implies that for any $x \in \mathcal{K}$,

$$\mathbb{P}_{\xi\sim\mathcal{N}(0,I_d)}(\langle\xi, \nabla^2\phi(x)^{-1/2}\nabla f(x)\rangle > t) \leq \exp\left(-\frac{t^2}{2\|\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}}}\right) \leq \exp\left(-\frac{t^2}{2\beta^2}\right)$$

where the final inequality is due to the fact that f is β-relatively Lipschitz with respect to ϕ. Therefore,

$$\mathbb{P}_{\xi\sim\mathcal{N}(0,I_d)}(\langle\xi, \nabla^2\phi(x)^{-1/2}\nabla f(x)\rangle > \beta\cdot\mathsf{I}_\varepsilon) \leq \frac{\varepsilon}{8}\;,$$

where l_ε was previously defined in eq. (11). Therefore, with probability at least $1 - \frac{\varepsilon}{4}$,

$$\begin{aligned}\|\nabla\phi(x) - \nabla\phi(z)\|^2_{\nabla^2\phi(x)^{-1}} &\le h^2 \cdot \|\nabla f(x)\|^2_{\nabla^2\phi(x)^{-1}} + 2h \cdot d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h} \cdot \beta \cdot \mathsf{l}_\varepsilon \\ &\le h^2 \cdot \beta^2 + 2h \cdot d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h} \cdot \beta \cdot \mathsf{l}_\varepsilon \,.\end{aligned}$$

We hence pick the event $\mathfrak{E}$ to be

$$\mathfrak{E} \overset{\text{def}}{=} \left\{\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi(x)^{-1}} \le \sqrt{h^2 \cdot \beta^2 + 2h \cdot d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h} \cdot \beta \cdot \mathsf{l}_\varepsilon}\right\}.$$

Using the simple algebraic fact that $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$, and from the choice of h in the lemma it can be verified that

$$h \le \frac{1}{10 \cdot \beta \cdot \alpha}; \quad h \le \frac{1}{200 \cdot d \cdot \mathsf{N}_\varepsilon \cdot \alpha^2}; \quad h \le \frac{1}{50 \cdot \mathsf{l}_\varepsilon^{2/3} \cdot \alpha^{4/3} \cdot \beta^{2/3}} \,,$$

and we are assured that $\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi(x)^{-1}} \le \frac{3}{10\cdot\alpha}$ as required.

D.3.2.2. Lower bound on the T_i^A quantities In this part, we condition on the event $\mathfrak{E}$ which implies that $\|\nabla\phi(z) - \nabla\phi(x)\|_{\nabla^2\phi(x)^{-1}} \le \frac{3}{10\cdot\alpha}$ to obtain lower bounds on $\mathcal{A}(x, z)$ defined previously. We do so by giving lower bounds for each of the T_i^A quantities individually defined in eq. (13).

In this part of the proof, we will make use of the identity

$$\nabla^2\phi^*(\nabla\phi(x)) = \nabla^2\phi(x)^{-1} \qquad \forall\, x \in \mathcal{K}\,. \tag{14}$$

This can be derived by differentiating the identity $\nabla\phi^*(\nabla\phi(x)) = x$ by the invertibility of $\nabla\phi(x)$.

T_1^A: We use the shorthand notation $v_{x,z} = \nabla\phi(x) - \nabla\phi(z)$ for convenience.

$$\begin{aligned}T_1^A &= \frac{1}{4h}\left(\|v_{x,z}\|^2_{\nabla^2\phi(x)^{-1}} - \|v_{x,z}\|^2_{\nabla^2\phi(z)^{-1}}\right) \\ &\overset{(i)}{=} \frac{1}{4h}\left(\|v_{x,z}\|^2_{\nabla^2\phi^*(\nabla\phi(x))} - \|v_{x,z}\|^2_{\nabla^2\phi^*(\nabla\phi(z))}\right) \\ &\overset{(ii)}{\ge} \frac{1}{4h}\left(\|v_{x,z}\|^2_{\nabla^2\phi^*(\nabla\phi(x))} - \frac{\|v_{x,z}\|^2_{\nabla^2\phi^*(\nabla\phi(x))}}{(1 - \alpha \cdot \|v_{x,z}\|_{\nabla^2\phi^*(\nabla\phi(x))})^2}\right) \\ &= \frac{1}{h \cdot \alpha^2} \cdot \frac{(\alpha \cdot \|v_{x,z}\|_{\nabla^2\phi^*(\nabla\phi(x))})^2}{4} \cdot \left(1 - \frac{1}{(1 - \alpha \cdot \|v_{x,z}\|_{\nabla^2\phi^*(\nabla\phi(x))})^2}\right) \\ &\overset{(iii)}{\ge} -\frac{3\alpha}{2h} \cdot \|v_{x,z}\|^3_{\nabla^2\phi^*(\nabla\phi(x))} \,.\end{aligned}$$

Step (i) uses eq. (14). Step (ii) uses the fact that ϕ^* is self-concordant with parameter α from Proposition 21(3), and the result of Proposition 21(1) to change the metric from $\nabla^2\phi^*(\nabla\phi(z))$ to $\nabla^2\phi^*(\nabla\phi(x))$ since $\|v_{x,z}\|_{\nabla^2\phi^*(\nabla\phi(x))} < 1/\alpha$. Step (iii) uses an algebraic lemma Lemma 23.

T_2^A**:** We can use eq. (14) and the product property of determinants to rewrite T_2^A as follows.

$$\begin{aligned}T_2^A &= \frac{3}{2}\log\det \nabla^2\phi(z)^{-1}\nabla^2\phi(x)\\ &= \frac{3}{2}\log\det \nabla^2\phi^*(\nabla\phi(x))^{-1/2}\ \nabla^2\phi^*(\nabla\phi(z))\ \nabla^2\phi^*(\nabla\phi(x))^{-1/2}\\ &= \frac{3}{2}\sum_{i=1}^{d}\log\lambda_i(M^*)\end{aligned}$$

where $M^* = \nabla^2\phi^*(\nabla\phi(x))^{-1/2}\nabla^2\phi^*(\nabla\phi(z))\nabla^2\phi^*(\nabla\phi(x))^{-1/2}$, and $\{\lambda_i(M^*)\}$ is the sequence of its eigenvalues. These eigenvalues lie in an interval due to ϕ^* being self-concordant, and the associated Hessian ordering as given in Proposition 21(1,3) courtesy of $\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi^*(\nabla\phi(x))} < 1/\alpha$ as guaranteed by $\mathfrak{E}$. This yields

$$\lambda_i(M^*) \geq (1 - \alpha\cdot\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi^*(\nabla\phi(x))})^2\ ,$$

and we use this to obtain the lower bound

$$\begin{aligned}T_2^A &\geq 3d\cdot\log(1 - \alpha\cdot\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi^*(\nabla\phi(x))})\\ &\geq -\frac{9d}{2}\cdot\alpha\cdot\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi^*(\nabla\phi(x))}\ .\end{aligned}$$

The final inequality is due to Lemma 24.

$T_3^A + T_4^A$**:** Let $\psi = f\circ\nabla\phi^* : \mathbb{R}^d \to \mathbb{R}$. Then,

$$\begin{aligned}D_\psi(\bar{x};\bar{z}) &= \psi(\bar{x}) - \psi(\bar{z}) - \langle\nabla\psi(\bar{z}), \bar{x} - \bar{z}\rangle\\ &= \psi(\bar{x}) - \psi(\bar{z}) - \langle\nabla^2\phi^*(\bar{z})\nabla f(\nabla\phi^*(\bar{z})), \bar{x} - \bar{z}\rangle\ .\end{aligned}$$

The Bregman commutator $\zeta_\psi : \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ of ψ (Wibisono et al., 2022) is defined as

$$\zeta_\psi(\bar{x},\bar{z}) = \frac{1}{2}(D_\psi(\bar{x};\bar{z}) - D_\psi(\bar{z};\bar{x}))\ .$$

Substituting $\bar{x} = \nabla\phi(x)$ and $\bar{z} = \nabla\phi(z)$ gives

$$\begin{aligned}D_\psi(\nabla\phi(x);\nabla\phi(z)) &= f(x) - f(z) - \langle\nabla f(z), \nabla\phi(x) - \nabla\phi(z)\rangle_{\nabla^2\phi(z)^{-1}} = T_3^A\\ D_\psi(\nabla\phi(z);\nabla\phi(x)) &= f(z) - f(x) - \langle\nabla f(x), \nabla\phi(z) - \nabla\phi(x)\rangle_{\nabla^2\phi(x)^{-1}} = -T_4^A\ .\end{aligned}$$

Consequently, the sum of T_3^A and T_4^A is

$$T_3^A + T_4^A = \frac{1}{2}\left(D_\psi(\nabla\phi(x);\nabla\phi(z)) - D_\psi(\nabla\phi(z);\nabla\phi(x))\right) = \zeta_\psi(\nabla\phi(x),\nabla\phi(z))\ .$$

For convenience, we use the shorthand notation $\bar{x}$ and $\bar{z}$ for $\nabla\phi(x)$ and $\nabla\phi(z)$ respectively. Define $p_t = \bar{z} + t(\bar{x} - \bar{z})$ for $t\in[0,1]$. We work with the following identity from Wibisono et al. (2022, Eq. 15)

$$\zeta_\psi(\bar{x},\bar{z}) = \frac{1}{2}\int_0^1 (1-2t)\nabla^2\psi(p_t)[\bar{x}-\bar{z},\bar{x}-\bar{z}]dt\ .$$

The Hessian of ψ is

$$\begin{aligned}\nabla^2\psi(p_t) &= \nabla^3\phi^*(p_t)[\nabla f(\nabla\phi^*(p_t))] + \nabla^2\phi^*(p_t)\nabla^2 f(\nabla\phi^*(p_t))\nabla^2\phi^*(p_t)\\ &= \nabla^3\phi^*(p_t)[\nabla f(\nabla\phi^*(p_t))] + \nabla^2\phi(\nabla\phi^*(p_t))^{-1}\nabla^2 f(\nabla\phi^*(p_t))\nabla^2\phi(\nabla\phi^*(p_t))^{-1}\end{aligned}$$

Consequently,

$$\begin{aligned}|\zeta_\psi(\bar{x},\bar{z})| \le \frac{1}{2}\underbrace{\int_0^1 |1-2t|\cdot|\nabla^3\phi^*(p_t)[\bar{x}-\bar{z},\bar{x}-\bar{z},\nabla f(\nabla\phi^*(p_t))]|dt}_{I_1}\\ +\frac{1}{2}\underbrace{\int_0^1 |1-2t|\cdot|\langle \bar{x}-\bar{z},\nabla^2\phi(\nabla\phi^*(p_t))^{-1}\nabla^2 f(\nabla\phi^*(p_t))\nabla^2\phi(\nabla\phi^*(p_t))^{-1}[\bar{x}-\bar{z}]\rangle|dt}_{I_2}\ .\end{aligned}$$

We can bound I_1 and I_2 using properties of ϕ, f and the fact that $\bar{x} = \nabla\phi(x), \bar{z} = \nabla\phi(z)$. From Proposition 21(3), we know that ϕ^* is self-concordant with parameter α. This implies

$$\begin{aligned}I_1 &\le \int_0^1 |1-2t|\left(2\alpha\cdot\|\bar{x}-\bar{z}\|^2_{\nabla^2\phi^*(p_t)}\cdot\|\nabla f(\nabla\phi^*(p_t))\|_{\nabla^2\phi^*(p_t)}\right)dt\\ &\le \int_0^1 |1-2t|\left(2\alpha\cdot\|\bar{x}-\bar{z}\|^2_{\nabla^2\phi^*(p_t)}\cdot\beta\right)dt\end{aligned}$$

The last inequality uses the fact that f is β-relatively Lipschitz with respect to ϕ and by writing $\nabla^2\phi^*(p_t) = \nabla^2\phi(\nabla\phi^*(p_t))^{-1}$ from eq. (14).

Since f is λ-relatively smooth with respect to ϕ, we have for any $w \in \mathcal{K}$,

$$\nabla^2 f(w) \preceq \lambda\cdot\nabla^2\phi(w) \Leftrightarrow \nabla^2\phi(w)^{-1/2}\nabla^2 f(w)\nabla^2\phi(w)^{-1/2} \preceq \lambda\cdot I\ .$$

With this, we can bound I_2 as

$$\begin{aligned}I_2 &\le \int_0^1 |1-2t|\cdot\lambda\cdot|\langle\nabla^2\phi(\nabla\phi^*(p_t))^{-1/2}(\bar{x}-\bar{z}),\nabla^2\phi(\nabla\phi^*(p_t))^{-1/2}(\bar{x}-\bar{z})\rangle|dt\\ &= \int_0^1 |1-2t|\cdot\lambda\cdot\|\bar{x}-\bar{z}\|^2_{\nabla^2\phi^*(p_t)}dt\ .\end{aligned}$$

Collecting the bounds,

$$\begin{aligned}|\zeta_\psi(\bar{x},\bar{z})| &\le \left(\alpha\cdot\beta+\frac{\lambda}{2}\right)\cdot\int_0^1 |1-2t|\cdot\|\bar{x}-\bar{z}\|^2_{\nabla^2\phi^*(p_t)}dt\\ &\overset{(i)}{\le} \left(\alpha\cdot\beta+\frac{\lambda}{2}\right)\cdot\int_0^1 |1-2t|\cdot\frac{\|\bar{x}-\bar{z}\|^2_{\nabla^2\phi^*(\bar{x})}}{(1-\alpha\cdot(1-t)\|\bar{x}-\bar{z}\|_{\nabla^2\phi^*(\bar{x})})^2}dt\\ &\overset{(ii)}{\le} \left(\alpha\cdot\beta+\frac{\lambda}{2}\right)\cdot\|\bar{x}-\bar{z}\|^2_{\nabla^2\phi^*(\bar{x})}\cdot\int_0^1\frac{100\cdot|1-2t|}{(7+3t)^2}dt\\ &\le \left(\alpha\cdot\beta+\frac{\lambda}{2}\right)\cdot\|\nabla\phi(x)-\nabla\phi(z)\|^2_{\nabla^2\phi^*(\nabla\phi(x))}\ .\end{aligned}$$

Recall that $\|\bar{x} - \bar{z}\|_{\nabla^2 \phi^*(\bar{x})} < 1/\alpha$ and $\|p_t - \bar{x}\|_{\nabla^2 \phi^*(\bar{x})} = (1-t)\|\bar{z} - \bar{z}\|_{\nabla^2 \phi^*(\bar{x})} < 1/\alpha$ when $t \in [0,1]$. In Step (i), we use this fact to use Proposition 21(1) to get

$$\nabla^2 \phi^*(p_t) \preceq \frac{1}{(1 - \alpha \cdot (1-t)\|\bar{x} - \bar{z}\|_{\nabla^2 \phi^*(\bar{x})})^2} \cdot \nabla^2 \phi^*(\bar{x})$$

Step (ii) uses the fact that $\|\bar{x} - \bar{z}\|_{\nabla^2 \phi^*(\bar{x})} \leq 3/10 \cdot \alpha$ and that $(1-x)^{-2}$ is increasing for $x < 1$. The final inequality is an upper bound on the integral which can be computed as a closed form expression with value ≈ 0.73. This gives us the final bound

$$|\zeta_\psi(\nabla\phi(x), \nabla\phi(z))| \leq \gamma \cdot \|\nabla\phi(x) - \nabla\phi(z)\|^2_{\nabla^2\phi(x)^{-1}} ,$$

where $\gamma = \frac{\lambda}{2} + \alpha \cdot \beta$, which implies

$$T_3^A + T_4^A \geq -\gamma \cdot \|\nabla\phi(x) - \nabla\phi(z)\|^2_{\nabla^2\phi(x)^{-1}} .$$

T_5^A: Since f is β-relatively Lipschitz with respect to ϕ, this can be simply bounded from below as

$$T_5^A \geq -\frac{h \cdot \beta^2}{4}.$$

For convenience, we will use the shorthand notation $\ell_{x,z}$ to denote $\|\nabla\phi(x) - \nabla\phi(z)\|_{\nabla^2\phi(x)^{-1}}$. Collating these lower bounds, we get the net lower bound on $\mathcal{A}(x,z)$ as

$$\mathcal{A}(x,z) \geq -\frac{3\bar{\alpha}}{2h} \cdot \ell^3_{x,z} - \frac{9d \cdot \bar{\alpha}}{2} \cdot \ell_{x,z} - \gamma \cdot \ell^2_{x,z} - \frac{h \cdot \beta^2}{4} .$$

D.3.2.3. Choosing t in the conditional probability quantity Under the event $\mathfrak{E}$, we can substitute the upper bound on $\ell_{x,z}$ in the lower bound for $\mathcal{A}(x,z)$.

This gives us

$$\begin{aligned}
\mathcal{A}(x,z) \geq &-\frac{3\alpha}{2h} \cdot (h^2 \cdot \beta^2 + 2h \cdot d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h} \cdot \beta \cdot \mathsf{I}_\varepsilon)^{3/2} \\
&- \frac{9d \cdot \alpha}{2} \cdot (h^2 \cdot \beta^2 + 2h \cdot d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h} \cdot \beta \cdot \mathsf{I}_\varepsilon)^{1/2} \\
&- \gamma \cdot (h^2 \cdot \beta^2 + 2h \cdot d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}h\sqrt{h} \cdot \beta \cdot \mathsf{I}_\varepsilon) - \frac{h \cdot \beta^2}{4} \\
= &-\frac{3\alpha \cdot \sqrt{h}}{2}(h \cdot \beta^2 + 2d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2h} \cdot \beta \cdot \mathsf{I}_\varepsilon)^{3/2} \\
&- \frac{9\sqrt{h} \cdot \alpha}{2}(h \cdot d^2 \cdot \beta^2 + 2d^3 \cdot \mathsf{N}_\varepsilon + 2\sqrt{2h} \cdot \beta \cdot \mathsf{I}_\varepsilon)^{1/2} \\
&- \gamma \cdot h \cdot (h \cdot \beta^2 + 2d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}\sqrt{h} \cdot \beta \cdot \mathsf{I}_\varepsilon) - \frac{h \cdot \beta^2}{4} .
\end{aligned}$$

Define the following constants.

$$C_\varepsilon^{(1)} = \left(\frac{\varepsilon}{24}\right)^{2/3}; \quad C_\varepsilon^{(2)} = \left(\frac{\varepsilon}{72}\right)^2; \quad C_\varepsilon^{(3)} = \frac{\varepsilon}{16}$$

$$C_\varepsilon^{(i,1)} = \frac{C_\varepsilon^{(i)}}{3}; \quad C_\varepsilon^{(i,2)} = \frac{C_\varepsilon^{(i)}}{6 \cdot \mathsf{N}_\varepsilon}; \quad C_\varepsilon^{(i,3)} = \frac{C_\varepsilon^{(i)}}{6\sqrt{2} \cdot \mathsf{I}_\varepsilon} \qquad \forall\, i \in [3] .$$

When $\varepsilon < 1$, note that all of these constants are strictly less than 1, since $\mathsf{N}_\varepsilon \geq 1$ and $\mathsf{I}_\varepsilon \geq 1$. With this we define some limits on h.

$$\begin{aligned} C^{(1)}_{\max}(\varepsilon) &= \min\left\{\left(C^{(1,1)}_\varepsilon\right)^{3/4} \cdot \frac{1}{\beta^{3/2}} \cdot \frac{1}{\sqrt{\alpha}}\,,\ \left(C^{(1,2)}_\varepsilon\right)^{3} \cdot \frac{1}{\alpha^{2}} \cdot \frac{1}{d^3}\,,\ \left(C^{(1,3)}_\varepsilon\right)^{6/5} \cdot \frac{1}{\alpha^{4/5}} \cdot \frac{1}{\beta^{6/5}}\right\}\,, \\ C^{(2)}_{\max}(\varepsilon) &= \min\left\{\left(C^{(2,1)}_\varepsilon\right)^{1/2} \cdot \frac{1}{d} \cdot \frac{1}{\alpha} \cdot \frac{1}{\beta}\,,\ \left(C^{(2,2)}_\varepsilon\right) \cdot \frac{1}{\alpha^{2}} \cdot \frac{1}{d^3}\,,\ \left(C^{(2,3)}_\varepsilon\right)^{2/3} \cdot \frac{1}{\beta^{2/3}} \cdot \frac{1}{\alpha^{4/3}}\right\}\,, \\ C^{(3)}_{\max}(\varepsilon) &= \min\left\{(C^{(3,1)}_\varepsilon)^{1/2} \cdot \frac{1}{\beta} \cdot \frac{1}{\sqrt{\gamma}}\,,\ (C^{(3,2)}_\varepsilon) \cdot \frac{1}{d} \cdot \frac{1}{\gamma}\,,\ (C^{(3,3)}_\varepsilon) \cdot \frac{1}{\beta^{2/3}} \cdot \frac{1}{\gamma^{2/3}}\right\}\,, \\ C^{(4)}_{\max}(\varepsilon) &= \frac{\varepsilon}{4 \cdot \beta^2}\,. \end{aligned}$$

When $h \leq \min\{C^{(1)}_{\max}(\varepsilon), C^{(2)}_{\max}(\varepsilon), C^{(3)}_{\max}(\varepsilon), C^{(4)}_{\max}(\varepsilon)\}$, it can be verified that

$$\begin{aligned} \frac{3\alpha \cdot \sqrt{h}}{2}(h \cdot \beta^2 + 2d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2h} \cdot \beta \cdot \mathsf{I}_\varepsilon)^{3/2} &\leq \frac{\varepsilon}{16} \\ \frac{9\sqrt{h} \cdot \alpha}{2}(h \cdot d^2 \cdot \beta^2 + 2d^3 \cdot \mathsf{N}_\varepsilon + 2\sqrt{2h} \cdot \beta \cdot \mathsf{I}_\varepsilon)^{1/2} &\leq \frac{\varepsilon}{16} \\ \gamma \cdot h \cdot (h \cdot \beta^2 + 2d \cdot \mathsf{N}_\varepsilon + 2\sqrt{2}\sqrt{h} \cdot \beta \cdot \mathsf{I}_\varepsilon) &\leq \frac{\varepsilon}{16} \\ \frac{h \cdot \beta^2}{4} &\leq \frac{\varepsilon}{16} \end{aligned}$$

and consequently,

$$\mathcal{A}(x, z) \geq -\frac{\varepsilon}{4}\,.$$

It can be verified that when $h \leq b_{\max}(\varepsilon; d, \alpha, \beta, \lambda)$ as defined in eq. (7), with $\mathcal{C}_i(\varepsilon)$ defined in eq. (12) that when $h \leq b_{\max}(\varepsilon)$, $h \leq \min\left\{C^{(1)}_{\max}(\varepsilon), C^{(2)}_{\max}(\varepsilon), C^{(3)}_{\max}(\varepsilon), C^{(4)}_{\max}(\varepsilon)\right\}$. $b_{\max}(\varepsilon)$ is a condensed version where we have used the facts

- $\beta\sqrt{\gamma} \geq \beta^{3/2}\sqrt{\alpha}$,
- $d\gamma \geq d\alpha\beta$,
- if $\gamma \geq \beta^2$, then $\beta^{2/3}\gamma^{2/3} \geq \beta\sqrt{\gamma} \geq \beta^2$, and when $\gamma \leq \beta^2$, then $\beta^{2/3}\gamma^{2/3} \leq \beta\sqrt{\gamma} \leq \beta^2$, and
- if $\alpha \geq \beta$, then $\alpha^{4/3}\beta^{2/3} \geq \alpha^{4/5}\beta^{6/5} \geq \beta^2$, and when $\alpha \leq \beta$, then $\alpha^{4/3}\beta^{2/3} \leq \alpha^{4/5}\beta^{6/5} \leq \beta^2$.

to reduce the cases.

This implies that when $t = e^{-\varepsilon/4}$ which is less than 1,

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)p_z(x)}{\pi(x)p_x(z)} \geq t \;\middle|\; \mathfrak{E}\right] = 1\,.$$

To conclude the proof, from part D.3.2.1, we have that $\mathbb{P}_{z\sim\mathcal{P}_z}(\mathfrak{E}) \geq 1 - \frac{\varepsilon}{4}$. From part D.3.2.3, setting $t = e^{-\varepsilon/4}$ ensures that when conditioning on $\mathfrak{E}$, $\mathbb{P}_{z\sim\mathcal{P}_x}\left[\exp(\mathcal{A}(x, z)) \geq t \mid \mathfrak{E}\right] = 1$. This finally results in

$$\begin{aligned}\mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z)p_z(x)}{\pi(x)p_x(z)}\right\}\right] &\geq t\,\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)p_z(x)}{\pi(x)p_x(z)} \geq t \;\middle|\; \mathfrak{E}\right]\cdot\mathbb{P}_{z\sim\mathcal{P}_x}[\mathfrak{E}] \\ &= \exp(-\varepsilon/4)\cdot(1-\varepsilon/4) \geq (1-\varepsilon/4)^2 \geq 1-\varepsilon/2\end{aligned}$$

and recalling the form of $\mathrm{d}_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{P}_x)$, we get the bound

$$\mathrm{d}_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{P}_x) \leq \frac{\varepsilon}{2}\,.$$

■

D.4. Proofs of corollaries in Section 4

In this section, we provide the proofs for the corollaries in Section 4.

D.4.1. PROOFS OF COROLLARY 7

Proof We begin with the following fact, as stated in Nesterov (2018, Ex. 5.1.1(4)). Let $P \in \mathbb{R}^{d\times d}$ and $P \succeq 0$, $q \in \mathbb{R}^d, r \in \mathbb{R}$. Then, $x \mapsto -\log(r + q^\top x + x^\top P x)$ is a self-concordant function with parameter 1. For $\mathsf{Polytope}(A, b)$, the log-barrier is the sum of m such functions where $r = b_i$, $q = -a_i$ and $P = 0$ for $i \in [m]$. For $\mathsf{Ellipsoids}(\{(c_i, M_i)\}_{i=1}^m)$, the log-barrier is the sum of m such functions as well, where $r = 1 - \|c_i\|_{M_i}^2$, $q = 2M_ic_i$, $P = M_i$ for $i \in [m]$. By Nesterov (2018, Thm. 5.1.1), this is a self-concordant function with parameter 1 as well.

For uniform sampling, since f is a constant function, $\nabla f(x) = 0$ for $x \in \mathrm{int}(\mathcal{K})$. Hence, $\mu = \lambda = \beta = 0$ in assumptions $(\mathbf{A}_2)$ and $(\mathbf{A}_3)$. For assumption $(\mathbf{A}_4)$, Lemmas 10 and 11 states that $\nu = m$ for $\mathcal{K} = \mathsf{Polytope}(A, b)$, and $\nu = 2m$ for $\mathcal{K} = \mathsf{Ellipsoids}(\{(c_i, M_i)\}_{i=1}^m)$. Substituting these values for the parameters in the assumptions in the mixing time bound for the (weakly) convex case in Theorem 6 recovers the statements of the lemmas. ■

D.4.2. PROOF OF COROLLARY 8

Proof The simplex $\mathcal{K}$ is a special polytope, and since ϕ is the log-barrier of $\mathcal{K}$, the self-concordance parameter is 1, as previously discussed in Section D.4.1.

We have the following explicit expression for the gradient of f and Hessians of ϕ and f.

$$\begin{aligned}\nabla f(x) &= \left[-\frac{a_1}{x_1} + \frac{a_{d+1}}{1 - \mathbf{1}^\top x}, \cdots, \frac{a_d}{x_d} + \frac{a_{d+1}}{1 - \mathbf{1}^\top x}\right] \\ \nabla^2 f(x) &= \mathrm{diag}\left(\frac{a_1}{x_1^2}, \cdots, \frac{a_d}{x_d^2}\right) + \frac{a_{d+1}}{(1 - \mathbf{1}^\top x)^2}\mathbf{1}_{d\times d}\,, \\ \nabla^2 \phi(x) &= \mathrm{diag}\left(\frac{1}{x_1^2}, \cdots, , \frac{1}{x_d^2}\right) + \frac{1}{(1 - \mathbf{1}^\top x)^2}\mathbf{1}_{d\times d}\,.\end{aligned}$$

The smallest λ such that $\lambda \cdot \nabla^2\phi(x) - \nabla^2 f(x) \succeq 0$ is $\boldsymbol{a}_{\max}$. The largest μ such that $\nabla^2 f(x) - \mu \cdot \nabla^2\phi(x) \succeq 0$ is $\boldsymbol{a}_{\min}$. Consequently, f is $\boldsymbol{a}_{\max}$-relatively smooth and $\boldsymbol{a}_{\min}$-relatively convex with respect to ϕ.

From the expressions above, for any $v \in \mathbb{R}^d$,

$$\langle \nabla f(x), v\rangle = -\sum_{i=1}^{d} \frac{a_i v_i}{x_i} + \frac{a_{d+1}}{1 - \mathbf{1}^\top x} \cdot \sum_{i=1}^{d} v_i \ , \ \ \langle v, \nabla^2\phi(x) v\rangle = \sum_{i=1}^{d} \frac{v_i^2}{x_i^2} + \frac{1}{(1-\mathbf{1}^\top x)^2} \cdot \left(\sum_{i=1}^{d} v_i\right)^2 .$$

From Lemma 25 with $z_i \leftarrow \frac{v_i}{x_i}$, $w_i \leftarrow v_i$, and $c \leftarrow \frac{1}{1-\mathbf{1}^\top x}$, we have the inequality for any $x \in \mathrm{int}(\mathcal{K})$

$$\langle \nabla f(x), v\rangle^2 \leq \|\boldsymbol{a}\|^2 \cdot \langle v, \nabla^2\phi(x) v\rangle \Leftrightarrow \langle \nabla f(x), \nabla^2\phi(x)^{-1}\nabla f(x)\rangle \leq \|\boldsymbol{a}\|^2 \ ,$$

which shows that f is $\|\boldsymbol{a}\|$-relatively Lipschitz with respect to ϕ.

The constant $\gamma = \frac{\boldsymbol{a}_{\max}}{2} + \|\boldsymbol{a}\|$ is at most $\frac{3\|\boldsymbol{a}\|}{2}$. Since $a_i > 0$ for all i, we instantiate the $\mu > 0$ case of Theorem 6 to get the mixing time guarantee

$$\mathcal{O}\left(\frac{1}{\boldsymbol{a}_{\min}} \cdot \max\left\{d^3,\ d \cdot \|\boldsymbol{a}\|,\ \|\boldsymbol{a}\|^{2/3},\ \|\boldsymbol{a}\|^{3/2},\ \|\boldsymbol{a}\|^2\right\} \cdot \log\left(\frac{\sqrt{M}}{\delta}\right)\right) \tag{15}$$

When $\|\boldsymbol{a}\| \geq 1$, $\|\boldsymbol{a}\|^2$ dominates $\|\boldsymbol{a}\|^{3/2}$ and $\|\boldsymbol{a}\|^{2/3}$, thus recovering the statement of the lemma. ■

D.5. Proofs of other technical lemmas

D.5.1. PROOF OF LEMMA 11

Proof The log-barrier of this set is defined by

$$\phi(x) = -\sum_{i=1}^{m} \log(1 - \|x - c_i\|_{M_i}^2) \ .$$

The Hessian of ϕ is given by

$$\nabla^2\phi(x) = \sum_{i=1}^{m} \frac{2M_i(1 - \|x - c_i\|_{M_i}^2) + 4M_i(x - c_i)(x - c_i)^\top M_i}{(1 - \|x - c_i\|_{M_i}^2)^2} \ .$$

For any $y \in \mathcal{E}_x^{\phi}(\sqrt{r})$,

$$\|y - x\|_{\nabla^2\phi(x)}^2 \leq r \Leftrightarrow \sum_{i=1}^{m} \frac{2\|y - x\|_{M_i}^2(1 - \|x - c_i\|_{M_i}^2) + 4\langle y - x, x - c_i\rangle_{M_i}^2}{(1 - \|x - c_i\|_{M_i}^2)^2} \leq r \ . \tag{16}$$

Let $\mathcal{S}_i$ be the ellipsoid defined by the pair (c_i, M_i). To show that ϕ is a symmetric barrier over $\mathcal{K} = \mathsf{Ellipsoids}(\{(c_i, M_i)\}_{i=1}^m)$, we need to show that there exists $r > 0$, such that for all $x \in \mathrm{int}(\mathcal{K})$,

$$\mathcal{E}_x^{\phi}(1) \subseteq \mathcal{K} \cap (2x - \mathcal{K}) \subseteq \mathcal{E}_x^{\phi}(r) \ . \tag{17}$$

We begin by giving an equivalent algebraic statement for $y \in \mathcal{K} \cap (2x - \mathcal{K})$. Note that $\mathcal{K} \cap (2x - \mathcal{K}) = \bigcap_{i=1}^{m} \{\mathcal{S}_i \cap (2x - \mathcal{S}_i)\}$.
For any $i \in [m]$, let $y \in \mathcal{S}_i \cap (2x - \mathcal{S}_i)$ which is equivalent to $y \in \mathcal{S}_i \wedge (2x - y) \in \mathcal{S}_i$.

$$\begin{aligned}
y \in \mathcal{S}_i &\Leftrightarrow \|y - c_i\|_{M_i}^2 \leq 1 \\
&\Leftrightarrow \|y - x + x - c_i\|_{M_i}^2 \leq 1 \\
&\Leftrightarrow \|y - x\|_{M_i}^2 + \|x - c_i\|^2 + 2\langle y - x, x - c_i \rangle_{M_i} \leq 1 \\
&\Leftrightarrow 2\langle y - x, x - c_i \rangle_{M_i} \leq 1 - \|x - c_i\|_{M_i}^2 - \|y - x\|_{M_i}^2 . \\
2x - y \in \mathcal{S}_i &\Leftrightarrow \|2x - y - c_i\|_{M_i}^2 \leq 1 \\
&\Leftrightarrow \|x - c_i + x - y\|_{M_i}^2 \leq 1 \\
&\Leftrightarrow \|x - c_i\|_{M_i}^2 + \|y - x\|_{M_i}^2 + 2\langle x - c_i, x - y \rangle_{M_i} \leq 1 \\
&\Leftrightarrow -2\langle y - x, x - c_i \rangle_{M_i} \leq 1 - \|x - c_i\|_{M_i}^2 - \|y - x\|_{M_i}^2 .
\end{aligned}$$

Therefore, $y \in \mathcal{S}_i \cap (2x - \mathcal{S}_i)$ if and only if

$$2\,|\langle y - x, x - c_i \rangle_{M_i}| \leq 1 - \|x - c_i\|_{M_i}^2 - \|y - x\|_{M_i}^2 .$$

Squaring both sides, and moving terms around we have that $y \in \mathcal{K} \cap (2x - \mathcal{K})$ if and only if

$$\underbrace{\frac{2(1 - \|x - c_i\|_{M_i}^2)\|y - x\|_{M_i}^2 + 4\langle y - x, x - c_i \rangle_{M_i}^2}{(1 - \|x - c_i\|_{M_i}^2)^2}}_{a_i} \leq \underbrace{1 + \frac{\|y - x\|_{M_i}^4}{(1 - \|x - c_i\|_{M_i}^2)^2}}_{b_i} \quad \forall\, i \in [m] .$$

By the definition as stated in eq. (16), we have following equivalence

$$y \in \mathcal{E}_x^{\phi}(1) \Rightarrow y \in \mathcal{K} \cap (2x - \mathcal{K}) \quad \Leftrightarrow \quad \sum_{i=1}^{m} a_i \leq 1 \Rightarrow a_i \leq b_i \quad \forall\, i \in [m]$$

Suppose there exists $i \in [m]$ such that $a_i > b_i$, then $a_i > 1$ since $b_i \geq 1$. By definition, $a_i \geq 0$ for all $i \in [m]$, and this in conjuction with the assumption implies that $\sum_{i=1}^{m} a_i > 1$. By contraposition, this is equivalent to stating that $\sum_{i=1}^{m} a_i \leq 1$ implies $a_i \leq b_i$ for all $i \in [m]$. This proves that $\mathcal{E}_x^{\phi}(1) \subseteq \mathcal{K} \cap (2x - \mathcal{K})$ as $y \in \mathcal{E}_x^{\phi}(1)$ was chosen arbitrarily.

We now show the second statement in the containment chain (eq. (17)). If $y \in \mathcal{S}_i \cap (2x - \mathcal{S}_i)$ for some $i \in [m]$,

$$\begin{aligned}
\|y - x\|_{M_i}^2 &= \|y - 2x + c_i + x - c_i\|_{M_i}^2 \\
&= \|2x - y - c_i\|_{M_i}^2 + \|x - c_i\|_{M_i}^2 - 2\langle 2x - y - c_i, x - c_i \rangle_{M_i} \\
&\overset{(a)}{\leq} 1 + \|x - c_i\|_{M_i}^2 - 2\langle x - c_i + x - y, x - c_i \rangle_{M_i} \\
&= 1 - \|x - c_i\|_{M_i}^2 + 2\langle y - x, x - c_i \rangle_{M_i} \\
&\overset{(b)}{\leq} 1 - \|x - c_i\|_{M_i}^2 + 1 - \|x - c_i\|_{M_i}^2 - \|y - x\|_{M_i}^2 \\
&\leq 2 - 2\|x - c_i\|_{M_i}^2 - \|y - x\|_{M_i}^2
\end{aligned}$$

Step (a) uses the fact that $y \in 2x - \mathcal{S}_i$, and step (b) uses the equivalence for $y \in \mathcal{S}_i$ shown above. This concludes that if $y \in \mathcal{K} \cap (2x - \mathcal{K})$, $\|y - x\|_{M_i}^2 \leq 1 - \|x\|_{M_i}^2$ for all $i \in [m]$.

We can use the above assertion to bound the b_i quantities by 2. In summary, if $y \in \mathcal{K} \cap (2x - \mathcal{K})$, then $a_i \leq b_i \leq 2$ for all $i \in [m]$. Note that this implies $\sum_{i=1}^m a_i \leq 2m$, which is equivalent to stating that if $y \in \mathcal{K} \cap (2x - \mathcal{K})$, then $y \in \mathcal{E}_x^{\phi}(\sqrt{2m})$. Due to y being arbitrary, we have shown that $\mathcal{K} \cap (2x - \mathcal{K}) \subseteq \mathcal{E}_x^{\phi}(\sqrt{2m})$, thus completing the proof. ∎

Appendix E. Miscellaneous algebraic lemmas

Lemma 22 *Let $f : (0, \infty) \to (0, \infty)$ be defined as $f(x) = x - 1 - \log(x)$. Then for all $x > 0$,*

$$f(x) \leq \frac{(x-1)^2}{x}.$$

Proof We begin with an algebraic manipulation.

$$\begin{aligned} g(x) &= f(x) - \frac{(x-1)^2}{x} \\ &= x - 1 - \log(x) - \frac{(x-1)^2}{x} \\ &= \frac{x^2 - x - (x-1)^2}{x} - \log(x) \\ &= \frac{x-1}{x} - \log(x). \end{aligned}$$

The function $g(x) = \frac{x-1}{x} - \log(x)$ has derivative $g'(x) = \frac{(1-x)}{x^2}$. The only solution to $g'(x) = 0$ for $x > 0$ is $x = 1$. Moreover, $g''(x) = -\frac{2}{x^3} + \frac{1}{x^2}$, and $g''(1) = -1 < 0$. This implies that $g(x)$ attains its maximum at $x = 1$, and this maximum value is $1 - 1 - \log(0) = 0$. As a result of the calculation above, for any $x > 0$, $f(x) - \frac{(x-1)^2}{x} \leq \max\limits_{x>0} g(x) = 0$, which concludes the proof. ∎

Lemma 23 *Let $t \in [0, 0.5]$. Then,*

$$\frac{t^2}{4}\left(1 - \frac{1}{(1-t)^2}\right) \geq -\frac{3}{2}t^3$$

Proof We begin with the quantity

$$\frac{t^2}{4}\left(1 - \frac{1}{(1-t)^2}\right) + \frac{3}{2}t^3 = \frac{t^3(6t^2 - 11t + 4)}{4(1-t)^2} = \frac{t^3(3t-4)(2t-1)}{4(1-t)^2}\ .$$

When $t \in [0, 0.5]$, $t^3, (1-t)^2 \geq 0$, and $(3t-4), (2t-1) \leq 0$. This implies that

$$\frac{t^3(3t-4)(2t-1)}{4(1-t)^2} \geq 0 \Leftrightarrow \frac{t^2}{4}\left(1 - \frac{1}{(1-t)^2}\right) + \frac{3}{2}t^3 \geq 0\ ,$$

which completes the proof. ∎

Lemma 24 *Let $t \in [0, 0.5]$. Then,*

$$\log(1-t) \geq -\frac{3}{2}t$$

Proof Consider the function $f(t) = \log(1-t) + \frac{3}{2}t$. The second derivative of f is $f''(t) = -\frac{1}{(1-t)^2}$. Note that $f''(t) < 0$ for $t \in [0, 0.5]$, which implies that f is strictly concave in $[0, 0.5]$. Consequently, the minimum of f restricted to $[0, 0.5]$ is attained at either 0 or 0.5. Hence, for any $t \in [0, 0.5]$,

$$f(t) \geq \min\{f(0), f(0.5)\} = \min\{0, 0.75 + \log(0.5)\} = 0\ .$$

■

Lemma 25 *Let $a \in \mathbb{R}^{d+1}$. Then, for any $w, z \in \mathbb{R}^d$, $c \in \mathbb{R}$,*

$$\left(-\sum_{i=1}^{d} a_i z_i + c \cdot a_{d+1} \cdot \sum_{i=1}^{d} w_i\right)^2 \leq \|a\|^2 \cdot \left(\sum_{i=1}^{d} z_i^2 + c^2 \cdot \left\{\sum_{i=1}^{d} w_i\right\}^2\right)$$

Proof Construct two vectors of length $d+1$ as shown below.

$$\mathsf{A} = [-a_1, \cdots, -a_d, a_{d+1}]\ ; \qquad \mathsf{B} = \left[z_1, \cdots, z_d, c \cdot \sum_{i=1}^{d} w_i\right]\ .$$

The LHS of the inequality in the statement is $(\mathsf{A}^\top \mathsf{B})^2$. Using the Cauchy-Schwarz inequality,

$$(\mathsf{A}^\top \mathsf{B})^2 \leq \|\mathsf{A}\|^2 \|\mathsf{B}\|^2 = \|a\|^2 \cdot \left(\sum_{i=1}^{d} z_i^2 + c^2 \cdot \left\{\sum_{i=1}^{d} w_i\right\}^2\right)$$

thus obtaining the statement of the lemma. ■