

On a Class of Gibbs Sampling over Networks

Bo Yuan

Georgia Institute of Technology

BYUAN48@GATECH.EDU

Jiaojiao Fan

Georgia Institute of Technology

JIAOJIAOFAN@GATECH.EDU

Jiaming Liang

Yale University

JIAMING.LIANG@YALE.EDU

Andre Wibisono

Yale University

ANDRE.WIBISONO@YALE.EDU

Yongxin Chen

Georgia Institute of Technology

YONGCHEN@GATECH.EDU

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the sampling problem from a composite distribution whose potential (negative log density) is $\sum_{i=1}^n f_i(x_i) + \sum_{j=1}^m g_j(y_j) + \sum_{i=1}^n \sum_{j=1}^m \sigma_{ij}/2\eta \|x_i - y_j\|_2^2$ where each of x_i and y_j is in \mathbb{R}^d , $f_1, f_2, \dots, f_n, g_1, g_2, \dots, g_m$ are strongly convex functions, and $\{\sigma_{ij}\}$ encodes a network structure. Building on the Gibbs sampling method, we develop an efficient sampling framework for this problem when the network is a bipartite graph. More importantly, we establish a non-asymptotic linear convergence rate for it. This work extends earlier works that involve only a graph with two nodes [Lee et al. \(2021\)](#). To the best of our knowledge, our result represents the first non-asymptotic analysis of a Gibbs sampler for structured log-concave distributions over networks. Our framework can be potentially used to sample from the distribution $\propto \exp(-\sum_{i=1}^n f_i(x_i) - \sum_{j=1}^m g_j(y_j))$ in a distributed manner.

1. Introduction

Sampling has been increasingly important in the areas of computer science and Bayesian inference as it is often the critical element for parameter estimations, simulations, and numerical approximations. Designing and analysis of sampling for log-concave distributions whose negative log density is convex have become an active research field since the randomized convex body volume approximation algorithm proposed by [Dyer et al. \(1991\)](#). Depending on various structured distributions including non-convex potentials and associated efficient sampling algorithms, recently a large number of non-asymptotic results are established under different scenarios, *e.g.*, Langevin Monte Carlo ([Dalalyan, 2017](#); [Cheng et al., 2018](#); [Dalalyan and Riou-Durand, 2018](#); [Li et al., 2021](#); [Zhang et al., 2023](#)), proximal samplers ([Lee et al., 2021](#); [Chen et al., 2022](#); [Gopi et al., 2022](#); [Fan et al., 2023](#); [Altschuler and Chewi, 2023](#)), lower bounds of convergence rates ([Chewi et al., 2022](#), [2023a,b](#)), etc.

In this work, we establish a linear convergence rate of a Gibbs sampler for the following unnormalized target distribution \propto

$$\exp\left(-\sum_{i=1}^n f_i(x_i) - \sum_{j=1}^m g_j(y_j) - \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - y_j\|_2^2\right). \quad (1)$$

Here, both f_i and g_j functions are strongly convex (see Section 2.2 for definitions), and x_i and y_j represent random variables on vertices on the underlying network. Our formalization indicates the network considered in this work is a bipartite graph where the two disjoint sets are expressed as $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$, respectively. In the expression, the parameter σ_{ij} encodes the graph structure, *i.e.*, $0 < \sigma_{ij} \leq 1$ if there is an edge between vertex x_i and vertex y_j . Otherwise, $\sigma_{ij} = 0$. We also utilize a positive coefficient η to control the dependency between variables.

The structured sampling problem (1) has potential applications in the fields of graphical models, distributed sampling, federated learning, etc. For instance, the joint distribution over Hidden Markov models or spanning trees with edge potentials being Gaussian distributions can be formulated as our target distribution, where $\eta/\sigma_{ij}\mathbf{I}$ is the covariance matrix of edge potential for the edge between vertices x_i and y_j . Our analysis can be straightforwardly generalized to the case that the covariance matrix is not diagonal. Other applications include estimation problems for robotics where the quadratic terms correspond to odometry measurements and the other terms f_i, g_j correspond to other measurements, *e.g.*, perception.

In proximal samplers, for a target distribution $\exp(-h(x))$, a Gibbs sampling framework is utilized to sample from the joint distribution $\exp(-h(x) - 1/2\eta\|x - y\|_2^2)$. Since the X -marginal distribution of the joint one is exactly $\exp(-h(x))$, it is sufficient to generate samples from the joint one efficiently. In brief, as the conditional distribution $\exp(-h(x) - 1/2\eta\|x - y_0\|_2^2)$ for any given y_0 is a better-conditioned distribution than the target, the proximal sampler can be proved to have stronger convergence rate under various assumptions on $h(x)$ and implementations of sampling from the conditional measure (Chen et al., 2022). To improve the performance of proximal samplers, the key is to efficiently implement the sampler for conditional distributions, namely, Restricted Gaussian Oracle (RGO) (Lee et al., 2021). Among recent works in this direction, Gopi et al. (2022); Fan et al. (2023) utilized an approximate Rejection sampling framework and Altschuler and Chewi (2023) adopted a well-designed warm start. The proximal samplers achieve state-of-the-art results in a variety of settings of sampling (Fan et al., 2023; Altschuler and Chewi, 2023). Our target is a generalization of both proximal samplers and the distributed variable splitting MCMC (Vono et al., 2022) for large-scale Bayesian inference problems. Similarly, in our case, if $n = m = 1$, our target (1) given a sufficiently small η approximates a lower-dimensional composite distribution, $\exp(-f(x) - g(x))$. More precisely, we have shown that the distribution of samples converges in terms of total variation distance (see Section 5 for details).

Due to the structure of our target distribution (1), we adopt Gibbs Sampling in this work as it can be potentially used to realize a fully distributed algorithm for sampling as in Vono et al. (2022). In addition to the distributed nature and promising theoretical properties of Gibbs samplers, another reason to consider Gibbs samplers is the widely-used applications (Daphne et al., 2009). Analyzing its non-asymptotic behaviors is of great importance but challenging. A pioneering work that establishes the quantitative convergence rate for general Gibbs samplers was proposed by Rosenthal (1995) via the drift and minorization conditions. A long line of works was proposed for the mixing time of Gibbs samplers over graphs. (Hrycej, 1990; Venugopal and Gogate, 2012; De Sa et al., 2015; Zhang and De Sa, 2019; Vono et al., 2022). In recent works, Zhang and De Sa (2019) analyzed the measured of a mini-batched Gibbs sampler measured by the spectral gap (De Sa et al., 2018), which has been used for analyzing the plain Gibbs sampling (Levin and Peres, 2017). A recently developed class of samplers for

large-scale graphical models is distributed variable splitting MCMC. Among them, sampling is performed on an artificial hierarchical Bayesian model akin to the alternating direction method of multipliers (ADMM) in optimization (Vono et al., 2019; Rendell et al., 2020; Vono et al., 2020, 2022). To sample from $\exp(-\sum_{i=1}^n h_i(x))$, Vono et al. (2022) considers an augmented distribution $\exp(-\sum_{i=1}^n h_i(x_i) - 1/2\eta\|x_i - x\|^2)$ and shows the augmented one approximates the target by a sufficiently small η . Hence it is sufficient to sample from the augmented one with a Gibbs sampler. In Vono et al. (2022), the mixing time measured in 1-Wasserstein and TV distances have been given by Ricci curvature and coupling techniques. Inspired by the recent progress in proximal sampling, we give the first non-asymptotic analysis of a Gibbs sampler for (1) by a new strong data-processing inequality on diffusion processes. Moreover, in Vono et al. (2022), only the exact implementation of RGO was studied. In contrast, in Section 3, we discuss the overall mixing time of our algorithm with the inexact implementations of RGO in Fan et al. (2023).

Section 2 presents basic definitions and notation used throughout the paper. In Section 3, we formally prove the linear convergence rate of a Gibbs sampler for a two-node graph with a newly-established strong data processing inequality (Cover, 1999). In Section 4, we generalize the result to distributions over the general bipartite graph. Section 5 gives the analysis of convergence to the composite distribution $\exp(-f(x) - g(x))$, where we prove the asymptotic convergence and give a non-asymptotic convergence rate for special cases. Section 6 presents some concluding remarks and possible extension. Appendices A-D provide a few useful technical results and missing proofs in the paper.

2. Preliminaries

2.1. Notations

A Gaussian distribution is denoted by \mathcal{N} and the notation \mathbf{I} represents an identity matrix whose dimension is clear from the context. Assume π^{XY} denote the joint distribution of (X, Y) , then the X -marginal distribution is represented by π^X , and the X -conditional distribution given y is $\pi^{X|Y=y}$. Moreover, the joint distribution of multivariable \mathbf{X} and \mathbf{Y} is $\pi^{\mathbf{XY}}$, and similarly, its \mathbf{X} -marginal distribution and \mathbf{X} -conditional distribution are denoted by $\pi^{\mathbf{X}}$ and $\pi^{\mathbf{X}|\mathbf{Y}}$, respectively.

2.2. Sampling Background

We start by introducing strong convexity and smoothness, which are the main assumptions in this paper. In what follows, we use ∇f to represent the subgradient of f . A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is α -strongly convex if and only if $f(x) - f(y) \geq \nabla f(y)^T(x - y) + \alpha/2\|x - y\|_2^2$ holds for all $x, y \in \mathbb{R}^d$. Furthermore, if the strongly convex function f is twice-differentiable, then $\alpha\mathbf{I} \preceq \nabla^2 f$. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -smooth if and only if $f(x) - f(y) \leq \nabla f(y)^T(x - y) + \beta/2\|x - y\|_2^2$ is true for all $x, y \in \mathbb{R}^d$. Similarly, for a twice-differentiable β -smooth function f , one has $\nabla^2 f \preceq \beta\mathbf{I}$.

Next we introduce the functional inequalities used in this work. A probability distribution ν satisfies logarithmic Sobolev inequality with a positive constant ρ (in short LSI(ρ)) if, for any probability distribution μ such that $\mu \ll \nu$, the Kullback–Leibler (KL) divergence $D_{\text{KL}}(\mu||\nu) = \int \log \frac{\mu}{\nu} \mu$ and Fisher information $I(\mu||\nu) = \int \|\nabla \log \frac{\mu}{\nu}\|^2 \mu$ satisfy $D_{\text{KL}}(\mu||\nu) \leq$

$\frac{1}{2\rho}I(\mu||\nu)$. A well-known result that connects strongly convexity and logarithmic Sobolev inequality is that if μ is a strongly log-concave density with a constant α , then μ satisfies LSI(α). Moreover, Pinsker's inequality is $\text{TV}(\mu, \nu) \leq \sqrt{\frac{1}{2}D_{\text{KL}}(\mu||\nu)}$. Here the total variation between μ and ν satisfies $\text{TV}(\mu, \nu) = 1/2\|\mu - \nu\|_1$. In addition, we say a distribution ν satisfies Poincaré inequality with a constant $\rho > 0$ (in short PI(ρ)) if for any functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with $\nabla f \in L^2(\nu)$, $\text{Var}_\nu(f) \leq 1/\rho \mathbb{E}_\nu(\|\nabla f\|_2^2)$. It is well-known that LSI(ρ) implies PI(ρ).

3. Improved analysis of Composite Sampling

Assumption 1 *Let $f(x)$ be an α_f -strongly convex function and $g(y)$ be an α_g -strongly convex function.*

We are interested in the sampling problem from target distribution (1). For the sake of presentation, we start presenting the analysis with the two-node case as the proof of the bipartite graph is essentially the same as this one. In this case, the target distribution

$$\pi^{XY} \propto \exp\left(-f(x) - g(y) - \frac{1}{2\eta}\|x - y\|_2^2\right) \quad (2)$$

under Assumption 1. Based on the Gibbs sampling framework, our sampling algorithm 1 runs as follows.

Algorithm 1 A Gibbs sampler for a two-node graph

Input: π^{XY} : the target distribution
 x^0 : an initial sample drawn from μ_0^X
Output: (x^K, y^K) : samples generated in the k -th iteration
for $k \leftarrow 0, \dots, K$ **do**
 Step 1: Sample $y^k \sim \pi^{Y|X=x^k} \propto \exp(-g(y) - 1/2\eta\|x^k - y\|_2^2)$.
 Step 2: Sample $x^{k+1} \sim \pi^{X|Y=y^k} \propto \exp(-f(x) - 1/2\eta\|x - y^k\|_2^2)$.
end for

The critical step in Algorithm 1 is to sample from the conditional distribution, *i.e.*, Step 1 and 2. To analyze these steps, we construct an auxiliary diffusion process to represent the sampling from conditional distributions in Lemma 4. Note these processes are constructed merely for analysis. There exist efficient implementations of the RGO; see the discussions at the end of this section. To understand the dynamics of the diffusion process, we establish a strong data processing inequality on it by Lemma 2 and Theorem 3. In this way, we are able to show the sample distribution converges to the target one (2) in terms of KL divergence in Theorem 5.

Lemma 2 *Given an arbitrary diffusion*

$$d\mathbf{Z}_t = b_t(\mathbf{Z}_t)dt + \boldsymbol{\sigma}_t d\mathbf{W}_t,$$

where $\mathbf{Z}_t \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_t$ is a $d \times d$ **diagonal** matrix. Furthermore denote the i -th entry of $\boldsymbol{\sigma}_t$ as $\sigma_{t,i}$ and assume $\lambda_t^2 \leq \min_i \sigma_{t,i}^2$. Let μ_t, π_t be the distributions of \mathbf{Z}_t initialized with $\mathbf{Z}_0 \sim \mu_0, \mathbf{Z}_0 \sim \pi_0$ respectively. Then we have

$$\frac{\partial D_{\text{KL}}(\mu_t||\pi_t)}{\partial t} \leq -\frac{\lambda_t^2}{2} I(\mu_t||\pi_t). \quad (3)$$

Proof The proof is based on the Fokker–Planck equation. See the full proof in Appendix A.1. \blacksquare

Lemma 2 is different from the classical result describing the convergence to the stationary distribution of a Langevin diffusion. In Lemma 2, neither μ_t nor π_t needs to be the stationary distribution; they can both be time-varying. To our best knowledge, only a special case of it involving simultaneous heat flow (*i.e.*, $b_t(Z_t) \equiv 0$) was recently used in a few works (Lee et al., 2021; Chen et al., 2022).

Theorem 3 (Strong data processing inequality with LSI) *Under the condition of Lemma 2, if we further assume π_t satisfies the LSI with a coefficient α_t , $\forall t \in [0, T]$, then*

$$D_{\text{KL}}(\mu_T \parallel \pi_T) \leq \exp\left(-\int_0^T \alpha_t \lambda_t^2 dt\right) D_{\text{KL}}(\mu_0 \parallel \pi_0).$$

Proof With the logarithmic Sobolev inequality of π_t and (3), we have

$$\frac{\partial D_{\text{KL}}(\mu_t \parallel \pi_t)}{\partial t} \leq -\alpha_t \lambda_t^2 D_{\text{KL}}(\mu_t \parallel \pi_t).$$

Solving the above differential inequality gives

$$D_{\text{KL}}(\mu_T \parallel \pi_T) \leq \exp\left(-\int_0^T \alpha_t \lambda_t^2 dt\right) D_{\text{KL}}(\mu_0 \parallel \pi_0).$$

\blacksquare

Note that Theorem 3 is tight if the elements in σ_t is the same. This can be seen from a simple example $b_t(Z_t) = -Z_t$, $\sigma_t = \sqrt{2}\mathbf{I}$, $\mu_0 = \mathcal{N}(1, 1)$ and $\nu_0 = \mathcal{N}(2, 1)$ in Lemma 2. Lemma 2 is a generalization of Lemma 16 in Vempala and Wibisono (2019) and Lemma 12 in Chen et al. (2022). In Lemma 2, we assume each entry of the diagonal matrix σ_t can be different. It is worth mentioning that similar results can be established with other probability divergences. Under the LSI assumption, one can obtain the contraction property with Rényi divergence (Section A.4 in Chen et al. (2022)). Moreover, similar results still hold with different assumptions for π_t . For instance, in Appendix C, we show a data processing inequality with the convexity assumption. We refer the reader to Chen et al. (2022) for a comprehensive study. Next we show how to construct the diffusion process for a given joint distribution of two endpoints.

Lemma 4 *Define a distribution $P(z(\cdot))$ over the space of trajectories $z(\cdot) \in C([0, 1]; \mathbb{R}^d)$ as follows: $P(z(0), z(1)) \propto \exp(-f(z(0)) - g(z(1)) - 1/2\eta\|z(0) - z(1)\|^2)$ is the joint distribution of Z_0 and Z_1 , and $P(z(\cdot)|z(0), z(1))$ is the process of $dZ_t = \sqrt{\eta}dW_t$ conditioning on $Z_0 = z(0), Z_1 = z(1)$ (*a.k.a.*, the Brownian bridge). Then P has the diffusion representation*

$$dZ_t = \eta \nabla \log \phi_t(Z_t) dt + \sqrt{\eta} dW_t, \quad Z_0 \sim P(z(0)) \tag{4}$$

in the forward direction and

$$dZ_t = -\eta \nabla \log \hat{\phi}_t(Z_t) dt + \sqrt{\eta} dW_t, \quad Z_1 \sim P(z(1)) \tag{5}$$

in the backward direction. Here

$$\begin{aligned}\phi_t(z) &= \int \exp[-g(y) - \frac{1}{2\eta(1-t)}\|z - y\|_2^2] dy \\ \hat{\phi}_t(z) &= \int \exp[-f(x) - \frac{1}{2\eta t}\|z - x\|_2^2] dx.\end{aligned}\tag{6}$$

Moreover, the marginal distribution of P satisfies $P(z(t)) \propto \phi_t(z(t))\hat{\phi}_t(z(t))$.

Proof We show that $P(Z_1|Z_0, Z_1)$, the conditional distribution of (4), coincides with the Brownian bridge. See the full proof in Appendix A.2. \blacksquare

In Lemma 4, the forward diffusion (4) only relies on the conditional distribution $P(z(1)|z(0))$. It is also worth pointing out that the diffusion process in Lemma 4 is exactly the solution of the Schrödinger bridge problem. Meanwhile, (6) is the solution of the Schrödinger system, one way to solve the Schrödinger bridge problem. We refer the reader to Léonard (2014); Chen et al. (2016, 2021); Pavon et al. (2021); Vargas (2021) for more detailed explanations. Theorem 5 is the main result for the two-node case. With Theorem 3 and Lemma 4, we have the linear convergence rate for the Gibbs sampling framework shown as follows.

Theorem 5 (Convergence result with strong-convexity) *Denote the distribution of the k -th samples (x^k, y^k) by μ_k^{XY} . For the target distribution (2), under Assumption 1,*

$$D_{\text{KL}}(\mu_k^{XY} \parallel \pi^{XY}) \leq \frac{D_{\text{KL}}(\mu_0^X \parallel \pi^X)}{(1 + \eta\alpha_f)^{2k}(1 + \eta\alpha_g)^{2k}}.\tag{7}$$

Proof In Lemma 4, the forward diffusion process (4) only depends on the the conditional distribution $P(Z(1)|Z(0))$, so we can construct the same diffusion process for both (π^X, π^Y) and (μ_k^X, μ_k^Y) , i.e.,

$$dZ_t = \eta \nabla \log \phi_t(Z_t) dt + \sqrt{\eta} dW_t.$$

As $\pi^{XY} \propto \exp(-f(x) - g(y) - \frac{1}{2\eta}\|x - y\|^2)$, we have the explicit expression for the distribution of Z_t by (6)

$$P(z_t) \propto \int \exp[-g(y) - \frac{1}{2\eta(1-t)}\|z - y\|_2^2] dy \int \exp[-f(x) - \frac{1}{2\eta t}\|z - x\|_2^2] dx.\tag{8}$$

Under Assumption 1, $f(x)$ is α_f -strongly convex and $g(y)$ is α_g -strongly convex. Hence, by Theorem 3.7.b and 3.7.d in Saumard and Wellner (2014), we get $P(z_t)$ is strongly log-concave with a coefficient $\frac{\alpha_f}{1 + \eta t \alpha_f} + \frac{\alpha_g}{1 + \eta(1-t) \alpha_g}$. Therefore, $P(z_t)$ satisfies $\text{LSI}(\frac{\alpha_f}{1 + \eta t \alpha_f} + \frac{\alpha_g}{1 + \eta(1-t) \alpha_g})$. So, for Step 1 in Algorithm 1, we can employ the strong data processing inequality in Theorem 3, which yields

$$D_{\text{KL}}(\mu_k^Y \parallel \pi^Y) \leq D_{\text{KL}}(\mu_k^X \parallel \pi^X) \frac{1}{(1 + \eta\alpha_f)(1 + \eta\alpha_g)}, \forall k.$$

Similarly, for Step 2 in Algorithm 1 we can perform a symmetric analysis, which directly gives

$$D_{\text{KL}}(\mu_{k+1}^X \parallel \pi^X) \leq D_{\text{KL}}(\mu_k^Y \parallel \pi^Y) \frac{1}{(1 + \eta\alpha_f)(1 + \eta\alpha_g)}, \forall k.$$

Therefore,

$$D_{\text{KL}}(\mu_k^X \parallel \pi^X) \leq D_{\text{KL}}(\mu_0^X \parallel \pi^X) \frac{1}{(1 + \eta\alpha_f)^{2k}(1 + \eta\alpha_g)^{2k}}.$$

Lastly according to Lemma 15 we have

$$D_{\text{KL}}(\mu_k^{XY} \parallel \pi^{XY}) = D_{\text{KL}}(\mu_k^X \parallel \pi^X) \leq D_{\text{KL}}(\mu_0^X \parallel \pi^X) \frac{1}{(1 + \eta\alpha_f)^{2k}(1 + \eta\alpha_g)^{2k}}.$$

■

Our result indicates that the convergence is faster with larger η , which means Gibbs sampling may have good performance for slightly correlated variables. For more discussion of the impact of correlated structure on Gibbs sampling, see e.g., [Roberts and Sahu \(1997\)](#). Notice that our result is still meaningful, even one of f or g is convex. However, if the strong convexity constants α_f and α_g both approach 0, *i.e.*, f and g are convex instead of strongly convex, the convergence fails. Hence we discuss how to relax the strong-convexity assumption. The strong convexity in Assumption 1 can be relaxed to convexity. Following similar techniques in section A.3 in [Chen et al. \(2022\)](#), we establish a data processing inequality with the convexity assumption and thereby convergence rate of the two-node case.

Theorem 6 (Convergence rate with convexity) *Denote the distribution of the k -th samples (x^k, y^k) by μ_k^{XY} . Assume f and g are convex. For target distribution (2), we have*

$$D_{\text{KL}}(\mu_k^{XY} \parallel \pi^{XY}) \leq \frac{W_2^2(\mu_0^X, \pi^X)}{k\eta}.$$

Proof See Appendix C for the proof. ■

Remark 7 (LSI assumption) *The Assumption 1 can be potentially relaxed to LSI conditions. Note that we only need LSI conditions in Theorem 3. However, to establish the convergence rate, we need to compute the LSI constant for (6), which requires the LSI constant for the product of two densities that satisfy LSI. To the best of our knowledge, it is an open problem without assuming stronger conditions. We refer the reader to [Villani \(2021\)](#) for more discussions.*

Next, we consider the mixing time of Algorithm 1 by combining Theorem 5 and the implementation of RGO. For the exact implementation, one can obtain the sampling complexity of (2) with additional conditions in Assumption 8.

Assumption 8 *Assume f and g are smooth with coefficients β_f and β_g , respectively.*

Under Assumption 1 and 8, one can employ the rejection sampling framework described in Section 4.2 in [Chen et al. \(2022\)](#) to have an **exact** implementation of Step 1 and 2. In short, for a strongly convex and smooth potential, one can find the minimizer of the potential and then use rejection sampling with a Gaussian distribution centered at the minimizer. Then, the expected number of iterations for Step 1 and Step 2 is $\left(\frac{\beta_g + 1/\eta}{\alpha_g + 1/\eta}\right)^{d/2}$ and $\left(\frac{\beta_f + 1/\eta}{\alpha_f + 1/\eta}\right)^{d/2}$,

respectively. (Theorem 7 in [Chewi et al. \(2021\)](#)). Hence, given the condition $\eta = \mathcal{O}(\frac{1}{(\beta_f + \beta_g)d})$, sampling from conditional distributions only requires $\mathcal{O}(1)$ steps in expectation. Then with Theorem 5 and Pinsker's inequality, to have δ TV distance to the target (2), the number of Gibbs sampling iterations satisfies

$$K = \mathcal{O}\left(\frac{(\beta_f + \beta_g)d}{\alpha_f + \alpha_g} \log \frac{D_{\text{KL}}(\mu_0^X || \pi^X)}{\delta^2}\right).$$

Under the same assumptions, Assumption 1 and 8, we can also use the **inexact** implementation in [Fan et al. \(2023\)](#) for Step 1 and 2. By Theorem 3 in [Fan et al. \(2023\)](#), the expected number of iterations for Step 1 and Step 2 are $\mathcal{O}(1)$ given $\eta = \mathcal{O}(\frac{\log^{-1}(1/\epsilon_{\text{RGO}})}{(\beta_f + \beta_g)d^{1/2}})$ where ϵ_{RGO} is the TV distance between the conditional distribution and the distribution of samples. The error ϵ_{RGO} can accumulate at each iteration of the Gibbs sampler but can be controlled due to the triangular inequality for total variation and the logarithmic dependence of η over ϵ_{RGO} . Following a similar argument as in Proposition 1 in [Fan et al. \(2023\)](#), we can establish the number of iterations is

$$K = \tilde{\mathcal{O}}\left(\frac{(\beta_f + \beta_g)d^{1/2}}{\alpha_f + \alpha_g}\right).$$

Note that the smoothing assumption is not required in the proof of convergence rate. It is only necessary to establish the bound of the number of iterations when sampling from the conditional distributions. Therefore, Theorem 5 is still applicable to strongly convex but non-smooth functions.

4. Gibbs sampling over bipartite graphs

Assumption 9 *Let both f_i and g_j be strongly convex functions, and denote the strong convexity constants of f_i and g_j by α_{f_i} and α_{g_j} , respectively. We also impose two regularity conditions for the graph structure, i.e., $\sum_{i=1}^n \sigma_{ij} > 0$, $\forall j$ and $\sum_{j=1}^m \sigma_{ij} > 0$, $\forall i$.*

In this section, we consider sampling over bipartite graphs with a Gibbs sampler under Assumption 9. A bipartite graph G is a graph whose vertices can be divided into two disjoint sets U and V with the constraint that each edge in G connects one vertex in U and another one in V . Note that bipartite graphs include a large number of graph structures, e.g., a tree is always a bipartite graph. The target distribution we sample from is of the form

$$\pi^{\mathbf{XY}} \propto \exp\left[-\sum_{i=1}^n f_i(x_i) - \sum_{j=1}^m g_j(y_j) - \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - y_j\|_2^2\right]$$

where each x_i and y_j represents a vertex in U and V , respectively. If there is an edge between vertex i in U and vertex j in V , then $0 < \sigma_{ij} \leq 1$. Otherwise $\sigma_{ij} = 0$. Note that the two regularity conditions in Assumption 9 imply that for any vertex in G , at least one edge connects it with the rest part of G . Next we present a blocked Gibbs sampling algorithm over a bipartite graph G . In Algorithm 2, \mathbf{X}_i^k denotes the sample on vertex i in the k -th iteration.

Algorithm 2 A Gibbs sampler over a bipartite graph

Input: $\pi^{\mathbf{XY}} \propto \exp[-\sum_{i=1}^n f_i(x_i) - \sum_{j=1}^m g_j(y_j) - \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - y_j\|_2^2]$: the target distribution
X⁰: an initial sample of \mathbf{X} drawn from $\mu_0^{\mathbf{X}}$
Output: $(\mathbf{X}^K, \mathbf{Y}^K)$: samples that approximate the target distribution $\pi^{\mathbf{XY}}$

for $k \leftarrow 0, \dots, K$ **do**

for $j \leftarrow 1, \dots, m$ **do**

Step 1: Sample $\mathbf{Y}_j^k \sim \pi^{\mathbf{Y}_j | \mathbf{X} = \mathbf{X}^k} \propto \exp\left[-g_j(y_j) - \sum_{i=1}^n \frac{\sigma_{ij}}{2\eta} \|\mathbf{X}_i^k - y_j\|_2^2\right]$.

end for

for $i \leftarrow 1, \dots, n$ **do**

Step 2: Sample $\mathbf{X}_i^{k+1} \sim \pi^{\mathbf{X}_i | \mathbf{Y} = \mathbf{Y}^k} \propto \exp\left[-f_i(x_i) - \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - \mathbf{Y}_j^k\|_2^2\right]$.

end for

end for

One can see that Step 1 and Step 2 in the above Algorithm 2 can be executed in a distributed manner thanks to the bipartite graph structure and Gibbs samplers. Notice that in Step 1,

$$\exp\left[-g_j(y_j) - \sum_{i=1}^n \frac{\sigma_{ij}}{2\eta} \|\mathbf{X}_i^k - y_j\|_2^2\right] \propto \exp\left[-g_j(y_j) - \frac{\sum_{i=1}^n \sigma_{ij}}{2\eta} \left\| y_j - \frac{\sum_{i=1}^n \sigma_{ij} \mathbf{X}_i^k}{\sum_{i=1}^n \sigma_{ij}} \right\|_2^2\right], \quad (9)$$

so we can employ a similar framework in Section 3 to analyze this general case, which also holds for Step 2. More specifically, in Section 3, we construct a diffusion process whose initial distribution is the marginal of $\pi^{\mathbf{XY}}$, while in this section, the initial distribution is the pushforward measure of marginals.

Theorem 10 Denote the probability distribution of $(\mathbf{X}^k, \mathbf{Y}^k)$ by $\mu_k^{\mathbf{XY}}$. Then under Assumption 9, for the target distribution (1), we have

$$D_{\text{KL}}(\mu_k^{\mathbf{XY}} \parallel \pi^{\mathbf{XY}}) \leq \exp(-kC) D_{\text{KL}}(\mu_0^{\mathbf{X}} \parallel \pi^{\mathbf{X}})$$

with C being defined as

$$\begin{aligned} C := & \frac{\eta}{n} \int_0^1 \left[\frac{\eta t(1-t)}{\min_j \sum_{i=1}^n \sigma_{ij}} + \frac{mn(1-t)^2}{\min_i \alpha_{f_i}} + \frac{t^2}{\min_j \alpha_{g_j}} \right]^{-1} dt \\ & + \frac{\eta}{m} \int_0^1 \left[\frac{\eta t(1-t)}{\min_i \sum_{j=1}^m \sigma_{ij}} + \frac{mn(1-t)^2}{\min_j \alpha_{g_j}} + \frac{t^2}{\min_i \alpha_{f_i}} \right]^{-1} dt. \end{aligned} \quad (10)$$

Proof The proof of Theorem 10 is essentially a generalization of the proof of the two-node case in Section 3. Inspired by (9), we compare $D_{\text{KL}}(\mu_k^{A\mathbf{X}} \parallel \pi^{A\mathbf{X}})$ and $D_{\text{KL}}(\mu_k^{\mathbf{X}} \parallel \pi^{\mathbf{X}})$ where the j -th entry of $A\mathbf{X} = (\sum_{i=1}^n \sigma_{ij} \mathbf{X}_i^k) / (\sum_{i=1}^n \sigma_{ij})$. According to (20) in Lemma 15, we have

$$D_{\text{KL}}(\mu_k^{A\mathbf{X}} \parallel \pi^{A\mathbf{X}}) \leq D_{\text{KL}}(\mu_k^{\mathbf{X}} \parallel \pi^{\mathbf{X}}). \quad (11)$$

Then by Lemma 4 and (9), the transition occurring in Step 1 can be described by the following diffusion processes,

$$dZ_t^j = \eta_j \nabla \log \phi_t^j(Z_t^j) dt + \sqrt{\eta_j} dW_t, \quad Z_0^j = \frac{\sum_{i=1}^n \sigma_{ij} \mathbf{X}_i^k}{\sum_{i=1}^n \sigma_{ij}}, \quad \forall j = 1, 2, \dots, m.$$

Here $\eta_j = \eta / \sum_{i=1}^n \sigma_{ij}$. Let $\mathbf{Z}_t = (Z_t^1, \dots, Z_t^m)$. Denote the distribution of \mathbf{Z}_t as π_t if $\mathbf{X}^k \sim \pi^{\mathbf{X}}$. Clearly, $\pi_0 \sim \pi^{A\mathbf{X}}$ and $\pi_1 \sim \pi^{\mathbf{Y}}$. It is noteworthy to mention that typically $A\mathbf{X}$ is not absolutely continuous w.r.t. Lebesgue measure, but π_t still has a density representation. By Lemma 4, the density of π_t satisfies

$$\begin{aligned} \pi_t &= \int \pi(\mathbf{Z}_t | \mathbf{X}, \mathbf{Y}) d\pi^{\mathbf{XY}} = \int \pi(\mathbf{Z}_t | A\mathbf{X}, \mathbf{Y}) d\pi^{A\mathbf{X}} \\ &\propto \int \exp \left[- \sum_{i=1}^n f_i(x_i) - \sum_{j=1}^m g_j(y_j) - \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - y_j\|_2^2 \right] \\ &\quad \exp \left[- \sum_{j=1}^m \frac{\sum_{i=1}^n \sigma_{ij}}{2t(1-t)\eta} \left\| z_t^j - ty_j - (1-t) \frac{\sum_{i=1}^n \sigma_{ij} x_i}{\sum_{i=1}^n \sigma_{ij}} \right\|_2^2 \right] d\mathbf{X} d\mathbf{Y}. \end{aligned}$$

By Proposition 16, we have π_t is strongly log-concave with a coefficient

$$c_t = \left[\frac{\eta t(1-t)}{\min_j \sum_{i=1}^n \sigma_{ij}} + \frac{mn(1-t)^2}{\min_i \alpha_{f_i}} + \frac{t^2}{\min_j \alpha_{g_j}} \right]^{-1}. \quad (12)$$

Observe that $\eta_j \geq \frac{\eta}{n}$, $\forall j$. Hence, by Theorem 3 and a lower bound of c_t in (12) we obtain

$$\begin{aligned} D_{\text{KL}}(\mu_k^{\mathbf{Y}} \| \pi^{\mathbf{Y}}) &\leq \exp \left(-\frac{\eta}{n} \int_0^1 c_t dt \right) D_{\text{KL}}(\mu_k^{A\mathbf{X}} \| \pi^{A\mathbf{X}}) \\ &\leq \exp \left(-\frac{\eta}{n} \int_0^1 c_t dt \right) D_{\text{KL}}(\mu_k^{\mathbf{X}} \| \pi^{\mathbf{X}}) \end{aligned}$$

where the last inequality is from (11). By a symmetry analysis, one can get

$$D_{\text{KL}}(\mu_{k+1}^{\mathbf{X}} \| \pi^{\mathbf{X}}) \leq \exp \left(-\frac{\eta}{m} \int_0^1 \left[\frac{\eta t(1-t)}{\min_i \sum_{j=1}^m \sigma_{ij}} + \frac{mn(1-t)^2}{\min_j \alpha_{g_j}} + \frac{t^2}{\min_i \alpha_{f_i}} \right]^{-1} dt \right) D_{\text{KL}}(\mu_k^{\mathbf{Y}} \| \pi^{\mathbf{Y}}).$$

Lastly, with (18) in Lemma 15,

$$D_{\text{KL}}(\mu_k^{\mathbf{XY}} \| \pi^{\mathbf{XY}}) \leq \exp(-kC) D_{\text{KL}}(\mu_0^{\mathbf{X}} \| \pi^{\mathbf{X}})$$

where the linear convergence rate C is defined in 10. ■

This convergence rate implies that the convergence depends on the worst strongly-convexity constant. For the impact of η , we have the same observation as in the two-node case, *i.e.*, larger η implies faster convergence speed. The convergence rate in Theorem 10 is not tight. For users who know the exact structure of the bipartite graph, the strong convexity constant c_t can be further improved and thereby the convergence rate $\exp(-C)$. The tight log-Sobolev constant under the strong convexity assumption is the convexity constant of \mathbf{Z} -marginal distribution in Proposition 16. A straightforward way to find the tight constant is to compute the Hessian matrix of the potential of joint density $\pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z})$. The lack of tightness is partially due to the matrix inequalities in the proof of Proposition 16.

5. Convergence analysis with small η for composite sampling

Inspired by the success of the proximal sampling framework, we postulate that if η is sufficient small, our target distribution (1) would converge to

$$\exp\left(-\sum_{i=1}^n f_i(x) - \sum_{j=1}^m g_j(x)\right).$$

Due to the difficulty of the non-asymptotic behavior of (1), in this section, we focus on the two-node case. In this case, with the additional Gaussian component, (2) is a better-conditioned distribution than $\exp(-f(x) - g(x))$. A similar observation was initially proposed by [Lee et al. \(2021\)](#) and induces efficient sampling algorithms for structured log-concave distributions. The section is also a direct generalization of the one-node case in [Vono et al. \(2022\)](#). For (2), we will show the marginal of our target density (2) actually converges to the composite density with a sufficiently small η , i.e.,

$$\lim_{\eta \rightarrow 0} \text{TV}(\pi_\eta^X, \nu) = 0$$

where π_η^X is the X -marginal distribution of (2) and $\nu \propto \exp[-f(x) - g(x)]$. Intuitively, if η is small, then the density of π_η will concentrate on the area, $x = y$. Therefore the marginal distribution of π_η would approximately be ν . To begin with, we prove the asymptotic property in the following Proposition 11.

Proposition 11 *Assume $\exp(-g) \in L_1(\mathbb{R}^d)$ and $\exp(-f)$ is essentially bounded from above, then for*

$$\pi_\eta^X \propto \int \exp\left[-f(x) - g(y) - \frac{1}{2\eta} \|x - y\|_2^2\right] dy$$

and

$$\nu \propto \exp[-f(x) - g(x)],$$

we have

$$\lim_{\eta \rightarrow 0} \text{TV}(\pi_\eta^X, \nu) = 0.$$

Proof Under the condition that $\exp(-g) \in L_1(\mathbb{R}^d)$, one can show the convolution $\exp(-g) * \mathcal{N}(0, \eta \mathbf{I})$ converges to $\exp(-g)$ almost everywhere (Theorem 1 in [Bear \(1979\)](#)), which yields

$$\lim_{\eta \rightarrow 0} \frac{1}{\sqrt{2\pi\eta}} \int \exp\left[-f(x) - g(y) - \frac{1}{2\eta} \|x - y\|_2^2\right] dy = \exp[-f(x) - g(x)] \quad a.e. \quad (13)$$

Meanwhile, as $\exp(-g) \in L_1(\mathbb{R}^d)$, by a standard argument (Theorem 9.1.11 in [Heil \(2019\)](#)), one has $\|\exp(-g) * \mathcal{N}(0, \eta \mathbf{I}) - \exp(-g)\|_1 \rightarrow 0$. As $\exp(-f)$ is essentially bounded from above, it follows that

$$\lim_{\eta \rightarrow 0} \frac{1}{\sqrt{2\pi\eta}} \int \exp\left[-f(x) - g(y) - \frac{1}{2\eta} \|x - y\|_2^2\right] dy dx = \int \exp[-f(x) - g(x)] dx. \quad (14)$$

With (13), (14) and Scheffè's lemma (Scheffé, 1947), we have

$$\lim_{\eta \rightarrow 0} \text{TV}(\pi_\eta^X, \nu) = 0.$$

■

However, the non-asymptotic behaviour of $\pi_\eta^X \rightarrow \nu$ is challenging. In this section, we summarize and analyze four special cases to promote the study of it. The idea of bounding the distance of π_η^X and ν is initially proposed in Lee et al. (2021), which we summarize in Example 1.

Example 1 Consider a joint distribution,

$$\pi_\eta^{XY} \propto \exp[-f(x) - g(y) - 1/2\eta\|x - y\|_2^2].$$

Assume $g(y) = 0$. Then $\pi_\eta^X \propto \exp[-f(x)]$. In this case, the X -marginal distribution is exactly ν .

To proceed, we analyze a symmetric case where $f(x) = 0$. This example resembles the setting of Vono et al. (2022), and we establish a better dimension dependency than Vono et al. (2022): from d to $d^{1/2}$.

Example 2 Consider a joint distribution,

$$\pi_\eta^{XY} \propto \exp[-f(x) - g(y) - 1/2\eta\|x - y\|_2^2]$$

where $x, y \in \mathbb{R}^d$. Assume $f(x) = 0$ and g is α -strongly convex and β -smooth. Then

$$D_{\text{KL}}(\pi_\eta^X \parallel \nu) \leq \frac{d\alpha}{2\beta} \left[\exp\left(\frac{2\beta^2\eta}{\alpha}\right) - 1 \right].$$

Moreover, with Pinsker's inequality, we obtain

$$\text{TV}(\pi_\eta^X, \nu) \leq \sqrt{\frac{d\alpha}{4\beta} \left[\exp\left(\frac{2\beta^2\eta}{\alpha}\right) - 1 \right]}.$$

Proof Appendix D.1. ■

Even though we already find the convergence rate for the particular cases with only one node, *i.e.*, Example 1 and 2, the analysis for the multi-variable case is demanding. It is worth noting that instead of using the potential $f(x) + g(y) + \frac{1}{2\eta}\|x - y\|_2^2$, Lee et al. (2021) constructs a new potential with an additional quadratic term, $f(x) + g(y) + \frac{1}{2\eta}\|x - y\|_2^2 + \frac{\eta\alpha_f^2}{2}\|x - x^*\|_2^2$, where x^* is the minimizer of $f + g$. Our analysis is different from this previous work in two aspects. First, our target density does not include the L_2 distance of x and x^* . Secondly, we find that without the additional term, it is impossible to uniformly bound the total variation distance for strongly convex and smooth potentials by a single η . Thus, for the non-asymptotic bound, we illustrate a necessary condition by Example 3.

Example 3 Consider a joint distribution,

$$\pi_\eta^{XY} \propto \exp[-f(x) - g(y) - 1/2\eta(x - y)^2].$$

Let $f(x) = (x - u_1)^2$ and $g(x) = (x - u_2)^2$. Then

$$\begin{aligned}\pi_\eta^X &= \mathcal{N}\left(\frac{u_2 + (2\eta + 1)u_1}{2\eta + 2}, \frac{2\eta + 1}{4\eta + 4}\right), \\ \nu &= \mathcal{N}\left(\frac{u_1 + u_2}{2}, \frac{1}{4}\right).\end{aligned}$$

By Theorem 1.3 in [Devroye et al. \(2018\)](#), we obtain

$$\text{TV}(\pi_\eta^X, \nu) \geq \min\left\{\frac{\eta|u_1 - u_2|}{5\eta + 5}, 1/200\right\}.$$

This means the convergence rate relies on $|u_1 - u_2|$. Note that even with other metrics, the term $|u_1 - u_2|$ would affect the non-asymptotic behavior. For instance,

$$D_{\text{KL}}(\nu || \pi_\eta^X) = \frac{1}{2} \log \frac{2\eta + 1}{\eta + 1} - \frac{\eta}{4\eta + 2} + \frac{\eta^2(u_1 - u_2)^2}{(2\eta + 1)(2\eta + 2)},$$

and

$$W_2^2(\pi_\eta^X, \nu) = \left(\sqrt{\frac{2\eta + 1}{4\eta + 4}} - \frac{1}{2}\right)^2 + \frac{\eta^2(u_1 - u_2)^2}{(2\eta + 2)^2}.$$

Hence, to bound the distance of π_η^X and ν , we make Assumption 12.

Assumption 12 Assume f and g share the same minimizer, i.e., $\nabla f(x^*) = \nabla g(x^*) = 0$.

Remark 13 The condition that f and g share the same minimizer can be achieved by shifting f and g by linear terms ([Lee et al., 2021](#)), i.e., $\bar{f}(x) = f(x) - \langle \nabla f(x^*), x \rangle$ and $\bar{g}(x) = g(x) + \langle \nabla g(x^*), x \rangle$ where x^* is the stationary point of $f + g$. Notice that $\exp[-f(x) - g(x)] = \exp[-\bar{f}(x) - \bar{g}(x)]$ and $\nabla \bar{f}(x^*) = \nabla \bar{g}(x^*)$.

Proposition 14 Consider a joint distribution,

$$\pi_\eta^{XY} \propto \exp[-f(x) - g(y) - 1/2\eta\|x - y\|_2^2].$$

Assume $g(y) = \frac{1}{2\sigma^2}\|y - u\|_2^2$ and $f(x)$ is α_f -strongly convex. Then with Assumption 12, we have

$$D_{\text{KL}}(\nu || \pi_\eta^X) \leq \frac{d\eta^2}{2(\alpha_f\sigma^4 + \sigma^2)^2}.$$

Moreover, by Pinsker's inequality, we have

$$\text{TV}(\pi_\eta^X, \nu) \leq \frac{\eta\sqrt{d}}{2(\alpha_f\sigma^4 + \sigma^2)}.$$

Proof We build a continuous path to connect π_η^X and ν and then establish an inequality for the first-order and second-order time derivatives of $D_{\text{KL}}(\pi_0 || \pi_t)$. See Appendix D.2 for the full proof. \blacksquare

6. Discussion

In this work, we establish the first non-asymptotic bound of a Gibbs sampler for structured log-concave distributions over bipartite graphs. For the two-node graph, we further prove its asymptotic property to a lower-dimensional composite distribution and show the convergence rate for special cases. Here are some directions for future investigation.

- How to compute the non-asymptotic rate for (1) for non-convex potentials? It is possible to generalize to the LSI assumption with the same techniques in this work. However, this requires a tight LSI constant as discussed in Remark 7.
- How to analyze the non-asymptotic bound of Gibbs sampling on general graphs? The analysis for the bipartite graphs is a natural generalization of the two-node case due to (9). Whether the same techniques can be adopted for general graphs is not clear.
- How to find the convergence rate of general distributions in the form of (2) as η approaches 0? Solving this question will lead to the non-asymptotical bounds for much more general distributions.

7. Acknowledgments

We extend our gratitude to the anonymous reviewers for their invaluable feedback that enhanced this manuscript. Financial support from NSF under grants 1942523, 2008513, and 2206576 is greatly acknowledged.

References

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *arXiv preprint arXiv:2302.10249*, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Herbert Bear. Approximate identities and pointwise convergence. *Pacific Journal of Mathematics*, 81(1):17–27, 1979.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrodinger bridge. *SIAM Review*, 63(2):249–313, 2021.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

Sinho Chewi, Patrik Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. *arXiv preprint arXiv:2105.14163*, 2021.

Sinho Chewi, Patrik R Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. In *Conference on Learning Theory*, pages 2041–2059. PMLR, 2022.

Sinho Chewi, Patrik Gerber, Holden Lee, and Chen Lu. Fisher information lower bounds for sampling. In *International Conference on Algorithmic Learning Theory*, pages 375–410. PMLR, 2023a.

Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. *arXiv preprint arXiv:2304.02599*, 2023b.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

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, 2017.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

Koller Daphne, Friedman Nir, and Bach Francis. Probabilistic graphical models: principles and techniques. *Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA*, 2009.

Christopher De Sa, Vincent Chen, and Wing Wong. Minibatch Gibbs sampling on large graphical models. In *International Conference on Machine Learning*, pages 1173–1181, 2018.

Christopher M De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Rapidly mixing Gibbs sampling for a class of factor graphs using hierarchy width. *Advances in neural information processing systems*, 28, 2015.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. *arXiv preprint arXiv:2302.10081*, 2023.

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

Christopher Heil. *Introduction to real analysis*, volume 280. Springer, 2019.

Tomas Hrycej. Gibbs sampling in Bayesian networks. *Artificial Intelligence*, 46(3):351–363, 1990.

Benton Jamison. Reciprocal processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 30(1):65–86, 1974.

Yin Tat Lee, Ruqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

Christian Léonard. A survey of the schrodinger problem and some of its connections with optimal transport. *DYNAMICAL SYSTEMS*, 34(4):1533–1574, 2014.

DA Levin and Y Peres. Markov chains and mixing times, volume 107 (American Mathematical Soc.). 2017.

Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt (d) dimension dependence of langevin monte carlo. *arXiv preprint arXiv:2109.03839*, 2021.

Michele Pavon, Giulio Trigila, and Esteban G Tabak. The Data-Driven Schrödinger Bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573, 2021.

Lewis J Rendell, Adam M Johansen, Anthony Lee, and Nick Whiteley. Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, 30(2):249–259, 2020.

Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

Henry Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947.

Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019.

Francisco Vargas. Machine-learning approaches for the empirical Schrödinger bridge problem. Technical report, University of Cambridge, Computer Laboratory, 2021.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Deepak Venugopal and Vibhav Gogate. On lifting the Gibbs sampling algorithm. *Advances in Neural Information Processing Systems*, 25, 2012.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Split-and-augmented Gibbs sampler—Application to large-scale inference problems. *IEEE Transactions on Signal Processing*, 67(6):1648–1661, 2019.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Asymptotically exact data augmentation: Models, properties, and algorithms. *Journal of Computational and Graphical Statistics*, 30(2):335–348, 2020.

Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting. *Journal of Machine Learning Research*, 23:1–69, 2022.

Matthew Zhang, Sinho Chewi, Mufan Bill Li, Krishnakumar Balasubramanian, and Murat A Erdogdu. Improved discretization analysis for underdamped Langevin Monte Carlo. *arXiv preprint arXiv:2302.08049*, 2023.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022.

Ruqi Zhang and Christopher M De Sa. Poisson-minibatching for Gibbs sampling with convergence rate guarantees. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix A. Supplementary lemmas

A.1. Proof of Lemma 2

Proof We start with the time derivative of $D_{\text{KL}}(\mu_t \parallel \pi_t)$. With the Fokker–Planck equation, we have

$$\begin{aligned} \frac{\partial D_{\text{KL}}(\mu_t \parallel \pi_t)}{\partial t} &= \int \partial_t \mu_t \log \frac{\mu_t}{\pi_t} + \partial_t \left(\frac{\mu_t}{\pi_t} \right) \pi_t \\ &= \int \partial_t \mu_t \log \frac{\mu_t}{\pi_t} + \int \partial_t \mu_t - \int \partial_t \pi_t \frac{\mu_t}{\pi_t} \\ &= \int \left[-\nabla \cdot (b_t \mu_t) + \sum_i \frac{\sigma_{t,i}^2}{2} \partial_i^2 \mu_t \right] \log \frac{\mu_t}{\pi_t} \\ &\quad - \int \left[-\nabla \cdot (b_t \pi_t) + \sum_i \frac{\sigma_{t,i}^2}{2} \partial_i^2 \pi_t \right] \frac{\mu_t}{\pi_t}, \end{aligned} \tag{15}$$

where the symbol ∂_i^2 is the second-order derivative w.r.t. the i -th coordinate. By integration by parts, we obtain

$$\begin{aligned} \int \nabla \cdot (b_t \mu_t) \log \frac{\mu_t}{\pi_t} &= - \int \mu_t b_t \cdot \nabla \log \frac{\mu_t}{\pi_t} \\ &= - \int \pi_t b_t \cdot \nabla \frac{\mu_t}{\pi_t} \\ &= \int \nabla \cdot (b_t \pi_t) \frac{\mu_t}{\pi_t}. \end{aligned} \tag{16}$$

Combining (15) and (16) yields

$$\begin{aligned} \frac{\partial D_{\text{KL}}(\mu_t \parallel \pi_t)}{\partial t} &= \sum_i \frac{\sigma_{t,i}^2}{2} \int \partial_i^2 \mu_t \log \frac{\mu_t}{\pi_t} - \partial_i^2 \pi_t \frac{\mu_t}{\pi_t} \\ &= - \sum_i \frac{\sigma_{t,i}^2}{2} \int \partial_i \mu_t \left(\partial_i \log \frac{\mu_t}{\pi_t} \right) - \partial_i \pi_t \left(\partial_i \frac{\mu_t}{\pi_t} \right) \\ &= - \sum_i \frac{\sigma_{t,i}^2}{2} \int \left\| \partial_i \log \frac{\mu_t}{\pi_t} \right\|_2^2 \mu_t \\ &\leq - \frac{\lambda_t^2}{2} I(\mu_t \parallel \pi_t) \end{aligned} \tag{17}$$

where $\lambda_t^2 \leq \min_i \sigma_{t,i}^2$. ■

A.2. Proof of Lemma 4

Proof We show $P(z(t)) \propto \phi_t(z(t)) \hat{\phi}_t(z(t))$ first. By definition of Brownian bridge, the conditional distribution of $z(t)$ on $z(0), z(1)$ is

$$P(z(t) | z(0), z(1)) \propto \exp\left(-\frac{1}{2\eta t(1-t)} \|z(t) - (1-t)z(0) - tz(1)\|^2\right).$$

It follows that

$$\begin{aligned} P(z(t)) &\propto \int \exp(-f(x) - g(y) - \frac{1}{2\eta} \|x - y\|^2 - \frac{1}{2\eta t(1-t)} \|z(t) - (1-t)x - ty\|^2) dx dy \\ &\propto \phi_t(z(t)) \hat{\phi}_t(z(t)). \end{aligned}$$

Now consider the forward process (4). It can be shown (Tzen and Raginsky, 2019; Zhang and Chen, 2022) that the conditional distribution of Z_1 on Z_0 is

$$P(Z_1|Z_0) \propto \exp(-g(Z_1) - \frac{1}{2\eta} \|Z_1 - Z_0\|^2),$$

and (4) belongs to the same reciprocal class (Jamison, 1974) as the Brownian bridge, that is, $P(Z_1|Z_0, Z_1)$ coincides with that of the Brownian bridge $P(z(\cdot)|z(0), z(1))$. The drift ϕ_t is known to be the Follmer drift (Jamison, 1974). Taking into account the initial condition $Z_0 \sim P(z(0))$, we conclude that the diffusion process (4) is a representation of the trajectory distribution we want.

By Doob's h -transform (Anderson, 1982; Chen et al., 2016), (4) has an equivalent backward representation

$$dZ_t = [\eta \nabla \log \phi_t(Z_t) - \eta \nabla \log P(Z_t)] dt + \sqrt{\eta} dW_t.$$

Invoking the fact that $P(z(t)) \propto \phi_t(z(t)) \hat{\phi}_t(z(t))$, we arrive at the backward representation (5). \blacksquare

A.3. KL divergence contraction

Lemma 15 *Let π^{XY} and μ^{XY} be any two joint distributions of (x, y) . Then we have*

$$D_{\text{KL}}(\mu^{XY} || \pi^{XY}) = D_{\text{KL}}(\mu^X || \pi^X) + \mathbb{E}_{\mu^X}(D_{\text{KL}}(\mu^{Y|X} || \pi^{Y|X})) \quad (18)$$

Note that if we further assume $\mu^{Y|X} = \pi^{Y|X}$, then

$$D_{\text{KL}}(\mu^{XY} || \pi^{XY}) = D_{\text{KL}}(\mu^X || \pi^X) \quad (19)$$

A more general form is $A \in \mathbb{R}^{m \times n}$ and $X \in \mathbb{R}^n$. For any two distributions μ^X and π^X , and their corresponding pushforward measures μ^{AX} and π^{AX} , we have

$$D_{\text{KL}}(\mu^{AX} || \pi^{AX}) \leq D_{\text{KL}}(\mu^X || \pi^X) \quad (20)$$

Proof

$$\begin{aligned} D_{\text{KL}}(\mu^{XY} || \pi^{XY}) &= \int \log \frac{\mu^{XY}}{\pi^{XY}} \mu^{XY} \\ &= \int \log \left(\frac{\mu^X}{\pi^X} \frac{\mu^{Y|X}}{\pi^{Y|X}} \right) \mu^X \mu^{Y|X} \\ &= \int \log \frac{\mu^X}{\pi^X} \mu^X \mu^{Y|X} + \int \log \frac{\mu^{Y|X}}{\pi^{Y|X}} \mu^X \mu^{Y|X} \\ &= \int \log \frac{\mu^X}{\pi^X} \mu^X + \int D_{\text{KL}}(\mu^{Y|X} || \pi^{Y|X}) \mu^X \\ &= D_{\text{KL}}(\mu^X || \pi^X) + \mathbb{E}_{\mu^X}(D_{\text{KL}}(\mu^{Y|X} || \pi^{Y|X})). \end{aligned}$$

We note that the decomposition above still holds if we treat $\frac{\mu^{XY}}{\pi^{XY}}$ as the Radon-Nikodym derivative instead of the ratio of density functions. See Lemma 1 in [Vargas \(2021\)](#) for details.

Next we prove (20). For the more general case, initially, we assume A is full row rank, which implies $m \leq n$. Then let \hat{A} be an $n \times n$ invertible matrix whose first m rows exactly comprise A . It follows that

$$\begin{aligned} D_{\text{KL}}(\mu^{\hat{A}X} \parallel \pi^{\hat{A}X}) &= \int \log \frac{\mu^{\hat{A}X}}{\pi^{\hat{A}X}} \mu^{\hat{A}X} \\ &= \int \log \frac{\mu^X(\hat{A}^{-1}x)}{\pi^X(\hat{A}^{-1}x)} |\det(\hat{A}^{-1})| \mu^X(\hat{A}^{-1}x) \\ &= \int \log \frac{\mu^X(y)}{\pi^X(y)} \mu^X(y) \\ &= D_{\text{KL}}(\mu^X \parallel \pi^X). \end{aligned}$$

According to Equation (18), since AX is just the first m components of $\hat{A}X$, we have

$$D_{\text{KL}}(\mu^{AX} \parallel \pi^{AX}) \leq D_{\text{KL}}(\mu^{\hat{A}X} \parallel \pi^{\hat{A}X}) = D_{\text{KL}}(\mu^X \parallel \pi^X).$$

Next consider the case where A is not full row rank. One can show, even in this case, AX is still a well-defined random vector, but μ^{AX} and π^{AX} are not absolutely continuous with respect to the Lebesgue measure. Define the rank of A by r . One can always pick up r independent rows from A which we denote by \bar{A} . Notice that the probability measures over $\bar{A}X$ is essentially the same as the measure over AX , as all other coordinates are uniquely determined by $\bar{A}X$. Hence, we have

$$D_{\text{KL}}(\mu^{AX} \parallel \pi^{AX}) = D_{\text{KL}}(\mu^{\bar{A}X} \parallel \pi^{\bar{A}X}).$$

Because \bar{A} is full row rank, according to the previous case, we eventually have

$$D_{\text{KL}}(\mu^{AX} \parallel \pi^{AX}) \leq D_{\text{KL}}(\mu^X \parallel \pi^X).$$

■

Appendix B. Strong-concavity constant

Proposition 16 *With Assumption 9, define*

$$\begin{aligned} \pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z}) &\propto \exp \left[- \sum_{i=1}^n f_i(x_i) - \sum_{j=1}^m g_j(y_j) - \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - y_j\|_2^2 \right] \\ &\quad \exp \left[- \sum_{j=1}^m \frac{\sum_{i=1}^n \sigma_{ij}}{2t(1-t)\eta} \|z_t^j - ty_j - (1-t) \frac{\sum_{i=1}^n \sigma_{ij} x_i}{\sum_{i=1}^n \sigma_{ij}}\|_2^2 \right]. \end{aligned} \tag{21}$$

Then its \mathbf{z} -marginal distribution, $\pi_t(\mathbf{z}) \propto \int \pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{x} d\mathbf{y}$, is strongly log-concave with a coefficient $c_t = \left[\frac{\eta t(1-t)}{\min_j \sum_{i=1}^n \sigma_{ij}} + \frac{mn(1-t)^2}{\min_i \alpha_{f_i}} + \frac{t^2}{\min_j \alpha_{g_j}} \right]^{-1}$.

Proof We will begin with proving the joint distribution $\pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is strongly concave. As each f_i and each g_j are strongly convex with coefficients α_{f_i} and α_{g_j} , respectively, we can rewrite $f_i(x_i) = \hat{f}_i(x_i) + \frac{\alpha_{f_i}}{2} \|x_i\|_2^2$ and $g_j(y_j) = \hat{g}_j(y_j) + \frac{\alpha_{g_j}}{2} \|y_j\|_2^2$ where \hat{f}_i and \hat{g}_j are convex functions. Then we can decompose $\pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z})$ as $-\log \pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z}) = F(\mathbf{x}, \mathbf{y}, \mathbf{z}) + G(\mathbf{x}, \mathbf{y}, \mathbf{z})$ where

$$F(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i=1}^n \hat{f}_i(x_i) + \sum_{j=1}^m \hat{g}_j(y_j) + \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_{ij}}{2\eta} \|x_i - y_j\|_2^2 \quad (22)$$

and

$$G(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_i \frac{\alpha_{f_i}}{2} \|x_i\|_2^2 + \sum_j \frac{\alpha_{g_j}}{2} \|y_j\|_2^2 + \sum_{j=1}^m \frac{\sum_{i=1}^n \sigma_{ij}}{2t(1-t)\eta} \left\| z_t^j - ty_j - (1-t) \frac{\sum_{i=1}^n \sigma_{ij} x_i}{\sum_{i=1}^n \sigma_{ij}} \right\|_2^2. \quad (23)$$

By definition, $F(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is convex jointly in $\mathbf{x}, \mathbf{y}, \mathbf{z}$ while $G(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is strongly-convex jointly in $\mathbf{x}, \mathbf{y}, \mathbf{z}$. Hence $\pi_t(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is strongly log-concave. By theorem 3.8 in (Saumard and Wellner, 2014), the \mathbf{z} -marginal distribution is also strongly log-concave. Then we will compute the strongly-convexity constant for $\pi_t(\mathbf{z})$. Define

$$M = \left(\begin{array}{c|c} \mathbf{I}_{m \times m} & B \\ \mathbf{0}_{(m+n) \times m} & \mathbf{I}_{(m+n) \times (m+n)} \end{array} \right)$$

where

$$B = (-(1-t)A^T \mid -t\mathbf{I}_{m \times m}).$$

and the entry of A at the i -th row and j -th column is $\sigma_{ij}/\sum_{i=1}^n \sigma_{ij}$. Furthermore, define a diagonal matrix Λ as

$$\Lambda = \begin{pmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \Lambda_3 \end{pmatrix}$$

with

$$\begin{aligned} \Lambda_1 &= \frac{1}{t(1-t)\eta} \text{diag} \left(\sum_{i=1}^n \sigma_{i1}, \dots, \sum_{i=1}^n \sigma_{im} \right), \\ \Lambda_2 &= \text{diag}(\alpha_{f_1}, \dots, \alpha_{f_n}), \\ \Lambda_3 &= \text{diag}(\alpha_{g_1}, \dots, \alpha_{g_m}). \end{aligned}$$

Then clearly,

$$G(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{2} (\mathbf{z}, \mathbf{x}, \mathbf{y})^T M^T \Lambda M (\mathbf{z}, \mathbf{x}, \mathbf{y})^T.$$

By Theorem 3.8 in (Saumard and Wellner, 2014), it is sufficient to compute the smallest eigenvalue of the inverse of the upper left block of $(M^T \Lambda M)^{-1}$. The inverse of the blocked matrix can be given by Woodbury matrix identity. Combining the two results, we only need to find $c_t > 0$ such that

$$c_t \mathbf{I} \preceq (\Lambda_1^{-1} + (1-t)^2 A^T \Lambda_2^{-1} A + t^2 \Lambda_3^{-1})^{-1}.$$

Notice that the largest eigenvalue of $A^T \Lambda_2^{-1} A$ is the squared operator norm of $\sqrt{\Lambda_2^{-1}} A$, which satisfies

$$\left\| \sqrt{\Lambda_2^{-1}} A \right\|_{\text{op}}^2 \leq mn \left\| \sqrt{\Lambda_2^{-1}} A \right\|_{\max}^2 \leq \frac{mn}{\min_i \alpha_{f_i}}.$$

Therefore,

$$c_t = \left[\frac{\eta t(1-t)}{\min_j \sum_{i=1}^n \sigma_{ij}} + \frac{mn(1-t)^2}{\min_i \alpha_{f_i}} + \frac{t^2}{\min_j \alpha_{g_j}} \right]^{-1}.$$

■

Appendix C. Convergence rate for the two-node case under the convexity assumption

To begin with, we introduce the 2-Wasserstein distance. The 2-Wasserstein distance for two distributions μ and ν with finite second-order moments is defined as follows,

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{(x,y) \sim \gamma} \|x - y\|_2^2 d\gamma(x, y)$$

where $\Gamma(\mu, \nu)$ is the set of all couplings for μ and ν .

Proposition 17 (Data processing inequality with log-concavity) *Under the condition of Lemma 2, if we further assume π_t is log-concave and $W_2^2(\mu_t, \pi_t) \leq C$, $\forall t \in [0, T]$, then*

$$\frac{1}{D_{\text{KL}}(\mu_T \parallel \pi_T)} \geq \frac{1}{D_{\text{KL}}(\mu_0 \parallel \pi_0)} + \frac{\int_0^T \lambda_t^2 dt}{2C}.$$

Proof Because of the log-concavity of π_t , one can show the convexity of $D_{\text{KL}}(\cdot \parallel \pi_t)$ along Wasserstein geodesics (see Theorem 9.4.11 in [Ambrosio et al. \(2005\)](#)), which follows that

$$0 = D_{\text{KL}}(\pi_t \parallel \pi_t) \geq D_{\text{KL}}(\mu_t \parallel \pi_t) + \mathbb{E}_{x_t, y_t \sim \pi^*(\mu_t, \pi_t)} \langle \nabla \log \frac{\mu_t}{\pi_t}(x_t), x_t - y_t \rangle.$$

where $\pi^*(\mu_t, \pi_t)$ is the optimal coupling in the sense of 2-Wasserstein distance. Hence,

$$D_{\text{KL}}(\mu_t \parallel \pi_t)^2 \leq \mathbb{E}_{\mu_t} \|\nabla \log \frac{\mu_t}{\pi_t}\|_2^2 W_2^2(\mu_t, \pi_t) = I(\mu_t \parallel \pi_t) W_2^2(\mu_t, \pi_t).$$

Combining with (17) yields

$$\frac{\partial D_{\text{KL}}(\mu_t \parallel \pi_t)}{\partial t} \leq -\frac{\lambda_t^2}{2} \frac{D_{\text{KL}}(\mu_t \parallel \pi_t)^2}{W_2^2(\mu_t, \pi_t)} \leq -\frac{\lambda_t^2 D_{\text{KL}}(\mu_t \parallel \pi_t)^2}{2C},$$

which gives

$$\frac{1}{D_{\text{KL}}(\mu_T \parallel \pi_T)} \geq \frac{1}{D_{\text{KL}}(\mu_0 \parallel \pi_0)} + \frac{\int_0^T \lambda_t^2 dt}{2C}.$$

■

We are ready to show the proof of Theorem 6.

Proof We construct the diffusion process for both (π^X, π^Y) and (π_k^X, π_k^Y) as in the proof of 5,

$$dZ_t = -\eta \nabla \log \phi_t(Z_t) dt + \sqrt{\eta} dW_t. \quad (19)$$

Then we have the explicit expression of π_t , the distribution of Z_t given $\pi_0 = \pi^X$, satisfies,

$$\pi_t \propto \int \exp[-g(y) - \frac{1}{2\eta(1-t)} \|z - y\|_2^2] dy \int \exp[-f(x) - \frac{1}{2\eta t} \|z - x\|_2^2] dx.$$

Notice that the drift term $-\eta \nabla \log \phi_t$ in (19) is convex. Thus, with W_2^2 -contraction, we have

$$W_2^2(\mu_t, \pi_t) \leq W_2^2(\mu_0, \pi_0) = W_2^2(\mu_0^X, \pi^X).$$

As π_t is convex, which is due to the convexity assumption for f and g (see Theorem 3.3 in [Saumard and Wellner \(2014\)](#)), we can apply Theorem 17 with $C \equiv W_2^2(\mu_0^X, \pi^X)$ on (19), which yields

$$\frac{1}{D_{\text{KL}}(\mu_k^Y || \pi^Y)} \geq \frac{1}{D_{\text{KL}}(\mu_k^X || \pi^X)} + \frac{\eta}{2W_2^2(\mu_0^X, \pi^X)}. \quad (20)$$

We can analyze Step 2 in Algorithm 1 with the same framework, which gives

$$\frac{1}{D_{\text{KL}}(\mu_{k+1}^X || \pi^X)} \geq \frac{1}{D_{\text{KL}}(\mu_k^Y || \pi^Y)} + \frac{\eta}{2W_2^2(\mu_0^X, \pi^X)}. \quad (21)$$

With (20), (21) and Lemma 15, one has for any k

$$\frac{1}{D_{\text{KL}}(\mu_k^{XY} || \pi^{XY})} \geq \frac{1}{D_{\text{KL}}(\mu_0^X || \pi^X)} + \frac{k\eta}{W_2^2(\mu_0^X, \pi^X)}.$$

It follows that

$$D_{\text{KL}}(\mu_k^{XY} || \pi^{XY}) \leq \frac{W_2^2(\mu_0^X, \pi^X)}{k\eta}.$$

■

Appendix D. Non-asymptotic convergence rate to composite distributions

D.1. Proof of Example 2

Proof Note that with the condition $f = 0$, $\pi_\eta^X = \nu * \mathcal{N}(0, \eta\mathbf{I})$. Thus, $\forall t \in (0, \eta]$, we define $\pi_t = \pi_0 * \mathcal{N}(0, t\mathbf{I})$. Then it is sufficient to bound the distance between π_t and π_0 . By definition, $\frac{\partial \pi_t}{\partial t} = \frac{1}{2} \Delta \pi_t$. It follows that

$$\begin{aligned} \frac{\partial D_{\text{KL}}(\pi_t || \pi_0)}{\partial t} &= \int \frac{\partial \pi_t}{\partial t} + \log \frac{\pi_t}{\pi_0} \frac{\partial \pi_t}{\partial t} \\ &= -\frac{1}{2} \int \pi_t \langle \nabla \log \pi_t, \nabla \log \frac{\pi_t}{\pi_0} \rangle \\ &= -\frac{1}{2} \int \pi_t \|\nabla \log \frac{\pi_t}{\pi_0}\|_2^2 - \frac{1}{2} \int \pi_t \langle \nabla \log \pi_0, \nabla \log \frac{\pi_t}{\pi_0} \rangle \\ &= -\frac{1}{2} \int \pi_t \|\nabla \log \frac{\pi_t}{\pi_0}\|_2^2 + \frac{1}{2} \int \pi_t \|\nabla \log \pi_0\|^2 - \frac{1}{2} \int \langle \nabla \log \pi_0, \nabla \pi_t \rangle \\ &\leq 0 + \frac{1}{2} \mathbb{E}_{\pi_t} (\|\nabla \log \pi_0\|_2^2 + \Delta \log \pi_0) \end{aligned}$$

With the assumption that g is α -strongly convex and β -smooth, basing on Lemma 12 in [Vempala and Wibisono \(2019\)](#), we could obtain

$$\begin{aligned}\frac{\partial D_{\text{KL}}(\pi_t \parallel \pi_0)}{\partial t} &\leq \frac{1}{2} \mathbb{E}_{\pi_t} (\|\nabla \log \pi_0\|^2 + \Delta \log \pi_0) \\ &\leq \frac{2\beta^2}{\alpha} D_{\text{KL}}(\pi_t \parallel \pi_0) + d\beta.\end{aligned}$$

Solving this differential inequality with the boundary condition $D_{\text{KL}}(\pi_0 \parallel \pi_0) = 0$ yields

$$D_{\text{KL}}(\pi_t \parallel \pi_0) \leq \frac{d\alpha}{2\beta} \left[\exp\left(\frac{2\beta^2 t}{\alpha}\right) - 1 \right].$$

By Pinsker's inequality, we finally have

$$\text{TV}(\pi_0, \pi_t) \leq \sqrt{\frac{d\alpha}{4\beta} \left[\exp\left(\frac{2\beta^2 t}{\alpha}\right) - 1 \right]}.$$

■

D.2. Proof of Proposition 14

Proof For ease of notation, in the proof, for any function h_t , we denote its first-order and second-order time derivative by \dot{h}_t and \ddot{h}_t , respectively.

For $t \in (0, \eta]$, let

$$\begin{aligned}\pi_t &= \exp[-f(x) - g_t(x) - \phi_t], \\ \pi_0 &= \lim_{t \rightarrow 0} \pi_t = \exp[-f(x) - g(x) - \phi_0]\end{aligned}$$

where

$$\begin{aligned}g_t(x) &= -\log \int \exp\left[-g(y) - \frac{1}{2t}\|x - y\|^2\right] dy, \\ \phi_t &= \log \int \exp[-f(x) - g_t(x)] dx, \\ \phi_0 &= \log \int \exp[-f(x) - g(x)] dx.\end{aligned}$$

It follows that the time derivatives of their terms are

$$\begin{aligned}\dot{\phi}_t &= -\mathbb{E}_{\pi_t}(\dot{g}_t) \\ \dot{\pi}_t &= -\pi_t \left[\dot{g}_t + \dot{\phi}_t \right]\end{aligned}\tag{22}$$

And

$$\dot{g}_t = -\frac{1}{2t^2} \mathbb{E}_{r_t(\cdot|x)} (\|x - y\|^2) \tag{23}$$

with

$$r_t(\cdot|x) = \exp\left[-g(y) - \frac{1}{2t}\|x - y\|^2 + g_t(x)\right]. \tag{24}$$

Combining with (22) and (23), we have

$$\begin{aligned}
 K(t) &:= \frac{\partial D_{\text{KL}}(\pi_0 \parallel \pi_t)}{\partial t} = \int \pi_0(\dot{g}_t + \dot{\phi}_t) \\
 &= \mathbb{E}_{\pi_0}(\dot{g}_t) - \mathbb{E}_{\pi_t}(\dot{g}_t) \\
 &= \frac{1}{2t^2} \mathbb{E}_{\pi_t} [\mathbb{E}_{r_t(\cdot|x)}(\|x - y\|^2)] - \frac{1}{2t^2} \mathbb{E}_{\pi_0} [\mathbb{E}_{r_t(\cdot|x)}(\|x - y\|^2)].
 \end{aligned} \tag{25}$$

To perform a more detailed analysis, we consider the second-order time derivative of $D_{\text{KL}}(\pi_0 \parallel \pi_t)$, which satisfies

$$\dot{K}(t) = \int \pi_0 \ddot{g}_t - \int \pi_t \ddot{g}_t - \int \dot{\pi}_t \dot{g}_t.$$

Plugging (22) into $\int \dot{\pi}_t \dot{g}_t$ yields

$$\int \dot{\pi}_t \dot{g}_t = -\text{Var}_{\pi_t}(\dot{g}_t). \tag{26}$$

Based on (23) and (24), we have

$$\ddot{g}_t = -\frac{1}{4t^4} \text{Var}_{r_t(\cdot|x)}[\|x - y\|^2] + \frac{1}{t^3} \mathbb{E}_{r_t(\cdot|x)}[\|x - y\|^2]. \tag{27}$$

Hence, with (23), (26) and (27),

$$\begin{aligned}
 \dot{K}(t) &= \mathbb{E}_{\pi_0} \left[\frac{1}{t^3} \mathbb{E}_{r_t(\cdot|x)} \|x - y\|^2 - \frac{1}{4t^4} \text{Var}_{r_t(\cdot|x)} \|x - y\|^2 \right] \\
 &\quad - \mathbb{E}_{\pi_t} \left[\frac{1}{t^3} \mathbb{E}_{r_t(\cdot|x)} \|x - y\|^2 - \frac{1}{4t^4} \text{Var}_{r_t(\cdot|x)} \|x - y\|^2 \right] \\
 &\quad + \frac{1}{4t^4} \text{Var}_{\pi_t} [\mathbb{E}_{r_t(\cdot|x)} \|x - y\|^2].
 \end{aligned} \tag{28}$$

In what follows, we plug in $g(y) = \frac{1}{2\sigma^2} \|y - u\|_2^2$. From (24),

$$\begin{aligned}
 r_t(\cdot|x) &= \exp \left[-g(y) - \frac{1}{2t} \|x - y\|^2 + g_t(x) \right] \\
 &= \mathcal{N}(t(\sigma^2 + t)^{-1}u + \sigma^2(\sigma^2 + t)^{-1}x, t\sigma^2(\sigma^2 + t)^{-1}\mathbf{I}).
 \end{aligned} \tag{29}$$

Hence,

$$\begin{aligned}
 \mathbb{E}_{r_t(\cdot|x)} \|x - y\|^2 &= \|t(\sigma^2 + t)^{-1}(u - x)\|^2 + t\sigma^2(\sigma^2 + t)^{-1}d \\
 \text{Var}_{r_t(\cdot|x)} \|x - y\|^2 &= 4t\sigma^2(\sigma^2 + t)^{-1} \|t(\sigma^2 + t)^{-1}(u - x)\|^2 + 2t^2\sigma^2(\sigma^2 + t)^{-2}d.
 \end{aligned}$$

It follows that

$$\begin{aligned}
 &\mathbb{E}_{\pi_0} \left[-\frac{1}{4t^4} \text{Var}_{r_t(\cdot|x)} \|x - y\|^2 \right] - \mathbb{E}_{\pi_t} \left[-\frac{1}{4t^4} \text{Var}_{r_t(\cdot|x)} \|x - y\|^2 \right] \\
 &= -\frac{\sigma^2(\sigma^2 + t)^{-1}}{t^3} (\mathbb{E}_{\pi_0} [\mathbb{E}_{r_t(\cdot|x)} \|x - y\|^2] - \mathbb{E}_{\pi_t} [\mathbb{E}_{r_t(\cdot|x)} \|x - y\|^2]) \\
 &= \frac{2\sigma^2(\sigma^2 + t)^{-1}}{t} K(t),
 \end{aligned}$$

and

$$\begin{aligned} K(t) &= \frac{1}{2t^2} \mathbb{E}_{\pi_t} [\mathbb{E}_{r_t(\cdot|x)} (\|x - y\|^2)] - \frac{1}{2t^2} \mathbb{E}_{\pi_0} [\mathbb{E}_{r_t(\cdot|x)} (\|x - y\|^2)] \\ &= \frac{1}{2} \mathbb{E}_{\pi_t} [\|(\sigma^2 + t)^{-1}(u - x)\|^2] - \frac{1}{2} \mathbb{E}_{\pi_0} [\|(\sigma^2 + t)^{-1}(u - x)\|^2] \end{aligned}$$

which implies $\lim_{t \rightarrow 0} K(t) = 0$ under mild conditions. With (25) and (28), we have

$$\dot{K}(t) \leq -\frac{2}{\sigma^2 + t} K(t) + \frac{1}{4t^4} \text{Var}_{\pi_t} [\|t(\sigma^2 + t)^{-1}(u - x)\|^2].$$

We also assume $f(x)$ is α_f -strongly convex, so π_t is $\alpha_f + \frac{1}{\sigma^2 + t}$ -strongly log-concave. Thus, π_t satisfies $\alpha_f + \frac{1}{\sigma^2 + t}$ -Poincaré inequality. This follows that

$$\frac{1}{4t^4} \text{Var}_{\pi_t} [\|t(\sigma^2 + t)^{-1}(u - x)\|^2] \leq \frac{1}{(\sigma^2 + t)^4 (\alpha_f + \frac{1}{\sigma^2 + t})} \mathbb{E}_{\pi_t} \|x - u\|_2^2.$$

As we assume the minimizer of $f(x)$ is u , the potential of π_t is minimized at u as well. By Theorem 1 in Durmus and Moulines (2016), we have

$$\mathbb{E}_{\pi_t} \|x - u\|_2^2 \leq \frac{d}{\alpha_f + \frac{1}{\sigma^2 + t}}.$$

Combining all the above inequality yields

$$\dot{K}(t) \leq -\frac{2}{\sigma^2 + t} K(t) + \frac{d}{(\sigma^2 + t)^4 (\alpha_f + \frac{1}{\sigma^2 + t})^2} \leq -\frac{2}{\sigma^2 + t} K(t) + \frac{d}{(\alpha_f \sigma^4 + \sigma^2)^2}$$

with the boundary condition $\lim_{t \rightarrow 0} K(t) = 0$. Solving the previous inequalities gives

$$K(t) \leq \frac{dt}{3(\alpha_f \sigma^4 + \sigma^2)^2} \left(1 + \left(\frac{\sigma^2}{\sigma^2 + t} \right)^2 + \frac{\sigma^2}{\sigma^2 + t} \right) \leq \frac{dt}{(\alpha_f \sigma^4 + \sigma^2)^2}.$$

Thus, we have

$$D_{\text{KL}}(\pi_0 || \pi_t) \leq \frac{dt^2}{2(\alpha_f \sigma^4 + \sigma^2)^2}.$$

By Pinsker's inequality, we have

$$\text{TV}(\pi_0, \pi_t) \leq \frac{t\sqrt{d}}{2(\alpha_f \sigma^4 + \sigma^2)}.$$

■