Minimizing Convex Functionals over Space of Probability Measures via KL Divergence Gradient Flow

Rentian Yao
UIUC

Linjun Huang
UIUC

Yun Yang
UIUC

Abstract

Motivated by the computation of the non-parametric maximum likelihood estimator (NPMLE) and the Bayesian posterior in statistics, this paper explores the problem of convex optimization over the space of all probability distributions. We introduce an implicit scheme, called the implicit KL proximal descent (IKLPD) algorithm, for discretizing a continuous-time gradient flow relative to the Kullback–Leibler (KL) divergence for minimizing a convex target functional. We show that IKLPD converges to a global optimum at a polynomial rate from any initialization; moreover, if the objective functional is strongly convex relative to the KL divergence, for example, when the target functional itself is a KL divergence as in the context of Bayesian posterior computation, IKLPD exhibits globally exponential convergence. Computationally, we propose a numerical method based on normalizing flow to realize IKLPD. Conversely, our numerical method can also be viewed as a new approach that sequentially trains a normalizing flow for minimizing a convex functional with a strong theoretical guarantee.

1 INTRODUCTION

Many problems in statistics and machine learning can be formulated as minimizing a functional, denoted as $\mathcal{F}$, over the space of all probability distributions $\mathscr{P}(\Theta)$ on a (parameter) space $\Theta \subset \mathbb{R}^d$. Examples include approximate Bayesian computation (Dai et al., 2016; Yao and Yang, 2022), non-parametric estimation (Yan et al., 2023), deep learning (Chizat, 2022; Chizat and Bach, 2018; Nitanda et al., 2022), and single-cell analysis in mathematical biology (Lavenant et al., 2021). Many recent studies consider addressing this optimization problem by numerically realizing the so-called Wasserstein gradient flow (WGF) for minimizing $\mathcal{F}$, a continuous dynamics on $\mathscr{P}(\Theta)$ that evolves in the steepest descent direction of $\mathcal{F}$ in the Wasserstein metric. WGFs inherit many appealing geometric interpretations from the conventional gradient flows in the Euclidean space $\mathbb{R}^d$ and extend them to $\mathscr{P}(\Theta)$. However, a fast (e.g., exponential) convergence guarantee for WGF usually requires the displacement convexity of the objective functional $\mathcal{F}$ (i.e., convexity of $\mathcal{F}$ along Wasserstein geodesics, see Appendix A.1 for a precise definition), which may impose more stringent conditions than the usual L_2 convexity of $\mathcal{F}$ and therefore may not hold in many applications (such as Examples 1 and 2 below). Several works instead consider relaxing the displacement convexity condition to a PL-type inequality on $\mathcal{F}$ (Bolley et al., 2012, 2013; Cattiaux et al., 2010; Chewi et al., 2020). However, a typical PL-type inequality can only hold with some impractical assumptions being imposed (Nitanda et al., 2022); moreover, it can be fairly difficult to verify even for simple problems in the Euclidean setting (Wensing and Slotine, 2020) as the inequality requires prior knowledge on the global optimum.

In this paper, we instead explore the use of Kullback–Leibler (KL) divergence gradient flow (KLGF) to minimize an L_2 convex target functional $\mathcal{F}$ over the space of all probability distributions $\mathscr{P}(\Theta)$. In particular, we will focus on the following two motivating examples, where the target functionals are L_2 convex but not necessarily displacement convex.

Example 1 (Non-parametric maximum likelihood estimation): The computation of the non-parametric maximum likelihood estimator (NPMLE) naturally arises in estimating the mixing distributions of mixture models and in using empirical Bayes methods to address compound decision problems. Concretely, we assume that the conditional distribution

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

of a random variable X given a parameter θ is $p(\cdot \,|\, \theta)$, where $\theta \in \Theta$ is drawn from an unknown mixing distribution P^* in $\mathscr{P}(\Theta)$. Given n i.i.d. copies $X^n = (X_1, \cdots, X_n)$ of X, the NPMLE of P^* is defined as

$$\begin{aligned} \widehat{P}_n &= \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \mathcal{L}_n(\rho), \quad \text{with} \\ \mathcal{L}_n(\rho) &\coloneqq \frac{1}{n} \sum_{i=1}^{n} -\log\Big(\int_\Theta p(X_i \,|\, \theta) \,\mathrm{d}\rho(\theta) \Big), \end{aligned} \tag{1}$$

which minimizes the (averaged) negative log-likelihood functional $\mathcal{L}_n : \mathscr{P}(\Theta) \to \mathbb{R}$. $\mathcal{L}_n$ is obviously L_2 convex on $\mathscr{P}(\Theta)$ but not generally displacement convex (see Appendix A.1 for an example).

The concept of NPMLE was initially introduced by Kiefer and Wolfowitz (1956). In this approach, the mixing distribution is not confined to a discrete measure with a predetermined number of atoms; instead, it is treated and estimated as an infinite-dimensional object. For NPMLE, Koenker and Mizera (2014) show that when the parameter space Θ is the one-dimensional real line, $\widehat{P}_n$ is a discrete probability measure with no more than n atoms. Moreover, they propose a numerical method for solving NPMLE in $\mathbb{R}$ by utilizing a space discretization scheme to reformulate (1) into a convex optimization problem. The resulting convex problem can then be efficiently solved using modern interior point methods. However, their computational approach is susceptible to the curse of dimensionality, making it computationally intensive when handling multivariate parameters, as demonstrated in our numerical comparison in Section 5. When the true mixing distribution P^* is sub-Gaussian on $\mathbb{R}$ and $p(\cdot \,|\, \theta) = \mathcal{N}(\theta, 1)$, Polyanskiy and Wu (2020) show that the number of atoms in $\widehat{P}_n$ decreases from $O(n)$ to $O(\log n)$. Soloff et al. (2021) further extend the optimality analysis of NPMLE to the multivariate and heteroscedastic normal observation model (with known heteroscedasticity), showing that, despite potential non-uniqueness when $d \geq 2$, a solution with at most n atoms exists. For the multivariate Gaussian location mixture model where $p(\cdot \,|\, \theta) = \mathcal{N}(\theta, I_d)$, Yan et al. (2023) propose an algorithm to solve (1) based on discretizing a Wasserstein–Fisher–Rao (WFR) gradient flow (Chizat et al., 2018; Gallouët and Monsaingeon, 2017). This method is numerically implemented using particle approximation.

Example 2 (Bayesian posterior sampling): In Bayesian statistics, a fundamental challenge involves sampling from the posterior distribution to estimate unknown parameters through the posterior mean and to construct credible intervals. This computational problem is especially relevant when exact computation of the posterior distribution is impractical due to non-conjugacy. Specifically, given a prior probability density function $\pi(\theta)$ for the unknown parameter $\theta \in \Theta$, and n independent and identically distributed samples $X^n = (X_1, \ldots, X_n)$ drawn from a conditional distribution $p(\cdot \,|\, \theta)$ that defines the likelihood function, the posterior density is $\pi_n(\theta) \propto \pi(\theta) \prod_{i=1}^n p(X_i \,|\, \theta)$. This posterior density admits a variational characterization

$$\begin{aligned} \pi_n &= \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \int V_n(\theta) \,\mathrm{d}\rho(\theta) + \int \rho \log \rho, \\ &\text{with} \quad V_n(\theta) = -\log \pi(\theta) - \sum_{i=1}^{n} \log p(X_i \,|\, \theta) \end{aligned} \tag{2}$$

denoting the effective potential function. In other words, the posterior can be identified as the global minimizer of the KL divergence functional $D_{\mathrm{KL}}(\cdot \,\|\, \pi_n)$ up to an additive constant. The KL functional is always L_2 convex on $\mathscr{P}(\Theta)$; but the displacement convexity requires more stringent conditions such as the convexity of V_n.

In addition to classical Markov Chain Monte Carlo (MCMC) algorithms (Tierney, 1994), recent advancements in sampling from Bayesian posterior distributions have focused on discretizing certain gradient flows in the space of all probability distributions. One such method is based on the WGF, involving the discretization of its associated stochastic differential equation, specifically, the Langevin dynamics. For instance, Dalalyan (2017b) introduced the (unadjusted) Langevin Monte Carlo algorithm, which employs an explicit scheme, the Euler–Maruyama method, to discretize the Langevin dynamics. However, this algorithm is known to produce a non-vanishing asymptotic bias due to the explicit discretization (Dalalyan, 2017a; Wibisono, 2018), and was later improved to an unbiased version using a forward-backward discretization scheme by (Wibisono, 2018). Nevertheless, the rapid convergence of these iterative algorithms, which are based on discretizing the WGF, requires stringent conditions on π_n, such as log-concavity, isoperimetry, or log-Sobolev inequalities (Chewi et al., 2021; Dalalyan, 2017b; Wibisono, 2018). Along a different track, Dai et al. (2016) proposed a stochastic particle mirror descent algorithm to iteratively approximate the Bayesian posterior density.

Our contributions. In this work, we introduce an implicit scheme, called the implicit KL proximal descent (IKLPD) algorithm, for discretizing a continuous-time gradient flow relative to the Kullback–Leibler divergence for minimizing a general L_2 convex functional $\mathcal{F}$. We show that, under the L_2 convexity condition alone, IKLPD converges to a global optimum at a polynomial rate from any initialization that admits a density; moreover, if $\mathcal{F}$ is strongly convex relative to the KL divergence, for example, when $\mathcal{F}$ itself is a KL divergence as in the

context of Bayesian posterior computation, IKLPD exhibits globally exponential convergence. The proposed implicit scheme thus eliminates the need for any smoothness condition on the L_2-gradient of $\mathcal{F}$, a condition typically needed by an explicit discretization scheme. In particular, in explicit discretization schemes, a low level of smoothness considerably limits the maximum permissible learning rate (or step size) of the algorithm. It is also crucial to note that while a Lipschitz L_2-gradient condition is generally not overly stringent for functions over Euclidean space, it can either exclude many commonly used functionals, such as the KL divergence, or requires substantial effort to verify. Additionally, our development can also be extended to a general (implicit) proximal mirror descent algorithm with a Bregman divergence beyond the KL; see Appendix A.2 for more details.

Computationally, we propose a numerical method based on normalizing flow (Dinh et al., 2016; Kingma and Dhariwal, 2018; Papamakarios et al., 2021; Rezende and Mohamed, 2015) to implement IKLPD. The compositional structure of normalizing flow is well-suited to the iterative nature of our time-discretization algorithm. In particular, we sequentially stack local short normalizing flows, each learned within an IKLPD iteration, to construct a global, layered normalizing flow, to approximate a minimizer of $\mathcal{F}$. Alternatively, our algorithm can be viewed as a novel method that sequentially trains a normalizing flow to minimize a convex functional $\mathcal{F}$ over $\mathscr{P}(\Theta)$ with a strong theoretical guarantee. When applied to computing the NPMLE and Bayesian posteriors, our numerical study suggests that the proposed method demonstrates promising performance, outperforming explicit schemes and other specialized competing algorithms.

We also consider two extensions of our development. In the first extension, we allow nonzero numerical error to occur when solving each implicit step. We examine how these errors accumulate (Theorem 4), offering insights for designing stopping criteria for the implicit step. In the second extension, we introduce and analyze the convergence of a stochastic variant of IKLPD (Theorem 5). This is particularly relevant in practical scenarios with large sample sizes. To our knowledge, this is the first study that analyzes a stochastic proximal type algorithm for optimizing functionals on the space of all probability distributions.

More related works. Ying (2020) applies a mirror descent algorithm for minimizing an interacting free energy over $\mathscr{P}(\Theta)$ composed of a potential energy, a KL divergence, and a self-interaction energy; however, they do not provide any convergence analysis. Aubin-Frankowski et al. (2022); Chizat (2021) prove the explicit convergence rate of the mirror descent for minimizing general (strongly) convex functionals over the space of all probability distributions. Chizat (2021) studies the convergence of the mirror descent algorithm for minimizing a special class of composite convex targets $\mathcal{F}$ whose primary component depends on $\rho \in \mathscr{P}(\Theta)$ through a linear functional. When specializing the Bregman divergence to the KL, their algorithm can be viewed as an explicit scheme to discretize the KL gradient flow. As a result, their theory requires $\mathcal{F}$ to have a Lipschitz L_2-gradient and does not cover common f-divergences (Rényi, 1961) such as the KL. Aubin-Frankowski et al. (2022) propose a different smoothness characterization called relative smoothness, which is analogous to the Euclidean case smoothness characterization via quadratic bounds. However, they only verify their conditions for the KL functional, with applications to the entropic optimal transport and Expectation Maximization (EM). In addition, their convergence bound diverges to infinity as the (global) minimizer of $\mathcal{F}$ becomes singular (i.e., does not admit a density). Last but not least, these two papers (Aubin-Frankowski et al., 2022; Chizat, 2021) do not provide concrete numerical methods to implement their algorithms.

As previously mentioned, Koenker and Mizera (2014), Soloff et al. (2021), and Yan et al. (2023) have proposed some advanced algorithms for numerically computing the NPMLE. The WFR based method by Yan et al. (2023) is limited to the Gaussian location mixture model and lacks an explicit convergence rate guarantee. The convex optimization based algorithms by Koenker and Mizera (2014); Soloff et al. (2021) approximate the target distribution through histograms by space discretization, and therefore suffers from the curse of dimensionality. In contrast, our method offers a worst-case $O(k^{-1})$ convergence guarantee after k iterations and tends to scale better to higher dimensions. For Bayesian posterior computation, MCMC is known for slow mixing in complex or high-dimensional problems. Most existing numerical algorithms based on Langevin dynamics require stringent conditions such as log-concavity, isoperimetry, or log-Sobolev inequalities (Chewi et al., 2021; Dalalyan, 2017b; Wibisono, 2018) for fast convergence. In comparison, our algorithm guarantees exponential convergence without imposing any conditions on the target posterior, provided the implicit step can be efficiently implemented, which holds true in our examples. Further literature review on mirror descent and stochastic (proximal) mirror descent in the Euclidean space, as well as optimization algorithms on the space of all probability distributions, is available in the Appendix.

2 KL DIVERGENCE GRADIENT FLOW AND IMPLICIT TIME DISCRETIZATION

To begin with, we briefly introduce our notation and some useful definitions. Let $\mathcal{F} : \mathscr{P}(\Theta) \to \mathbb{R}$ be a lower semi-continuous functional and $\mathscr{P}^r(\Theta)$ denote the set of all probability distributions admitting a density on Θ. Under mild conditions (see Appendix A.1 for details), one can define the *first variation* of $\mathcal{F}$ at $\rho \in \mathscr{P}(\Theta)$ as a map $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) : \Theta \to \mathbb{R}$ such that for any perturbation $\chi = \rho' - \rho$ with $\rho' \in \mathscr{P}^r(\Theta)$,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{F}(\rho + \varepsilon\chi)\Big|_{\varepsilon=0} = \int_\Theta \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\,\mathrm{d}\chi.$$

Note that $\frac{\delta \mathcal{F}}{\delta \rho}(\rho)$ is only uniquely defined up to an additive constant. The first variation can be viewed as the L_2-gradient of $\mathcal{F}$ in $\mathscr{P}(\Theta)$. A functional $\mathcal{F}$ is called λ-*relative strongly convex* (relative to KL) if for any pair of regular probability measures ρ, $\rho' \in \mathscr{P}(\Theta)$ ($\mathscr{P}^r(\Theta)$ when $\lambda > 0$) such that $\mathcal{F}(\rho)$ is finite,

$$\mathcal{F}(\rho') \geq \mathcal{F}(\rho) + \int \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\,\mathrm{d}(\rho' - \rho) + \lambda D_{\mathrm{KL}}(\rho' \,\|\, \rho).$$

We simply say $\mathcal{F}$ to be (L_2-)convex if $\mathcal{F}$ satisfies the above inequality with $\lambda = 0$. Note that the NPMLE example in Section 1 has a convex $\mathcal{F}$, and the Bayesian posterior example therein has a 1-relative strongly convex $\mathcal{F}$; see Appendix D.7 for a proof.

Remark 1. *The L_2 convexity and the displacement convexity are not directly comparable. For example, the KL divergence functional in* (2) *is always L_2 convex but not displacement convex unless potential V_n is a convex function over Θ. Conversely, the self-interaction energy functional $\mathcal{W}(\rho) = \int_{\mathbb{R}^2}(x - y)^2\,\mathrm{d}\rho(x)\mathrm{d}\rho(y)$ is displacement convex due to the convexity of the square function (McCann, 1997); however, direct calculation yields $\frac{1}{2}\big(\mathcal{W}(\rho) + \mathcal{W}(\rho')\big) - \mathcal{W}\big(\frac{1}{2}(\rho+\rho')\big) = -\frac{1}{2}\big(\int_{\mathbb{R}} x\,\mathrm{d}(\rho-\rho')(x)\big)^2 \leq 0$, indicating that $\mathcal{W}$ is instead L_2-concave.*

Given an initialization $\rho_0 \in \mathscr{P}^r(\Theta)$, we consider the following iterative scheme for minimizing $\mathcal{F}$ with step size $\{\tau_k : k \geq 1\}$,

$$\rho_k = \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \mathcal{F}(\rho) + \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1}),\ k \geq 1, \quad (3)$$

which will be referred to as the implicit KL proximal descent (IKLPD) algorithm. In Section 4, we propose using a normalizing flow (Kobyzev et al., 2020) to numerically optimize the objective in the implicit step (3). Note that this implicit step optimization problem becomes easier as the step size τ_k becomes smaller, as the optimal solution ρ_k is expected to become closer to the previous iterate ρ_{k-1} (e.g., $D_{\mathrm{KL}}(\rho_k \,\|\, \rho_{k-1}) = O(\tau_k)$), so that a few (stochastic) gradient iterations are sufficient to produce a relatively good solution. In contrast, as $\tau_k \to \infty$, implementing the implicit step becomes as hard as solving the original problem of minimizing $\mathcal{F}$. We conduct a numerical experiment in Section 5 to explore the impact of the step size τ_k on the implicit step computation and the overall convergence of the IKLPD algorithm.

It is worth noting that IKLPD extends the implicit gradient descent method for minimizing a function f on Θ under Euclidean ℓ_2 metric $\|\cdot\|$,

$$x_k = \operatorname*{argmin}_{x \in \Theta}\ f(x) + \frac{1}{2\tau_k}\|x - x_{k-1}\|^2,\ k \geq 1,$$

which is also the proximal point method (Boyd and Vandenberghe, 2004; Rockafellar, 1997) with convex function $\frac{1}{2}\|\cdot\|^2$; in particular, IKLPD changes the discrepancy measure $\frac{1}{2}\|x - x_{k-1}\|^2$ with $D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1})$. More generally, we may also consider a broader class of implicit mirror descent algorithms by substituting the KL with a general Bregman divergence, such as L_2 distance, Itakura–Saito divergence (Savchenko, 2019), and hyperbolic divergence (Ghai et al., 2020). The key property of Bregman divergences used in the proof is the "three-points identity" (e.g., Lemma 3.1 in (Chen and Teboulle, 1993)), which connects the first variation of the objective (3) with the Bregman divergence. However, our considered KL is often better aligned with the information geometry inherent to statistical problems. In contrast, other common divergences in statistics, such as the χ^2 divergence and the Rényi divergence are not Bregman divergences (see Appendix A.2).

Analogous to gradient (or mirror) descent in the Euclidean space (Krichene et al., 2015), which can be interpreted as discretizing a continuous-time gradient flow on Θ, IKLPD also corresponds to employing an implicit discretization scheme for the KL gradient flow (KLGF) on $\mathscr{P}(\Theta)$, which is described by the ordinary differential equation (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\log \rho_t\big) = -\frac{\delta \mathcal{F}}{\delta \rho}(\rho_t) + \int_\Theta \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t)(\theta)\,\mathrm{d}\rho_t(\theta). \quad (4)$$

This dynamic is also known as the Fisher–Rao gradient flow (Bauer et al., 2016; Yan et al., 2023). Let ρ^* be a global minimum of $\mathcal{F}$. The following theorem shows the convergence of KLGF for an L_2 convex functional $\mathcal{F}$.

Theorem 1. *Assume $\mathcal{F}$ to be a λ-relative strongly convex functional with respect to the KL divergence*

and $\rho^* \in \mathscr{P}^r(\Theta)$. If $\lambda > 0$, then ρ_t satisfies

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t) \le e^{-\frac{\lambda t}{2}} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0);$$

if $\lambda = 0$, then $\bar{\rho}_t = \frac{1}{t}\int_0^t \rho_s \,\mathrm{d}s$ satisfies

$$\mathcal{F}(\bar{\rho}_t) - \mathcal{F}(\rho^*) \le \frac{1}{t} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

3 THEORETICAL RESULTS

We analyze the convergence of IKLPD and two variants: an inexact IKLPD that permits non-zero numerical errors when solving the implicit step (3), and a stochastic version of IKLPD.

3.1 Convergence of IKLPD

We make the following assumptions.

Assumption 1 (Existence of IKLPD iterates)**.** *For each $k \ge 1$, the solution ρ_k in the k-th iteration of IKLPD algorithm as defined by* (3) *exists.*

Assumption 1 is typically verifiable by applying Prokhorov's Theorem (Prokhorov, 1956) when $\mathcal{F}$ is continuous with respect to the weak topology of $\mathscr{P}(\Theta)$. The continuity of $\mathcal{F}$ holds for many models, including the NPMLE discussed in Section 1.

Assumption 2 (Relative strong convexity)**.** *$\mathcal{F}$ is λ-relative strongly convex on $\mathscr{P}(\Theta)$ for $\lambda \ge 0$.*

Assuming some convexity condition is standard and necessary in the convergence analysis of proximal type algorithms (Aubin-Frankowski et al., 2022; Dragomir et al., 2021).

Theorem 2. *Suppose Assumptions 1 and 2 hold and $\rho^* \in \mathscr{P}^r(\Theta)$.*
(1) If $\lambda > 0$ and $\tau_k \equiv \tau > 0$, then we have

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k) \le \Big(1 + \frac{\lambda\tau}{2}\Big)^{-k} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

(2) If $\lambda = 0$, then we have

$$\min_{1\le\ell\le k} \mathcal{F}(\rho_\ell) - \mathcal{F}(\rho^*) \le \frac{1}{\sum_{\ell=1}^k \tau_\ell} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

Remark 2. *Several remarks are in order. First, when $\rho^* \in \mathscr{P}^r(\Theta)$, we do not need any extra condition beyong the convexity to guarantee the convergence of IKLPD. As we discussed in the introduction, this is different from the explicit discretization scheme considered in (Aubin-Frankowski et al., 2022; Chizat, 2021), which require additional smoothness conditions. Second, our proof for the $\lambda = 0$ case also implies the same convergence bound to hold for the last iterate ρ_k and the weighted trajectory average $\bar{\rho}_k = \big(\sum_{\ell=1}^k \tau_\ell\big)^{-1} \sum_{\ell=1}^k \tau_\ell \rho_\ell$. Third, if $\lambda = 0$ and $\tau_k = \tau$, then the IKLPD exhibits an $O(k^{-1})$ convergence rate after k iterations, which matches the convergence rate of the Euclidean proximal mirror descent algorithm for minimizing a smooth and convex function (e.g. Theorem 10.81 in (Beck, 2017)).*

Theorem 2 requires ρ^* to admit a density, so that the initial KL divergence $D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0)$ is finite. However, in many applications, such as the NPMLE computation, ρ^* can contain singular components or can even be a discrete measure (Polyanskiy and Wu, 2020). In these cases, Assumption 2 can only hold with $\lambda = 0$. To see this, we can apply the λ-relative strong convexity to $\rho = \rho_0$ for any $\rho_0 \in \mathscr{P}^r(\Theta)$ with a bounded $\frac{\delta\mathcal{F}}{\delta\rho}(\rho_0)$, and $\rho' = \rho^\sigma = \rho^* * \mathcal{N}(0, \sigma^2 I_d) \in \mathscr{P}^r(\Theta)$, the convolution of ρ^* with a normal distribution. This yields $\lambda D_{\mathrm{KL}}(\rho^\sigma \,\|\, \rho_0) \le \mathcal{F}(\rho^\sigma) - \mathcal{F}(\rho_0) - \int \frac{\delta\mathcal{F}}{\delta\rho_0}(\rho_0)\,\mathrm{d}(\rho^\sigma - \rho_0)$. As we let $\sigma \to 0^+$, the right-hand side of this inequality is finite, while the KL term $D_{\mathrm{KL}}(\rho^\sigma \,\|\, \rho_0)$ diverges when $\rho^* \notin \mathscr{P}^r(\Theta)$, indicating that $\lambda = 0$. To extend the convergence result to such ρ^* that does not admit a density, we need an additional assumption about the continuity of $\mathcal{F}$ around ρ^*. Let W_1 denote the 1-Wasserstein metric; see Appendix A.1 for a precise definition.

Assumption 3 (Local W_1-continuity)**.** *There exists a constant $L > 0$ such that*

$$|\mathcal{F}(\rho) - \mathcal{F}(\rho^*)| \le L W_1(\rho, \rho^*), \quad \forall\, \rho \in \mathscr{P}(\Theta).$$

This local continuity condition on $\mathcal{F}$ is less stringent than a typical smoothness condition assumed in the analysis of explicit schemes that involves the first variation, and it is satisfied in our examples.

Theorem 3. *If Assumptions 1, 2, and 3 hold with $\lambda = 0$ and $\rho^* \in \mathscr{P}(\Theta)$, then for any $\rho \in \mathscr{P}^r(\Theta)$,*

$$\min_{1\le\ell\le k} \mathcal{F}(\rho_\ell) - \mathcal{F}(\rho^*) \le \frac{D_{\mathrm{KL}}(\rho \,\|\, \rho_0)}{\sum_{\ell=1}^k \tau_\ell} + \frac{L}{2} W_1(\rho, \rho^*).$$

Remark 3. *Theorem 3 suggests that when ρ^* contains singular components, the convergence rate of IKLPD may depend on finer structures on the singularity of ρ^* as we want to construct some ρ to compensate for the singularity. For example, if ρ^* is a discrete measure, then the convergence rate is $O(\frac{d\log k}{k})$; generally, if ρ^* is supported on a d'-dimensional hyperplane in the ambient space $\Theta \subset \mathbb{R}^d$ with $d' < d$, then the convergence rate becomes $O(\frac{(d-d')\log k}{k})$ (the support of a discrete measure has an effective dimension $d' = 0$); see Appendix D.3 for a proof, where we choose ρ in the theorem as the convolution of ρ^* and a $(d-d')$-dim Gaussian distribution whose variance is optimized to make the upper bound smallest, so that the convoluted distribution admits a density.*

3.2 Convergence of Inexact IKLPD

For inexact IKLPD, we allow non-zero numerical errors when solving the implicit step (3), and study their impact on overall convergence and the design of the implicit step stopping criterion. In practice, one can use the first-order optimality condition $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) + \frac{1}{\tau_k} \log \frac{\rho}{\rho_{k-1}} = \text{constant}$ to design stopping criterion and monitor the convergence of the implicit step optimization subproblem (3). Specifically, let $\{\rho_k^{\mathrm{err}} : k \geq 0\}$ denote the iterates from an inexact IKLPD, and let

$$\eta_k(\cdot) \coloneqq \frac{\delta \mathcal{F}}{\delta \rho}(\rho_k^{\mathrm{err}})(\cdot) + \frac{1}{\tau_k} \log \frac{\rho_k^{\mathrm{err}}}{\rho_{k-1}^{\mathrm{err}}}(\cdot) \tag{5}$$

denote the first variation (as a function over Θ) of the target functional in the implicit step (3) evaluated at ρ_k^{err}. Let $\{\varepsilon_k : k \geq 1\}$ denote a generic sequence of error tolerance levels. For technical convenience, we characterize the convergence of each implicit step optimization via the oscillation of η_k, and make the following assumption.

Assumption 4 (Uniform error control)**.** *For each $k \geq 1$ and $\varepsilon_k \geq 0$, we have $\mathrm{Osc}_\Theta(\eta_k) \leq \varepsilon_k$, where*

$$\mathrm{Osc}_\Theta(\eta_k) \coloneqq \sup_{\theta, \theta' \in \Theta} \|\eta_k(\theta) - \eta_k(\theta')\|$$

is the oscillation of η_k over Θ.

There are also other types of inexact algorithms for optimizing functionals on the space of all probability distributions (Dai et al., 2016; Kent et al., 2021), some of which are not implementable since they require knowledge of unknown quantities, such as the exact solution of the subproblem, to evaluate the tolerance metric. In our context, one can also use other characterizations, such as the variance of η_k under $\rho_{k-1}^{\mathrm{err}}$ that is easier to compute in practice.

The following theorem illustrates the impact of error tolerance level on the convergence rate of inexact IKLPD when $\mathcal{F}$ is λ-relative strongly convex for $\lambda > 0$. In particular, we consider two regimes: ε_k has either an exponential decay or a polynomial decay in k; and the inexact IKLPD exhibits different convergence patterns under the two regimes.

Theorem 4. *Suppose Assumption 2 holds with $\lambda > 0$ and Assumption 4 also holds, and consider $\tau_k \equiv \tau$.*
(1) If $\varepsilon_k \leq \kappa \varepsilon^k$ for some $\kappa > 0$ and $0 < \varepsilon < 1$ satisfying $\varepsilon \sqrt{1 + \lambda\tau/2} \neq 1$, then there exists a constant $C = C(\tau, \lambda, \varepsilon) > 0$ such that

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k^{\mathrm{err}}) \leq \frac{C\kappa^2 + 2D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0)}{(\min\{\varepsilon^{-2}, 1 + \lambda\tau/2\})^k};$$

(2) If $\varepsilon_k \leq \varepsilon k^{-\alpha}$ for some $\varepsilon, \alpha > 0$, then there exists a constant $C = C(\tau, \lambda, \alpha) > 0$ such that

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k^{\mathrm{err}}) \leq \frac{2D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0)}{(1 + \lambda\tau/2)^k} + \frac{C\varepsilon^2}{k^{2\alpha}}$$

Remark 4. *Similar to Theorem 3, Theorem 4 requires ρ^* to have a density when $\lambda > 0$. Our theorem cannot cover the $\lambda = 0$ case, since in order to show $D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k^{\mathrm{err}})$ is decreasing in k, we need the relative strong convexity to contribute a term that compensates for the error caused by η_k. In addition, the current proof of Theorem 4 can only be extended to cover a Bregman divergence that dominates the L_1 distance, such as any divergences stronger than the KL, since we need to use it to address an additional error term that depends on the L_1 distance between ρ_k^{err} and ρ^*.*

3.3 Convergence of Stochastic IKLPD

In this section, we propose and analyze a stochastic version of IKLPD, whose k-th iterate is given by

$$\rho_k^{\mathrm{stoc}} = \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \mathcal{F}_{\xi_k}(\rho) + \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1}^{\mathrm{stoc}}). \tag{6}$$

Here $\mathcal{F}_{\xi_k}$ is an unbiased estimator of $\mathcal{F}$ for any fixed input in $\mathscr{P}(\Theta)$, with ξ_k indicating the source of randomness in iteration k. For example, in a statistical setting such as NPMLE, $\mathcal{F}_{\xi_k}$ can be the negative log-likelihood functional over a random selected mini-batch. To prove the convergence, we make the following Assumption.

Assumption 5 (Stochastic IKLPD)**.** *The stochastic objective functional $\mathcal{F}_\xi$ satisfies:*
(1) (Unbiasedness) $\mathbb{E}_\xi\big[\mathcal{F}_\xi(\rho)\big] = \mathcal{F}(\rho)$.
(2) (Solution existence) A solution of (6) exists.
(3) (Randomness condition) $\{\xi_k : k \geq 1\}$ are independently and identically distributed.
(4) (One-sided relative Lipschitz continuity) For some $L(\xi)$ with a finite second-order moment,

$$\mathcal{F}_\xi(\rho) - \mathcal{F}_\xi(\rho') \leq L(\xi) \sqrt{D_{\mathrm{KL}}(\rho' \,\|\, \rho)}$$

holds for every $\rho, \rho' \in \mathscr{P}(\Theta)$.

The one-sided relative Lipschitz continuity condition is also considered by Bertsekas (2011); Davis et al. (2018), which was utilized to analyze the convergence of stochastic proximal descent and stochastic proximal mirror descent in the Euclidean space. In our proof, this condition is used to bound the difference of $\mathcal{F}_{\xi_k}(\rho_k^{\mathrm{stoc}})$ and $\mathcal{F}_{\xi_k}(\rho_{k-1}^{\mathrm{stoc}})$.

Theorem 5. *Assume that $\mathcal{F}_\xi$ is λ-relative strongly convex for $\lambda \geq 0$. Suppose Assumption 5 holds and $\rho^* \in \mathscr{P}^r(\Theta)$. Let $\tau > 0$ be a constant.*

(1) If $\lambda = 0$, then by taking $\tau_k = \frac{\tau}{\sqrt{k+1}}$ we have

$$\min_{0\le\ell\le k-1} \mathbb{E}\big[\mathcal{F}(\rho_\ell^{\mathrm{stoc}})\big] - \mathcal{F}(\rho^*) \le \frac{4D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0) + \tau^2 \log(k+1)\mathbb{E}[L(\xi_1)^2]}{8\tau(\sqrt{k+1}-1)};$$

(2) If $\lambda > 0$, then by taking $\tau_k = \frac{2}{\lambda(k+1)}$ we have

$$\min_{0\le\ell\le k-1} \mathbb{E}\mathcal{F}(\rho_\ell^{\mathrm{stoc}}) - \mathcal{F}(\rho^*) \le \frac{2\lambda^2 D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0) + \log(k+1)\mathbb{E}[L(\xi_1)^2]}{2\lambda k}.$$

Remark 5. *The convergence rates in our theorem match those of stochastic gradient descent (Nemirovski et al., 2009; Rakhlin et al., 2011) and stochastic (proximal) mirror descent (Davis et al., 2018; Lan, 2020) for minimizing (strongly) convex functions in the Euclidean space. Additionally, when Assumption 3 holds, the same smoothing argument as in the proof of Theorem 3 can be carried over to deal with a singular $\rho^* \notin \mathscr{P}^r(\Theta)$.*

4 COMPUTATION VIA NORMALIZING FLOW

We use normalizing flow (NF) to solve the implicit step optimization problem (3). Normalizing flows (Dinh et al., 2016; Kingma and Dhariwal, 2018; Papamakarios et al., 2021; Rezende and Mohamed, 2015) offer a general mechanism for defining expressive probability distributions through transforming a simple probability distribution into a complex one using compositions of invertible and differentiable transformations. For simplicity, we will refer to the IKLPD steps as the outer loop (iterations), and the (stochastic) gradient steps for optimizing the NF parameters in the implicit scheme problem (3) as the inner loop (iterations).

Given the shared compositional structure between our iterative IKLPD algorithm and the NF, we propose sequentially stacking the local, short normalizing flows, learned within each inner-loop iteration, to form a global, layered normalizing flow for approximating ρ^*. Concretely, we use $T_{\#}\rho$ to denote the pushforward distribution of a distribution $\rho \in \mathscr{P}(\Theta)$ through a transport map $T : \Theta \to \Theta$, and use $\widehat{T}^{(k)}$ to denote the local normalizing flow learned through solving (3), yielding $\rho_k = \widehat{T}^{(k)}_{\#}\rho_{k-1}$, where

$$\widehat{T}^{(k)} = \operatorname*{argmin}_{T\in\mathcal{T}} \ \mathcal{F}\big(T_{\#}\rho_{k-1}\big) + \frac{1}{\tau_k} D_{\mathrm{KL}}\big(T_{\#}\rho_{k-1} \,\big\|\, \rho_{k-1}\big),$$

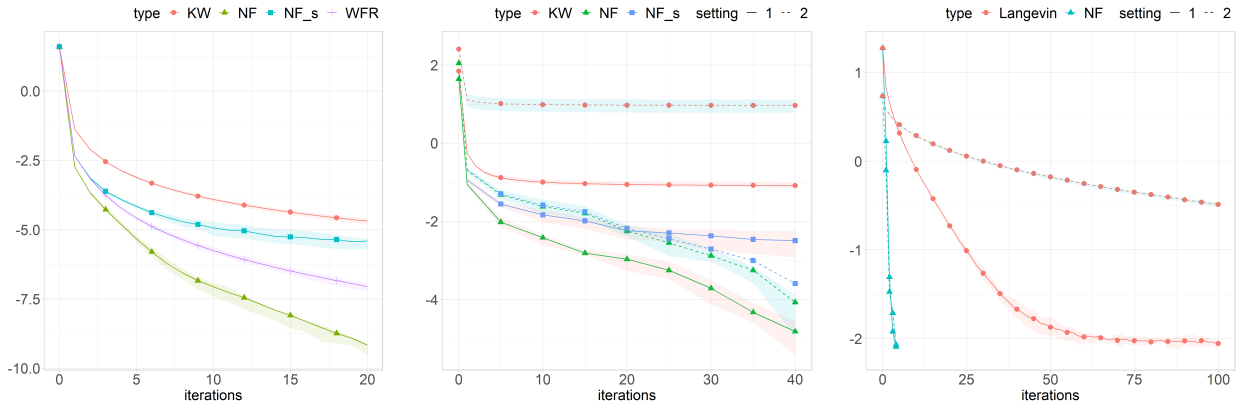
Here, $\mathcal{T}$ denotes a generic normalizing flow class.

Note that another benefit of using NF here is that the KL term can be directly computed in terms of a closed form expression of the log-density of $T_{\#}\rho_{k-1}$, whereas other numerical methods based on particle approximation require the use of kernel density estimation; further details of its numerical computation using (stochastic) gradient descent and the reparametrization trick are provided in Appendix C. With these local NF maps, we can use the telescoping trick to express $\rho_k = \widehat{T}^{(k)}_{\#} \circ \cdots \circ \widehat{T}^{(1)}_{\#}\rho_0$, which defines a generative process for sampling from ρ_k. As k increases, to maintain a fixed storage budget (e.g., keep at most k_0 local NFs), one may employ a teacher-student architecture (Hinton et al., 2015; Hu et al., 2022) to distill knowledge by utilizing a single NF to compress all historical local NFs beyond the most recent (k_0-1) ones; see Appendix C for an illustration.

5 NUMERICAL RESULTS

For the implementation, we use the Python `normflows` package (Stimper et al., 2023) based on PyTorch to implement the real-valued non-volume preserving (real-NVP) normalizing flow (Dinh et al., 2016) for our method. We consider three examples: NPMLE for Gaussian location mixture model, NPMLE for Gaussian location scale mixture model, and sampling from a distribution known up to a constant (Bayesian computation). For NPMLE, we also consider two state-of-the-art competing methods, the Wasserstein–Fisher–Rao (`WFR`) gradient flow (Yan et al., 2023) and a convex optimization based method (Koenker and Mizera, 2014) (referred to as the `KW` method). For the Bayesian computation example, we compare our method with the (unadjusted) Langevin Monte Carlo algorithm (`Langevin`), which corresponds to an explicit discretization scheme to the Wasserstein gradient flow. Due to space constraints, we defer the details about the implementations and setup of each example below, as well as additional plots and results, to Appendix C.

Gaussian location mixture model. We consider a two-dimensional Gaussian location mixture model, where for $\theta \in \Theta = \mathbb{R}^2$, the conditional distribution $p(\,\cdot\,|\,\theta)$ in the NPMLE formulation (1) is the density of $\mathcal{N}(\theta, I_2)$. We set the true (mixing) distribution of θ to be a bimodal two moon distribution (Stimper et al., 2023) and use a sample size of $n = 5000$. For our method (`NF`), we also implement the stochastic variant (`NF_s`) by using a randomly subsampled mini-batch of size $m = 500$ to compute the stochastic gradient during the training of the normalizing flow. We compare our method with the previously mentioned `WFR` and `KW` methods. Figure 1(a) displays the difference between $\mathcal{L}_n(\rho_k) - \mathcal{L}_n(\widehat{\rho})$ (in a logarithmic scale) as a function of the iteration count k, where $\widehat{\rho}$ is a numerically optimal solution obtained by running our method



(a) Location Mixture NPMLE (b) Location Scale Mixture NPMLE (c) KL Divergence for Posteriors

Figure 1: Numerical accuracy (with error bars) versus the iteration count k. For plots (a) and (b) (NPMLE), we report $\log\left(\mathcal{L}_n(\rho_k) - \mathcal{L}_n(\widehat{\rho})\right)$, where $\widehat{\rho}$ is the (numerically) optimal solution; and for (c), we report $\log W_1(\rho_k, \pi)$, with π denoting the target distribution. All results are based on 10 independent trials.

for a sufficient number of iterations. As can be seen, for this relatively simple problem, all methods exhibit rapid convergence. Our method with exact gradient descent (`NF`) achieves the fastest convergence, while our stochastic variant (`NF_s`) shows slower convergence compared to WFR.

Gaussian location scale mixture model. Our second example is a d-dimensional Gaussian location scale mixture model, where for $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}^d \times \mathbb{R}_+^d$, the conditional distribution $p(\cdot \,|\, \theta)$ is the density of $\mathcal{N}(\mu, \Sigma)$, with $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_d^2)$; the true (mixing) distribution $P^* = P_\mu^* \otimes P_{\sigma^2}^*$. We consider two settings, one with $d = 2$ and the other with $d = 3$, both with a sample size of $n = 5000$. Since `WFR` is applicable only to the Gaussian location mixture model, our comparison is limited to `NF`, `NF_s` and `KW`. Figure 1(b) shows the results. As we can observe, the necessity for KW to discretize the parameter space into equally spaced grids results in a non-vanishing bias term attributed to this discretization. This bias becomes larger as the dimensionality increases, owing to the curse of dimensionality. In contrast, our methods, including the stochastic variants, are relatively robust against dimensionality increase, with the numerical error keeps decreasing as the iteration count k increases.

Bayesian sampling. In this example, we set the true target distribution to have density $\pi(\theta) \propto e^{-\frac{1}{2\alpha}\|\theta\|^{2\alpha}}$ for $\theta \in \Theta = \mathbb{R}^2$, which is known only up to a normalization constant. The corresponding objective functional is $\mathcal{F}(\rho) = \int \frac{1}{2\alpha}\|\theta\|^{2\alpha} \,\mathrm{d}\rho(\theta) + \int \rho \log \rho$. We consider two settings: $\alpha = 2$ and $\alpha = 3$. Note that $\alpha = 2$ corresponds to a Lipschitz continuous potential function $\frac{1}{2\alpha}\|\cdot\|^{2\alpha}$, as required by explicit discretization methods, while $\alpha = 3$ violates this condition. We compare our method (`NF`) with the (unadjusted) `Langevin` method as a representative explicit discretization method. As illustrated in Figure 1(c), `NF` converges very rapidly for both values of α, in line with the prediction of our Theorem 2 under $\lambda = 1$. In contrast, `Langevin` exhibits significantly slower convergence, especially when $\alpha = 3$. For `Langevin`, we manually selected a best step size without divergence, ensuring the fastest convergence possible.

Impact of IKLPD step size τ_k. To examine the impact of the step size τ_k on the IKLPD algorithm, we conduct additional numerical experiments to compare the inner loop iterations using a first-order optimization algorithm and the outer loop iterations of IKLPD under varying constant step sizes $\tau_k \equiv \tau$. Specifically, in this experiment setting, we study the impact of step size $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.50\}$ and total inner loop iteration $N_2 \in \{4000, 6000\}$ on the overall convergence of IKLPD with $N_1 = 15$ total outer loop iterations. Further implementation details are provided in Appendix C.4. From Figure 2, we observe that under a fixed inner iteration budget, increasing the step size τ beyond certain threshold (i.e., $\tau = 0.20$) will make the IKLPD algorithm fails to converge to a good solution or even diverge. Upon closer examination of these non-converging cases, we identified two primary reasons for this failure, either the inner loop is trapped in a local minimum that is not global for problem (3), or it is unable to find a reasonably good solution given a limited inner loop iterations. Figure 2 also shows that for small values of τ below this non-convergent threshold, the subproblem (3) becomes easier as the inner loop needs fewer iterations to converge. However, since the progress made by each IKLPD step is smaller, the outer loop iterations converges slower. Moreover, we also observed that for a smaller τ, the normalizing flow architecture requires fewer parame-

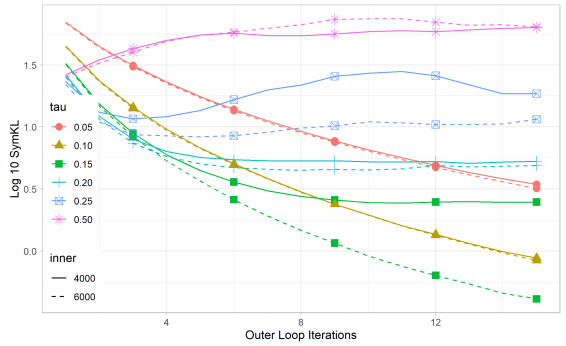


Figure 2: We study the impact of step size $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.50\}$ and total inner loop iteration $N_2 \in \{4000, 6000\}$ on the convergence of IKLPD with $N_1 = 15$ total outer loop iterations. Target distribution ρ^* is a mixture of $M = 10$ Gaussians in $\mathbb{R}^{12}$ as described in Appendix C.3. A symmetric KL divergence $\mathrm{SymKL}(\rho_k, \rho^*)$, defined in equation (14), is used as the error metric, and is plotted against the outer loop iterations (in the log-scale).

ters to approximately compute the minimum of subproblem (3); e.g., see Figures 6 and 7 in Appendix C.4. For the fixed budget of outer iterations $N_1 = 15$ used in this experiment, a medium step size ($\tau = 0.15$) with $N_2 = 6000$ inner loop iterations leads to a best performance, effectively balancing the convergence of the inner and outer loops under a limited computational budget. A more comprehensive study on the impact of the step size τ_k is available in Appendix C.4.

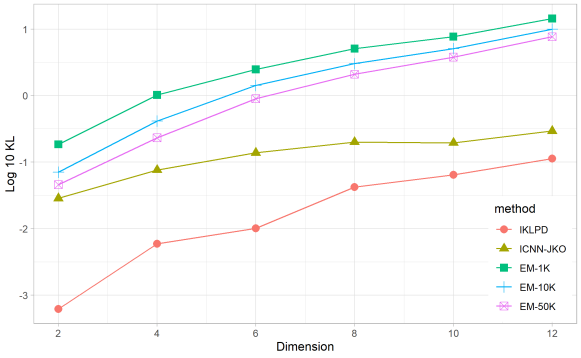


Figure 3: KL divergence between computed solutions and ρ^* versus dimension $D \in \{2, 4, \cdots, 12\}$. All numbers are averaged based on 3 independent trials.

Comparison with Wassersten gradient flows. Finally, we compare the performance of our method and recent methods based on discretizing the WGF. These include the ICNN-JKO method (Mokrov et al., 2021), which approximates the convex potential function defining an optimal transport map using input convex neural networks (ICNN), and the Euler-Maruyama (EM) method for discretizing the Langevin dynamics associated with the WGF. We tested the EM method using 10^3, 10^4, $5{\cdot}10^4$ particles (denoted as EM-1K, EM-10K, EM-50K). We follow the same experiment setting as Section 4.1 of (Mokrov et al., 2021), where the target functional is $\mathcal{F}(\rho) = D_{\mathrm{KL}}(\rho \,\|\, \rho^*)$ with $\rho^* = M^{-1} \sum_{m=1}^{M} \mathcal{N}(\mu_m, I_D)$ being a Gaussian mixtures with M components in $\mathbb{R}^D$. Here, we generate $\mu_1, \cdots, \mu_M \sim \mathrm{Uniform}([-5, 5]^D)$, under $M = D/2 + 4$ with $D \in \{2, 4, \ldots, 12\}$. As shown in Figure 3, a larger D makes the problem harder; nevertheless, IKLPD consistently achieves the best performance. Further details and results are provided in Appendix C.3.

6 DISCUSSION

In this work, we proposed an implicit KL proximal descent (IKLPD) algorithm, which discretized a continuous-time gradient flow relative to the Kullback–Leibler divergence for minimizing a convex functional defined over the space of all probability distributions. We utilized the proposed method to address two statistical applications, specifically, nonparametric maximum likelihood estimation (NPMLE) and Bayesian posterior computation. We demonstrated that our implicit method has multiple advantages compared to its explicit counterpart: 1. it did not require a Lipschitz L_2-gradient, thus allowing for larger step sizes and fewer iterations to converge; 2. it was more robust and did not need kernel density estimation in order to approximately compute the L_2-gradient as in the explicit method, making the explicit method suffer from the curse of dimensionality. Computationally, we proposed a numerical method based on normalizing flow to implement IKLPD, and utilized a teacher-student architecture to maintain constant space complexity. Conversely, our numerical method could also be viewed as a novel approach that sequentially trains a normalizing flow for minimizing a convex functional with strong theoretical guarantees. Some potential future directions include: 1. applying and analyzing IKLPD for other more complicated statistical applications, such as training Bayesian neural networks and variational inference with structural constraints; 2. extending the KL to a general Bregman divergence and identifying examples where using a particular Bregman divergence is beneficial; 3. analyzing the optimization landscape of the normalizing flow for solving each implicit step optimization problem in the IKLPD.

Acknowledgements

Yun Yang was partially supported by NSF DMS-2210717.

References

Asi, H. and Duchi, J. C. (2019). Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290.

Aubin-Frankowski, P.-C., Korba, A., and Léger, F. (2022). Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275.

Bauer, M., Bruveris, M., and Michor, P. W. (2016). Uniqueness of the fisher–rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506.

Bauschke, H. H., Bolte, J., and Teboulle, M. (2017). A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195.

Bolley, F., Gentil, I., and Guillin, A. (2012). Convergence to equilibrium in wasserstein distance for fokker–planck equations. *Journal of Functional Analysis*, 263(8):2430–2457.

Bolley, F., Gentil, I., and Guillin, A. (2013). Uniform convergence to equilibrium for granular media. *Archive for Rational Mechanics and Analysis*, 208:429–445.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cattiaux, P., Guillin, A., and Wu, L.-M. (2010). A note on talagrand's transportation inequality and logarithmic sobolev inequality. *Probability theory and related fields*, 148:285–304.

Chen, G. and Teboulle, M. (1993). Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., and Zhang, M. (2021). Analysis of langevin monte carlo from poincar\'e to log-sobolev. *arXiv preprint arXiv:2112.12662*.

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. (2020). Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR.

Chizat, L. (2021). Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368*.

Chizat, L. (2022). Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18:1–44.

Chizat, L., Zhang, S., Heitz, M., and Schiebinger, G. (2022). Trajectory inference via mean-field langevin in path space. *Advances in Neural Information Processing Systems*, 35:16731–16742.

Dai, B., He, N., Dai, H., and Song, L. (2016). Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994. PMLR.

Dalalyan, A. (2017a). Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR.

Dalalyan, A. S. (2017b). 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.

Davis, D., Drusvyatskiy, D., and MacPhee, K. J. (2018). Stochastic model-based minimization under high-order growth. *arXiv preprint arXiv:1807.00255*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Dragomir, R. A., Even, M., and Hendrikx, H. (2021). Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. In *International Conference on Machine Learning*, pages 2815–2825. PMLR.

Gallouët, T. O. and Monsaingeon, L. (2017). A jko splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130.

Ghai, U., Hazan, E., and Singer, Y. (2020). Exponentiated gradient meets gradient descent. In *Algorithmic learning theory*, pages 386–407. PMLR.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, C., Li, X., Liu, D., Chen, X., Wang, J., and Liu, X. (2022). Teacher-student architecture for

knowledge learning: A survey. *arXiv preprint arXiv:2210.17332*.

Kent, C., Blanchet, J., and Glynn, P. (2021). Frank-wolfe methods in probability space. *arXiv preprint arXiv:2105.05352*.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Krichene, W., Bayen, A., and Bartlett, P. L. (2015). Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28.

Lan, G. (2020). *First-order and stochastic optimization methods for machine learning*, volume 1. Springer.

Lavenant, H., Zhang, S., Kim, Y.-H., and Schiebinger, G. (2021). Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. *Advances in neural information processing systems*, 29.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179.

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J. M., and Burnaev, E. (2021). Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Nitanda, A., Wu, D., and Suzuki, T. (2022). Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680.

Polyanskiy, Y. and Wu, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214.

Rakhlin, A., Shamir, O., and Sridharan, K. (2011). Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Rockafellar, R. T. (1997). *Convex analysis*, volume 11. Princeton university press.

Ryu, E. K. and Boyd, S. (2014). Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Savchenko, V. (2019). Itakura–saito divergence as an element of the information theory of speech perception. *Journal of Communications Technology and Electronics*, 64:590–596.

Soloff, J. A., Guntuboyina, A., and Sen, B. (2021). Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *arXiv preprint arXiv:2109.03466*.

Stimper, V., Liu, D., Campbell, A., Berenz, V., Ryll, L., Schölkopf, B., and Hernández-Lobato, J. M. (2023). normflows: A pytorch package for normalizing flows. *Journal of Open Source Software*, 8(86):5361.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Wensing, P. M. and Slotine, J.-J. (2020). Beyond convexity—contraction and global convergence of gradient descent. *Plos one*, 15(8):e0236661.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Wright, S. J. and Recht, B. (2022). *Optimization for data analysis*. Cambridge University Press.

Yan, Y., Wang, K., and Rigollet, P. (2023). Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow. *arXiv preprint arXiv:2301.01766*.

Yao, R. and Yang, Y. (2022). Mean field variational inference via wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*.

Ying, L. (2020). Mirror descent algorithms for minimizing interacting free energy. *Journal of Scientific Computing*, 84(3):51.

APPENDIX

In this Appendix, we provide more background knowledge and useful results about optimizing a functional over the space of all probability distributions, including some of their connections with optimal transport (e.g., displacement convexity) and a more broader framework of the proximal mirror descent algorithm that allows an extension from the KL divergence to general Bregman divergences. We also review additional literature on mirror descent and stochastic (proximal) mirror descent in the Euclidean space, along with some further optimization algorithms on the space of all probability distributions. Moreover, we detail the implementation of the algorithms and the numerical experiments showcased in the main paper, and we provide additional numerical results. Finally, this Appendix includes all the proofs related to the main theoretical results presented in the paper, including the verification of the L_2-convexity of the target functionals in both the NPMLE and the Bayesian posterior computation examples.

A BACKGROUNDS AND FACTS

In this appendix, we provide additional background and facts related to optimization over the space of probability distributions, the non-parametric maximum likelihood estimation (NPMLE), and extensions of our developments to more general proximal mirror descent algorithms. However, as we mentioned in the main paper, our considered KL is often better aligned with the information geometry inherent to statistical problems; see, for example, the two motivating examples of NPMLE and Bayesian posterior computation considered in the paper.

A.1 Some Definitions and Consequences

First variation. We first provide a formal definition of first variations; more details can be found, e.g., in Section 7.2 of (Santambrogio, 2015). Let $\mathcal{F} : \mathscr{P}(\Theta) \to \mathbb{R}$ be a lower semi-continuous functional and $\mathscr{P}^r(\Theta)$ denote the set of all probability measures absolutely continuous with respect to the Lebesgue measure on Θ. A measure $\rho \in \mathscr{P}(\Theta)$ is called *regular* for $\mathcal{F}$ if $\mathcal{F}\big(\varepsilon\rho + (1-\varepsilon)\rho'\big) < \infty$ for all $\varepsilon \in (0,1)$ and any $\rho' \in \mathscr{P}^r(\Theta)$ that has compact support and bounded density. If ρ is regular for $\mathcal{F}$, one can define the *first variation* of $\mathcal{F}$ at ρ as a map $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) : \Theta \to \mathbb{R}$ such that for any perturbation $\chi = \rho' - \rho$, where $\rho' \in \mathscr{P}^r(\Theta)$ has bounded density and compact support,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{F}(\rho + \varepsilon\chi)\bigg|_{\varepsilon=0} = \int_\Theta \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\,\mathrm{d}\chi.$$

Pushforward. Let $\mathcal{X}$ be a measurable space, and $T : \mathcal{X} \to \Theta$ be a measurable function. The *pushforward* ρ of a measure $\mu \in \mathscr{P}(\mathcal{X})$ under T, denoted by $\rho = T_{\#}\mu$, is a measure on Θ defined as

$$\rho(A) = T_{\#}\rho(A) = \rho\big(T^{-1}(A)\big), \quad \forall A \subset \Theta \text{ is measurable.}$$

$\mathbf{W_p}$ distance and coupling. Let $\pi_0, \pi_1 : \Theta \times \Theta \to \Theta$ be the projection functions defined as $\pi_0(\theta, \theta') = \theta$ and $\pi_1(\theta, \theta') = \theta'$, and define $\pi_t = (1-t)\pi_0 + t\pi_1$. For any $\rho, \rho' \in \mathscr{P}(\Theta)$, γ is called a *coupling* of ρ and ρ', denoted by $\Pi(\rho, \rho')$, if $(\pi_0)_{\#}\gamma = \rho$ and $(\pi_1)_{\#}\gamma = \rho'$. Then, the W_p distance between ρ and ρ' is defined as

$$W_p^p(\rho, \rho') = \inf_{\gamma \in \Pi(\rho,\rho')} \int_{\Theta\times\Theta} \|\theta - \theta'\|^p \,\mathrm{d}\gamma(\theta, \theta') \tag{7}$$

By the above definition, it is clear that the W_p distance can also be defined through

$$W_p^p(\rho, \rho') = \inf_{\theta\sim\rho, \theta'\sim\rho'} \mathbb{E}\big[\|\theta - \theta'\|^p\big].$$

We say γ^* is an optimal coupling of ρ and ρ', denoted by $\Pi_o(\rho, \rho')$, if

$$W_2^2(\rho, \rho') = \int_{\Theta\times\Theta} \|\theta - \theta'\|^2 \,\mathrm{d}\gamma^*(\theta, \theta'),$$

i.e. the infimum in (7) is achieved at γ^*.

Wasserstein Geodesics, and (strong) convexity along geodesics. A (constant-speed Wasserstein) geodesics connecting ρ_0 and ρ_1 is a curve $\{\rho_t : 0 \le t \le 1\}$ on $\mathscr{P}(\Theta)$, such that there exists $\gamma^* \in \Pi_o(\rho_0, \rho_1)$ satisfying $\rho_t = (\pi_t)_\# \gamma^*$. A functional $\mathcal{F}$ is λ-*strongly convex along geodesics* if

$$\mathcal{F}(\rho_t) \le (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \frac{\lambda}{2} t(1-t)\, W_2^2(\rho_0, \rho_1)$$

holds for any geodesics $\{\rho_t : 0 \le t \le 1\}$ and $t \in [0,1]$.

Non-convexity along geodesics of NPMLE. Recall that given n observations $X^n = (X_1, \cdots, X_n)$, NPMLE is defined as

$$\widehat{P}_n = \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \mathcal{L}_n(\rho), \quad \text{where} \quad \mathcal{L}_n(\rho) \coloneqq -\frac{1}{n}\sum_{i=1}^n \log\Big(\int_\Theta p(X_i \,|\, \theta)\, \mathrm{d}\rho(\theta)\Big) \tag{8}$$

The objective functional $\mathcal{L}_n$ may not be geodesically convex. Consider $p(\cdot \,|\, \theta) = \mathcal{N}(\theta, 1)$, $\rho_0 = \mathcal{N}(0,1)$ and $\rho_1 = \mathcal{N}(0, 25)$. Since both ρ_0 and ρ_1 are Gaussian distributions, the optimal transport map from ρ_0 to ρ_1 is $T(\theta) = 5\theta$, and thus the geodescis connecting ρ_0 and ρ_1 is $\rho_t = \mathcal{N}(0, (1+4t)^2)$. In this case, we have

$$\mathcal{L}_n(\rho_t) = \frac{\sum_{i=1}^n X_i^2}{2[1+(1+4t)^2]} + \frac{n}{2}\log\big[2\pi\big(1+(1+4t)^2\big)\big].$$

When $\rho^* = \delta_0$ is the point mass, $\frac{1}{n}\mathcal{L}_n(\rho_t) \to \frac{1}{2}\log[2\pi(1+(1+4t)^2)] + \frac{1}{2[1+(1+4t)^2]}$ by law of large numbers. This function is not convex on $[0,1]$. Similar result of non-convexity is numerically verified by Yan et al. (2023).

A.2 Extension from KL to General Bregman Divergences

Let $\Phi : \mathscr{P}(\Theta) \to \mathbb{R} \cup \{+\infty\}$ be a (L_2-)convex functional with first variation $\frac{\delta\Phi}{\delta\rho}$. Define the associated Bregman divergence as

$$D_\Phi(\rho, \rho') \coloneqq \Phi(\rho) - \Phi(\rho') - \int_\Theta \frac{\delta\Phi}{\delta\rho}(\rho')\, \mathrm{d}(\rho - \rho').$$

Bregman divergence is always nonnegative due to the convexity of Φ. In the implicit proximal mirror descent algorithm (with respect to the Bregman function Φ) on the space of all probability distributions, given $\rho_0 \in \mathscr{P}(\Theta)$ such that $\Phi(\rho_0)$ is finite, we iteratively solve

$$\rho_k = \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \mathcal{F}(\rho) + \frac{1}{\tau_k} D_\Phi(\rho, \rho_{k-1}), \quad k \ge 1.$$

When $\Phi(\rho) = \int \rho \log \rho$ for $\rho \in \mathscr{P}^r(\Theta)$ and $\Phi(\rho) = +\infty$ for $\rho \notin \mathscr{P}^r(\Theta)$, it is easy to check that $D_\Phi(\rho, \rho') = D_{\mathrm{KL}}(\rho \,\|\, \rho')$.

χ^2-divergence is not a Bregman divergence. Recall that the χ^2-divergence between two probability distributions are

$$\chi^2(\rho, \rho') = \int_\Theta \Big(\frac{\rho(\theta)}{\rho'(\theta)} - 1\Big)^2 \mathrm{d}\rho(\theta).$$

If there exists a convex functional Φ, such that $\chi^2(\rho, \rho') = D_\Phi(\rho, \rho')$ for all $\rho, \rho' \in \mathscr{P}^r(\Theta)$. Let $\rho' = \mathcal{N}(0,1)$, we have

$$\begin{aligned}\Phi(\rho) &= \chi^2(\rho, \rho') + \Phi(\rho') + \int_\Theta \frac{\delta\Phi}{\delta\rho}(\rho')\, \mathrm{d}(\rho - \rho') \\ &= \Phi(\rho') + \int_\Theta \frac{\delta\Phi}{\delta\rho}(\rho')\, \mathrm{d}(\rho - \rho') + \int_\Theta \Big(\frac{\rho(\theta)}{\rho'(\theta)} - 1\Big)^2 \frac{\rho(\theta)}{\rho'(\theta)}\, \mathrm{d}\rho'.\end{aligned}$$

Note that the first two terms are linear in ρ, while the last term is not convex with respect to ρ. Therefore, there is no convex functional Φ such that $\chi^2(\rho, \rho') = D_\Phi(\rho, \rho')$.

Renyi's α-divergence is not a Bregman divergence. Recall that Renyi's α-divergence (Li and Turner, 2016) is defined as

$$R_\alpha(\rho,\rho') = \frac{1}{\alpha-1}\log\int_\Theta \rho(\theta)^\alpha \rho'(\theta)^{1-\alpha}\,\mathrm{d}\theta, \alpha\in(0,1).$$

If there exists a convex functional Φ, such that $R_\alpha(\rho,\rho') = D_\Phi(\rho,\rho')$ for all $\rho,\rho'\in\mathscr{P}^r(\Theta)$. Then

$$\Phi(\rho) = \Phi(\rho') + \int_\Theta \frac{\delta\Phi}{\delta\rho}(\rho')\,\mathrm{d}(\rho-\rho') + \frac{1}{\alpha-1}\log\int_\Theta \rho(\theta)^\alpha\rho'(\theta)^{1-\alpha}\,\mathrm{d}\theta. \tag{9}$$

Taking the first variation on both sides of (9) yields

$$\frac{\delta\Phi}{\delta\rho}(\rho)(\theta) = \frac{\delta\Phi}{\delta\rho}(\rho')(\theta) + \frac{1}{\alpha-1}\cdot\frac{\alpha\rho(\theta)^{\alpha-1}\rho'(\theta)^{1-\alpha}}{\int_\Theta \rho(\theta)^\alpha\rho'(\theta)^{1-\alpha}\,\mathrm{d}\theta}.$$

Taking this expression of $\frac{\delta\Phi}{\delta\rho}(\rho)$ back to (9) yields contradiction.

B MORE LITERATURE REVIEW

Mirror descent. Mirror descent for convex optimization in the Euclidean space was originally proposed by Nemirovskij and Yudin (1983). It is established that the mirror descent algorithm achieves a $O(k^{-1/2})$ convergence rate when dealing with a non-smooth convex objective function that possesses a uniformly bounded subgradient; this rate can be enhanced to $O(k^{-1})$ when the function is relatively smooth with respect to the Bregman divergence (Bauschke et al., 2017). When the objective function is convex and has Lipschitz gradients, Krichene et al. (2015) demonstrate that the accelerated mirror descent converges at a rate of $O(k^{-2})$. For additional details on mirror descent algorithms in the Euclidean space, we refer the reader to the monographs (Beck, 2017; Lan, 2020; Wright and Recht, 2022).

Stochastic proximal (mirror) descent. Stochastic proximal descent type algorithms have been shown to be more stable than stochastic gradient type algorithms (Ryu and Boyd, 2014) when optimizing a function in the Euclidean space. However, they have been less extensively studied compared to the latter. Considering a scenario where the random objective function is restricted strongly convex, Ryu and Boyd (2014) demonstrate that the expected L_2-distance between each iterate and the minima of the objective function converges exponentially fast, up to a constant factor. In cases where the objective function is convex, Asi and Duchi (2019) establish that the expected value of the objective function evaluated at each iterate approaches its global minimum at a polynomial rate. This is under the condition that the L_2-norm of the derivative of the stochastic objective function has uniformly bounded expected values. In contrast, Bertsekas (2011) shows that in a bounded search space with a one-sided Lipschitz continuous objective function, the expected number of iterations needed to achieve ε-accuracy, up to a fixed constant, is of the order $O(\varepsilon^{-1})$. When it comes to stochastic proximal mirror descent, Davis et al. (2018) prove a polynomial convergence rate for the expected value of the objective function across iterations, given a similar condition of one-sided Lipschitz continuity with respect to the square root of Bregman divergence.

Algorithms for optimizing functional on the space of probability distributions. Assuming the log-Sobolev inequality is satisfied, Chizat (2022) and Nitanda et al. (2022) demonstrate an exponential convergence rate for minimizing the entropic regularized objective functional across the space of probabilities using mean-field Langevin dynamics. Chizat (2022) further shows that the unregularized objective functional approaches its minimum at a rate of $O(\frac{\log\log t}{\log t})$, achieved by decreasing the regularization parameter at a rate of $O(\frac{1}{\log t})$ through an annealing argument. A similar annealing approach is employed in (Chizat et al., 2022), transforming trajectory inference problems into functional optimization problems. In a different vein, Kent et al. (2021) introduce the Frank–Wolfe algorithm in the space of probabilities, inspired by distributionally robust optimization approaches.

C MORE COMPUTATIONAL DETAILS AND NUMERICAL RESULTS

In this appendix, we provide more details about our use of the normalizing flow for implementing the proposed IKLPD algorithm and the setup of the three numerical examples in the main paper. We also provide additional

numerical results about: 1. the impact of step size τ_k on the inner/outer loop convergence of the IKLPD algorithm; 2. the teacher-student architecture for maintaining a fixed storage budget when composing short normalizing flows as the number of (outer loop) iterations increases. We conducted all experiments using the NVIDIA Tesla T4 GPU available on Google Colab.

C.1 Implementation via Normalizing Flows

Recall that the implicit KL proximal descent (IKLPD) algorithm minimizes the objective functional $\mathcal{F}$ by iteratively solving the subproblem

$$\rho_k = \operatorname*{argmin}_{\rho \in \mathscr{P}(\Theta)} \mathcal{F}(\rho) + \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1}),\ k \geq 1, \tag{10}$$

with an initialization $\rho_0 \in \mathscr{P}^r(\Theta)$ and step size $\{\tau_k : k \geq 1\}$. The main idea of using normalizing flow (NF) to solve (10) is to express ρ_k through a map $T^{(k)} : \Theta \to \Theta$ and the initialization ρ_0, which can be easily sampled from, by letting $\rho_k = T^{(k)}_{\#}\rho_0$. A closed-form expression of the density of $\log \rho_k$ derived from the normalizing flow enables exact computation of $D_{\mathrm{KL}}(\rho_k \,\|\, \rho_{k-1})$ through $T^{(k)}$ and $T^{(k-1)}$. Specifically, when the map $T^{(k)}$ is invertible and differentiable (which is satisfied by NF), if we denote the Jacobian matrix of $T^{(k)}$ by $J_{T^{(k)}}$, then the change of variable formula implies

$$\rho_k(\theta) = \rho_0\big((T^{(k)})^{-1}(\theta)\big)\big|\det J_{T^{(k)}}\big((T^{(k)})^{-1}(\theta)\big)\big|^{-1} \tag{11}$$

In practice, the reparametrization trick can be employed to simplify the numerical computation. Concretely, let $\tilde{\rho}_k$ be empirical distribution of M particles $\theta_1^{(k)}, \cdots, \theta_M^{(k)}$ sampled from $\rho_k = T^{(k)}_{\#}\rho_0$. By applying (11), the objective functional in (10) can be approximated by

$$\mathcal{F}_k \coloneqq \mathcal{F}(\tilde{\rho}_k) + \frac{1}{M\tau_k}\sum_{j=1}^{M}\left[\log\frac{\rho_0\big((T^{(k)})^{-1}(\theta_j^{(k)})\big)}{\rho_0\big((T^{(k-1)})^{-1}(\theta_j^{(k)})\big)} - \log\frac{\big|\det J_{T^{(k)}}\big((T^{(k)})^{-1}(\theta_j^{(k)})\big)\big|}{\big|\det J_{T^{(k-1)}}\big((T^{(k-1)})^{-1}(\theta_j^{(k)})\big)\big|}\right]. \tag{12}$$

Computing (12) requires efficient computation of the inverse maps of $T^{(k)}$ and $T^{(k-1)}$, which makes NF an appropriate choice for modeling these maps. An NF model with length L is a map composed of L invertible transformations $T_1, \cdots, T_L$, the inverse of which can be easily calculated. The invertibility of the NF model is guaranteed by the invertibility of these transformations. We choose the NF model with Real-NVP architecture (Dinh et al., 2016), where the transformations $\{T_l : 1 \leq l \leq L\}$ are affine coupling blocks. See Algorithm 1 for a summary of this straightforward implementation of IKLPD using NF via the Adam optimizer. In Appendix C.5 below, we present a computationally efficient method for sequentially stacking local, short normalizing flows, learned within each inner-loop iteration, to form a global, layered normalizing flow for approximating the target solution ρ^*. Additionally, we conduct a numerical experiment to compare this compositional scheme via a teacher-student architecture with Algorithm 1, which re-trains a long normalizing flow for each subproblem.

C.2 More Implementation Details of Examples in the Paper

Gaussian location mixture model. We consider a two-dimensional Gaussian location mixture model with the parameter space $\Theta = \mathbb{R}^2$. The true distribution P^* of the parameter θ is a bimodal two moon distribution (Stimper et al., 2023). The conditional distribution in NPMLE is $p(\cdot \,|\, \theta) = \mathcal{N}(\theta, I_2)$, and $n = 5000$ samples are generated from the model

$$\theta_i \overset{\text{i.i.d.}}{\sim} P^* \quad \text{and} \quad X_i \,|\, \theta_i \sim \mathcal{N}(\theta_i, I_2), \quad i = 1, 2, \cdots, n. \tag{13}$$

In our method, the NF model consists of 30 affine coupling blocks and each block contains two hidden layers with width 256. The initialization is $\rho_0 = \mathcal{N}(0, 4I_2)$, and $M = 3000$ particles are generated to approximate the probability measure ρ_k in each iteration. The outer iteration is run $N_1 = 25$ times with the step size $\tau_k = 5 \times \beta_2^{k-1}$ where the increase factor is $\beta_2 = 1.15$. In the k-th outer iteration, the subproblem (12) with $\mathcal{F} = \mathcal{L}_n$ defined in (8) is optimized via Adam optimizer with the initialized learning rate $\gamma_k = 10^{-4}\beta_1^{k-1}$ and the rate decay factor $\beta_1 = 0.912$ for at most $N_2 = 1000$ inner iterations. The inner loop stops early if the L_2-norm of the gradient of the parameters in the NF model $T^{(k)}$ reaches the threshold 10^{-4} or stops decreasing for 200 consecutive inner iterations.

Algorithm 1 Implementing IKLPD with Normalizing Flows

Require: data $X^n = (X_1, \cdots, X_n)$; initialized NF model $T^{(0)}$; number of particles M; initialization ρ_0; learning rate of Adam optimizer γ; step size τ; number of outer iterations N_1; number of inner iterations N_2; the decay factor of the learning rate in Adam β_1; the increase factor of the step size β_2

Sample M particles $\underset{\sim}{\theta}^{(0)} = [\theta_1^{(0)}, \cdots, \theta_M^{(0)}]$ from ρ_0.
for $k = 1$ to N_1 **do**
  Initialize $T^{(k)}$ as $T^{(k-1)}$.
  Compute the learning rate $\gamma_k = \gamma \cdot \beta_1^{k-1}$.
  Compute the step size $\tau_k = \tau \cdot \beta_2^{k-1}$.
  for $r = 1$ to N_2 **do**
    $\underset{\sim}{\theta}^{(k)} = T^{(k)}(\underset{\sim}{\theta}^{(0)})$.
    Compute the loss $\mathcal{F}_k = \mathcal{F}_k(\underset{\sim}{\theta}^{(k)}, \tau_k)$ in (12) with $\underset{\sim}{\theta}^{(k)}$.
    Update $T^{(k)}$ based on the loss $\mathcal{F}_k$ using Adam optimizer with learning rate γ_k.
  end for
end for

In the stochastic variant of NF, a randomly subsampled mini-batch of size $m = 500$ from samples X^n is used to compute the stochastic gradient during the training of the NF. Different from the deterministic NF, the increase factor is $\beta_2 = 1$, and the initialized learning rate in Adam optimizer is $1/(1 + k/27)$, which decays along outer iterations. All other settings are same as the ones in the deterministic NF model.

In this experiment, our methods are compared with the KW method (Koenker and Mizera, 2014) and the Wasserstein–Fisher–Rao (WFR) gradient flow (Yan et al., 2023). In the KW method, the probability measure is approximated by a discrete probability distribution supported on a fixed grid. Each grid point can be viewed as a particle with a fixed location, and the goal is to minimize $\mathcal{L}_n$ by finding the optimal weights of these particles, which can be achieved by applying Algorithm 2 in (Yan et al., 2023); this algorithm updates the weights of the particles by explicitly discretizing the Fisher–Rao gradient flow. On the other hand, both the locations and the weights are updated in WFR method by discretizing the WFR gradient flow through particles.

In both of these two methods, the step size is $\tau = 1$, as it is the largest step size to ensure that these methods converge. In the KW method, by letting $L = \|X^n\|_\infty$, probability distributions are approximated by a discrete probability distribution supported on a fixed and equally spaced grid on $[-L, L]^2$ with total 3025 grid points, and the mass on each grid is updated to minimize the functional loss $\mathcal{L}_n$ via Algorithm 2 in (Yan et al., 2023). In the WFR method, we directly use Algorithm 1 in (Yan et al., 2023) with the same initialization ρ_0 and the number of particles M as in our method.

Gaussian location scale mixture model. We consider a d-dimensional Gaussian location scale mixture model with parameters $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}^d \times \mathbb{R}_+^d$. The conditional distribution is $p(\cdot \,|\, \theta) = \mathcal{N}(\mu, \Sigma)$ with $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_d^2)$, and the true joint mixing distribution is $P^* = P_\mu^* \otimes P_{\sigma^2}^*$. We consider two settings. In Setting 1, we let $d = 2$ and P_μ^* be the bimodal two moon distribution. In Setting 2, we let $d = 3$ and P_μ^* be the tensor product of a bimodal two moon distribution for the first two coordinates of μ and a standard normal distribution for the last coordinate of μ. In both settings, we set $P_{\sigma^2}^*$ to be the joint distribution of d independent χ^2 distributions with degree of freedom 1, and the sample size is $n = 5000$.

In our methods, we use an NF model with $2d$-dimensional inputs and outputs, where the first d dimensions represent location parameters and the last d-dimensions represent scale parameters. The NF model consists of 30 affine coupling blocks and each block contains two hidden layers with width 64. The initialization is $\rho_0 = \mathcal{N}(0, 4I_d)$. M particles are generated to approximate the probability measure ρ_k in each iteration, and we choose $M = 2041$ in Setting 1 and $M = 4096$ in Setting 2. The outer iteration is run $N_1 = 50$ times. All other hyperparameters and the stopping criterion of the inner loop in deterministic NF and stochastic NF are same as in the experiments of Gaussian location mixture models.

In the experiment, our methods are compared with the KW method. By letting $L = \|X^n\|_\infty$, probability distributions are approximated by a discrete probability distribution supported on a fixed and equally spaced grid on $[-L, L]^d \times [0.01, 4]^d$ with total $M = 2041$ grid points in Setting 1 and $M = 4096$ grid points in Setting

2. The mass on each grid is updated to minimize the functional loss $\mathcal{L}_n$ via Algorithm 2 in (Yan et al., 2023) with step size 1.

Bayesian posterior sampling. The goal is to minimize the KL divergence $\mathcal{F}(\rho) = D_{\mathrm{KL}}(\rho \,\|\, \pi)$, where the target distribution $\pi(\theta) \propto e^{-\frac{1}{2\alpha}\|\theta\|^{2\alpha}}$ is known up to a normalization constant and $\theta \in \Theta = \mathbb{R}^2$. We consider two settings with $\alpha = 2$ in Setting 1 and $\alpha = 3$ in Setting 2. In our method, the NF model consists of 20 affine coupling blocks and each block contains two hidden layers with width 64. With initialization $\rho_0 = \mathcal{N}(0, 9I_2)$ in Setting 1 and $\rho_0 = \mathcal{N}(0, 4I_2)$ in Setting 2, $M = 1000$ particles are generated to approximate the probability measure ρ_k in each iteration. The outer iteration is run $N_1 = 25$ times with the step size $\tau_k = 5$ for all $k \geq 1$ (i.e. the increase factor is $\beta_2 = 1$). In the k-th outer iteration, the subproblem (12) with $\mathcal{F} = D_{\mathrm{KL}}(\cdot \,\|\, \pi)$ is optimized via Adam optimizer with the initialized learning rate $\gamma_k = 10^{-4}\beta_1^{k-1}$ and the rate decay factor $\beta_1 = 0.912$ for $N_2 = 1000$ inner iterations. Our method is compared with Langevin dynamics, where $M = 1000$ particles are generated from the same initialization as in our method and updated by an explicit discretization of Langevin dynamics,

$$\theta_j^{(k)} = \theta_j^{(k-1)} - \Delta t \big\|\theta_j^{(k-1)}\big\|^{(2\alpha-2)}\theta_j^{(k-1)} + \sqrt{2\Delta t}\, u_j^{(k)}, \quad j = 1, \cdots, M,$$

where $u_j^{(k)}$ are i.i.d. samples generated from $\mathcal{N}(0, I_2)$. We set $\Delta t = 10^{-2}$ in Setting 1 and $\Delta t = 4 \cdot 10^{-4}$ in Setting 2, as they are the largest Δt to ensure that the discretized Langevin dynamics does not diverge and therefore leads to the fastest convergence possible.

C.3 Comparisons with Wasserstein gradient flows

In this subsection, we compare the numerical performance between our method and some representative methods based on discretizing a WGF, namely, the ICNN-JKO method (Mokrov et al., 2021) and the Euler–Maruyama (EM) method with 10^3, 10^4, $5 \cdot 10^4$ particles (EM-1K, EM-10K, EM-50K shown in the figures). The ICNN-JKO method numerically solves the JKO scheme by using input-convex neural networks (ICNNs) to approximate the optimal transport map between two consecutive iterates. The EM method approximates the solution of Langevin dynamics by time discretization. We follow the same experiment setting as the Section 4.1 in (Mokrov et al., 2021). Precisely, let ρ^* be the target distribution and D be the dimension. We consider a random Gaussian mixture model $\frac{1}{M}\sum_{m=1}^{M}\mathcal{N}(\mu_m, I_D)$ with M mixtures, where $\mu_1, \cdots, \mu_M \sim \mathrm{Uniform}([-5, 5]^D)$. We let the dimension D take all even numbers from 2 to 10 and let M increases with D. Specifically we take $M = 5$ when $D = 2$, $M = 6$ when $D = 4$, $\cdots$, and $M = 10$ when $D = 12$. The goal is to minimize the KL divergence $\mathcal{F}(\rho) = D_{\mathrm{KL}}(\rho \,\|\, \rho^*)$. To qualitatively compare numerical results, we use the KL divergence and symmetric KL divergence defined as follows

$$\mathrm{SymKL}(\rho_k, \rho^*) = D_{\mathrm{KL}}(\rho_k \,||\, \rho^*) + D_{\mathrm{KL}}(\rho^* \,||\, \rho_k), \tag{14}$$

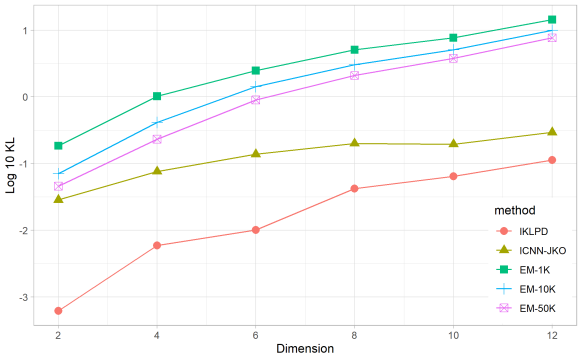
where ρ_k is distribution learned by the NF model after finishing the k-th outer step.

In our method, the NF model consists of 30 affine coupling blocks and each block contains three hidden layers with width 256. The initialization is $\rho_0 = \mathrm{Uniform}([-10, 10]^D)$. In each iteration, we use $M = 10000$ particles to approximte the probability measure ρ_k. The numbers of outer iterations and inner iterations are $N_1 = 20$ and $N_2 = 6000$ respectively. For the first 10 outer iterations $k \in \{1, 2, \cdots, 10\}$, we set the step size $\tau_k = 0.1$ (the increase factor $\beta_2 = 1$). For the last 10 outer iterations $k \in \{11, 12, \cdots, 20\}$, we set the step size $\tau_k = 0.1 \cdot 1.3^{k-11}$ (the increase factor $\beta_2 = 1.3$). In the k-th outer iteration, the subproblem (12) with $\mathcal{F} = D_{\mathrm{KL}}(\cdot \,\|\, \rho^*)$ is optimized via Adam optimizer with the initial learning rate $\gamma_k = 10^{-5}$ (the increase factor $\beta_1 = 1$) and the learning rate reduce by 0.5 every 2000 inner iteration steps for Adam optimizer. The ICNN-JKO method and the EM method follow the same setting of the experiment 4.1 in (Mokrov et al., 2021). Figure 4 reports the comparing results. As we can see, although all methods tend to incur a higher approximation error as the dimension D grows, our method (IKLPD) consistently attained the best accuracy.

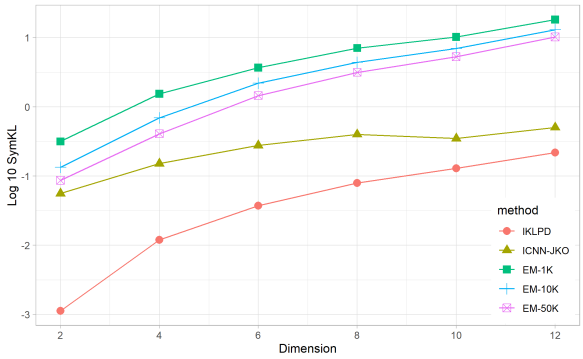
C.4 Impact of IKLPD Step Size τ_k

Recall that the first variation of $\mathcal{L}_n$ in the NPMLE problem (8) at a probability measure ρ is the map

$$\frac{\delta \mathcal{L}_n}{\delta \rho}(\rho) : \theta \mapsto -\frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i \,|\, \theta)}{\int_\Theta p(X_i \,|\, \theta)\,\mathrm{d}\rho(\theta)}. \tag{15}$$



(a) KL divergence between the computed and the target measure in $D = 2, 4, \cdots, 12$.



(b) SymKL between the computed and the target measure in $D = 2, 4, \cdots, 12$.

Figure 4: Performance metrics across dimensions for different methods (over 3 independent trials), highlighting the superiority of our approach.

For any (local) minimum ρ of $\mathcal{L}_n$, the first-order optimality condition (FOC) implies that $\frac{\delta \mathcal{L}_n}{\delta \rho}(\rho)$ is a constant on the support of ρ almost everywhere. In the experiments, we use the variance of first variation $\mathrm{Var}_{\theta \sim \rho}\left(\frac{\delta \mathcal{L}_n}{\delta \rho}(\rho)(\theta)\right)$ to characterize the closeness of $\frac{\delta \mathcal{L}_n}{\delta \rho}(\rho)$ to a constant. In the k-th iteration, this variance at ρ_k can be approximated by the sample variance of

$$\left\{ -\frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i \,|\, \theta_j^{(k)})}{\frac{1}{M}\sum_{j=1}^{M} p(X_i \,|\, \theta_j^{(k)})} : 1 \le j \le M \right\}$$

given M particles $\theta_1^{(k)}, \cdots, \theta_M^{(k)}$ generated from ρ_k. When the sample variance is smaller than a threshold ζ at some iteration k, we choose ρ_k as the final solution of the NPMLE problem.

Similarly, since the first variation $\mathcal{L}_n(\rho) + \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1})$ is

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i \,|\, \theta)}{\int_\Theta p(X_i \,|\, \theta)\, \mathrm{d}\rho(\theta)} + \frac{1}{\tau_k} \log \frac{\rho(\theta)}{\rho_{k-1}(\theta)},$$

the variance of this first variation of the subproblem at ρ_k can be approximated by the sample variance of

$$\left\{ -\frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i \,|\, \theta_j^{(k)})}{\frac{1}{M}\sum_{j=1}^{M} p(X_i \,|\, \theta_j^{(k)})} + \frac{1}{\tau_k} \log \left(\frac{\rho(\theta_j^{(k)})}{\rho_{k-1}(\theta_j^{(k)})} \right) : 1 \le j \le M \right\}.$$

When this sample variance is smaller than a threshold ζ_k, the inner loop stops and the current $T^{(k)}$ is used to construct the solution ρ_k of the subproblem through $\rho_k = T_{\#}^{(k)} \rho_0$.

Figure 5 summarizes our numerical results, illustrating the impact of the step size τ_k on the IKLPD algorithm. Here, we report the number of inner loop iterations executed using the Adam optimizer and the outer loop iterations of IKLPD under various constant step sizes $\tau_k \equiv \tau$, employing the stopping criterion based on the aforementioned variance of the first variation. Note that in the implementation, we designate the first two outer iterations as a burn-in period, applying the stopping criterion only after this burn-in; moreover, we include the burn-in period in the total count of outer iterations, resulting in 3 as the smallest possible number of outer iterations. We observe that for small (large) τ values, the subproblem (3) becomes simpler (more complex), requiring fewer (more) inner loop iterations to meet the stopping criterion. However, since each IKLPD step results in smaller (larger) progress, the total number of outer loop iterations correspondingly rises (falls). We note that since the minimal outer loop iteration is 3, the (averaged) outer loop iterations tend to stabilize within the interval $[3, 5]$ for those relatively large τ values with convergent inner loop iterations. Additionally, when τ exceeds a particular threshold (which is 8), the inner loop does not converge within the prescribed upper limit of 5000 iterations. Overall, $\tau = 3$ seems to be the optimal step size that balances the inner and outer loop

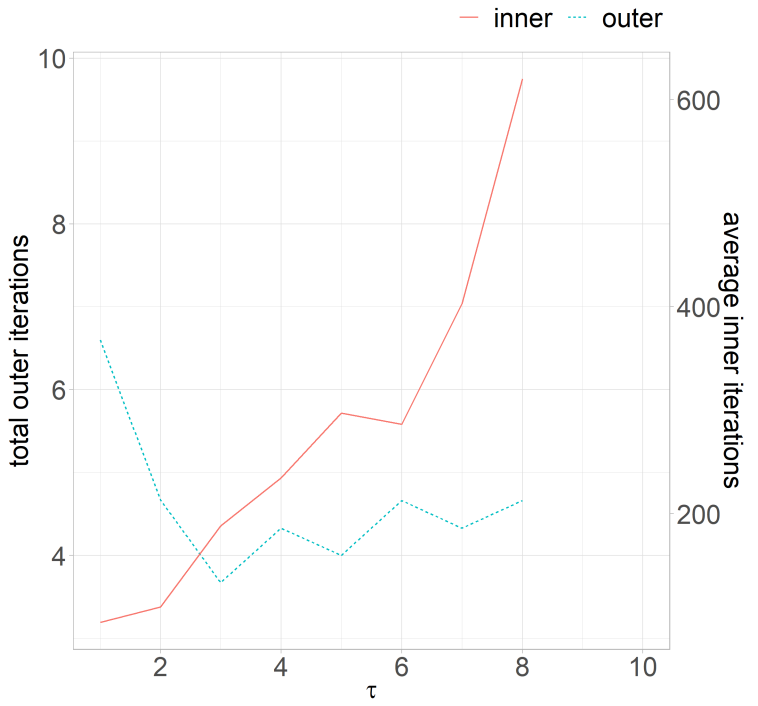


Figure 5: Outer and inner loop iterations (over 5 independent trials) versus step size τ. Here, we designate the first two outer iterations as a burn-in period, applying the stopping criterion only after this burn-in; moreover, we include the burn-in period in the total count of outer iterations, resulting in 3 as the smallest possible number of outer iterations. As we can observe, the implicit step optimization problem (12) becomes easier as the step size τ decreases. This trend is reflected in the lower average number of inner iterations and higher number of outer iterations when the step size τ is smaller. We note that since the minimal outer loop iteration is 3, the (averaged) outer loop iterations tend to stabilize within the interval $[3, 5]$ for those relatively large τ values with convergent inner loop iterations. However, when τ exceeds the threshold 8, the inner loop is unable to converge within a prescribed number of 5000 iterations. For these instances, we have chosen not to plot the corresponding inner and outer iterations.

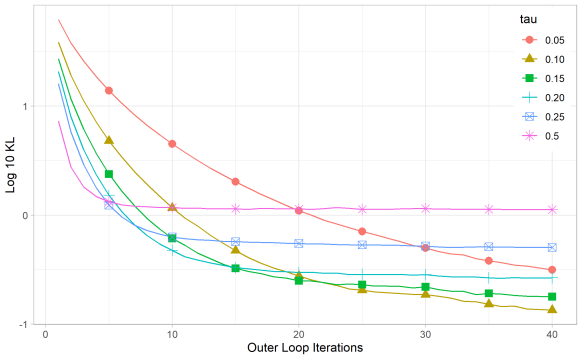
convergences for this particular example. This empirical finding is consistent with the discussion that follows equation (3).

The detailed experiment setup for this numerical study is described as follows. We consider a similar experiment setting as the Gaussian location mixture model. The true mixing distribution P^* is a bimodal two moon distribution with 1.4 times larger distance between two modes than the setting in Appendix C.2. $n = 1000$ samples are generated through the data generating process (13), and the step size $\tau_k = \tau$ is a constant. We use an NF model consisting of 10 affine coupling blocks, and each block contains two hidden layers with width 64. The initialization is $\rho_0 = \mathcal{N}(0, I_2)$, and $M = 1000$ particles are generated to approximate the probability distribution ρ_k at each iteration.

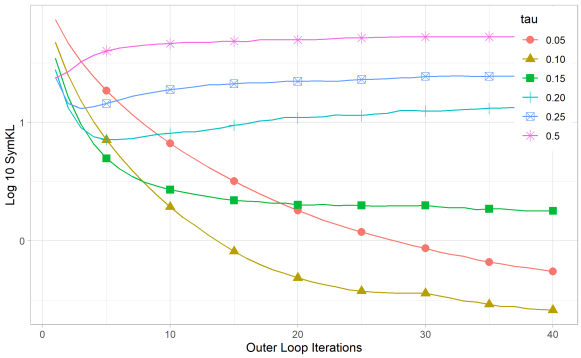
The threshold for the outer iterations is $\zeta = 0.05$. In the k-th outer iteration, the threshold for the inner loop is $\zeta_k = \frac{0.07 \cdot 20}{19+k}$. If this convergence condition of the inner loop is not met within 5000 iterations at the k-th iteration, we claim that the NF model fails find ρ_k due to a overly large choice of the step size τ. The maximum outer iteration is 50. We set the first two outer iterations as burn-in iterations, where the NF model will not be considered to fail to converge for the first two outer iterations if the convergence condition is not met.

We select the outer step size $\tau \in \{1, 2, \ldots, 9, 10\}$. For each τ, the learning rate γ for the Adam optimizer is selected to make the algorithm converge in smallest number of outer iteration. In the k-th outer iteration, the Adam optimizer with the initialized learing rate is $\gamma_k = \frac{20 \cdot \gamma}{19+k}$, which decays along outer iterations. When $\tau = 9$ or 10, for various choices of learning rate γ, the NF model fails to find ρ_k at some iteration k after the burn-in

period.

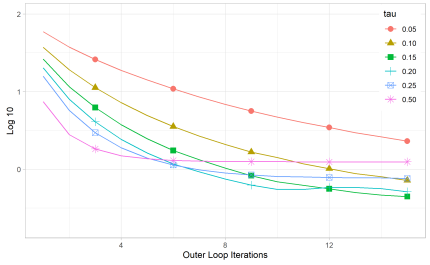


(a) KL divergence between the computed and the target measure for different τ.

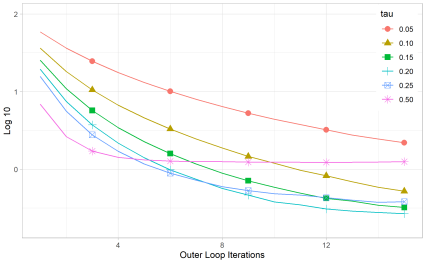


(b) SymKL between the computed and the target measure for different τ.

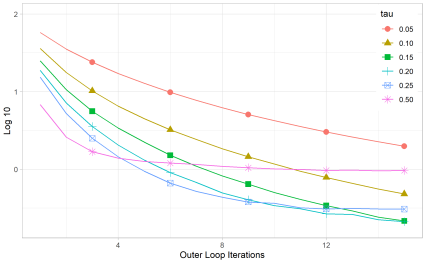
Figure 6: We investigate the impact of the step size τ on the IKLPD algorithm (over 3 independent trials). We set $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.5\}$ while fixing the total number of outer iterations $N_1 = 40$ and the total number of inner iterations $N_2 = 6000$. As the step size τ increases, each outer iteration of the IKLPD method achieves greater progress. Specifically, $\tau = 0.10$ leads to faster convergence of IKLPD compared with $\tau = 0.05$. However, beyond a certain threshold (e.g., $\tau \geq 0.15$), convergence to the global minimum becomes challenging due to the increasing difficulty of solving the implicit step optimization problem (12).



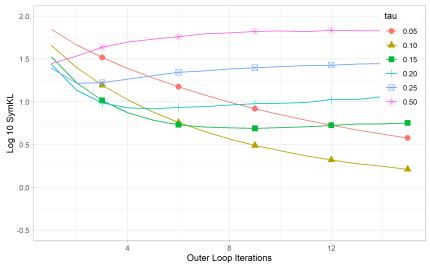
(a) $N_2 = 2000$, KL



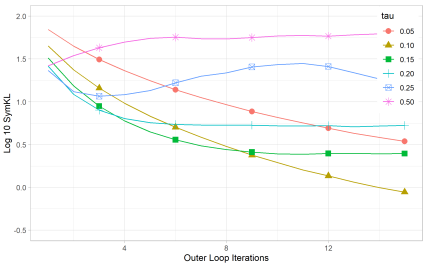
(b) $N_2 = 4000$, KL



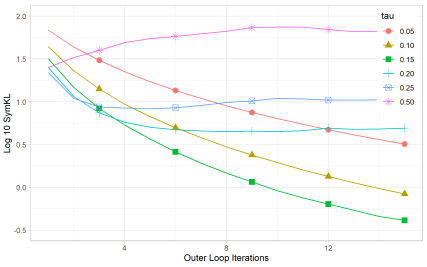
(c) $N_2 = 6000$, KL



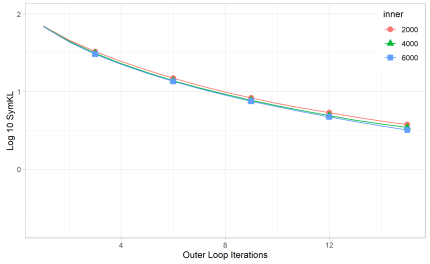
(d) $N_2 = 2000$, Symmetric KL



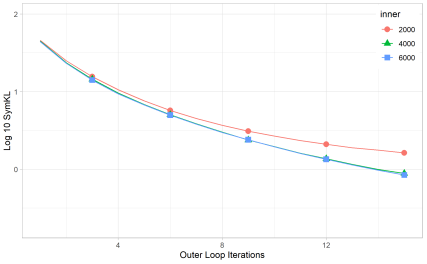
(e) $N_2 = 4000$, Symmetric KL



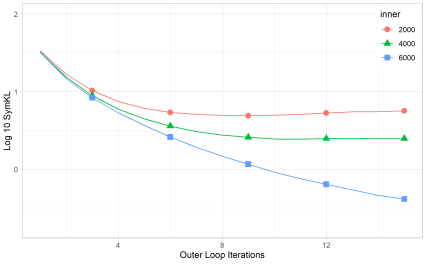
(f) $N_2 = 6000$, Symmetric KL



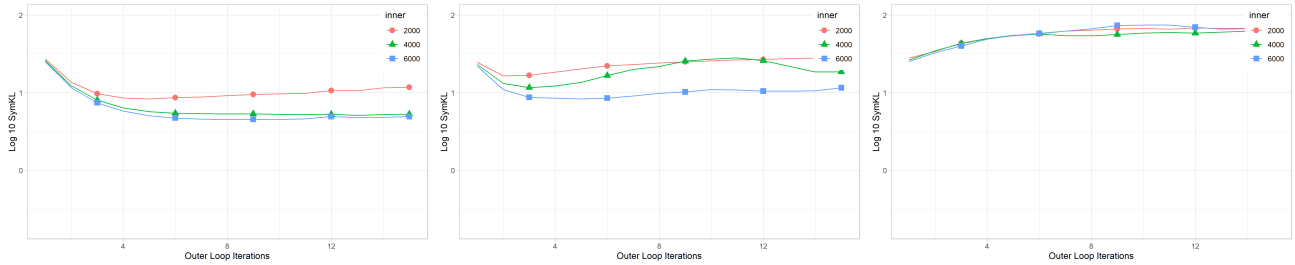
(g) $\tau = 0.05$, Symmetric KL



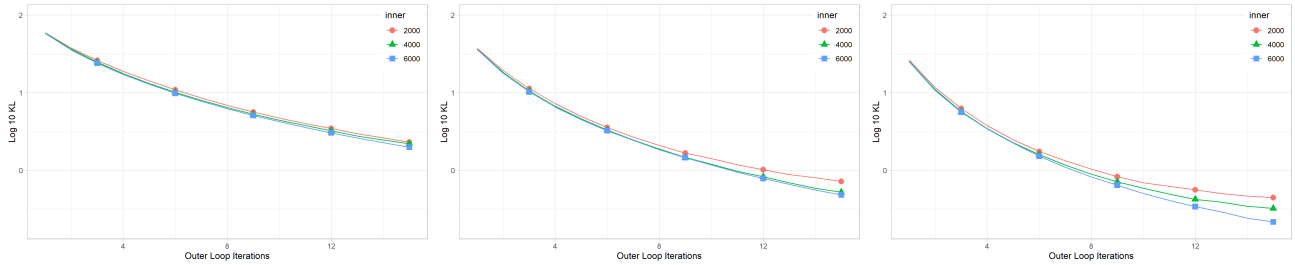
(h) $\tau = 0.10$, Symmetric KL



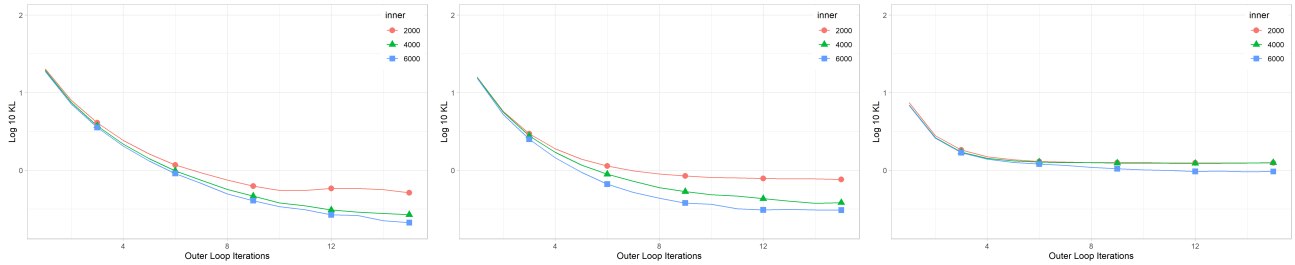
(i) $\tau = 0.15$, Symmetric KL



(j) $\tau = 0.20$, Symmetric KL (k) $\tau = 0.25$, Symmetric KL (l) $\tau = 0.5$, Symmetric KL



(m) $\tau = 0.05$, KL (n) $\tau = 0.10$, KL (o) $\tau = 0.15$, KL



(p) $\tau = 0.20$, KL (q) $\tau = 0.25$, KL (r) $\tau = 0.5$, KL

Figure 7: We explore the performance of IKLPD with a varying step size $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.5\}$ and number of inner loop iterations $N_2 \in \{2000, 4000, 6000\}$, while fixing the number of outer loop iterations $N_1 = 15$. The implicit step optimization problem (12) becomes easier (harder) as τ decreases (increases). The figures demonstrate that, to ensure the convergence of IKLPD to the global minimum, N_2 must increase as the step size τ increases. Specifically, for $\tau = 0.05$, 0.1, and 0.15, the required number of inner iterations for IKLPD to find the global minimum are $N_2 = 2000$, 4000, and 6000, respectively. When $\tau \geq 0.20$, the implicit step optimization problem (12) becomes significantly harder, preventing IKLPD from obtaining a reasonably good solution within a prescribed number of inner loop iterations.

To further explore the impact of the IKLPD step size τ on the convergence of the IKLPD algorithm, we conduct two additional experiments using the problem setting $D = 12$ and $M = 10$, as previously described in C.3. KL and symKL metric represent $D_{\mathrm{KL}}(\rho_k \,\|\, \rho^*)$ and $\mathrm{SymKL}(\rho_k, \rho^*)$ respectively. The IKLPD step size $\tau_k \equiv \tau$ is fixed during the training (i.e. the increase factor $\beta_2 = 1$) in both experiments. In the first experiment as in Figure 6, the training process is the same as the one stated in C.3 except that we choose $N_1 = 40$, $N_2 = 6000$ and explore different values of $\tau = 0.05$, 0.1, 0.15, 0.2, 0.25, 0.5. In the second experiment as in Figure 7, we keep the number of outer iterations $N_1 = 15$ and vary τ. For each τ, we consider three different choices of number of inner iterations $N_2 = 2000$, 4000, and 6000. The training process follows the same configurations as in the first experiment, but we adjust the width of the hidden layers to 512 instead of 256 for the NF model. The results are summarized in Figures 6 and 7, where all the patterns are consistent with our discussions and remarks made in the main paper.

Algorithm 2 IKLPD with composition of short flows and the teacher-student architecture

Require: data $X^n = (X_1, \cdots, X_n)$; initialized NF model $T^{(0)}$; number of particles M; initialization ρ_0; learning rate of Adam optimizer γ; learning rate of Adam optimizer used in the compression process γ'; step size τ; number of outer iterations N_1; number of inner iterations N_2; number of compression iterations N_3; the length of the compressed flow k_0; the maximum length of the flow before compression k_1; the length of the short flow k_2; the compression L_2 loss threshold ϵ; the decay factor of the learning rate in Adam β_1; the increase factor of the step size β_2

Sample M particles $\underset{\sim}{\theta}^{(0)} = [\theta_1^{(0)}, \cdots, \theta_M^{(0)}]$ from ρ_0.
for $k = 1$ to N_1 **do**
  Initialize a short flow $T'^{(k)}$ with length k_2 such that $T'^{(k)}_{\#}\rho_{k-1} = \rho_{k-1}$.
  Set RequiresGrad = False for all parameters in $T^{(k-1)}$. ▷ the parameters in $T^{(k-1)}$ are fixed
  Initialize $T^{(k)}$ as $T^{(k-1)} \circ T'^{(k)}$. ▷ only the parameters in $T'^{(k)}$ could be learned
  Compute the learning rate $\gamma_k = \gamma \cdot \beta_1^{k-1}$.
  Compute the step size $\tau_k = \tau \cdot \beta_2^{k-1}$.
  for $r = 1$ to N_2 **do**
    $\underset{\sim}{\theta}^{(k)} = T^{(k)}(\underset{\sim}{\theta}^{(0)})$.
    Compute the loss $\mathcal{F}_k = \mathcal{F}_k(\underset{\sim}{\theta}^{(k)}, \tau_k)$ in (12) with $\underset{\sim}{\theta}^{(k)}$.
    Update $T^{(k)}$ based on the loss $\mathcal{F}_k$ using Adam optimizer with learning rate γ_k.
  end for
  if Length of the NF model $T^{(k)} > k_1$ **then**
    Initialize a flow $T''^{(k)}$ with length k_0 as $T''^{(k)}_{\#}\rho_0 = \rho_0$.
    for $s = 1$ to N_3 **do**:
      Compute the L_2 loss $L_2(T^{(k)}, T''^{(k)}) := \ell_2(T^{(k)}(\underset{\sim}{\theta}^{(0)}), T''^{(k)}(\underset{\sim}{\theta}^{(0)}))$.
      Update $T''^{(k)}$ based on $L_2(T^{(k)}, T''^{(k)})$ using Adam optimizer with learning rate γ'.
      if $L_2(T^{(k)}, T''^{(k)}) \le \epsilon$ **then**
        Break the current loop.
      end if
    end for
    Let $T^{(k)} = T''^{(k)}$.
  end if
end for

C.5 Composition of Short Flows and Teacher-Student Architecture

Algorithm 2 summarizes the algorithm for the compositional scheme of sequentially stacking the local, short normalizing flows (each with length k_2), learned within each inner-loop iteration, to form a global, layered normalizing flow for minimizing $\mathcal{F}$. In the k-th outer iteration, the total length of the large NF model is $(k+1) \cdot k_2$ since we composite a new length-k_2 short flow with the original NF model with length $k \cdot k_2$. When the length of this compositional NF model exceeds the threshold of maximum length k_1, we employ a teacher-student architecture to distill knowledge from the compositional NF model to a shorter NF model of length k_0. This is achieved by minimizing (a sample version of) the L_2 distance between the larger (teacher) NF model and the smaller, length-k_0 (student) NF model.

Figure 8 provides a numerical comparison between Algorithm 2 that re-trains a long normalizing flow for each subproblem (indicated as `NF`) and Algorithm 2 that uses the compositional scheme and teacher-student architecture (indicated as `NF_ST`). As we can see, the expressive capability of the composited normalizing flow model is comparable to that of the computationally more expensive `NF` method, which re-trains a lengthy normalizing flow at each iteration of the IKLPD algorithm. In addition, by employing the teacher-student architecture, we can preserve a constant storage budget while maintaining the expressive capability of the compositional normalizing flow model.

We describe below the concrete setting of this numerical experiment. We use a similar objective functional as in the Gaussian location mixture models. The NF model consists of 30 affine coupling blocks, and each block contains two hidden layers having 256 units. The step size $\tau_k = \tau$ is fixed (i.e. the increase factor $\beta_2 = 1$),

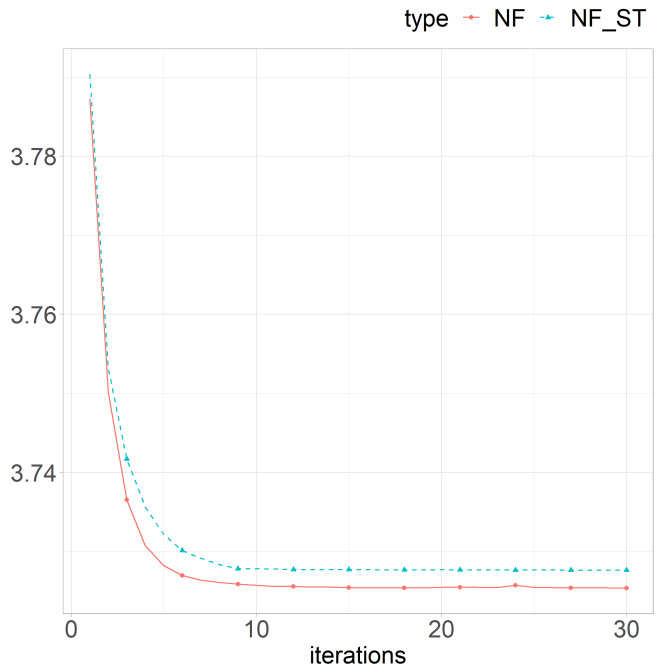


Figure 8: Optimized objective value $\mathcal{L}_n(\rho_k)$ versus iteration count k. We report the averaged $\mathcal{L}_n(\rho_k)$ over 5 independent trials. The results indicate that `NF_ST` and `NF` demonstrate very similar performance. The compression through the teacher-student architecture successfully maintains the expressive capability of the original model.

and no early stopping criterion is applied for the inner loop. All other hyperparameters are the same as in the deterministic NF model in Appendix C.2.

Algorithm 1 is compared with the composition of short NF model with teacher-student architecture (`NF_ST`) as shown in Algorithm 2. In each outer iteration, a short flow with length $k_2 = 4$ is composited with the original flow. Each short flow consists of 4 affine coupling blocks and each block contains two hidden layers with width 512. When the length of the composited flow exceeds the maximum length $k_1 = 40$, it will be compressed into a flow with length $k_0 = 20$. The initialized Adam learning rate is $\gamma = 8 \times 10^{-5}$. The compression process is run at most $N_3 = 3000$ iterations with the Adam learning rate $\gamma' = 10^{-5}$ and stops early if the L_2 distance between the compressed flow and the original composited flow is less than the threshold $\epsilon = 10^{-4}$. Other hyperparameters are the same as the NF model in Appendix C.2

D PROOFS OF THEORETICAL RESULTS

In this appendix, we provide all deferred proofs for the main theoretical results from the main paper.

D.1 Proof of Theorem 1

Taking derivative with respect to time yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t) &= -\frac{\mathrm{d}}{\mathrm{d}t} \int \log \rho_t \,\mathrm{d}\rho^* \\
&= \int_\Theta \Big(\frac{\delta \mathcal{F}}{\delta \rho}(\rho_t)(\theta) - \int_\Theta \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t)(\theta') \,\mathrm{d}\rho_t(\theta') \Big) \,\mathrm{d}\rho^*(\theta) \\
&= \int \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t) \,\mathrm{d}(\rho^* - \rho_t)(\theta) \\
&\le \mathcal{F}(\rho^*) - \mathcal{F}(\rho_t) - \frac{\lambda}{2} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t).
\end{aligned} \tag{16}$$

When $\lambda > 0$, since $\mathcal{F}(\rho^*) - \mathcal{F}(\rho_t) \le 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t) \le -\frac{\lambda}{2} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t).$$

By Gronwall's inequality, we have

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t) \le e^{-\frac{\lambda t}{2}} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

When $\lambda = 0$, (16) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t) \le \mathcal{F}(\rho^*) - \mathcal{F}(\rho_t).$$

This implies

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_t) - D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0) \le t\mathcal{F}(\rho^*) - \int_0^t \mathcal{F}(\rho_s)\,\mathrm{d}s.$$

By Jensen's inequality, we have

$$\mathcal{F}\Big(\frac{1}{t}\int_0^t \rho_s\,\mathrm{d}s\Big) - \mathcal{F}(\rho^*) \le \frac{1}{t}\int_0^t \mathcal{F}(\rho_s)\,\mathrm{d}s - \mathcal{F}(\rho^*) \le \frac{1}{t} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

D.2 Proof of Theorem 2

We need the following lemma to bound the functional value at each iterate, the proof of which is deferred to Section D.6 in this Appendix.

Lemma A1. *For any* $\rho \in \mathscr{P}^r(\Theta)$

$$\mathcal{F}(\rho_k) - \mathcal{F}(\rho) \le \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1}) - \Big(\frac{1}{\tau_k} + \frac{\lambda}{2}\Big) D_{\mathrm{KL}}(\rho \,\|\, \rho_k) - \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho_k \,\|\, \rho_{k-1}).$$

Applying Lemma A1 with $\rho = \rho^*$ yields

$$0 \le \mathcal{F}(\rho_k) - \mathcal{F}(\rho^*) \le \frac{1}{\tau_k} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_{k-1}) - \Big(\frac{1}{\tau_k} + \frac{\lambda}{2}\Big) D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k). \tag{17}$$

When $\lambda > 0$ and $\tau_k = \tau$ for all $k \ge 1$, the above inequality implies

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k) \le \Big(1 + \frac{\lambda\tau}{2}\Big)^{-1} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_{k-1}).$$

Therefore, we have

$$D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k) \le \Big(1 + \frac{\lambda\tau}{2}\Big)^{-k} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

When $\lambda = 0$, (17) implies

$$\tau_k \big[\mathcal{F}(\rho_k) - \mathcal{F}(\rho^*)\big] \le D_{\mathrm{KL}}(\rho^* \,\|\, \rho_{k-1}) - D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k).$$

Summing the above inequality from 1 to k, we have

$$\sum_{l=1}^{k} \tau_l \big[\mathcal{F}(\rho_l) - \mathcal{F}(\rho^*)\big] \le D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0) - D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k).$$

Therefore, we have

$$\min_{1 \le l \le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho^*) \le \frac{1}{\tau_1 + \cdots + \tau_k} D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0).$$

D.3 Proof of Theorem 3

Similar to the proof of Theorem 2, for any $\rho \in \mathscr{P}^r(\Theta)$ we have

$$\min_{1\le l\le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho) \le \frac{1}{\tau_1 + \cdots + \tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_0). \tag{18}$$

By Assumption 3, we have

$$|\mathcal{F}(\rho^*) - \mathcal{F}(\rho)| \le L W_1(\rho^*, \rho).$$

Therefore,

$$\min_{1\le l\le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho^*) \le \frac{1}{\tau_1 + \cdots + \tau_k} D_{\mathrm{KL}}(\rho \,\|\, \rho_0) + L W_1(\rho^*, \rho).$$

Special cases: discrete measures and singular measures supported on hyperplanes. Let ψ_{m_1} be a probability measure on Θ with first-order moment $m_1 = \mathbb{E}_{\theta\sim\psi_{m_1}}\|\theta\| < \infty$. Let $X \sim \rho$ and $Y \sim \psi_{m_1}$. Then $X + Y \sim \rho * \psi_{m_1}$. By definition, we have

$$W_1(\rho, \rho * \psi_{m_1}) = \inf_{\theta\sim\rho, \theta'\sim\rho*\psi_{m_1}} \mathbb{E}\|\theta - \theta'\| \le \mathbb{E}\|X - (X+Y)\| = \mathbb{E}\|Y\| = m_1. \tag{19}$$

This result helps control the smoothing error through W_1-distance.

Case 1: ρ^* is a discrete measure with bounded support. We need the following lemma to control the Gaussian smoothing error in KL divergence. The proof is deferred to Section D.6 in this Appendix.

Lemma A2 (KL divergence bound after Gaussian smoothing)**.** *Assume ρ^* is a discrete probability measure with bounded support. Let $R_\theta = \sup\{\|\theta\| : \theta \in \mathrm{supp}(\rho^*)\}$, and $\rho^\sigma = \rho^* * \mathcal{N}(0, \sigma^2 I_d)$. If $\rho_0 = \mathcal{N}(0, \beta^2 I_d)$, we have*

$$D_{\mathrm{KL}}(\rho^\sigma \,\|\, \rho_0) \le d \log \frac{\beta}{\sigma} + \frac{d\sigma^2 + R_\theta^2}{2\beta^2} - \frac{d}{2}.$$

Note that the first-order moment of $\mathcal{N}(0, \sigma^2 I_d)$ is smaller than $\sqrt{d\sigma^2}$. Applying Lemma A2 and Inequality (19) yields

$$0 \le \min_{1\le l\le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho^*) \le \frac{1}{\tau_1 + \cdots + \tau_k}\Big(d \log \frac{\beta}{\sigma} + \frac{d\sigma^2 + R_\theta^2}{2\beta^2} - \frac{d}{2}\Big) + L\sqrt{d\sigma^2}.$$

Since the above inequality holds for all $\sigma > 0$, by choosing $\sigma^2 = L^{-2}(\tau_1 + \cdots + \tau_k)^{-2}$, we have

$$0 \le \min_{1\le l\le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho^*) \le \frac{d \log[\beta L(\tau_1 + \cdots + \tau_k)] + \frac{d}{2\beta^2 L_2(\tau_1+\cdots+\tau_k)^2} + \frac{R_\theta^2}{2\beta^2} + \sqrt{d} - \frac{d}{2}}{\tau_1 + \cdots + \tau_k}.$$

When $\tau_1 = \cdots = \tau_k = \tau$, the upper bound has order $O(\frac{d \log k}{k})$.

Case 2: ρ^* is absolutely continuous with respect to the Lebesgue measure supported on a d'-dimensional hyperplane. Without loss of generality, assume ρ^* is supported on $\mathrm{supp}(\rho^*) = \{(\theta', 0, \cdots, 0) \in \mathbb{R}^d : \theta' \in \mathbb{R}^{d'}\}$. Let $\rho^*_{d'}$ denote the distribution of ρ^* restricted to the first d' coordinates. Then $\rho^*_{d'} \in \mathscr{P}^r(\mathbb{R}^{d'})$. Assume $Z = (X, Y)$ with $X \in \mathbb{R}^{d'}$ and $Y \in \mathbb{R}^{d-d'}$ such that $(X, 0_{d-d'}) \sim \rho^*$, $Y \sim \mathcal{N}(0, \sigma^2 I_{d-d'})$, and X is independent with Y. Then $X \sim \rho^*_{d'}$ is a continuous random variable in $\mathbb{R}^{d'}$. Similarly, let $Z_0 = (X_0, Y_0) \sim \rho_0 = \mathcal{N}(0, \beta^2 I_d)$, such that $X_0 \sim \mathcal{N}(0, \beta^2 I_{d'})$ and $Y_0 \sim \mathcal{N}(0, \beta^2 I_{d-d'})$. Then, we have

$$P_Z(z) = P_X(x) P_Y(y) \quad \text{and} \quad \rho_0(z) = P_{Z_0}(z) = P_{X_0}(x) P_{Y_0}(y).$$

Note that

$$D_{\mathrm{KL}}(P_Z \,\|\, \rho_0) = D_{\mathrm{KL}}(P_Z \,\|\, P_{Z_0}) = \int \log \frac{P_Z}{P_{Z_0}} \,\mathrm{d}P_Z$$

$$
\begin{aligned}
&= \iint \log \frac{P_X(x)P_Y(y)}{P_{X_0}(x)P_{Y_0}(y)} \,\mathrm{d}P_X(x)\,\mathrm{d}P_Y(y) \\
&= D_{\mathrm{KL}}(P_X \,\|\, \mathcal{N}(0,\beta^2 I_{d'})) + D_{\mathrm{KL}}\big(\mathcal{N}(0,\sigma^2 I_{d-d'}) \,\big\|\, \mathcal{N}(0,\beta^2 I_{d-d'})\big) \\
&= D_{\mathrm{KL}}(P_X \,\|\, \mathcal{N}(0,\beta^2 I_{d'})) + \frac{d-d'}{2}\Big(\log\frac{\beta^2}{\sigma^2} - 1 + \frac{\sigma^2}{\beta^2}\Big).
\end{aligned}
$$

By Theorem 3 and Inequality (19), for every $\sigma^2 > 0$ we have

$$
\min_{1\le l\le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho^*) \le \frac{1}{\tau_1+\cdots+\tau_k}\Big[D_{\mathrm{KL}}(P_X \,\|\, \mathcal{N}(0,\beta^2 I_{d'})) + \frac{d-d'}{2}\Big(\log\frac{\beta^2}{\sigma^2} - 1 + \frac{\sigma^2}{\beta^2}\Big)\Big] + L\sqrt{(d-d')\sigma^2}.
$$

Noting that $P_X = \rho^*_{d'}$, by choosing $\sigma^2 = L^{-2}(\tau_1+\cdots+\tau_k)^{-2}$, the above inequality implies

$$
\min_{1\le l\le k} \mathcal{F}(\rho_l) - \mathcal{F}(\rho^*) \le \frac{(d-d')\log[\beta L(\tau_1+\cdots+\tau_k)] + \frac{d-d'}{2\beta^2 L_2(\tau_1+\cdots+\tau_k)^2} + \sqrt{d-d'} - \frac{d-d'}{2} + D_{\mathrm{KL}}(\rho^*_{d'} \,\|\, \mathcal{N}(0,\beta^2 I_{d'}))}{\tau_1+\cdots+\tau_k}.
$$

When $\tau_1 = \cdots = \tau_k = \tau$, the upper bound has order $O\big(\frac{(d-d')\log k}{k}\big)$.

D.4 Proof of Theorem 4

Recall that

$$
\eta_k(\theta) = \frac{\delta\mathcal{F}}{\delta\rho}(\rho_k^{\mathrm{err}})(\theta) + \frac{1}{\tau_k}\log\frac{\rho_k^{\mathrm{err}}}{\rho_{k-1}^{\mathrm{err}}}(\theta).
$$

Let $\tilde{\eta}_k = \eta_k - \inf_{\theta\in\Theta}\eta_k(\theta)$. Therefore, we have $\|\tilde{\eta}_k\|_\infty \le \varepsilon_k$. Since $\mathcal{F}$ is λ-relative strongly convex, we have

$$
\begin{aligned}
\mathcal{F}(\rho) - \mathcal{F}(\rho_k^{\mathrm{err}}) &\ge \int_\Theta \frac{\delta\mathcal{F}}{\delta\rho}(\rho_k^{\mathrm{err}})(\theta)\,\mathrm{d}(\rho-\rho_k^{\mathrm{err}})(\theta) + \frac{\lambda}{2}D_{\mathrm{KL}}(\rho\,\|\,\rho_k^{\mathrm{err}}) \\
&= \int_\Theta \tilde{\eta}_k(\theta) - \frac{1}{\tau_k}\log\frac{\rho_k^{\mathrm{err}}}{\rho_{k-1}^{\mathrm{err}}}(\theta)\,\mathrm{d}(\rho-\rho_k^{\mathrm{err}})(\theta) + \frac{\lambda}{2}D_{\mathrm{KL}}(\rho\,\|\,\rho_k^{\mathrm{err}}) \\
&= -\frac{1}{\tau_k}D_{\mathrm{KL}}(\rho\,\|\,\rho_{k-1}^{\mathrm{err}}) + \Big(\frac{1}{\tau_k}+\frac{\lambda}{2}\Big)D_{\mathrm{KL}}(\rho\,\|\,\rho_k^{\mathrm{err}}) + \frac{1}{\tau_k}D_{\mathrm{KL}}(\rho_k^{\mathrm{err}}\,\|\,\rho_{k-1}^{\mathrm{err}}) + \int_\Theta \tilde{\eta}_k(\theta)\,\mathrm{d}(\rho-\rho_k^{\mathrm{err}})(\theta).
\end{aligned}
$$

Note that

$$
\int_\Theta \tilde{\eta}_k(\theta)\,\mathrm{d}(\rho-\rho_k^{\mathrm{err}})(\theta) \le \|\tilde{\eta}_k\|_\infty\|\rho-\rho_k^{\mathrm{err}}\|_1 \le \varepsilon_k\cdot\sqrt{2D_{\mathrm{KL}}(\rho\,\|\,\rho_k^{\mathrm{err}})},
$$

where the last inequality is due to Pinsker's inequality. Thus, we have

$$
0 \ge \mathcal{F}(\rho^*) - \mathcal{F}(\rho_k^{\mathrm{err}}) \ge -\frac{1}{\tau_k}D_{\mathrm{KL}}(\rho^*\,\|\,\rho_{k-1}^{\mathrm{err}}) + \Big(\frac{1}{\tau_k}+\frac{\lambda}{2}\Big)D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}}) - \varepsilon_k\sqrt{2D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}})}.
$$

This implies

$$
\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}})} \le \frac{\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_{k-1}^{\mathrm{err}})}}{\sqrt{1+\tau_k\lambda/2}} + \sqrt{2}\tau_k\varepsilon_k.
$$

Therefore,

$$
\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}})} \le \frac{\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0)}}{\prod_{l=1}^k\sqrt{1+\lambda\tau_l/2}} + \sum_{l=1}^k \frac{\sqrt{2}\tau_l\varepsilon_l}{\prod_{s=l+1}^k\sqrt{1+\lambda\tau_s/2}}. \tag{20}
$$

Case 1: When $\varepsilon_k \le \kappa\varepsilon^k$ for some $0<\varepsilon<1, \kappa>0$ and $\tau_k=\tau$,

$$\sum_{l=1}^{k}\frac{\sqrt{2}\tau_l\varepsilon_l}{\prod_{s=l+1}^{k}\sqrt{1+\lambda\tau_s/2}} \le \sum_{l=1}^{k}\frac{\sqrt{2}\tau\kappa\varepsilon^l(1+\lambda\tau/2)^{l/2}}{(1+\lambda\tau/2)^{k/2}} = \frac{\sqrt{2}\tau\kappa}{(1+\lambda\tau/2)^{k/2}}\cdot\sum_{l=1}^{k}\big[\varepsilon\sqrt{1+\lambda\tau/2}\big]^l.$$

We can always assume that $\varepsilon\sqrt{1+\lambda\tau/2}\neq 1$, since if $\varepsilon\sqrt{1+\lambda\tau/2}=1$, we can find $\varepsilon'\in(\varepsilon,1)$, so that $\varepsilon_k\le\kappa(\varepsilon')^k$. Note that

$$\sum_{l=1}^{k}\big[\varepsilon\sqrt{1+\lambda\tau/2}\big]^l \le \begin{cases}\frac{1}{1-\varepsilon\sqrt{1+\lambda\tau/2}}, & \varepsilon\sqrt{1+\lambda\tau/2}<1\\ \frac{[\varepsilon\sqrt{1+\lambda\tau/2}]^{k+1}}{\varepsilon\sqrt{1+\lambda\tau/2}-1}, & \varepsilon\sqrt{1+\lambda\tau/2}>1\end{cases}.$$

Therefore, we have

$$\sum_{l=1}^{k}\frac{\sqrt{2}\tau_l\varepsilon_l}{\prod_{s=l+1}^{k}\sqrt{1+\lambda\tau_s/2}} \le \begin{cases}\frac{\sqrt{2}\tau\kappa}{1-\varepsilon\sqrt{1+\lambda\tau/2}}\cdot\big(1+\frac{\lambda\tau}{2}\big)^{-\frac{k}{2}}, & \varepsilon\sqrt{1+\lambda\tau/2}<1\\ \frac{\sqrt{2}\tau\kappa\varepsilon}{\varepsilon\sqrt{1+\lambda\tau/2}-1}\varepsilon^k, & \varepsilon\sqrt{1+\lambda\tau/2}>1\end{cases}$$

Therefore, there exists $C=C(\tau,\lambda,\varepsilon)>0$, such that

$$\sum_{l=1}^{k}\frac{\sqrt{2}\tau_l\varepsilon_l}{\prod_{s=l+1}^{k}\sqrt{1+\lambda\tau_s/2}} \le C\kappa\max\{\varepsilon,(1+\lambda\tau/2)^{-1/2}\}^k.$$

Combining the above inequality with (20) yields

$$D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}}) \le \frac{C\kappa^2+2D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0)}{(\min\{\varepsilon^{-2},1+\lambda\tau/2\})^k}.$$

Case 2: When $\varepsilon_k=\varepsilon k^{-\alpha}$ for some $\alpha,\varepsilon>0$ and $\tau_k=\tau$ for every $k\ge 1$, we show that $D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k)\lesssim k^{-2\alpha}$. In fact, note that

$$\sum_{l=1}^{k}\frac{\sqrt{2}\tau_l\varepsilon_l}{\prod_{s=l+1}^{k}\sqrt{1+\lambda\tau_s/2}} \le \frac{\sqrt{2}\tau\varepsilon}{(1+\tau\lambda/2)^{k/2}}\sum_{l=1}^{k}\frac{(1+\tau\lambda/2)^{l/2}}{l^\alpha}.$$

We prove that there exists $C=C(\tau,\lambda,\alpha)>0$, such that

$$\sum_{l=1}^{k}\frac{(1+\tau\lambda/2)^{l/2}}{l^\alpha} \le \frac{C(1+\tau\lambda/2)^{k/2}}{k^\alpha}. \tag{21}$$

We use the induction to prove (21). If the statement is correct for k, then

$$\sum_{l=1}^{k+1}\frac{(1+\tau\lambda/2)^{l/2}}{l^\alpha} \overset{\text{(i)}}{\le} \frac{C(1+\tau\lambda/2)^{k/2}}{k^\alpha}+\frac{(1+\tau\lambda/2)^{(k+1)/2}}{(k+1)^\alpha} \overset{\text{(ii)}}{\le} \frac{C(1+\tau\lambda/2)^{(k+1)/2}}{(k+1)}.$$

Here, (i) is by the induction hypothesis, and (ii) is equivalent to

$$\Big(1+\frac{1}{k}\Big)^\alpha C+\sqrt{1+\frac{\tau\lambda}{2}} \le C\sqrt{1+\frac{\tau\lambda}{2}}.$$

The above inequality is true when

$$\Big(1+\frac{1}{k}\Big)^\alpha \le \sqrt{1+\frac{\tau\lambda}{4}}, \quad \text{and} \quad C\ge\frac{\sqrt{1+\tau\lambda/2}}{\sqrt{1+\tau\lambda/2}-\sqrt{1+\tau\lambda/4}}.$$

When $(1+k^{-1})^\alpha > \sqrt{1+\tau\lambda/4}$, i.e. $k < \big[(1+\tau\lambda/4)^{1/2\alpha}-1\big]^{-1}$, we can choose C large enough such that (21) holds. Therefore, by induction, we know (21) is true for all $k \geq 1$ when

$$C = \max\bigg\{\frac{\sqrt{1+\tau\lambda/2}}{\sqrt{1+\tau\lambda/2}-\sqrt{1+\tau\lambda/4}}, \max\Big\{\frac{k^\alpha}{(1+\tau\lambda/2)^{k/2}}\cdot\sum_{l=1}^{k}\frac{(1+\tau\lambda/2)^{l/2}}{l^\alpha} : k < \frac{1}{(1+\tau\lambda/4)^{1/2\alpha}-1}\Big\}\bigg\}.$$

Applying (21) to (20) yields

$$\begin{aligned}\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}})} &\leq \frac{\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0)}}{(1+\lambda\tau/2)^{k/2}} + \frac{\sqrt{2}\tau\varepsilon}{(1+\tau\lambda/2)^{k/2}}\cdot\frac{C(1+\tau\lambda/2)^{k/2}}{k^\alpha}\\ &= \frac{\sqrt{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0)}}{(1+\lambda\tau/2)^{k/2}} + \frac{\sqrt{2}C\tau\varepsilon}{k^\alpha}.\end{aligned}$$

Therefore, we have

$$D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{err}}) \leq \frac{2D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0)}{(1+\lambda\tau/2)^k} + \frac{C\varepsilon^2}{k^{2\alpha}}$$

for some $C = C(\tau,\lambda,\alpha)$.

D.5 Proof of Theorem 5

By applying Lemma A1, we have

$$\Big(\frac{1}{\tau_k}+\frac{\lambda}{2}\Big)D_{\mathrm{KL}}(\rho\,\|\,\rho_k^{\mathrm{stoc}}) - \frac{1}{\tau_k}D_{\mathrm{KL}}(\rho\,\|\,\rho_{k-1}^{\mathrm{stoc}}) \leq \mathcal{F}_{\xi_k}(\rho) - \mathcal{F}_{\xi_k}(\rho_k^{\mathrm{stoc}}) - \frac{1}{\tau_k}D_{\mathrm{KL}}(\rho_k^{\mathrm{stoc}}\,\|\,\rho_{k-1}^{\mathrm{stoc}}).$$

Note that

$$\begin{aligned}\mathbb{E}\big[\mathcal{F}_{\xi_k}(\rho) - \mathcal{F}_{\xi_k}(\rho_k^{\mathrm{stoc}})\,\big|\,\rho_{k-1}^{\mathrm{stoc}}\big] &= \mathbb{E}\big[\mathcal{F}_{\xi_k}(\rho) - \mathcal{F}_{\xi_k}(\rho_{k-1}^{\mathrm{stoc}})\,\big|\,\rho_{k-1}^{\mathrm{stoc}}\big] + \mathbb{E}\big[\mathcal{F}_{\xi_k}(\rho_{k-1}^{\mathrm{stoc}}) - \mathcal{F}_{\xi_k}(\rho_k^{\mathrm{stoc}})\,\big|\,\rho_{k-1}^{\mathrm{stoc}}\big]\\ &\overset{\text{(i)}}{=} \mathcal{F}(\rho) - \mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) + \mathbb{E}\big[\mathcal{F}_{\xi_k}(\rho_{k-1}^{\mathrm{stoc}}) - \mathcal{F}_{\xi_k}(\rho_k^{\mathrm{stoc}})\,\big|\,\rho_{k-1}^{\mathrm{stoc}}\big]\\ &\overset{\text{(ii)}}{\leq} \mathcal{F}(\rho) - \mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) + \mathbb{E}\big[L(\xi_k)\sqrt{D_{\mathrm{KL}}(\rho_k^{\mathrm{stoc}}\,\|\,\rho_{k-1}^{\mathrm{stoc}})}\,\big|\,\rho_{k-1}^{\mathrm{stoc}}\big]\\ &\overset{\text{(iii)}}{\leq} \mathcal{F}(\rho) - \mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) + \sqrt{\mathbb{E}L(\xi_k)^2}\cdot\sqrt{\mathbb{E}\big[D_{\mathrm{KL}}(\rho_k^{\mathrm{stoc}}\,\|\,\rho_{k-1}^{\mathrm{stoc}})\,\big|\,\rho_{k-1}^{\mathrm{stoc}}\big]}.\end{aligned}$$

Here, both (i) and (ii) are by Assumption 5, and (iii) is by Cauchy–Schwarz inequality. Therefore, we have

$$\begin{aligned}&\Big(\frac{1}{\tau_k}+\frac{\lambda}{2}\Big)\mathbb{E}D_{\mathrm{KL}}(\rho\,\|\,\rho_k^{\mathrm{stoc}}) - \frac{1}{\tau_k}\mathbb{E}D_{\mathrm{KL}}(\rho\,\|\,\rho_{k-1}^{\mathrm{stoc}})\\ &\qquad\leq \mathcal{F}(\rho) - \mathbb{E}\mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) + \sqrt{\mathbb{E}L(\xi_k)^2}\cdot\sqrt{\mathbb{E}D_{\mathrm{KL}}(\rho_k^{\mathrm{stoc}}\,\|\,\rho_{k-1}^{\mathrm{stoc}})} - \frac{1}{\tau_k}D_{\mathrm{KL}}(\rho_k^{\mathrm{stoc}}\,\|\,\rho_{k-1}^{\mathrm{stoc}})\\ &\qquad\leq \mathcal{F}(\rho) - \mathbb{E}\mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) + \frac{\tau_k}{4}\mathbb{E}L(\xi)^2. \qquad (22)\end{aligned}$$

When $\lambda = 0$, (22) implies

$$\tau_k\big[\mathbb{E}\mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) - \mathcal{F}(\rho^*)\big] \leq \mathbb{E}D_{\mathrm{KL}}(\rho^*\,\|\,\rho_{k-1}^{\mathrm{stoc}}) - \mathbb{E}D_{\mathrm{KL}}(\rho^*\,\|\,\rho_k^{\mathrm{stoc}}) + \frac{\tau_k^2}{4}\mathbb{E}L(\xi)^2.$$

Therefore, we have

$$\min_{0\leq l\leq k-1}\mathbb{E}\mathcal{F}(\rho_l^{\mathrm{stoc}}) - \mathcal{F}(\rho^*) \leq \frac{D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0)}{\tau_1+\cdots+\tau_k} + \frac{\tau_1^2+\cdots+\tau_k^2}{4(\tau_1+\cdots+\tau_k)}\mathbb{E}L(\xi)^2.$$

By taking $\tau_k = \frac{\tau}{\sqrt{k}}$ and using $k^{-1/2} \geq 2\sqrt{k+1}-2\sqrt{k}$, the above inequality implies that

$$\min_{0\leq l\leq k-1}\mathbb{E}\mathcal{F}(\rho_l^{\mathrm{stoc}}) - \mathcal{F}(\rho^*) \leq \frac{4D_{\mathrm{KL}}(\rho^*\,\|\,\rho_0) + \tau^2\log(k+1)\mathbb{E}L(\xi)^2}{8\tau(\sqrt{k+1}-1)}.$$

When $\lambda > 0$, (22) implies

$$\mathbb{E}\mathcal{F}(\rho_{k-1}^{\mathrm{stoc}}) - \mathcal{F}(\rho^*) \le \frac{1}{\tau_k}\mathbb{E}D_{\mathrm{KL}}(\rho^* \,\|\, \rho_{k-1}^{\mathrm{stoc}}) - \Big(\frac{1}{\tau_k} + \frac{\lambda}{2}\Big)\mathbb{E}D_{\mathrm{KL}}(\rho^* \,\|\, \rho_k^{\mathrm{stoc}}) + \frac{\tau_k}{4}\mathbb{E}L(\xi)^2.$$

By taking $\tau_k = \frac{2}{\lambda(k+1)}$, we have

$$\begin{aligned}\min_{0\le l\le k-1} \mathbb{E}\mathcal{F}(\rho_l^{\mathrm{stoc}}) - \mathcal{F}(\rho^*) &\le \frac{\lambda}{k}D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0) + \frac{\mathbb{E}L(\xi)^2}{4k}\sum_{l=1}^{k}\frac{2}{\lambda(l+1)} \\ &\le \frac{\lambda}{k}D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0) + \frac{\log(k+1)}{2\lambda k}\mathbb{E}L(\xi)^2.\end{aligned}$$

In the last inequality, we use $\frac{1}{l+1} \le \log\frac{l+1}{l}$ for all $l \ge 1$.

D.6 Proofs of Technical Results

Proof of Lemma A1. By first-order optimality condition of (3), we know that

$$\frac{\delta\mathcal{F}}{\delta\rho}(\rho_k) + \frac{1}{\tau_k}\log\frac{\rho_k}{\rho_{k-1}}$$

is a constant. Since $\mathcal{F}$ is λ-relative strongly convex, we have

$$\begin{aligned}\mathcal{F}(\rho) - \mathcal{F}(\rho_k) &\ge \int_\Theta \frac{\delta\mathcal{F}}{\delta\rho}(\rho_k)(\theta)\,\mathrm{d}(\rho-\rho_k)(\theta) + \frac{\lambda}{2}D_{\mathrm{KL}}(\rho \,\|\, \rho_k) \\ &= -\frac{1}{\tau_k}\int_\Theta \log\frac{\rho_k}{\rho_{k-1}}(\theta)\,\mathrm{d}(\rho-\rho_k)(\theta) + \frac{\lambda}{2}D_{\mathrm{KL}}(\rho \,\|\, \rho_k) \\ &= -\frac{1}{\tau_k}D_{\mathrm{KL}}(\rho \,\|\, \rho_{k-1}) + \Big(\frac{1}{\tau_k} + \frac{\lambda}{2}\Big)D_{\mathrm{KL}}(\rho \,\|\, \rho_k) + \frac{1}{\tau_k}D_{\mathrm{KL}}(\rho_k \,\|\, \rho_{k-1}).\end{aligned}$$

□

Proof of Lemma A2. Since ρ^* is discrete probability measure, $\rho^\sigma = \rho^* * \mathcal{N}(0, \sigma^2 I_d)$ is a Gaussian mixture distribution. The main step is to prove

$$D_{\mathrm{KL}}(\rho^\sigma \,\|\, \rho_0) \le \sup_{\theta\in\mathrm{supp}(\rho^*)} D_{\mathrm{KL}}\Big(\mathcal{N}\big(\theta, \sigma^2 I_d\big)\,\Big\|\,\rho_0\Big). \tag{23}$$

In fact, for any $\theta_j, \theta_l \in \mathrm{supp}(\rho^*)$ with $\theta_j \ne \theta_l$, assume $w_j = \rho^*(\theta_j)$ and $w_l = \rho^*(\theta_l)$. Let

$$\rho^*_{-jl} = \sum_{\substack{\theta\in\mathrm{supp}(\rho^*)\\ \theta\ne\theta_j,\theta_l}} \rho^*(\theta)\delta_\theta$$

be a measure by deleting the contribution of θ_j and θ_l in ρ^*. (Note that $\rho^*_{-jl}(\Theta) = 1 - w_j - w_l < 1$, so ρ^*_{-jl} is not a probability measure.) Consider the optimization problem

$$\begin{aligned}\max_{\substack{w+w'=w_j+w_l\\ w,w'\ge 0}} g_{jl}(w,w') &\coloneqq D_{\mathrm{KL}}\Big(\big(\rho^*_{-jl} + w\delta_{\theta_j} + w'\delta_{\theta_l}\big) * \mathcal{N}(0,\sigma^2)\,\Big\|\,\rho_0\Big) \\ &= D_{\mathrm{KL}}\Big(\rho^*_{-jl} * \mathcal{N}(0,\sigma^2) + w\mathcal{N}(\theta_j,\sigma^2 I_d) + w'\mathcal{N}(\theta_l,\sigma^2 I_d)\,\Big\|\,\rho_0\Big).\end{aligned}$$

It is easy to see that g_{jl} is a convex function on $\{(w,w') \subset \mathbb{R}^2_{\ge 0} : w + w' = w_j + w_l\}$. Therefore, g_{jl} achieves its maximum on the boundary $(w,w') = (w_j + w_l, 0)$ or $(w,w') = (0, w_j + w_l)$. The above argument indicates that we can always merge two mixtures of ρ^σ into one while the KL divergence is not decreasing. Therefore, the inequality (23) holds. Applying (23), we know

$$D_{\mathrm{KL}}(\rho^\sigma \,\|\, \rho_0) \le \sup_{\theta\in\mathrm{supp}(\rho^*)} D_{\mathrm{KL}}\Big(\mathcal{N}\big(\theta, \sigma^2 I_d\big)\,\Big\|\,\rho_0\Big)$$

$$
\begin{aligned}
&= \sup_{\theta\in\mathrm{supp}(\rho^*)} \frac{1}{2}\Big(\log\frac{\det(\beta^2 I_d)}{\det(\sigma^2 I_d)} - d + \mathrm{tr}\,(\beta^{-2}\sigma^2 I_d) + \theta^\top(\beta^2 I_d)^{-1}\theta\Big)\\
&= \frac{1}{2}\Big(2d\log\frac{\beta}{\sigma} - d + \frac{d\sigma^2}{\beta^2} + \frac{\|\theta\|^2}{\beta^2}\Big)\\
&\le d\log\frac{\beta}{\sigma} + \frac{d\sigma^2 + R_\theta^2}{2\beta^2} - \frac{d}{2}.
\end{aligned}
$$

□

D.7 Convexity of NPMLE and KL Divergence

NPMLE. Recall that the empirical loss function in NPMLE is

$$
\mathcal{L}_n(\rho) = -\frac{1}{n}\sum_{i=1}^n \log\Big(\int_\Theta p(X_i\,|\,\theta)\,\mathrm{d}\rho(\theta)\Big).
$$

Then, for any $\rho, \rho' \in \mathscr{P}(\Theta)$ and $t \in [0,1]$, we have

$$
\begin{aligned}
\mathcal{L}_n\big((1-t)\rho + t\rho'\big) &= -\frac{1}{n}\sum_{i=1}^n \log\Big((1-t)\int_\Theta p(X_i\,|\,\theta)\,\mathrm{d}\rho(\theta) + t\int_\Theta p(X_i\,|\,\theta)\,\mathrm{d}\rho'(\theta)\Big)\\
&\overset{\text{(i)}}{\le} -\frac{1-t}{n}\sum_{i=1}^n \log\Big(\int_\Theta p(X_i\,|\,\theta)\,\mathrm{d}\rho(\theta)\Big) - \frac{t}{n}\sum_{i=1}^n \log\Big(\int_\Theta p(X_i\,|\,\theta)\,\mathrm{d}\rho'(\theta)\Big)\\
&= (1-t)\mathcal{L}_n(\rho) + t\mathcal{L}_n(\rho').
\end{aligned}
$$

Here, (i) is due to the convexity of function $x \mapsto -\log x$. The above inequality implies

$$
\mathcal{L}_n(\rho') - \mathcal{L}_n(\rho) \ge \frac{\mathcal{L}_n\big(\rho + t(\rho'-\rho)\big) - \mathcal{L}_n(\rho)}{t}, \quad \forall\, t \in [0,1].
$$

By the definition of first-order variation and letting $t \to 0^+$ on the right-hand side yield

$$
\mathcal{L}_n(\rho') - \mathcal{L}_n(\rho) \ge \lim_{t\to 0^+}\frac{\mathcal{L}_n\big(\rho + t(\rho'-\rho)\big) - \mathcal{L}_n(\rho)}{t} = \int \frac{\delta\mathcal{F}}{\delta\rho}(\rho)\,\mathrm{d}(\rho'-\rho).
$$

Therefore, $\mathcal{L}_n$ is (L_2-)convex.

KL divergence For any $\pi \in \mathscr{P}^r(\Theta)$, we will show that $D_{\mathrm{KL}}(\cdot\,\|\,\pi)$ is 1-relative strongly convex. We provide the proof to make our paper self-contained. For any $\rho, \rho' \in \mathscr{P}^r(\Theta)$, we have

$$
\begin{aligned}
D_{\mathrm{KL}}(\rho'\,\|\,\pi) - D_{\mathrm{KL}}(\rho\,\|\,\pi) &= \int_\Theta -\log\pi\,\mathrm{d}(\rho'-\rho) + D_{\mathrm{KL}}(\rho'\,\|\,\rho) + \int_\Theta \log\rho\,\mathrm{d}(\rho'-\rho)\\
&= \int_\Theta \frac{\delta D_{\mathrm{KL}}(\cdot\,\|\,\pi)}{\delta\rho}(\rho)\,\mathrm{d}(\rho'-\rho) + D_{\mathrm{KL}}(\rho'\,\|\,\rho).
\end{aligned}
$$

In the last equation, we use the fact that

$$
\frac{\delta D_{\mathrm{KL}}(\cdot\,\|\,\pi)}{\delta\rho}(\rho) = \log\rho - \log\pi.
$$

In fact, Chizat (2022); Nitanda et al. (2022) show a stronger result that for any convex functional $\mathcal{H}$, the functional $\mathcal{F}(\rho) = \mathcal{H}(\rho) + \lambda\int\rho\log\rho$ is λ-relative strongly convex. In the KL divergence case, we can simply take $\mathcal{H}(\rho) = -\int_\Theta \log\pi\,\mathrm{d}\rho$ and $\lambda = 1$.