Riemannian Langevin Monte Carlo schemes for sampling PSD matrices with fixed rank *

Tianmin Yu†, Shixin Zheng‡, Jianfeng Lu§, Govind Menon†, AND Xiangxiong Zhang‡

Abstract. This paper introduces two explicit schemes to sample matrices from Gibbs distributions on $\mathcal{S}_+^{n,p}$, the manifold of real positive semi-definite (PSD) matrices of size $n \times n$ and rank p. Given an energy function $\mathcal{E} : \mathcal{S}_+^{n,p} \to \mathbb{R}$ and certain Riemannian metrics g on $\mathcal{S}_+^{n,p}$, these schemes rely on an Euler-Maruyama discretization of the Riemannian Langevin equation (RLE) with Brownian motion on the manifold. We present numerical schemes for RLE under two fundamental metrics on $\mathcal{S}_+^{n,p}$: (a) the metric obtained from the embedding of $\mathcal{S}_+^{n,p} \subset \mathbb{R}^{n\times n}$; and (b) the Bures-Wasserstein metric corresponding to quotient geometry. We also provide examples of energy functions with explicit Gibbs distributions that allow numerical validation of these schemes.

Key words. Langevin dynamics, sampling schemes, Bures-Wasserstein metric, Burer-Monteiro, embedded geometry, positive semi-definite matrices, Riemannian optimization

MSC codes.

1. Introduction.

1.1. Problem statement. Consider the space of real, symmetric positive semi-definite matrices with size $n \times n$ and rank p, denoted by

$$\mathcal{S}_+^{n,p} = \{X \in \mathbb{R}^{n\times n} | X = X^T, X \succeq 0, \operatorname{rank}(X) = p\}. \tag{1.1}$$

Given an energy $\mathcal{E} : \mathcal{S}_+^{n,p} \to \mathbb{R}$ and a parameter $\beta > 0$ referred to as the inverse temperature, our goal is to sample efficiently from the Gibbs distribution

$$\rho_\beta(X) = \frac{1}{Z_\beta} e^{-\beta\mathcal{E}(X)} \rho_{\mathrm{ref}}(X), \quad Z_\beta = \int_{\mathcal{S}_+^{n,p}} e^{-\beta\mathcal{E}(X')} \rho_{\mathrm{ref}}(X')\, dX'. \tag{1.2}$$

Gibbs measures must be defined with respect to a base measure. In this work, we equip the space $\mathcal{S}_+^{n,p}$ with a Riemannian metric g and choose $\rho_{\mathrm{ref}}(X)dX = \sqrt{\det g(X)}dX$ to be the canonical volume form associated to the metric g. This volume form is expressed in coordinates for the metrics studied in this paper in Section 4.

This sampling problem is related to the optimization problem $\min_{X\in\mathcal{S}_+^{n,p}} \mathcal{E}(X)$ since in the limit $\beta \to \infty$ the Gibbs distribution concentrates at the global minima of $\mathcal{E}(X)$. Minimization problems over the space $\mathcal{S}_+^{n,p}$ arise in many areas, especially semidefinite programming and machine learning, and have been studied extensively. Gibbs distributions originate in statistical physics, while the sampling problem may

* **Funding:** S.Z. and X.Z. are supported by NSF DMS-2208518. J.L. is supported in part by NSF DMS-2012286 and DMS-2309378. G.M. is supported in part by NSF DMS-2107205

†Division of Applied Mathematics, Brown University, Providence, RI (tianmin_yu@brown.edu, govind_menon@brown.edu).

‡Department of Mathematics, Purdue University, West Lafayette, IN (zheng513@purdue.edu, zhan1966@purdue.edu).

§Departments of Mathematics, Physics, and Chemistry, Duke University, Durham, NC (jianfeng@math.duke.edu).

also be seen as a stochastic variant of the optimization problem. For these reasons, the sampling problem has a broad range of applications; see Section 1.5 below.

The main contribution of this paper are efficient sampling schemes for ρ_β based on Langevin dynamics. Our approach builds on the geometric theory of optimization; in particular, we extend Riemannian optimization on $\mathcal{S}_+^{n,p}$ [34, 38] to Gibbs sampling as follows. In [34] it was recognized that two commonly used gradient descent schemes over $\mathcal{S}_+^{n,p}$ are time discretizations of *Riemannian* gradient flows, where $\mathcal{S}_+^{n,p}$ is equipped with the two natural Riemannian metrics listed below. We combine this observation with the theory of Brownian motion on Riemannian manifolds to obtain Riemannian Langevin equations and explicit sampling schemes.

The reader unfamiliar with these concepts should note that while the abstract theory serves to guide our work, the schemes presented in this paper may be implemented without requiring a complete understanding of the underlying theory. Further, while this paper is focused on the two numerical schemes below, the underlying framework can be used to extend other Riemannian gradient descent schemes to sampling schemes for the Gibbs measure. The new phenomenon that arises is the interplay between Brownian motion and curvature in the Riemannian Langevin equation. This interplay has been studied in depth by two of the authors (TY and GM) and their co-workers in recent papers for geometries used in optimization and physics [20, 28, 29].

1.2. Two Riemannian metrics on $\mathcal{S}_+^{n,p}$. Given $X \in \mathcal{S}_+^{n,p}$, let $X = YY^T$ be a low-rank decomposition where $Y \in \mathbb{R}^{n\times p}$. We use two fundamental metrics on $\mathcal{S}_+^{n,p}$ obtained from this parametrization, from the Euclidean metric for either the variable X or the variable Y through the use of Riemannian embedding and Riemannian submersion respectively. These are the two most natural ways of defining metrics on $\mathcal{S}_+^{n,p}$.

The flat metric for X corresponds to the embedded geometry of $\mathcal{S}_+^{n,p}$ in the Euclidean space $\mathbb{R}^{n\times n}$ [34]. Precisely, we consider the natural Riemannian embedding $\mathcal{S}_+^{n,p} \hookrightarrow \mathbb{R}^{n\times n}$ and use the Frobenius norm on $\mathbb{R}^{n\times n}$ to define a metric on $\mathcal{S}_+^{n,p}$. Denote it by g_E, then $g_E(A, B) = \mathrm{Tr}\,(A^T B)$ for any two square matrices A, B in the tangent space of $\mathcal{S}_+^{n,p}$, where Tr denotes the trace of a matrix.

On the other hand, we may also use the flat geometry on Y to define a metric on $\mathcal{S}_+^{n,p}$. We observe that if $YY^T = X$, then it is also true that $\tilde{Y}\tilde{Y}^T = X$ where $\tilde{Y} = YO$ and $O \in \mathcal{O}_p$, the orthogonal group of dimension p. Thus, we may identify $\mathcal{S}_+^{n,p} \approx \mathbb{R}_*^{n\times p}/\mathcal{O}_p$, as a quotient space, with a quotient map

$$\begin{aligned} \pi : \mathbb{R}_*^{n\times p} &\to \mathbb{R}_*^{n\times p}/\mathcal{O}_p \\ Y &\mapsto [Y] = \{YO \mid O \in \mathcal{O}_p\}. \end{aligned}$$

Here $\mathbb{R}_*^{n\times p}$ denotes full rank matrices.

The quotient space structure can be enhanced with a Riemannian metric through the use of Riemannian submersion. Roughly, the metric for X corresponds to the metric for Y in a manner that respects the splitting of the tangent space at Y into the space of the group action and its complement. If $\mathbb{R}_*^{n\times p}$ is equipped with Euclidean metric, then the metric induced by the submersion is often called the Bures-Wasserstein metric on $\mathcal{S}_+^{n,p} \approx \mathbb{R}_*^{n\times p}/\mathcal{O}_p$, denoted by g_{BW} (see [2, 26, 27]).

1.3. Langevin dynamics and the Riemannian Langevin equation. We now explain how Langevin equations may be defined intrinsically on $(\mathcal{S}_+^{n,p}, g)$.

Let us first recall the Langevin equation on $\mathbb{R}^n$. Assume given a potential or energy function $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ and let W_t denote the standard Wiener process on $\mathbb{R}^n$. The Langevin equation for the potential $\mathcal{E}$ is the Itô differential equation

$$dx_t = -\nabla\mathcal{E}(x_t)\,dt + \sqrt{\frac{2}{\beta}}\,dW_t. \tag{1.3}$$

The Fokker-Planck equation describes the evolution of the probability density of x_t. With $\rho(x,t)\,dx = \mathbb{P}(x_t \in (x, x+dx))$, we have

$$\partial_t\rho = \frac{1}{\beta}\triangle\rho + \nabla\cdot(\rho\nabla\mathcal{E})\,. \tag{1.4}$$

The Gibbs density (with reference density being uniform with respect to Lebesgue measure) is the unique equilibrium of equation (1.4) under natural growth assumptions on the energy $\mathcal{E}$ as $|x| \to \infty$.

The Langevin equation immediately yields a numerical scheme for (approximate) sampling from the Gibbs distribution. Fix a step size $\Delta t > 0$, let $t_k = k\Delta t$, $k = 0, 1, \ldots$, and let x_k denote the numerical approximation to (1.3) at time t_k. The Euler-Maruyama scheme to approximate equation (1.3), also known as Langevin Monte Carlo in the statistics literature, is

$$x_{k+1} = x_k - \Delta t\nabla\mathcal{E}(x_k) + \sqrt{\frac{2\Delta t}{\beta}}\xi_k, \tag{1.5}$$

where $\xi_k = (\xi_k^1, \ldots, \xi_k^n)$ is an i.i.d. sequence of standard Gaussian vectors in $\mathbb{R}^n$. This scheme is explicit. In order to extend it to sampling from (1.2) we must understand how to modify the Langevin equation on the Riemannian manifold $(\mathcal{S}_+^{n,p}, g)$.

First, the term $\nabla\mathcal{E}$ must be replaced by the Riemannian gradient, written as $\operatorname{grad}\mathcal{E}$. The more subtle modification of equation (1.3) concerns the noise. The natural analogy is to replace the Wiener process W_t on $\mathbb{R}^n$ with Brownian motion on the Riemannian manifold $(\mathcal{S}_+^{n,p}, g)$ at inverse temperature β, denoted $\boldsymbol{B}_t^{g,\beta}$. This yields the (formal) Riemannian Langevin equation on $(\mathcal{S}_+^{n,p}, g)$

$$\mathrm{d}\boldsymbol{X}_t = -\operatorname{grad}\mathcal{E}(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta}. \tag{1.6}$$

This equation is only formal because stochastic differential equations on manifolds must be defined using the Stratonovich formulation in order to ensure coordinate independence (Itô differentials do not satisfy the chain rule, while Stratonovich differentials do) [16, 19]. On the other hand, Itô differential equations are convenient for analysis as well as simulation. Thus, in formulating the Riemannian Langevin equation, it is necessary to first formulate the appropriate Stratonovich equation and then compute the deterministic Itô–Stratonovich correction. A central observation in our work is that this correction term is due to curvature and is explicitly computable for several Riemannian geometries relevant to optimization [17, 20, 28, 29].

1.4. Riemannian Langevin Monte Carlo sampling schemes. For the two metrics considered in this paper, the Itô–Stratonovich correction due to curvature may also be computed explicitly, yielding the SDEs in Section 2. The rigorous analysis of these SDEs is presented in the companion paper [36], and we focus on numerical algorithms in this paper. The Euler-Maruyama approximation to these SDEs yields the numerical sampling schemes listed below.

The SDEs also admit other numerical approximations. We have chosen the Euler-Maruyama schemes because these schemes are fully explicit, simple to state, implement and numerically validate. They are generalizations of the popular unadjusted Langevin Monte Carlo for sampling in Euclidean spaces. Further, these schemes reduce to deterministic Riemannian gradient descent methods in the limit $\beta \to \infty$.

1.4.1. Scheme E for the embedded geometry. For the embedded manifold $(\mathcal{S}_+^{n,p}, g_E)$, the scheme is

$$X_{k+1} = \mathrm{P}_{\mathcal{S}_+^{n,p}}\left[X_k - \Delta t \,\mathrm{grad}\,\mathcal{E}(X_k) + Q_k\left(\sqrt{\frac{2\Delta t}{\beta}}\begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & 0\end{bmatrix} + \frac{\Delta t}{\beta}\sum_{i=1}^{p}\frac{1}{\lambda_i}\begin{bmatrix}0 & 0\\ 0 & I_{n-p}\end{bmatrix}\right)Q_k^T\right], \tag{1.7}$$

where $\mathrm{P}_{\mathcal{S}_+^{n,p}}$ is the Euclidean projection to $\mathcal{S}_+^{n,p}$, and $X_k = Q_k\Lambda Q_k^T$ is the full SVD of $X_k = \mathcal{S}_+^{n,p}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p > 0$. The entries of B_{12} are i.i.d. drawn from $\sqrt{\frac{1}{2}}\mathcal{N}(0,1)$. The entries of the symmetric B_{11} are defined as follows: the diagonal entries are i.i.d. drawn from $\mathcal{N}(0,1)$, and off-diagonal entries are $b_{ij} = b_{ji} \sim \sqrt{\frac{1}{2}}\mathcal{N}(0,1)$. When $\beta = \infty$, equation (1.7) reduces to $X_{k+1} = \mathrm{P}_{\mathcal{S}_+^{n,p}}\left(X_k - \Delta t\,\mathrm{grad}\,\mathcal{E}(X_k)\right)$, which is the Riemannian gradient descent on $(\mathcal{S}_+^{n,p}, g_E)$, see [1, 38]. We refer to (1.7) as Scheme E.

In this scheme, the term $\dfrac{\Delta t}{\beta}\sum\limits_{i=1}^{p}\dfrac{1}{\lambda_i}\begin{bmatrix}0 & 0\\ 0 & I_{n-p}\end{bmatrix}$ in equation (1.7) is the correction due to the mean curvature of the embedding of $\mathcal{S}_+^{n,p} \hookrightarrow \mathbb{R}^{n\times n}$.

1.4.2. Scheme BW for the Bures-Wasserstein metric. For the quotient manifold $(\mathbb{R}_*^{n\times p}/\mathcal{O}_p, g_{BW})$, the scheme is

$$Y_{k+1} = Y_k - \Delta t 2\nabla\mathcal{E}(Y_kY_k^T)Y_k + \sqrt{\frac{2\Delta t}{\beta}}B_k + \frac{\Delta t}{\beta}U_k\left[\sum_{j:j\neq i}\frac{\sigma_i}{\sigma_i^2+\sigma_j^2}\right]_{ii}V_k^T, \tag{1.8}$$

where B_k is n-by-p matrix with entries being i.i.d. standard Gaussian, $Y_k = U_k\Sigma_kV_k^T \in \mathbb{R}^{n\times p}$ is the compact SVD with singular values σ_i, and $\left[\sum\limits_{j:j\neq i}\frac{\sigma_i}{\sigma_i^2+\sigma_j^2}\right]_{ii}$ is the diagonal matrix whose i-th diagonal entry is $\sum\limits_{j:j\neq i}\frac{\sigma_i}{\sigma_i^2+\sigma_j^2}$. We refer to (1.8) as Scheme BW. The Riemannian Langevin Monte Carlo scheme (1.8) can be viewed as a natural extension of Burer-Monteiro gradient descent method

$$Y_{k+1} = Y_k - \Delta t\, 2\nabla\mathcal{E}(Y_kY_k^T)Y_k, \tag{1.9}$$

which is the simplest low-rank gradient descent method for minimizing $\mathcal{E}(X)$ under the constraint $X \in \mathcal{S}_+^{n,p}$. It is clear that as $\beta \to \infty$, (1.8) reduces to (1.9). The Burer-Monteiro gradient descent method is equivalent to a Riemannian gradient descent method on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ with Bures-Wasserstein metric, see [38].

1.4.3. Gibbs distribution sampling and numerical validation. While the Gibbs distribution always has the same density function $e^{-\beta\mathcal{E}}$ with respect to ρ_{ref}, the reference density ρ_{ref} depends on the metric. Thus, the two schemes (1.7) and (1.8), generate samples for two different probability distributions. In order to validate our schemes, we choose energy functions that allow an explicit computation of these densities for both metric. These energy functions yield matrix integrals of independent analytic interest. They also allow side-to-side benchmarking for different Gibbs samplers on $\mathcal{S}_+^{n,p}$. We demonstrate the efficiency of sampling from these Gibbs distributions numerically. Further analysis on convergence to equilibrium as $t \to \infty$ using the Bakry-Emery criterion is considered in the companion paper [36].

Finally, while we do not discuss their convergence and efficiency approximating SDE as the step-size $\Delta t \to 0$; this is possible following existing approximation results [22, 8, 24].

1.5. Some applications and related work.

1.5.1. Applications of PSD matrices. Positive semi-definite (PSD) fixed rank matrices arise in many problems such as distance matrices [33] and covariance matrices in statistics, and have been used in applications including kernels in machine learning [30], semidefinite optimization [4], quantum information, etc. Riemannian optimization algorithms over $\mathcal{S}_+^{n,p}$ under different metrics have been well studied, e.g., see [34, 18, 27, 38] and references therein.

1.5.2. Langevin dynamics and Monte Carlo schemes on manifolds.

There is an extensive literature on Langevin dynamics in statistics and related areas, with interest in nonconvex optimization [6, 7], as well as machine learning such as generative models [12].

In recent years, there has been interest in studying Langevin diffusion and Monte Carlo Markov Chain (MCMC) schemes on manifolds [9, 10, 15, 3, 5, 37, 31, 13, 23, 24]. In this paper, we are interested in Riemannian Langevin Monte Carlo schemes on $\mathcal{S}_+^{n,p}$.

In the statistics literature, manifold Langevin schemes have been studied in [15, 5]. However, these schemes apply only to simpler embedded manifolds $\mathcal{M} \subset \mathbb{R}^n$ with explicit geodesics such as the sphere and Stiefel manifolds. The above schemes do not directly apply to the manifold $\mathcal{S}_+^{n,p}$, even for the embedded geometry. In [37], a sampling scheme using projection to surface is constructed; however, this is not a Langevin scheme.

In general, a Langevin scheme can be used for either optimization [35, 24], or Monte Carlo type numerical integration, which is common in Bayesian statistic. For optimization, stochastic optimization by Langevin dynamics with simulated annealing is an established approach [25]. In [6], underdamped Langevin schemes are shown to be much more efficient than the overdamped case (1.5). For sampling, Metropolis-adjusted Langevin algorithm [15] is often used. For simplicity, we focus on the simple schemes (1.7) and (1.8) without considering any of simulated annealing, underdamped Langevin, or Metropolis-adjustment, to which it is possible to extend our schemes. Though the Riemannian optimization on $\mathcal{S}_+^{n,p}$ can be easily extended to Hermitian PSD matrices of fixed rank [38], we remark that such an extension for Langevin dynamics would be significantly different.

1.6. Organization of the paper. In Section 2, we state the explicit formulae for the SDE (1.6) and Gibbs measure on the manifold $\mathcal{S}_+^{n,p}$ under two metrics g_E and g_{BW}. We then derive the schemes (1.7) and (1.8) in Section 3. The energy functions and Gibbs distributions used to benchmark the schemes are presented in Section 4. The numerical results are studied in Section 5.

2. Riemannian Langevin equations on $\mathcal{S}_+^{n,p}$. In this section, we state the Itô form of the Riemannian Langevin equation (1.6) for both Riemannian geometries studied in this paper. The theoretical basis for these SDEs is discussed at greater depth in [36]. The main ideas are as follows: (a) the abstract theory of Brownian motion on Riemannian manifolds is used to define the Riemannian Langevin equation in Stratonovich form for the metrics g_E and g_{BW} on $\mathcal{S}_+^{n,p}$; (b) the Itô-Stratonovich conversion rule is used to compute the associated Itô form of these SDEs and it is observed that the Itô-Stratonovich correction term corresponds to mean curvature. This approach yields the SDEs below. These SDEs are used to develop numerical schemes in Section 3.

2.1. The Riemannian Langevin equation for embedded geometry $(\mathcal{S}_+^{n,p}, g_E)$. Let $X \in S_+^{n,p}$ have the compact SVD $X = U\Lambda U^T$ with $U \in \mathbb{R}^{n\times p}$. Let $U_\perp \in \mathbb{R}^{n\times(n-p)}$ be a matrix with columns orthonormal to columns of U. The tangent space of $S_+^{n,p}$ at $X = U\Lambda U^T \in S_+^{n,p}$ is given by [34, 38]:

(2.1)
$$T_X S_+^{n,p} = \left\{ \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} H & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix} : \forall K \in \mathbb{R}^{(n-p)\times p}, \forall H \in \mathbb{R}^{p\times p}, H^T = H \right\}.$$

The induced metric g_E by the embedding $\mathcal{S}_+^{n,p} \hookrightarrow \mathbb{R}^{n\times n}$ is then defined as

$$g_E(A, B) = \mathrm{Tr}\,(A^T B), \quad \forall A, B \in T_X S_+^{n,p},$$

which is the Frobenius inner product for two matrices.

Equation (1.6) describes the evolution of a point $\boldsymbol{X}_t \in \mathcal{S}_+^{n,p}$ in abstract terms. We now rewrite it in a simpler equivalent form describing the evolution of the entries of the matrix entries $\{(X_t)_{ij}\}_{i,j=1}^n$ representing $\boldsymbol{X}_t$. Let us write $X = U\Lambda U^T$ for the compact singular value decomposition (SVD) of X. We further assume that the singular values $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ are written in decreasing order. We suppress the subscript t in the following equations, though the reader should note that U and Λ depend on X_t.

Then we find that the law of X_t is determined by the Itô differential equation

$$\mathrm{d}X_t = -\mathrm{grad}\,\mathcal{E}(X_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^{n,p,X_t} + \frac{1}{\beta}H(X_t)\mathrm{d}t. \tag{2.2}$$

In this equation, the stochastic forcing W_t^{n,p,X_t} is the orthogonal projection of white noise in $\mathbb{R}^{n\times n}$ onto $T_{X_t}\mathcal{S}_+^{n,p}$. Precisely, given W_t^i for $1 \le i \le n$ and $W_t^{i,j}$ for $1 \le i <$

$j \leq n$ independent standard one-dimensional Wiener process, we set

$$\mathrm{d}W_t^{n,p,X_t} = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \mathrm{d}W_t^1 & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p} & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p} & \cdots & \mathrm{d}W_t^{p} & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,n} \\ \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,p+1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,n} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,n} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix},$$

The term $H(X_t)$ is the mean curvature of the embedding $\mathcal{S}_+^{n,p} \to \mathbb{R}^{n\times n}$. We adopt the convention in geometric analysis: the mean curvature is defined as the trace of the second fundamental form of the embedding. Explicitly, we have

$$H(X_t) = \left(\sum_{i=1}^{p} \frac{1}{\lambda_i}\right) \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{n-p} \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix}. \tag{2.3}$$

The following feature of equation (2.2) is fundamental. The stochastic forcing is the naive projection of white noise in the ambient space $\mathbb{R}^{n\times n}$ onto $T_{X_t}\mathcal{S}_+^{n,p}$. Intuitively, when one uses the Euler-Maruyama discretization, the role of this term is to update X_t by taking unbiased random steps in any direction in the tangent space. However, Itô calculus has a subtle interplay with the geometry of the embedding, and in order to keep X_t on the manifold $\mathcal{S}_+^{n,p}$, it is necessary to include the correction term given by the mean curvature.

2.2. The Riemannian Langevin equation for Bures-Wasserstein geometry $(\mathcal{S}_+^{n,p}, g_{BW})$. The manifold $S_+^{n,p}$ can also be viewed as a quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, for which the noncompact Stiefel manifold $\mathbb{R}_*^{n\times p}$ is called the *total space*. Denote the natural projection as

$$\pi : \mathbb{R}_*^{n\times p} \to \mathbb{R}_*^{n\times p}/\mathcal{O}_p.$$

For any $Y \in \mathbb{R}_*^{n\times p}$, the equivalence class containing Y is

$$[Y] = \pi^{-1}(\pi(Y)) = \{YO \mid O \in \mathcal{O}_p\},$$

which is an embedded submanifold of $\mathbb{R}_*^{n\times p}$ (see e.g., [1, Prop. 3.4.4]). The tangent space of $[Y]$ at Y is therefore a subspace of $T_Y\mathbb{R}_*^{n\times p}$ called the *vertical space* at Y, and is denoted by $\mathcal{V}_Y = \left\{Y\Omega \mid \Omega^T = -\Omega, \Omega \in \mathbb{R}^{p\times p}\right\}$, see [38].

Define

$$\begin{aligned} \theta : \mathbb{R}_*^{n\times p} &\to \mathcal{S}_+^{n,p} \\ Y &\mapsto YY^T. \end{aligned}$$

Then θ is invariant under the equivalence relation and induces a bijection $\tilde{\theta}$ on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ such that $\theta = \tilde{\theta} \circ \pi$. For any function $\mathcal{E}(X)$ defined on $\mathcal{S}_+^{n,p}$, there is a function F defined on $\mathbb{R}_*^{n\times p}$ that induces $\mathcal{E}$: for any $X = YY^T \in \mathcal{S}_+^{n,p}$, $F(Y) := \mathcal{E} \circ \theta(Y) = \mathcal{E}(YY^T)$. This is summarized in the diagram below:

$$\begin{array}{ccccc} \mathbb{R}_*^{n\times p} & & & & \\ \downarrow{\scriptstyle \pi} & \searrow^{\theta := \tilde{\theta}\circ\pi} & & & \\ \mathbb{R}_*^{n\times p}/\mathcal{O}_p & \xleftrightarrow{\tilde{\theta}} & \mathcal{S}_+^{n,p} & \xrightarrow{\mathcal{E}} & \mathbb{R} \end{array}$$

In particular, $\mathcal{S}_+^{n,p}$ is diffeomorphic to $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ under $\tilde{\theta}$, see [38]. For any $Y \in \mathbb{R}_*^{n\times p}$, the flat metric for the total space $\mathbb{R}_*^{n\times p}$, correction term corresponds to mean curvature.

$$g(a,b) = \mathrm{Tr}\,(a^T b), \forall a, b \in T_Y\mathbb{R}_*^{n\times p} = \mathbb{R}^{n\times p}$$

induces a metric on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, which is called Bures-Wasserstein metric, see [27, 26, 38]. Another way to understand the Bures-Wasserstein metric at $X \in \mathcal{S}_+^{n,p} \backsimeq \mathbb{R}_*^{n\times p}/\mathcal{O}_p$ is via the map θ:

$$\begin{aligned} g_{BW}(A,B) = \mathrm{Tr}\,(ab^T) \qquad & \forall A, B \in T_X\mathcal{S}_+^{n,p}, a, b \in T_Y\mathbb{R}_*^{n\times p} \\ & \text{s.t. } \mathrm{d}\theta(Y)[a] = A, \mathrm{d}\theta(Y)[b] = B, a, b \in \mathrm{Ker}(\mathrm{d}\theta(Y))^\perp \end{aligned} \tag{2.4}$$

where X has decomposition $X = YY^T$, $\mathrm{d}\theta(Y)[a] = Ya^T + aY^T$ is the differential of θ at Y, and $a \in \mathrm{Ker}(\mathrm{d}\theta(Y))^\perp \Leftrightarrow Y^T a = a^T Y$.

The Riemannian Langevin equation is now determined by the geometry of Riemannian submersion. We must obtain an Itô differential equation for Y_t, such that $X_t = Y_t Y_t^T$ is a matrix that has the same law as the solution to (1.6) in $(\mathcal{S}_+^{n,p}, g_{BW})$.

In comparison with equation (2.2), we see that the natural choice for white noise driving Y_t is white noise in $\mathbb{R}^{n\times p}$. This is the stochastic differential $\mathrm{d}W_t$, where $W_t = \{W_t^{ij}\}_{1\le i\le n, 1\le j\le p}$ consists of np independent standard one-dimensional Wiener processes. However, as in equation (2.2) we must include a deterministic correction. This correction corresponds to mean curvature again, but in a more subtle way than (2.2). The equivalence class of Y such that $X = YY^T$ is a group orbit of $\mathcal{O}_p$ embedded within $\mathbb{R}^{n\times p}$. The logarithm of the volume of this group orbit constitutes a natural Boltzmann entropy. It may be computed explicitly, and we find

$$S(Y) = \frac{1}{2}\sum_{i=1}^{p}\sum_{j=i+1}^{p}\log(\sigma_i^2 + \sigma_j^2) \tag{2.5}$$

where $\{\sigma_i\}_{i=1}^p$ are singular values of Y. It is known that $\nabla S(Y)$ is the mean curvature of the group orbit in $\mathbb{R}^{n\times p}$ [32, p.3505].

We then have the following Itô differential equation for Y_t such that $X_t = Y_t Y_t^T$ has the same law as the solution to (1.6).

$$\mathrm{d}Y_{ij} = -\frac{\partial \mathcal{E}(YY^T)}{\partial Y_{ij}}\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^{ij} - \frac{1}{\beta}\frac{\partial S(Y)}{\partial Y_{ij}}\mathrm{d}t, \qquad 1 \le i \le n, 1 \le j \le p. \tag{2.6}$$

The correction term can be explicitly computed using the following

LEMMA 1. *If $Y \in R_*^{n\times p}$ has SVD as $Y = Q\Sigma P^T$ with singular values σ_i, then the gradient of the correction term S is given by $\nabla S(Y) = Q\tilde{\Sigma}P^T$ where $\tilde{\Sigma}$ is a diagonal matrix with diagonal entries $\sum_{j\neq 1}\frac{\sigma_1}{\sigma_1^2+\sigma_j^2}, \sum_{j\neq 2}\frac{\sigma_2}{\sigma_2^2+\sigma_j^2}, \cdots, \sum_{j\neq p}\frac{\sigma_p}{\sigma_p^2+\sigma_j^2}$.*

3. Two Riemannian Langevin Monte Carlo schemes. To get a simple Riemannian Langevin Monte Carlo sampling scheme, we only consider convenient discretization and approximation methods, which can be easily and efficiently implemented. For the Brownian motion term, we consider the most straightforward and simplest discretization of the SDEs (2.2) and (2.6), i.e., the Euler-Maruyama type discretization.

One extra complication from the manifold constraint is how to approximate the exponential map. For optimization algorithms on Riemannian manifolds [1], retraction, which is at least a first order approximation to the exponential map, is often used. For instance, for approximating an ODE $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X} = -\mathrm{grad}\,\mathcal{E}(\boldsymbol{X})$ on a manifold $\mathcal{M}$, with any retraction operator $\mathcal{R}_{\mathcal{M}}$ mapping to $\mathcal{M}$, a simple forward Euler type approximation, or equivalently the Riemannian gradient descent method, is given by

$$\boldsymbol{X}_{k+1} = \mathcal{R}_{\mathcal{M}}[\boldsymbol{X}_{k+1} - \Delta t\,\mathrm{grad}\,\mathcal{E}(\boldsymbol{X}_k)].$$

In particular, when combining the Euler-Maruyama type discretization for SDE and the simple Riemannian gradient descent by retraction, we get the two simple Riemmanian Langevin Monte Carlo schemes as follows.

3.1. Scheme E for the embedded geometry.

3.1.1. The Riemannian gradient. For a given energy function $\mathcal{E}(X)$, its Riemannian gradient $\mathrm{grad}\,\mathcal{E}(X)$ of at $X \in \mathcal{S}_+^{n,p}$, is the Euclidean projection of the Euclidean gradient $\nabla\mathcal{E}(X) \in \mathbb{R}^{n\times n}$ defined as $[\nabla\mathcal{E}(X)]_{ij} = \frac{\partial}{\partial X_i}\mathcal{E}(X)$, onto the tangent space $T_X\mathcal{S}_+^{n,p}$, see [1, 34, 38]. It is straightforward to verify that $\nabla\mathcal{E}(X)$ is a symmetric matrix for any differentiable $\mathcal{E}$ and any $X \in \mathcal{S}_+^{n,p}$. For any given $X \in \mathcal{S}_+^{n,p}$, let $X = U\Lambda U^T$ be its compact SVD. Let $P_U = UU^T$ and $P_{U_\perp} = U_\perp U_\perp^T = I - UU^T$. By derivations in [38], $\mathrm{grad}\,\mathcal{E}(X)$ can be computed and represented as

$$\begin{aligned}\mathrm{grad}\,\mathcal{E}(X) &= \begin{bmatrix}U & U_\perp\end{bmatrix}\begin{bmatrix}U^T\nabla\mathcal{E}(X)U & U^T\nabla\mathcal{E}(X)U_\perp \\ U_\perp^T\nabla\mathcal{E}(X)U & 0\end{bmatrix}\begin{bmatrix}U^T \\ U_\perp^T\end{bmatrix} \\ &= P_U\nabla\mathcal{E}(X)P_U + P_{U_\perp}\nabla\mathcal{E}(X)P_U + P_U\nabla\mathcal{E}(X)P_{U_\perp}.\end{aligned}$$

The compact implementation of computing $\mathrm{grad}\,\mathcal{E}(X)$ is given in Algorithm 3.1.

Algorithm 3.1 Compact computation of the Riemannian gradient $\mathrm{grad}\,\mathcal{E}(X)$

Require: The compact SVD of $X \in S_+^{n,p}$: $X = U\Lambda U^T$
Ensure: $\mathrm{grad}\,\mathcal{E}(X) = UHU^T + U_pU^T + UU_p^T \in T_XS_+^{n,p}$
$T \leftarrow \nabla\mathcal{E}(X)U$
$H \leftarrow U^TT$
$U_p \leftarrow T - UH$

3.1.2. The retraction by projection. Let $\mathcal{S}^{n\times n}$ denote symmetric matrices, then the Euclidean projection $\mathrm{P}_{S_+^{n,p}} : \mathcal{S}^{n\times n} \longrightarrow \mathcal{S}_+^{n,p}$ is a convenient retraction operator, see [1, 34, 38]. A straightforward implementation is given in Aglorithm 3.2.

Algorithm 3.2 Computation of the retraction $P_{S_+^{n,p}}(X + Z)$

Require: the compact SVD of X: $X = U\Lambda U^T \in S_+^{n,p}$, $Z \in \mathcal{S}^{n\times n}$.
Ensure: $P_{S_+^{n,p}}(X + Z) = Q_+\Lambda_+Q_+^T \in S_+^{n,p}$.
$(Q_+, \Lambda_+) = \mathrm{svd}(X + Z)$
$U_+ \leftarrow Q_+(:, 1:p) \quad \Lambda_+ \leftarrow \Lambda_+(1:p, 1:p)$

3.1.3. A Riemannian Langevin Monte Carlo scheme. For approximating the SDE (2.2) on $(S_+^{n,p}, g_E)$, with the retraction operator and Euler-Maruyama method for SDE, we have the scheme (1.7), which can be more explicitly written as

(3.1)
$$X_{k+1} = \mathrm{P}_{S_+^{n,p}}\left(\begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \Lambda - \Delta t U^T \nabla \mathcal{E}(X_k) U + \sqrt{\frac{2\Delta t}{\beta}} B_{11} & -\Delta t U^T \nabla \mathcal{E}(X_k) U_\perp + \sqrt{\frac{2\Delta t}{\beta}} B_{12} \\ -\Delta t U_\perp^T \nabla \mathcal{E}(X_k) U + \sqrt{\frac{2\Delta t}{\beta}} B_{12}^T & \frac{\Delta t}{\beta} \sum_{i=1}^{p} \frac{1}{\lambda_i} I_{n-p} \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix} \right),$$

where $X_k = U\Lambda U^T$ is the compact SVD of $X_k \in S_+^{n,p}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. The third term in the right hand side is the white noise term in the tangent space $T_{X_k}\mathcal{S}_+^{n,p}$. Entries of $B_{12} \in \mathbb{R}^{p\times(n-p)}$ are i.i.d drawn from $\sqrt{\frac{1}{2}}\mathcal{N}(0,1)$, and $B_{11} \in \mathbb{R}^{p\times p}$ are defined as follows.

(3.2)
$$B_{11} = \begin{bmatrix} \mathcal{N}(0,1) & & & \\ & \ddots & b_{ij} & \\ & b_{ji} & \ddots & \\ & & & \mathcal{N}(0,1) \end{bmatrix}$$

with $b_{ij} = b_{ji} \sim \sqrt{\frac{1}{2}}\mathcal{N}(0,1)$. The implementation details of the scheme (1.7) are given as follows in the Algorithm 3.3.

Algorithm 3.3 The Riemannian Langevin Monte Carlo scheme (1.7) for $(\mathcal{S}_+^{n,p}, g_E)$

Require: initial iterate $X_1 \in \mathcal{S}_+^{n,p}$; full SVD of X_1: $X_1 = Q_1\Lambda_1 Q_1^T$

1: **for** $k = 1, 2, \ldots, N$ **do**

2: Compute Riemannian gradient
$\xi_k := \operatorname{grad} \mathcal{E}(X_k)$ ▷ See Algorithm 3.1

3: Compute noise term
$$B = \sqrt{\frac{2\Delta t}{\beta}} \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & 0 \end{bmatrix} + \frac{\Delta t}{\beta} \sum_{i=1}^{p} \frac{1}{\lambda_i} \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix}$$

4: Obtain the new iterate by retraction
$X_{k+1} = \mathrm{P}_{\mathcal{S}_+^{n,p}}(X_k - \Delta t \xi_k + Q_k B Q_k^T)$ ▷ See Algorithm 3.2

5: **end for**

REMARK 2. *The mean curvature correction term is necessary for avoiding rank deficient samples in the following sense. A sampling scheme on $\mathcal{S}_+^{n,p}$ might generate a sample X with a rank numerically close to $p-1$, and the mean curvature correction term in the scheme* (1.7) *would be huge if $\lambda_p \to 0$, thus it will force iterate X_k to stay away from the boundary of $\mathcal{S}_+^{n,p}$.*

REMARK 3. *Notice that the complexity of computing SVD of $X + Z$ in Algorithm 3.2 would be $\mathcal{O}(n^3)$ in a naive implementation. For a Riemannian gradient method, if $Z \in T_{X_k}\mathcal{S}_+^{n,p}$, a compact implementation of computing $P_{\mathcal{S}_+^{n,p}}(X + Z)$ in [38] is only $\mathcal{O}(np^2) + \mathcal{O}(p^3)$, which is no longer possible for the Langevin Monte Carlo scheme* (1.7) *due to the mean curvature correction term in the normal space. On the other hand, if Lanczos type algorithm is used for computing to top p eigen-componenes of $X+Z$, it seems possible to explore the special structure in* (3.1) *to find a more efficient implementation, but we do not consider a more compact implementation in this paper.*

3.2. Scheme BW for the Bures-Wasserstein metric.

3.2.1. The Riemannian gradient and a simple retraction operator.

Given a smooth energy function $\mathcal{E}(X)$ defined on $\mathcal{S}_+^{n,p}$, the corresponding function h on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ satisfies

$$\begin{aligned} h : \mathbb{R}_*^{n\times p}/\mathcal{O}_p &\to \mathbb{R} \\ \pi(Y) &\mapsto \mathcal{E}(\tilde{\beta}(\pi(Y))) = \mathcal{E}(\beta(Y)) = \mathcal{E}(YY^T). \end{aligned} \tag{3.3}$$

Observe that the function $F(Y) := \mathcal{E}(YY^T)$ satisfies $F(Y) = h\circ\pi(Y) = \mathcal{E}\circ\beta(Y)$. The Riemannian gradient of h at $\pi(Y)$ is a tangent vector in $T_{\pi(Y)}\mathbb{R}_*^{n\times p}/\mathcal{O}_p$. The next theorem is given in [1, Section 3.6.2], showing that the horizontal lift of $\operatorname{grad} h(\pi(Y))$ can be obtained from the Riemannian gradient of F defined on $\mathbb{R}_*^{n\times p}$.

THEOREM 4. *The horizontal lift of the gradient of h at $\pi(Y)$ is the Riemannian gradient of F at Y. That is,*

$$\overline{\operatorname{grad} h(\pi(Y))}_Y = \operatorname{grad} F(Y).$$

For the Bures-Wasserstein metric, the following result is proven in [38]:

PROPOSITION 5. *Let $\mathcal{E}$ be a smooth real-valued function defined on $\mathcal{S}_+^{n,p}$ and let $F : \mathbb{R}_*^{n\times p} \to \mathbb{R} : Y \mapsto \mathcal{E}(YY^T)$. Assume $YY^T = X$. Then the Riemannian gradient of F is given by*

$$\operatorname{grad} F(Y) = 2\nabla\mathcal{E}(YY^T)Y$$

where $\nabla\mathcal{E}(\cdot)$ is the gradient of $\mathcal{E}$ w.r.t. X.

In [26, Prop. A.8], the relationship between the horizontal lifts of the quotient tangent vector $\xi_{\pi(Y)}$ lifted at different representatives in $[Y]$ is given:

LEMMA 6. *Let η be a vector field on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, and let $\bar{\eta}$ be the horizontal lift of η. Then for each $Y \in \mathbb{R}_*^{n\times p}$, we have*

$$\bar{\eta}_{YO} = \bar{\eta}_Y O$$

for all $O \in \mathcal{O}_p$.

The retraction on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ can be defined using the retraction on the total space $\mathbb{R}_*^{n\times p}$. For any $A \in T_Y\mathbb{R}_*^{n\times p}$ and a step size $\tau > 0$,

$$\overline{R}_Y(\tau A) := Y + \tau A,$$

is a retraction on $\mathbb{R}_*^{n\times p}$ if $Y + \tau A$ remains full rank, which is ensured for small enough τ. Then Lemma 6 indicates that $\overline{R}$ satisfies the conditions of [1, Prop. 4.1.3], which implies that

$$R_{\pi(Y)}(\tau\eta_{\pi(Y)}) := \pi(\overline{R}_Y(\tau\overline{\eta}_Y)) = \pi(Y + \tau\overline{\eta}_Y) \tag{3.4}$$

defines a retraction on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ for a small enough step size $\tau > 0$.

Finally, we give an example of what these results imply by considering the Riemannian gradient descent method for minimizing $\mathcal{E}(X)$ over $(\mathcal{S}_+^{n,p}, g_{BW})$. With the simple

retraction (3.4), the Riemannian gradient descent method for minimizing the function $h[\pi(Y)]$ on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ is given by

$$Y_{k+1} = Y_k - \Delta t 2\nabla\mathcal{E}(Y_kY_k^T)Y_k,$$

which is the simple Burer-Monteiro gradient descent method for minimizing $\mathcal{E}(X)$ over $\mathcal{S}_+^{n,p}$. See Section 5.1 in [38] for details.

3.2.2. A simple Riemannian Langevin Monte Carlo scheme. With the Euler-Maruyama discretization for SDE (2.6), and the simple retraction and Riemannian gradient as given previously, a simple Riemannian Langevin Monte Carlo scheme for approximating the Riemannian SDE (2.6) on the Riemannian manifold $(S_+^{n,p}, g_{BW})$ can be given as

$$Y_{k+1} = Y_k - \Delta t 2\nabla\mathcal{E}(Y_kY_k^T)Y_k + \sqrt{\frac{2\Delta t}{\beta}}B_k + \frac{\Delta t}{\beta}U\left[\sum_{j:j\neq i}\frac{\sigma_i}{\sigma_i^2+\sigma_j^2}\right]_{ii}V^T, \tag{3.5}$$

where B_k is n-by-p matrix with i.i.d. $\mathcal{N}(0,1)$ entries and $Y_k = U\Sigma V^T$ is the compact SVD of Y with singular values $\sigma_i > 0$ for $i = 1, 2, \cdots, p$.

Notice that all operations are performed in the space of size $n\times p$. For finding compact SVD of Y, one can first compute QR decomposition of Y, which costs $\mathcal{O}(np^2)+\mathcal{O}(p^3)$. Then compute SVD of size $p\times p$, which is $\mathcal{O}(p^3)$. So the complexity of this scheme is $\mathcal{O}(np^2) + \mathcal{O}(p^3)$ for each iteration. For large n and small p, Scheme BW should be cheaper than Scheme E in each iteration, but they generate different samples for different Gibbs distributions which depend on the metric, i.e., Scheme BW cannot replace Scheme E for generating Gibbs distribution defined by embedded geometry.

4. Examples with analytical formulae. In this section, we provide a few examples with analytical formulae so that they can be used in numerical experiments for testing the two schemes (3.1) and (3.5) on the Gibbs distribution.

For the rest of this section, $X = Q\Lambda Q^T \in \mathcal{S}_+^{n,p}$ denotes the full SVD with descending eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$.

4.1. Scalar random variables. Let X be a random variable satisfying the Gibbs distribution on $\mathcal{S}_+^{n,p}$ with dimension $N = np - \frac{p(p-1)}{2}$ under either metric, then X is a matrix-valued random variable. For convenience, we consider a scalar random variable $D = D(X)$ as a function of $X \in \mathcal{S}_+^{n,p}$, e.g., $D = \|X\|_F$ where $\|\cdot\|_F$ is the matrix Frobenius norm.

We consider the distribution function for the scalar random variable D:

$$\Pr[D < d] = \frac{1}{Z_\beta}\int_{U_d} \mathrm{e}^{-\beta\mathcal{E}}\mathrm{d}V, \quad Z_\beta = \int_{\mathcal{M}} \mathrm{e}^{-\beta\mathcal{E}}\mathrm{d}V, \tag{4.1}$$

where $U_d := \{X \in \mathcal{S}_+^{n,p} | D(X) < d\}$ is the domain of integral. For simplicity we only consider symmetric functions such that the random variable D, the energy function $\mathcal{E}$, and the volume form are all invariant under the group action by the orthogonal group $\mathcal{O}_n$. We consider an energy function $\mathcal{E}$ satisfying $\mathcal{E}(X) = \mathcal{E}(OXO^T)$, $\forall\, O \in \mathcal{O}_n$, so that Gibbs distribution function only depends on the spectrum of X when considering (4.1) with $D = ||X||_F = \sqrt{\lambda_1^2 + \cdots + \lambda_p^2}$. Since $\mathcal{O}_n$ is an isometry group for both

metrics g_E and g_{BW}, the volume form $\mathrm{d}V$ in the two cases is also invariant under $\mathcal{O}_n$ action.

Notice that Q and Λ can be used as coordinates of the manifold $\mathcal{S}_+^{n,p}$. The volume form expressed by coordinates Q and Λ is given by

$$\mathrm{d}V = \sqrt{\det g}(\prod_{i=1}^{p} \mathrm{d}\lambda_i)\mathrm{d}\mu_{\mathcal{O}_n}, \tag{4.2}$$

where $\mu_{\mathcal{O}_n}$ is the Haar measure on $\mathcal{O}_n$, and g is the matrix of metric g_E or g_{BW} expressed under coordinate Q and λ. For g_E its determinant $\det g$ is

$$\det g = \big(\prod_{1\le i<j\le p} |\lambda_i - \lambda_j|^2\big)\big(\prod_{1\le i\le p} \lambda_i^{2(n-p)}\big), \tag{4.3}$$

and for g_{BW} it is

$$\det g = \big(\prod_{1\le i<j\le p} \frac{|\lambda_i - \lambda_j|^2}{\lambda_i + \lambda_j}\big)\big(\prod_{1\le i\le p} \lambda_i^{(n-p)}\big). \tag{4.4}$$

So for g_E the distribution $\Pr[D < d]$ is expressed as

$$\begin{aligned}\Pr[D < d] =& \frac{1}{Z_\beta} \int_{||X||_F<d} \mathrm{e}^{-\beta\mathcal{E}}\mathrm{d}V \\ \propto& \int_{\substack{\sum_{i=1}^{p}\lambda_i^2<d^2 \\ \lambda_i>0,i=1,\dots,p}} \mathrm{e}^{-\beta\mathcal{E}(\lambda_1,\dots,\lambda_p)}\big(\prod_{1\le i<j\le p} |\lambda_i - \lambda_j|\big)\big(\prod_{1\le i\le p} \lambda_i^{n-p}\big)\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_p,\end{aligned} \tag{4.5}$$

where we have used the fact that the integrand does not depend on the coordinate $Q \in \mathcal{O}_n$, so the integral of $\mu_{\mathcal{O}_n}$ only provides a constant coefficient. As we could always renormalize $\Pr[D < d]$ by considering the quotient $\frac{\Pr[D<d]}{\Pr[D<\infty]}$, we only need the dependence of the integral on parameter d.

Similarly, for the Bures-Wasserstein metric g_{BW} we have

$$\Pr[D < d] \propto \int_{\substack{\sum_{i=1}^{p}\lambda_i^2<d^2 \\ \lambda_i>0,i=1,\dots,p}} \mathrm{e}^{-\beta\mathcal{E}(\lambda_1,\dots,\lambda_p)}\big(\prod_{1\le i<j\le p} \frac{|\lambda_i - \lambda_j|}{\sqrt{\lambda_i + \lambda_j}}\big)\big(\prod_{1\le i\le p} \lambda_i^{\frac{n-p}{2}}\big)\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_p \tag{4.6}$$

Next we give a few energy functions.

4.2. Example I: $\mathcal{E}(X) = \frac{1}{2}||X||_F^2$. This is the simplest example. Using the general expression above, for embedded geometry g_E we have

$$\begin{aligned}
\Pr[D < d] &\propto \int_{\substack{\sum_{i=1}^p \lambda_i^2 < d^2 \\ \lambda_i > 0, i=1,\dots,p}} \mathrm{e}^{-\frac{\beta}{2}\sum_{i=1}^p \lambda_i^2} \big(\prod_{1\le i<j\le p} |\lambda_i - \lambda_j| \big) \big(\prod_{1\le i\le p} \lambda_i^{n-p} \big) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_p \\
&= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \big(\int_{S_+^{p-1}} \prod_{1\le i<j\le p} |\omega_i - \omega_j| \prod_{i=1}^p |\omega_i|^{n-p} \prod_{i=1}^p \mathrm{d}\omega \big) \mathrm{d}\rho \\
&= \big(\int_{S_+^{p-1}} \prod_{1\le i<j\le p} |\omega_i - \omega_j| \prod_{i=1}^p |\omega_i|^{n-p} \prod_{i=1}^p \mathrm{d}\omega \big) \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \mathrm{d}\rho \\
&\propto \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \mathrm{d}\rho, \qquad (4.7)
\end{aligned}$$

where we have used the spherical coordinate for $(\lambda_1, ..., \lambda_p) = \rho\omega$, with $\rho = \sqrt{\sum_{i=1}^p \lambda_i^2}$ being the radius and $\omega \in S_+^{p-1} = S^{p-1} \cap \mathbb{R}_+^p$ being the coordinate on the positive orthant of unit sphere.

For g_{BW}, similarly we have

$$\Pr[D < d] \propto \int_0^d \mathrm{e}^{-\beta\rho^2} \rho^{\frac{N}{2}-1} \mathrm{d}\rho. \qquad (4.8)$$

Now we can see that $\beta D^2 = \beta||X||_F^2$ is subject to $\chi^2(N)$ distribution for the embedded metric g_E, and $\chi^2(\frac{N}{2})$ distribution for the Bures-Wasserstein metric.

4.3. Example II: $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$. We consider the von Neumann entropy

$$\mathcal{E}(X) = \mathrm{Tr}(X \log X) = \sum_{i=1}^p \lambda_i \log \lambda_i$$

and construct a more interesting example. The minimizers of $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$ on $\mathcal{S}_+^{n,p}$ are matrices $X \in \mathcal{S}_+^{n,p}$ with spectrum $\lambda_1 = \cdots = \lambda_p = \mathrm{e}^{-1}$.

The random variable we consider is still $D = ||X||_F$. Since $\mathcal{E}(X) = \mathrm{Tr}(X \log X) = \sum_{i=1}^p \lambda_i \log \lambda_i$ only depends on spectrum, the argument in the previous section about

integral on $\mathcal{O}_n$ still applies. Similar to (4.7), for g_E we have

$$\begin{aligned}\Pr(D<d) &= \int_{\substack{\sum_{i=1}^p \lambda_i^2<d^2 \\ \lambda_i>0, i=1,\dots,p}} \mathrm{e}^{-\beta\sum_{i=1}^p \lambda_i\log\lambda_i} \prod_{1\le i<j\le p}|\lambda_i-\lambda_j| \prod_{i=1}^p |\lambda_i|^{n-p} \prod_{i=1}^p \mathrm{d}\lambda_i \\ &= \int_{\substack{\sum_{i=1}^p \lambda_i^2<d^2 \\ \lambda_i>0, i=1,\dots,p}} \prod_{1\le i<j\le p}|\lambda_i-\lambda_j| \prod_{i=1}^p |\lambda_i|^{n-p-\beta\lambda_i} \prod_{i=1}^p \mathrm{d}\lambda_i,\end{aligned}$$

and for g_{BW} we have

$$\Pr(D<d) = \int_{\substack{\sum_{i=1}^p \lambda_i^2<d^2 \\ \lambda_i>0, i=1,\dots,p}} \prod_{1\le i<j\le p}\frac{|\lambda_i-\lambda_j|}{\sqrt{\lambda_i+\lambda_j}} \prod_{i=1}^p |\lambda_i|^{\frac{n-p}{2}-\beta\lambda_i} \prod_{i=1}^p \mathrm{d}\lambda_i.$$

Although we do not have a closed expression for both cases, such integrals can be easily approximated by an accurate quadrature.

4.4. Example III: $\mathcal{E}(X) = \frac{1}{2}||X-A||_F^2$. We consider a quadratic function $\mathcal{E}(X) = \frac{1}{2}||X-A||_F^2$ where $A \in \mathcal{S}_+^{n,p}$ with $D = ||X-A||_F$. In this example, $\mathcal{O}_n$ symmetry does not hold, and we can only make an estimate of the distribution function.

The random variable D we are considering now is $D = ||X-A||_F$, its distribution function is evaluated as

$$\Pr(D<d) \propto \int_{U_d} \mathrm{e}^{-\frac{\beta}{2}D^2}\mathrm{d}V \tag{4.9}$$

where $U_d = \{X \in \mathcal{S}_+^{n,p} | D(X) < d\}$. Using delta function, formally we can simplify the integral to

$$\begin{aligned}\Pr(D<d) &\propto \int_{\mathcal{M}} \mathbf{1}_{\{D<d\}}\mathrm{e}^{-\frac{\beta}{2}D^2}\mathrm{d}V \\ &= \int_{\mathcal{M}}\Big(\int_0^\infty \mathbf{1}_{\{\rho<d\}}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\delta(D-\rho)\mathrm{d}\rho\Big)\mathrm{d}V \\ &= \int_0^\infty \mathbf{1}_{\{\rho<d\}}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\Big(\int_{\mathcal{M}}\delta(D-\rho)\mathrm{d}V\Big)\mathrm{d}\rho \\ &= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\Big(\int_{\mathcal{M}}\frac{\mathrm{d}}{\mathrm{d}\rho}\mathbf{1}_{\{D-\rho\}}\mathrm{d}V\Big)\mathrm{d}\rho \\ &= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\frac{\mathrm{d}}{\mathrm{d}\rho}\Big(\int_{\mathcal{M}}\mathbf{1}_{\{D-\rho\}}\mathrm{d}V\Big)\mathrm{d}\rho \\ &= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\frac{\mathrm{d}}{\mathrm{d}\rho}V_D(\rho)\mathrm{d}\rho\end{aligned} \tag{4.10}$$

(4.11)

where $V_D(\rho) = \int_{\mathcal{M}}\mathbf{1}_{\{D<\rho\}}\mathrm{d}V = \int_{D<\rho}\mathrm{d}V$.

In general it is difficult to calculate $\int_{D<\rho} \mathrm{d}V$, but we consider the following approximation. Consider the volume of the ball $B_A^{n,p}(r) = B_A(r) \cap \mathcal{S}_+^{n,p}$, where

$$B_A(r) = \big\{X \in \mathcal{S}^{n\times n} : ||X - A||_F < r\big\}.$$

It is difficult to compute $\mathrm{Vol}(B_A^{n,p}(r))$, but we propose the following estimate, for fixed $A \in \mathcal{S}_+^{n,p}$:

$$\mathrm{Vol}(B_{cA}^{n,p}(r)) \approx \alpha r^N, \quad c \gg 1, \tag{4.12}$$

where α is a constant that does not depend on r, N is the dimension of $\mathcal{S}_+^{n,p}$. For g_E, α is exactly the volume of unit ball in $\mathbb{R}^N$, while for g_{BW}, α depends on dimension N and $A \in \mathcal{S}_+^{n,p}$.

For the embedded geometry, the approximation (4.12) can be justified by the following arguments:

1. The second fundamental form $\boldsymbol{I\!I}_{cA}$ of the manifold is vanishing for fixed A and $c \to \infty$. See [36].
2. The Riemannian curvature tensor of ambient space $\mathcal{S}^{n\times n}$ is 0. Applying the Gauss equation [11, Prop 3.1] we can express the Riemannian curvature tensor R of $(\mathcal{S}_+^{n,p}, g_E)$ in terms of its second fundamental form $\boldsymbol{I\!I}$:

$$\begin{aligned} &\langle R(\boldsymbol{x},\boldsymbol{y})\boldsymbol{z},\boldsymbol{w}\rangle = -\langle \boldsymbol{I\!I}(\boldsymbol{x},\boldsymbol{z}), \boldsymbol{I\!I}(\boldsymbol{y},\boldsymbol{w})\rangle + \langle \boldsymbol{I\!I}(\boldsymbol{x},\boldsymbol{w}), \boldsymbol{I\!I}(\boldsymbol{y},\boldsymbol{z})\rangle, \\ &\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{w} \in T\mathcal{S}_+^{n,p}, \qquad \langle U, V\rangle = \mathrm{Tr}(UV^T) \text{ is the metric in } \mathcal{S}^{n\times n}, . \end{aligned} \tag{4.13}$$

   Thus, with vanishing $\boldsymbol{I\!I}$ we have vanishing Riemannian curvature tensor, and zero sectional curvature.
3. Vanishing extrinsic curvature and intrinsic curvature means that the neighborhood is approximately an Euclidean space, so the ball $B_{cA}^{n,p}(r)$ is approximately just a ball in $\mathbb{R}^N$ and has volume αr^N, with α being the volume of a unit ball.

For the g_{BW} metric, following similar arugments, we can get the same approximation (4.12). We emphasize that the approximation (4.12) is accurate only if c is large enough. Putting all this together, when A has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \gg 1$, we have the following

$$\Pr(D < d) \propto \int_{D<d} \mathrm{e}^{-\frac{\beta}{2}D^2}\mathrm{d}V = \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\frac{\mathrm{d}}{\mathrm{d}\rho}\Big(\int_{D<\rho}\mathrm{d}V\Big)\mathrm{d}\rho \mathrel{\tilde{\propto}} \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho, \tag{4.14}$$

where $\mathrel{\tilde{\propto}}$ stands for *being approximately proportional to*.

4.5. MCMC numerical integration. It is well known that MCMC can be used for integrating a function numerically, and that one of the main advantages is that the convergence rate is independent of the dimension. Both schemes in this paper are MCMC type sampling schemes on the manifold. Suppose we have generated samples X_i satisfying the Gibbs distribution on the manifold, e.g.,

$$X_i \sim \frac{1}{Z_\beta} e^{-\beta\mathcal{E}(X)}\mathrm{d}V_g,$$

where $Z_\beta = \int\limits_{S_+^{n,p}} \mathrm{e}^{-\beta\mathcal{E}(X)}\mathrm{d}V$ is an unknown normalization factor and $\mathrm{d}V$ is the volume form depending on the metric. Then for approximating the integral of a nice function $f(X)$ on the same manifold $\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V$, we can use

$$\frac{1}{m}\sum_{i=1}^{m} f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)} \approx \frac{\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V}{\int\limits_{S_+^{n,p}} \mathrm{e}^{-\beta\mathcal{E}(X)}\mathrm{d}V} = \frac{1}{Z_\beta}\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V, \tag{4.15}$$

because each $f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}$ is a random variable with expectation

$$\mathbb{E}\left[f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}\right] = \frac{1}{Z_\beta}\int\limits_{S_+^{n,p}} f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}\mathrm{e}^{-\beta\mathcal{E}(X_i)}\mathrm{d}V,$$

and the left hand side is a random variable with expectation

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} f(X_i)\mathrm{e}^{\beta E(X_i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f(X_i)\mathrm{e}^{\beta E(X_i)}\right] = \frac{1}{Z_\beta}\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V,$$

where the expectation $\mathbb{E}[\cdot]$ is taken w.r.t. Gibbs distribution under corresponding metric.

So using the generated samples X_i, we can approximate the integral $\int\limits_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V$ up to a constant Z_β that does not depend on $f(X)$. Notice that the additional advantage of Monte Carlo type quadrature on a manifold is that we do not need to know what $\mathrm{d}V$ is. On the other hand, Z_β cannot be approximated by the same approach. Though we do not consider any specific application for numerical integration, equation (4.15) can be used as one way to validate the Riemannian Langevin Monte Carlo schemes.

For the following special functions, it is possible to calculate exact integrals. For the energy function $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$, and a special integrand $f(X) = \|X\|_F^k \mathrm{e}^{-\frac{\alpha}{m}\|X\|_F^m}$ with $k > -N, m > 2, \alpha > 0$, using the results in 4.2, the distribution of $D = ||X||_F$ is

$$\text{for metric } g_E: \quad \Pr[D < d] \propto \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho, \tag{4.16}$$

$$\text{for metric } g_{BW}: \quad \Pr[D < d] \propto \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{\frac{N}{2}-1}\mathrm{d}\rho, \tag{4.17}$$

so the integral on the manifold could be expressed by expectation of a random variable, which leads to

$$\text{for } g_E: \frac{1}{Z_\beta}\int_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V = \mathbb{E}[f(X)\mathrm{e}^{\frac{\beta}{2}||X||_F^2}] = \mathbb{E}[D^k \mathrm{e}^{-\frac{\alpha}{m}D^m}\mathrm{e}^{\frac{\beta}{2}D^2}]$$

$$= \frac{\int_0^\infty \rho^k \mathrm{e}^{-\frac{\alpha}{m}\rho^m+\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho}{\int_0^\infty \rho^{N-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho} = \frac{\frac{1}{m}(\alpha/m)^{-\frac{k+N}{m}}\Gamma((k+N)/m)}{\frac{1}{2}(\beta/2)^{-N/2}\Gamma(N/2)} \tag{4.18}$$

$$\text{for } g_{BW}: \frac{1}{Z_\beta}\int_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V = \mathbb{E}[f(X)\mathrm{e}^{\frac{\beta}{2}||X||_F^2}] = \mathbb{E}[D^k \mathrm{e}^{-\frac{\alpha}{m}D^m}\mathrm{e}^{\frac{\beta}{2}D^2}]$$

$$= \frac{\int_0^\infty \rho^k \mathrm{e}^{-\frac{\alpha}{m}\rho^m+\frac{\beta}{2}\rho^2}\rho^{\frac{N}{2}-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho}{\int_0^\infty \rho^{\frac{N}{2}-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho} = \frac{\frac{1}{m}(\alpha/m)^{-\frac{k+N/2}{m}}\Gamma((k+N/2)/m)}{\frac{1}{2}(\beta/2)^{-N/4}\Gamma(N/4)}. \tag{4.19}$$

5. Numerical tests. In this section we test the samples generated by the two Riemannian Langevin Monte Carlo schemes (3.1) and (3.5) on the examples constructed in the previous section. The samples are generated by the following procedure: we run the iterative schemes (3.1) or (3.5) for sufficiently many $\tilde{m}$ iterations then take the last m iterates as the samples for the Gibbs distribution. Both $\tilde{m}$ and m should be chosen such that the $(\tilde{m}-m)$-th iterate has already reached equilibrium e.g., $\tilde{m}$ is $6,000,000$ and m is $5,000,000$ for specially chosen energy functions and parameters β.

Now suppose we have generated samples $X_i \in \mathcal{S}_+^{n,p}$ $(i = 1, \cdots, m)$ for either metric. In order to test or show the numerical convergence to the Gibbs distribution, we will consider two kinds of numerical tests.

The first kind of tests is to test on the scalar random variable $D(X) = \|X\|_F$ or $D(X) = \|X - A\|_F$ as described in Section 4. Then we compare the cumulative distribution function (CDF) of the random variable D with its empirical CDF calculated from the MCMC samples.

Denote the true CDF of D by $F_D(t) := \Pr(D \le t)$. The empirical CDF of samples is

$$\hat{F}_D(t) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{D(X_i)\le t},$$

where $\mathbb{1}_{D(X_i)\le t}$ takes value 1 if $D(X_i) \le t$, and value 0 if otherwise. The Kolmogorov–Smirnov test statistic (K-S statistic) is defined by

$$KS_D := \sup_t \left| F_D(t) - \hat{F}_D(t) \right|. \tag{5.1}$$

In our numerical tests, we compute the KS statistic by taking the maximum difference of F_D and $\hat{F}_D$ at 100 equally spaced points in the interval $[0, t_{max}]$ where $F_D(t_{max}) \approx 1$.

The second kind of tests is on the integral examples in Section 4.5, let X be a random variable satisfying Gibbs distribution on the manifold $\mathcal{S}_+^{n,p}$ under either metric. Define

$$\mu := \mathbb{E}\left(f(X)\mathrm{e}^{\beta\mathcal{E}(X)}\right) = \frac{1}{Z_\beta}\int_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V.$$

Given m samples $X_i \in \mathcal{S}_+^{n,p}$, we define

$$\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^{m} f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}. \tag{5.2}$$

Notice that samples generated by MCMC are not independent. If we assume

$$\sigma^2 := \mathrm{var}\left(f(X_1)\mathrm{e}^{\beta\mathcal{E}(X_1)}\right) + 2\sum_{k=1}^{\infty}\mathrm{cov}\left(f(X_1)\mathrm{e}^{\beta\mathcal{E}(X_1)}, f(X_{1+k})\mathrm{e}^{\beta\mathcal{E}(X_{1+k})}\right) < \infty,$$

then by the Markov Chain Central Limit Theorem[21, 14], as $m \to \infty$, we have

$$\sqrt{m}(\hat{\mu}_m - \mu) \to \mathcal{N}(0, \sigma^2) \tag{5.3}$$

where the convergence is in the sense of distribution. Thus if $m \gg 1$, $\frac{\hat{\mu}_m - \mu}{\mu}$ roughly follows the distribution $\mathcal{N}(0, \mathcal{O}(\frac{1}{m}))$ and the relative error term $\left|\frac{\hat{\mu}_m - \mu}{\mu}\right|$ roughly follows the folded normal distribution with mean $\mathcal{O}(\frac{1}{\sqrt{m}})$ and variance $\mathcal{O}(\frac{1}{m})$. Hence we can use $\hat{\mu}_m$ defined in (5.2) to estimate $\mu = \frac{1}{Z_\beta}\int\limits_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V$, and the relative error is $\mathcal{O}(\frac{1}{\sqrt{m}})$.

5.1. Numerical validation of the scalar variable $D(X)$. The manifold $\mathcal{S}_+^{n,p}$ has dimension $N = np - p(p-1)/2$. For both metrics, we consider three examples in Section 4 with special energy functions $\mathcal{E}$ in the Gibbs distribution $\mathrm{e}^{-\beta\mathcal{E}}$ and the CDF for the scalar variable $D(X)$:

1. Example I: $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$ with the CDF for $D(X) = \|X\|_F$:

$$\text{For } g_E: \quad F_D(t) = \Pr(\|X\|_F \le t) \propto \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho,$$

$$\text{For } g_{BW}: \quad F_D(t) = \Pr(\|X\|_F \le t) \propto \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N/2-1}\mathrm{d}\rho.$$

2. Example II: $\mathcal{E}(X) = \mathrm{Tr}\,(X\log X)$ with the CDF $F_D(t) = \Pr(\|X\|_F \le t)$ for $D(X) = \|X\|_F$:

$$\text{For } g_E: \quad F_D(t) \propto \int\limits_{\substack{\sum_{i=1}^p \lambda_i^2 < t^2 \\ \lambda_i > 0, i=1,\ldots,p}} \prod_{1\le i<j\le p}|\lambda_i - \lambda_j| \prod_{i=1}^p |\lambda_i|^{n-p-\beta\lambda_i} \prod_{i=1}^p \mathrm{d}\lambda_i,$$

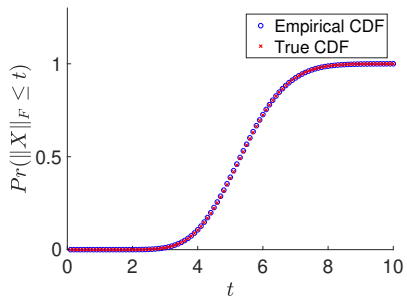
$$\text{For } g_{BW}: \quad F_D(t) \propto \int\limits_{\substack{\sum_{i=1}^p \lambda_i^2 < t^2 \\ \lambda_i > 0, i=1,\ldots,p}} \prod_{1\le i<j\le p}\frac{|\lambda_i - \lambda_j|}{\sqrt{\lambda_i+\lambda_j}} \prod_{i=1}^p |\lambda_i|^{\frac{n-p-1}{2}-\beta\lambda_i} \prod_{i=1}^p \mathrm{d}\lambda_i.$$

which is a p-fold integral and can be approximated accurately by quadrature such as Simpson's rule for relatively small values of p, e.g., $p = 2, 3$.

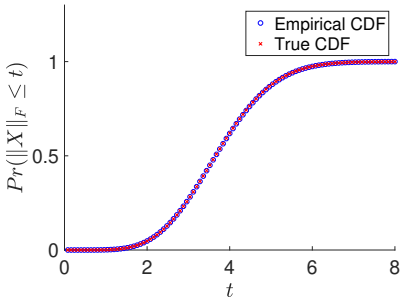
3. Example III: $\mathcal{E}(X) = \frac{1}{2}\|X - A\|_F^2$ where $A \in \mathcal{S}_+^{n,p}$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \gg 1$, with the CDF for $D(X) = \|X - A\|_F$:

$$\text{For both } g_E \text{ and } g_{BW}: \quad F_D(t) = \Pr(\|X - A\|_F \leq t) \propto \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \mathrm{d}\rho.$$

In implementation of the scheme, the step size Δt and β in schemes (3.1) and (3.5) are two parameters that need to be tuned to reach equilibrium with reasonable computing time. We first use a numerically stable Δt then adjust β so that the noise term has reasonable variance. And of course one needs a sufficient large number of iterations for schemes (3.1) and (3.5) to reach their equilibrium state, and a sufficient large number m of samples to observe numerical convergence toward the Gibbs distribution through the scalar random variable D, e.g., the KS statistic (5.1) should be small. See Figure 1, Figure 2, Figure 3, and Figure 4 for the numerical results.



(a) Scheme E (3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.4$. The error between two CDFs is $KS = 0.0054$.



(b) Scheme BW (3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\Delta t = 0.001$ and $\beta = 0.4$. The error between two CDFs is $KS = 0.0023$.

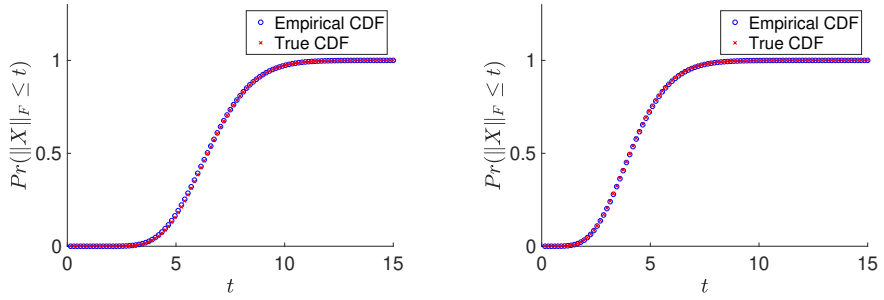
FIG. 1. *Example I:* $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2, n = 5, p = 3$ *and manifold dimension is* $N = 12$. *The empirical CDF is computed by* $5E6$ *MCMC samples generated after* $6E6$ *iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs of scheme E and scheme BW are evaluated at 100 equally spaced points on* $[0, 10]$ *and* $[0, 8]$*, respectively, and the difference can be measured by the KS statistic* (5.1).

5.2. MCMC numerical integration. We consider special cases $k = 0, m = 2$ in the examples (4.18) and (4.19), then (4.18) reduces to $(\frac{\beta}{\alpha})^{N/2}$ and (4.19) reduces to $(\frac{\beta}{\alpha})^{N/4}$. In other words, we may verify the numerical convergence of samples X_i to Gibbs distribution by verifying

$$\text{For } g_E: \quad \frac{1}{m}\sum_{i=1}^{m} \mathrm{e}^{-\frac{\alpha-\beta}{2}\|X_i\|_F^2} \to (\frac{\beta}{\alpha})^{N/2}, \tag{5.4}$$

$$\text{For } g_{BW}: \quad \frac{1}{m}\sum_{i=1}^{m} \mathrm{e}^{-\frac{\alpha-\beta}{2}\|X_i\|_F^2} \to (\frac{\beta}{\alpha})^{N/4}. \tag{5.5}$$

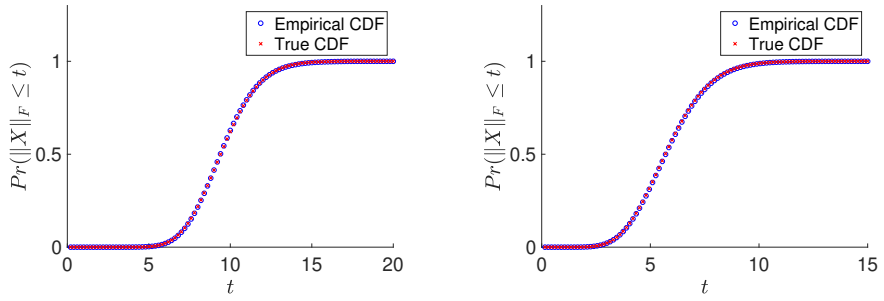
In Figure 5 we indeed observe the $\mathcal{O}(1/\sqrt{m})$ for the relative error of numerical integration.



(a) Scheme E (3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.0096$

(b) Scheme BW (3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.0043$.

FIG. 2. *Example II:* $\mathcal{E}(X) = \mathrm{Tr}\,(X \log X)$, $n = 5, p = 3$ *and manifold dimension is* $N = 12$. *The empirical CDF is computed by* $5E6$ *MCMC samples generated after* $6E6$ *iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs are evaluated at 100 equally spaced points on* $[0, 15]$, *and the difference can be measured by the KS statistic* (5.1).



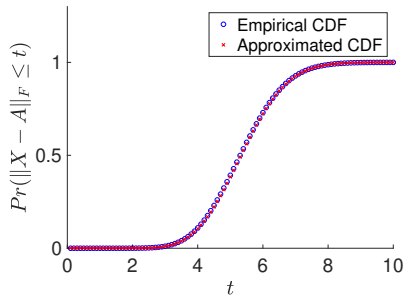
(a) Scheme E (3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.006$.

(b) Scheme BW (3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.0043$.

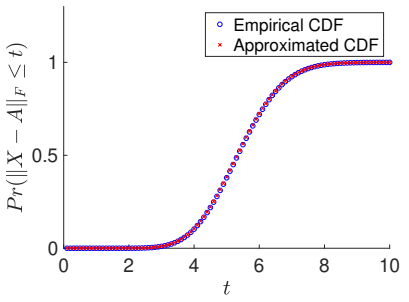
FIG. 3. *Example II:* $\mathcal{E}(X) = \mathrm{Tr}\,(X \log X)$, $n = 10, p = 2$ *and manifold dimension is* $N = 19$. *The empirical CDF is computed by* $5E6$ *MCMC samples generated after* $6E6$ *iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs of scheme E and scheme BW are evaluated at 100 equally spaced points on* $[0, 20]$ *and* $[0, 15]$,*respectively, and the difference can be measured by the KS statistic* (5.1).

5.3. A numerical study of the convergence to equilibrium. The general mathematical theory of convergence of a Langevin equation to its equilibrium measure has been well studied; we consider the specific case of the RLE studied here in the companion paper [36]. One particular application of the two Riemannian Langevin Monte Carlo schemes is to use them to numerically study the SDE solutions, e.g., by taking very small time steps, a Riemannian Langevin Monte Carlo scheme approximates the Riemannian Langevin equation on the manifold. We have shown comparison of the Langevin equation on $(\mathcal{S}_+^{n,p}, g_E)$, $(\mathcal{S}_+^{n,p}, g_{BW})$, $\mathbb{R}^{n\times n}$ in Figure 6, in which we can see interesting differences between two metrics. With all three figures in Figure 6, we can see that the SDE on $(\mathcal{S}_+^{n,p}, g_{BW})$ has a much faster convergence to its Gibbs measure than the SDE on $(\mathcal{S}_+^{n,p}, g_E)$.

6. Conclusion. We have constructed two efficient Riemannian Langevin Monte Carlo schemes for sampling PSD matrices of fixed rank from the Gibbs distribution

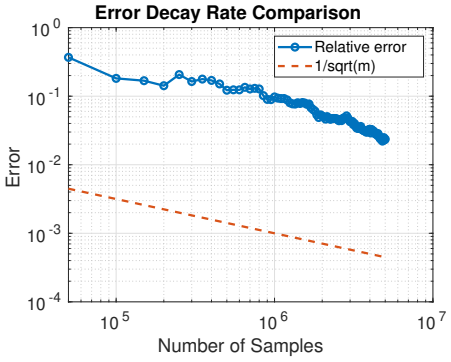


(a) Scheme E (3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.4$. The error between two CDFs is $KS = 0.0084$.

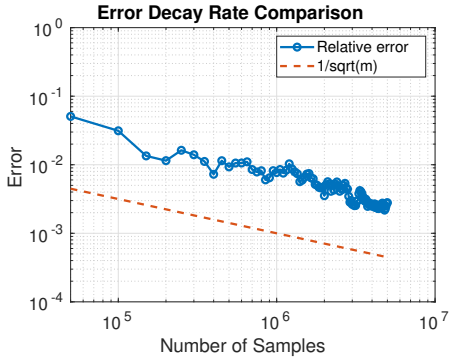


(b) Scheme BW (3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with Δt =2E-7 and $\beta = 0.4$. The error between two CDFs is $KS = 0.0052$.

FIG. 4. *Example III:* $\mathcal{E}(X) = \frac{1}{2}\|X - A\|_F^2$, $n = 5, p = 3$ *and manifold dimension is* $N = 12$. *The nonzero eigenvalues of* A *are equally spaced between 10000 and 20000. The empirical CDF is computed by* $5E6$ *MCMC samples generated after* $6E6$ *iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs are evaluated at 100 equally spaced points on* $[0, 10]$, *and the difference can be measured by the KS statistic* (5.1).



(a) Integration on $(\mathcal{S}_+^{n,p}, g_E)$ via samples generated by Scheme E (3.1) with $\Delta t = 0.001$ and $\beta = 0.4$.



(b) Integration on $(\mathcal{S}_+^{n,p}, g_{BW})$ via samples generated by Scheme BW (3.5) with $\Delta t = 0.001$ and $\beta = 0.4$.

FIG. 5. *Convergence rate of the relative error of* $\left|\frac{\hat{\mu}_m - \mu}{\mu}\right|$ *MCMC integration on the manifold with* $n = 10, p = 2$ *and dimension* $N = 19$. *Parameters are* $\alpha = 0.75, \beta = 0.4$, *for which it is a numerical integration of the function* $f(X) = \frac{1}{2}\|X\|_F^2$ *on the manifold* $\mathcal{S}_+^{n,p}$. *The error shown is the averaged one of 12 independent runs.*

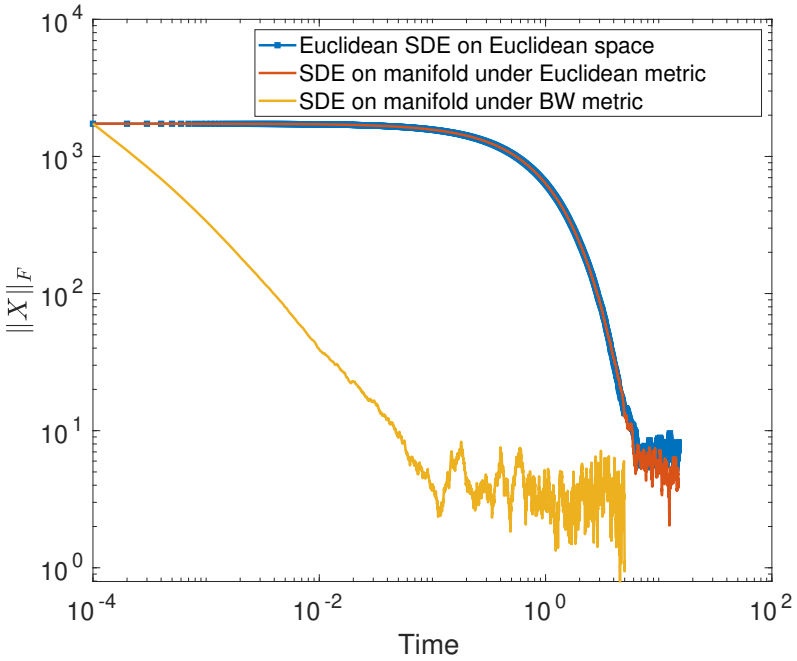
on the manifold $\mathcal{S}_+^{n,p}$ equipped with two fundamental metrics. We have also provided several examples for which these sampling schemes can be numerically validated.

REFERENCES

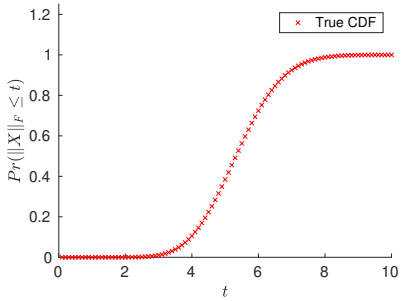
[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2008.

[2] R. Bhatia, T. Jain, and Y. Lim, *On the Bures-Wasserstein distance between positive definite matrices*, 2017, https://arxiv.org/abs/1712.01504.

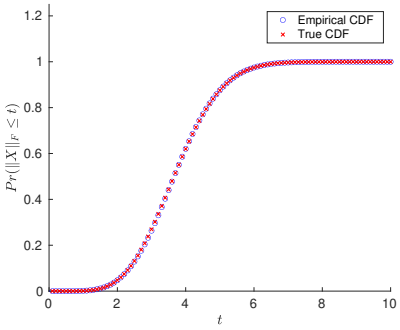
[3] M. Brubaker, M. Salzmann, and R. Urtasun, *A Family of MCMC Methods on Implicitly Defined Manifolds*, in Proceedings of the Fifteenth International Conference on Artificial



(a) Convergence of SDEs to the equilibrium.



(b) The true CDF of $D = \|X\|_F$ on $(\mathcal{S}_+^{n,p}, g_E)$. This implies the blue curve in Figure (a) has not reached equilibrium at Time=1.



(c) The true CDF of $D = \|X\|_F$ on $(\mathcal{S}_+^{n,p}, g_{BW})$, and empirical CDF of 5E6 samples from the Scheme BW after Time=5.

FIG. 6. *For $\mathcal{S}_+^{n,p}$ with $n = 5, p = 3$, the dimension is $N = 12$, and $\mathbb{R}^{n\times n}$ has dimension 25. The Gibbs measure is $e^{-\beta\mathcal{E}(x)}$ with $\beta = 0.4$ and $\mathcal{E} = \|X\|_F^2$. The time step size is $1e-4$. The initial guess is a random PSD matrix of rank-3 with eigenvalues = [1000,1000,1000].*

Intelligence and Statistics, N. D. Lawrence and M. Girolami, eds., vol. 22 of Proceedings of Machine Learning Research, La Palma, Canary Islands, 21–23 Apr 2012, PMLR, pp. 161–172.

[4] S. Burer and R. D. Monteiro, *Local minima and convergence in low-rank semidefinite programming*, Mathematical programming, 103 (2005), pp. 427–444.

[5] S. Byrne and M. Girolami, *Geodesic Monte Carlo on embedded manifolds*, Scandinavian Journal of Statistics, 40 (2013), pp. 825–845.

[6] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, in Conference on learning theory, PMLR, 2018, pp. 300–323.

[7] X. Cheng, D. Yin, P. Bartlett, and M. Jordan, *Stochastic gradient and Langevin processes*, in International Conference on Machine Learning, PMLR, 2020, pp. 1810–1819.
[8] X. Cheng, J. Zhang, and S. Sra, *Efficient Sampling on Riemannian Manifolds via Langevin MCMC*, in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., vol. 35, Curran Associates, Inc., 2022, pp. 5995–6006.
[9] G. Ciccotti, R. Kapral, and E. Vanden-Eijnden, *Blue moon sampling, vectorial reaction coordinates, and unbiased constrained dynamics*, ChemPhysChem, 6 (2005), pp. 1809–1814.
[10] G. Ciccotti, T. Lelievre, and E. Vanden-Eijnden, *Projection of diffusions on submanifolds: Application to mean force computation*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 61 (2008), pp. 371–408.
[11] M. P. Do Carmo and J. Flaherty Francis, *Riemannian geometry*, vol. 6, Springer, 1992.
[12] Y. Du and I. Mordatch, *Implicit generation and modeling with energy based models*, Advances in Neural Information Processing Systems, 32 (2019).
[13] R. Ge, H. Lee, J. Lu, and A. Risteski, *Efficient sampling from the bingham distribution*, in Algorithmic Learning Theory, PMLR, 2021, pp. 673–685.
[14] C. J. Geyer, *Markov chain monte carlo lecture notes*, Course notes, Spring Quarter, 80 (1998).
[15] M. Girolami and B. Calderhead, *Riemann manifold langevin and hamiltonian monte carlo methods*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 123–214.
[16] E. P. Hsu, *Stochastic analysis on manifolds*, vol. 38 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2002.
[17] C.-P. Huang, D. Inauen, and G. Menon, *Motion by mean curvature and Dyson Brownian motion*, Electron. Commun. Probab., 28 (2023), pp. 1–10, https://doi.org/10.1214/23-ECP540.
[18] W. Huang and X. Zhang, *Solving PhaseLift by low-rank Riemannian optimization methods for complex semidefinite constraints*, SIAM Journal on Scientific Computing, 39 (2017), pp. B840–B859.
[19] N. Ikeda and S. Watanabe, *Stochastic differential equations and diffusion processes*, vol. 24 of North-Holland Mathematical Library, North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second ed., 1989.
[20] D. Inauen and G. Menon, *Stochastic Nash evolution*, 2023, https://arxiv.org/abs/TBD.
[21] G. L. Jones, *On the Markov chain central limit theorem*, Probability Surveys, 1 (2004), pp. 299 – 320, https://doi.org/10.1214/154957804100000051, https://doi.org/10.1214/154957804100000051.
[22] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, vol. 23 of Applications of Mathematics (New York), Springer-Verlag, Berlin, 1992.
[23] J. Leake, C. McSwiggen, and N. K. Vishnoi, *Sampling matrices from harish-chandra–itzykson–zuber densities with applications to quantum inference and differential privacy*, in Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, 2021, pp. 1384–1397.
[24] M. B. Li and M. A. Erdogdu, *Riemannian langevin algorithm for solving semidefinite programs*, 2023, https://arxiv.org/abs/2010.11176.
[25] J. S. Liu and J. S. Liu, *Monte Carlo strategies in scientific computing*, vol. 75, Springer, 2001.
[26] E. Massart and P.-A. Absil, *Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices*, SIAM Journal on Matrix Analysis and Applications, 41 (2020), pp. 171–198.
[27] E. Massart, J. M. Hendrickx, and P.-A. Absil, *Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the bures–wasserstein metric*, in Geometric Science of Information: 4th International Conference, GSI 2019, Toulouse, France, August 27–29, 2019, Proceedings, Springer, 2019, pp. 739–748.
[28] G. Menon and T. Yu, *The Riemannian Langevin equation and conic programs*, 2023, https://arxiv.org/abs/2302.11653.
[29] G. Menon and T. Yu, *Siegel Brownian motion*, 2023, https://arxiv.org/abs/TBD.
[30] G. Meyer, S. Bonnabel, and R. Sepulchre, *Regression on fixed-rank positive semidefinite matrices: a Riemannian approach*, The Journal of Machine Learning Research, 12 (2011), pp. 593–625.
[31] A. Moitra and A. Risteski, *Fast Convergence for Langevin with Matrix Manifold Structure*, in ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations, 2020.

[32] T. Pacini, *Mean curvature flow, orbits, moment maps*, Trans. Amer. Math. Soc., 355 (2003), pp. 3343–3357, https://doi.org/10.1090/S0002-9947-03-03307-5, https://doi.org/10.1090/S0002-9947-03-03307-5.

[33] A. Tasissa and R. Lai, *Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion*, IEEE Transactions on Information Theory, 65 (2018), pp. 3124–3144.

[34] B. Vandereycken, P.-A. Absil, and S. Vandewalle, *Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank*, in 2009 IEEE/SP 15th Workshop on Statistical Signal Processing, IEEE, 2009, pp. 389–392.

[35] P. Xu, J. Chen, D. Zou, and Q. Gu, *Global convergence of langevin dynamics based algorithms for nonconvex optimization*, 2020, https://arxiv.org/abs/1707.06618.

[36] T. Yu, S. Zheng, J. Lu, G. Menon, and X. Zhang, *Riemannian Langevin equations for PSD matrices of fixed rank*, 2023, https://arxiv.org/abs/TBD.

[37] E. Zappa, M. Holmes-Cerfon, and J. Goodman, *Monte Carlo on manifolds: sampling densities and integrating functions*, Communications on Pure and Applied Mathematics, 71 (2018), pp. 2609–2647.

[38] S. Zheng, W. Huang, B. Vandereycken, and X. Zhang, *Riemannian optimization using three different metrics for Hermitian PSD fixed-rank constraints: an extended version*, 2022, https://arxiv.org/abs/2204.07830.