Sample complexity of Schrödinger potential estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We address the problem of Schrödinger potential estimation, which plays a crucial role in modern generative modelling approaches based on Schrodinger bridges and stochastic optimal control for SDEs. Given a simple prior diffusion process, these methods search for a path between two given distributions ρ_0 and ρ_T requiring minimal efforts. The optimal drift in this case can be expressed through a Schrödinger potential. In the present paper, we study generalization ability of an empirical Kullback-Leibler (KL) risk minimizer over a class of admissible log-potentials aimed at fitting the marginal distribution at time T. Under reasonable assumptions on the target distribution ρ_T and the prior process, we derive a non-asymptotic high-probability upper bound on the KL-divergence between ρ_T and the terminal density corresponding to the estimated log-potential. In particular, we show that the excess KL-risk may decrease as fast as $\mathcal{O}(\log n/n)$ when the sample size n tends to infinity even if both ρ_0 and ρ_T have unbounded supports.

1 Introduction

The Schrödinger Bridge problem (SBP) originates from a question posed by Erwin Schrödinger in 1932 [Schrödinger, 1932], seeking the most likely evolution of a probability distribution between two given endpoint distributions while minimizing relative entropy with respect to a prior stochastic process. This problem has deep connections with optimal transport [Leonard, 2014] and stochastic control [Dai Pra, 1991]. In its simplest continuous-time form, one aims to construct a so-called *Schrödinger Markov process* whose joint begin-end distribution $\pi(\mathrm{d}x, \mathrm{d}z)$ has the representation

$$\pi(\mathrm{d}x, \mathrm{d}z) = \mathsf{Q}(z, T \mid x, 0)\, \nu_0(\mathrm{d}x)\, \nu_T(\mathrm{d}z), \tag{1}$$

where $\mathsf{Q}(z, T \mid x, 0)$ is the transition kernel of a reference Markov process, and ν_0, ν_T are unknown "boundary potentials" to be determined. The desired marginals $\pi(\mathrm{d}x, \mathbb{R}^d)$ and $\pi(\mathbb{R}^d, \mathrm{d}z)$ are given, and one seeks ν_0 and ν_T that reproduce these marginals. In the rest of the paper, we assume that both $\pi(\mathrm{d}x, \mathbb{R}^d)$ and $\pi(\mathbb{R}^d, \mathrm{d}z)$ are absolutely continuous with respect to the Lebesgue measure and denote the corresponding densities by ρ_0 and ρ_T^*, respectively. Classical existence proofs for the SBP date back to Fortet [1940] (in 1D) and Beurling [1960], with a modern fixed-point approach in [Chen et al., 2016]. Recent extensions to the case of noncompactly supported marginal distributions can be found in [Conforti et al., 2024] and [Eckstein, 2025]. Recently, the problem attracted attention of machine learners in the context of generative modelling (see, for instance, [Tzen and Raginsky, 2019, De Bortoli et al., 2021, Shi et al., 2023, Korotin et al., 2024, Gushchin et al., 2024a, Rapakoulias et al., 2024] to name a few). It follows from Theorem 3.2 in [Dai Pra, 1991] that the optimal Markov process X_t^* solving the Schrödinger problem with marginals (ρ_0, ρ_T^*) can be constructed as a solution of the following SDE:

$$\mathrm{d}X_t^* = \left(b(X_t^*, t) + \sigma(X_t^*, t)\sigma(X_t^*, t)^\top \nabla \log h(X_t^*, t)\right)\, \mathrm{d}t + \sigma(X_t^*, t)\, \mathrm{d}W_t, \quad X_0 \sim \rho_0,$$

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

where

$$h(w,t) = \int_{\mathbb{R}^d} \mathsf{Q}(y, T \mid w, t)\, \nu_T(\mathrm{d}y)$$

and Q is the transition density of the reference (or base) diffusion process

$$\mathrm{d}X_t = b(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t, \quad X_0 \sim \rho_0.$$

The transition density Q^* of the reciprocal process X_t^* can be obtained from Q via the so-called Doob's h–transform:

$$\mathsf{Q}^*(y, T \mid x, t) = \mathsf{Q}(y, T \mid x, t)\, \frac{h(y,T)}{h(x,t)}. \tag{2}$$

This is precisely the law of the base process conditioned by the function h (see [Jamison, 1974]). In many presentations of the Schrödinger Bridge problem, one takes a very simple reference process (for instance, a Brownian motion) so that its transition kernel is straightforward to write down (see, for example, [Pooladian and Niles-Weed, 2024] and [Baptista et al., 2024]). However, there are several practical and theoretical advantages to considering more general (potentially higher-dimensional, or with domain constraints, or with a non-trivial drift/diffusion) reference processes.

In the present paper, we are interested in estimation of the Schrödinger potential ν_T from n i.i.d. samples $Y_1, \dots, Y_n \sim \rho_T^*$. Given a class of log-potentials Ψ, we study generalization ability of an empirical risk minimizer

$$\widehat{\psi} \in \operatorname*{argmin}_{\psi \in \Psi} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log \left(\int_{\mathbb{R}^d} \mathsf{Q}(Y_i, T \mid x, 0)\, \frac{h_\psi(Y_i, T)}{h_\psi(x, 0)}\, \rho_0(x)\mathrm{d}x \right) \right\}, \tag{3}$$

where

$$h_\psi(x,t) = \int \mathsf{Q}(y, T \mid x, t)\, e^{\psi(y)}\, \mathrm{d}y.$$

Let us note that, in view of (2),

$$\rho_T^\psi(y) = \int_{\mathbb{R}^d} \mathsf{Q}(y, T \mid x, 0)\, \frac{h_\psi(y,T)}{h_\psi(x,0)}\, \rho_0(x)\mathrm{d}x$$

is the marginal endpoint probability density of a diffusion process X_t^ψ corresponding to Doob's h_ψ-transform:

$$\mathrm{d}X_t^\psi = \Big(b(X_t^\psi, t) + \sigma(X_t^\psi, t)\sigma(X_t^\psi, t)^\top \nabla \log h_\psi(X_t^\psi, t) \Big)\, \mathrm{d}t + \sigma(X_t^\psi, t)\,\mathrm{d}W_t, \quad X_0 \sim \rho_0.$$

In other words, the estimate $\widehat{\psi}$ minimizes empirical Kullback-Leibler (KL) divergence between the actual target ρ_T^* and the marginal densities ρ_T^ψ over the class of admissible log-potentials Ψ. That is, we chose the log-potential ψ that makes the transformed reference diffusion hit the observed terminal law, and measure error only through KL of the marginals. Because h_ψ is used inside the Doob factor, the learnt potential is compatible with a single Markov process; one never risks obtaining mutually inconsistent forward/backward potentials. The method combines the full problem (the marginals, transition densities, and the potential function) into one single optimization framework. By doing so, it aims to directly minimize the objective of matching the marginals at time T without separating the problem into smaller subproblems. In contrast, the Sinkhorn algorithm, commonly used for optimal transport problems, approaches the problem by iteratively updating the potentials in a decoupled manner. At each iteration, a simpler least squares problem appears, which is linear in one potential function given that another one is fixed from the previous iteration. The Sinkhorn algorithm alternates between updating the potential functions to match the marginals of the distributions and adjusting the transport plan until convergence. We refer to Pooladian and Niles-Weed [2024], Chiarini et al. [2024] for recent results. The primary advantage of the Sinkhorn approach is its computational efficiency. By decoupling the optimization process into simpler, linear problems, the Sinkhorn method can handle large-scale problems effectively. This iterative procedure allows for faster updates, and it has become a popular method for many optimal transport applications, see Genevay et al. [2018], March and Henry-Labordere [2023] However, the approach presented in this paper differs in that it does not separate the problem into independent steps. Instead, it aims at solving the Schrödinger system approximately by formulating it as a single optimization problem involving Doob h-transform of the base process X parametrized by the Schrödinger potential. Unlike iterative proportional fitting (Sinkhorn), everything is learnt in one go, avoiding slow or unstable fixed-point cycles. This results in a more accurate and robust solution. The trade-off between the two methods lies in computational efficiency versus the quality of the solution. The Sinkhorn approach provides a quick and efficient solution by solving simpler problems at each iteration, but it may not achieve the best possible solution for the full problem. On the other hand, the method presented in this paper offers a more holistic approach, which could lead to a more accurate matching of the marginal distributions but might require more computational resources.

The approach presented in this paper can also be compared to methods that rely on optimization over transport maps, see Korotin et al. [2024], Gushchin et al. [2024a]. In transport map-based approaches, the goal is to find a map $\mathcal{T}$ that transports one probability distribution to another. The optimization typically focuses on minimizing a quadratic cost functional that penalizes the difference between the target distribution and the transformed distribution under the transport map. These methods are often framed as optimal transport problems, where the map $\mathcal{T}$ is determined by solving an optimization problem that involves the marginal distributions. The advantage of optimization over transport maps lies in its clear geometric interpretation, where the transport map provides a direct way to relate the two distributions. This can lead to efficient algorithms, especially when the transport map can be parametrized in a way that allows for fast computations, such as in the case of certain neural network architectures or simple affine transformations, Rapakoulias et al. [2024].

However, transport map-based approaches are typically constrained to quadratic costs, which may limit their applicability in some cases. Specifically, quadratic cost functionals, such as the 2-Wasserstein distance, often assume a certain structure or symmetry that may not be ideal for more general or complex problems.

In contrast, the approach discussed in this paper is not limited to quadratic costs. It allows for more general cost structures and is based on minimizing the Kullback-Leibler divergence (KL-divergence), which can accommodate a wider range of problem types. This flexibility is particularly valuable when dealing with more complex distributions or when the underlying problem involves non-quadratic costs that capture other aspects of the distribution, such as entropy regularization or non-linear interactions between variables.

Contribution The main contribution of the present paper a sharper non-asymptotic high-probability upper bound on generalization error of the empirical risk minimizer $\widehat{\psi}$ defined in (3).

- Taking a multivariate Ornstein-Uhlenbeck process as the reference one, we show that (see Theorem 1), with probability at least $(1 - 2\delta)$, the excess KL-risk of the marginal endpoint density $\widehat{\rho}_T$ corresponding to $\widehat{\psi}$ satisfies the inequality

$$\mathsf{KL}(\rho_T^*, \widehat{\rho}_T) - \inf_{\psi \in \Psi} \mathsf{KL}(\rho_T^*, \rho_T^{\psi}) \lesssim \sqrt{\Upsilon(n,\delta) \inf_{\psi \in \Psi} \mathsf{KL}(\rho_T^*, \rho_T^{\psi})} + \Upsilon(n,\delta),$$

  where

$$\Upsilon(n,\delta) \lesssim \frac{\log^2 n + \log(1/\delta)\log n}{n}.$$

  Here and further in the paper, the sign $\lesssim$ stands for an inequality up to a multiplicative constant. The derived upper bound has several advantages over the existing results. First, in contrast to Korotin et al. [2024], the excess risk may decrease as fast as $\mathcal{O}(\log^2 n/n)$ provided that the class of log-potentials Ψ is rich enough to approximate the target density ρ_T^*. Second, unlike theoretical guarantees for Sinkhorn-based approaches (see e.g. Pooladian and Niles-Weed [2024]), we are able to relate the endpoint marginal densities ρ_T^* and $\widehat{\rho}_T$.
- We impose very mild assumptions on the target density ρ_T^*. We only require ρ_T^* to be bounded and sub-Gaussian. On the other hand, the available convergence proofs for the Sinkhorn algorithm rely on the stronger assumption that the marginals are log-concave, see Conforti et al. [2024]. We also avoid the so-called strong density assumptions like boundedness from below often used in nonparametric statistics in the context of log-density estimation.
- The assumptions on the class of log-potentials Ψ are also reasonable. We support our claim with several examples.

Paper structure The rest of the paper is organized as follows. Section 2 is devoted to a short review of related work. In Section 3, we introduce necessary definitions and notations. After that, we present our main result (Theorem 1) in Section 4 and discuss main ideas of its proof in Section 5. Rigorous derivations as well as auxiliary technical results are deferred to the supplementary material.

2 Related work

Here is a short review of methods used in the literature to compute Schrödinger potentials, including the Sinkhorn algorithm. The Schrödinger potential, which arises in optimal transport problems, represents a key component in the solution of transport problems involving marginal distributions. Over time, several methods have been proposed to compute these potentials efficiently, with applications in areas ranging from statistical mechanics to machine learning. Here, we review some of the most prominent methods used in the literature.

Sinkhorn algorithm The Sinkhorn algorithm Sinkhorn [1967] is one of the most widely used methods for computing Schrödinger potentials in the context of optimal transport. It is based on iterative scaling and aims to solve the optimal transport problem by alternating between updating two potentials ν_0 and ν_T to enforce marginal constraints. The key advantage of the Sinkhorn approach is its computational efficiency, particularly when the transport problem is framed with a quadratic cost (such as the 2-Wasserstein distance), see Pavon et al. [2021], Chen et al. [2021], Stromme [2023] for reference. In each iteration, the algorithm solves a simpler problem that involves scaling the potentials in a way that brings the marginals of the transformed distribution closer to the target. Although Sinkhorn's algorithm is efficient and widely applicable, it is often limited by its assumption of quadratic costs. Additionally, the algorithm does not directly handle more complex cost structures, such as non-quadratic costs or non-linear dynamics, which can be a limitation in some applications.

Sinkhorn bridge The Sinkhorn Bridge proposed by Pooladian and Niles-Weed [2024], provides a way to estimate the Schrödinger bridge using Sinkhorn's algorithm in an efficient manner. The key insight of this method is that the potentials obtained from the static entropic optimal transport problem can be modified to yield a natural plug-in estimator for the drift function that defines the Schrödinger bridge. However, this work does not provide bounds on the distance between marginal distributions at time $T = 1$ because there is an exploding term $(1-\tau)^{k+2}$ as $\tau \to 1$ where k is the dimension of the underlying manifold. This term leads to a "curse of dimensionality" where the error grows rapidly as τ approaches 1, especially in high-dimensional settings. As a result, the estimation error increases significantly when attempting to estimate the Schrödinger bridge at the terminal time, making it difficult to obtain precise bounds for $T = 1$.

Dual Formulation of the Schrödinger Problem In the dual formulation of the Schrödinger problem, the Schrödinger potential is computed by solving a convex optimization problem. This approach reformulates the problem in terms of a dual objective that involves the Kullback-Leibler (KL) divergence between the target and predicted distributions. The dual problem is then solved using optimization techniques such as gradient descent or variational methods, see Zhang and Chen [2022], Tzen and Raginsky [2019] for reference. This formulation is more flexible than the Sinkhorn algorithm, as it can accommodate more general cost functions and is not limited to quadratic losses.

While the dual approach is flexible, it is often computationally more demanding than Sinkhorn's method due to the need for iterative optimization over high-dimensional spaces. This makes the dual formulation suitable for smaller or more specialized problems, but it can become computationally expensive in large-scale applications.

Approximate Solutions Using Monte Carlo Methods Monte Carlo methods, particularly those relying on reverse diffusion processes, have also been employed to approximate Schrödinger potentials. In these methods, a reverse process is simulated, and the potential is iteratively refined to minimize the discrepancy between the predicted and target marginals, see Korotin et al. [2024] for reference. These methods are often used when the problem involves complex dynamics that are difficult to capture using direct optimization techniques.

Monte Carlo methods are particularly useful when dealing with high-dimensional problems, as they allow for the sampling of large spaces. However, they can be computationally expensive and may require a significant number of samples to achieve an accurate solution.

In addition, there are approaches that rely heavily on Monte Carlo approximations of intermediate values rather than the Schrödinger potentials themselves, among which the following should be noted De Bortoli et al. [2021], Vargas et al. [2021], Peluchetti [2023].

Neural Network-Based Approaches Recent advancements in deep learning have led to the use of neural networks to approximate Schrödinger potentials. These approaches treat the potential function as a parameterized neural network and use gradient-based optimization techniques to learn the potential that best matches the marginals. The use of neural networks offers a flexible and powerful way to model complex non-linear potentials, making these methods well-suited for problems with intricate dynamics or non-quadratic costs.While neural network-based approaches are highly flexible, they require large amounts of data and computational resources to train the network, and they are often prone to overfitting if not regularized appropriately. Despite these challenges, they represent a promising direction for future research, especially when the problem at hand involves complex and high-dimensional systems. We refer to Liu et al. [2023], Wang et al. [2021] for recent results.

Iterative Markovian Fitting The Iterative Markovian Fitting (IMF) method, introduced in the recent work by Shi et al. [2023], offers an approach to solving Schrödinger Bridge (SB) problems. Unlike previous methods, such as Iterative Proportional Fitting (IPF), IMF guarantees the preservation of both the initial and terminal distributions in each iteration, which is a key advantage over IPF where these marginals are not always preserved. IMF alternates between two types of projections: Markovian projections and reciprocal projections, ensuring that the resulting distribution remains within the correct class (Markovian or reciprocal) while progressively approximating the Schrödinger Bridge. We refer to Gushchin et al. [2024b] for recent results.

In Silveri et al. [2024], the authors provide the convergence analysis for diffusion flow matching (DFM), a method used to generate approximate samples from a target distribution by bridging it with a base distribution through diffusion dynamics. Their theoretical work includes non-asymptotic bounds on the Kullback-Leibler (KL) divergence between the true target distribution and the distribution generated by the DFM model. A key insight from this paper is the incorporation of two sources of error: drift-estimation and time-discretization errors. However, while the convergence analysis offers theoretical guarantees, the statistical error is not explicitly addressed in this paper. The analysis assumes that all expectations are exact, which might not hold in practical settings where samples are finite, and statistical errors could arise due to the approximations involved in the generative process. Thus, future work will need to extend this analysis to quantify the impact of statistical approximations in finite-sample settings.

3 Preliminaries and notations

This section collects necessary definitions and notations. As we announced in the contribution paragraph, we are going to consider a multivariate Ornstein-Uhlenbeck process as a reference one. For this reason, we elaborate on its basic properties in this section.

Multivariate Ornstein-Uhlenbeck process To be more specific, we will consider the base process X_t^0 solving the SDE

$$\mathrm{d}X_t^0 = b\left(m - X_t^0\right)\mathrm{d}t + \Sigma^{1/2}\mathrm{d}W_t, \quad 0 \leqslant t \leqslant T,$$

where $b > 0$ controls the drift rate, $m \in \mathbb{R}^d$ represents the mean-reversion level, $\Sigma \in \mathbb{R}^{d\times d}$ is a positive definite symmetric matrix, and W_t is a standard d-dimensional Wiener process. It is known that the conditional distribution of X_t^0 given $X_0^0 = x$ is Gaussian $\mathcal{N}\big(m_t(x), \Sigma_t\big)$ with

$$m_t(x) = (1 - e^{-bt})m + e^{-bt}x \quad \text{and} \quad \Sigma_t = \frac{1 - e^{-2bt}}{2b}\Sigma. \tag{4}$$

This implies that the corresponding Doob's h-transform can be expressed through the Ornstein-Uhlenbeck operator

$$\mathcal{T}_t g(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_t)}} \int\limits_{\mathbb{R}^d} \exp\left\{-\frac{1}{2}\|\Sigma_t^{-1/2}(y - m_t(x))\|^2\right\} g(y)\,\mathrm{d}y.$$

Indeed, it holds that $h_\psi(x,t) = \mathcal{T}_{T-t}e^{\psi(x)}$. Then, introducing

$$\mathsf{q}(y \,|\, x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_T)}} \exp\left\{-\frac{1}{2}\|\Sigma_T^{-1/2}(y - m_T(x))\|^2\right\},$$

we note that

$$\rho_T^\psi(y) = \int\limits_{\mathbb{R}^d} \frac{\mathsf{q}(y \,|\, x) e^{\psi(y)}}{\mathcal{T}_T e^{\psi(x)}}\, \rho_0(x)\,\mathrm{d}x \tag{5}$$

is the marginal density of X_T^ψ, the endpoint of a random process X_t^ψ governed by h_ψ:

$$\mathrm{d}X_t^\psi = b\left(m - X_t^\psi\right)\mathrm{d}t + \nabla \log\left(\mathcal{T}_{T-t} e^{\psi(X_t^\psi)}\right)\mathrm{d}t + \Sigma^{1/2}\mathrm{d}W_t, \quad X_0^\psi \sim \rho_0.$$

If the Schrödinger potential ν_T admits a density e^{ψ^*} with respect to the Lebesgue measure, then the optimally controlled process X_t^* solves the SDE

$$\mathrm{d}X_t^* = b\left(m - X_t^*\right)dt + \nabla \log\left(\mathcal{T}_{T-t} e^{\psi^*(X_t^*)}\right)\mathrm{d}t + \Sigma^{1/2}\mathrm{d}W_t, \quad X_0^* \sim \rho_0.$$

Finally, it is well known that the unique stationary (invariant) distribution of X_t^0 is Gaussian, that is, X_t^0 converges to X_∞^0 in distribution as $t \to \infty$ with $X_\infty \sim \mathcal{N}(m, \Sigma/(2b))$. Since the parameters of the limiting distribution do not depend on the starting point, $\mathcal{T}_\infty g(x) \equiv \mathcal{T}_\infty g$ is a constant.

Other notations The notation $f \lesssim g$ or $g \gtrsim f$ means that $f = \mathcal{O}(g)$. Besides, we often replace $\max\{a, b\}$ and $\min\{a, b\}$ by shorter expressions $a \vee b$ and $a \wedge b$, respectively. For any $s \geqslant 1$, the Orlicz ψ_s-norm of a random variable ξ is defined as

$$\|\xi\|_{\psi_s} = \inf\left\{u > 0 : \mathbb{E}e^{|\xi|^s/u^s} \leqslant 2\right\}.$$

Finally, given $p \geqslant 1$ and a probability density ρ, the weighted L_p-norm of a function f is defined as $\|f\|_{L_p(\rho)} = \left(\mathbb{E}_{\xi\sim\rho}|f(\xi)|^p\right)^{1/p}$. Given two probability densities $\rho_0 \ll \rho_1$ on $\mathbb{R}^d$, the Kullback-Leibler divergence between them is defined as $\mathsf{KL}(\rho_0, \rho_1) = \mathbb{E}_{\xi\sim\rho_0} \log\left(\rho_0(\xi)/\rho_1(\xi)\right)$.

4 Main result

In the present section, we discuss statistical properties of the empirical risk minimizer $\widehat{\psi}$ defined in (3). In particular, Theorem 1 provides a Bernstein-type upper bound on its excess KL-risk. We impose the following assumptions. First, as we announced before, we use the Ornstein-Uhlenbeck process X_t^0 as the reference one.

Assumption 1. *The base process X^0 solves the following SDE*

$$\mathrm{d}X_t^0 = b\left(m - X_t^0\right)\mathrm{d}t + \Sigma^{1/2}\,\mathrm{d}W_t, \quad 0 \leqslant t \leqslant T.$$

where $b > 0$, $m \in \mathbb{R}^d$, Σ is a positive definite symmetric matrix of size $d \times d$, and W is a d-dimensional Brownian motion.

Main properties of the Ornstein-Uhlenbeck process were discussed in the previous section. Second, we suppose that the target density ρ_T^* meets the following requirements.

Assumption 2. *The target distribution at time T admits a bounded density ρ_T^* with respect to the Lebesgue measure such that*

$$\rho_T^*(x) \leqslant \rho_{\max} \quad \textit{for all } x \in \mathbb{R}^d.$$

Moreover, the target distribution ρ_T^* is sub-Gaussian with variance proxy $\mathtt{v}^2$, that is,

$$\mathbb{E}_{Y\sim\rho_T^*} e^{u^\top Y} \leqslant e^{\mathtt{v}^2\|u\|^2/2} \quad \textit{for any } u \in \mathbb{R}^d. \tag{6}$$

Assumption 2 is very mild. Despite the fact that we deal with logarithmic loss, we do not require ρ_T^* to be bounded away from zero. We do not even require its support to be compact. This significantly complicates the proof of the excess KL-bound and poses nontrivial technical challenges. Let us note that the condition 6 yields that $\mathbb{E}_{Y \sim \rho_T^*} Y = 0$. However, it does not diminish generality of our setup.

The remaining assumptions concern properties of the class of log-potentials Ψ. First, we assume that admissible log-potentials $\psi(x)$ are bounded from above and behave as $\mathcal{O}(\|x\|^2)$ as x tends to infinity.

Assumption 3. *There exist non-negative constants Λ and M such that*

$$-\Lambda \left\| \Sigma^{-1/2}(x - m) \right\|^2 - M \leqslant \psi(x) \leqslant M \quad \textit{for all } x \in \mathbb{R}^d \textit{ and } \psi \in \Psi.$$

Moreover, for any $\psi \in \Psi$, it holds that $\mathcal{T}_\infty \psi = \mathbb{E}\psi(X_\infty) = 0$.

The condition $\mathcal{T}_\infty \psi = 0$ appears because of the fact that the Schrödinger potentials ν_0 and ν_T (see (1)) are defined up to a multiplicative constant. The requirement $\mathcal{T}_\infty \psi = 0$ is nothing but a normalization. Second, we assume that Ψ is parametrized by a finite-dimensional parameter $\theta \in \mathbb{R}^D$:

$$\Psi = \{\psi_\theta : \theta \in \Theta\},$$

where Θ is a subset of a D-dimensional cube $[-R, R]^D$ and each function ψ_θ maps $\mathbb{R}^d$ onto $\mathbb{R}$. We suppose that the parametrization is sufficiently smooth in the following sense.

Assumption 4. *There exists $L \geqslant 0$ such that*

$$|\psi_\theta(x) - \psi_{\theta'}(x)| \leqslant L \left(1 + \|x\|^2\right) \|\theta - \theta'\|_\infty \quad \textit{for all } \theta, \theta' \in \Theta \textit{ and all } x \in \mathbb{R}^d.$$

Assumptions 3 and 4 are quite general. We provide two examples when they hold. First, in a recent paper [Korotin et al., 2024], the authors model $e^{\psi(x)}$ as a Gaussian mixture. Let $\alpha_1, \dots, \alpha_K$ be non-negative numbers such that $\alpha_1 + \ldots + \alpha_K = 1$ and consider

$$e^{\psi(x)} = e^{-C} \sum_{k=1}^K \alpha_k \varphi_{m_k, \Sigma_k}(x), \quad \text{where} \quad \varphi_{m_k, \Sigma_k}(x) = \frac{e^{-\|\Sigma_k^{-1/2}(x - m_k)\|^2/2}}{(2\pi)^{d/2} \det(\Sigma_k)^{1/2}}.$$

Here C is a normalizing constant which ensures that $\mathcal{T}_\infty \psi = 0$. In this situation, the parameter θ consists of all α_k's and all components of m_k's and Σ_k's, $k \in \{1, \dots, K\}$. If the smallest eigenvalues of $\Sigma_1, \dots, \Sigma_K$ are bounded away from zero uniformly over $k \in \{1, \dots, K\}$, then $e^{\psi(x)}$ is bounded. On the other hand, if K is fixed, there is a component with a weight at least $1/K$. Without loss of generality, we assume that it is the first one. Then

$$\psi(x) \geqslant -C + \log\left(\alpha_1 \varphi_{m_1, \Sigma_1}(x)\right) \geqslant -C - \log K - \frac{1}{2} \left\| \Sigma_1^{-1/2}(x - m_1) \right\|^2,$$

and we conclude that Assumption 3 is satisfied. Verification of the Assumption 4 is straightforward once we assume that the weight of each component is bounded away from zero, and the norms $\|m_k\|$, $\|\Sigma_k\|$, and $\|\Sigma_k^{-1}\|$ are bounded uniformly over $k \in \{1, \dots, K\}$ (which is the case in [Korotin et al., 2024]). Second, Assumptions 3 and 4 will be fulfilled if one deals, for example, with a class of truncated feedforward neural networks with bounded weights and ReLU activations. It is known that (see [Schmidt-Hieber, 2020, Lemma 5]) they are Lipschitz with respect to each weight, and the Lipschitz constant grows linearly with $\|x\|$. More generally, Conforti [2024] analyzed semiconvexity properties of the Schrödinger potentials under rather mild assumptions on the marginals.

We are ready to formulate the main result of this section.

Theorem 1. *Let ρ_0 be the density of the standard Gaussian distribution $\mathcal{N}(0, I_d)$. Grant Assumptions 1, 2, 3, and 4. Assume that T is sufficiently large in a sense that*

$$bT \geqslant (5 + \log d) \vee \log\left(160b\,(\mathrm{v}^2 \vee 1) \left\|\Sigma^{-1}\right\|\right).$$

Let $\widehat{\psi}$ be defined in (3) and let $\widehat{\rho}_T$ be the corresponding density of $X_T^{\widehat{\psi}}$. Then, for any $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$, it holds that

$$\mathsf{KL}(\rho_T^*, \widehat{\rho}_T) - \inf_{\psi \in \Psi} \mathsf{KL}(\rho_T^*, \rho_T^\psi) \lesssim \sqrt{\Upsilon(n, \delta) \inf_{\psi \in \Psi} \mathsf{KL}(\rho_T^*, \rho_T^\psi)} + \Upsilon(n, \delta),$$

where

$$\Upsilon(n, \delta) = (\Lambda d + M + d) \left(d + \log\frac{RLn}{\delta} + (M \vee \log \Lambda)\sqrt{d} e^{-bT}\right) \frac{D \log n}{n}.$$

The hidden constant behind $\lesssim$ depends on Σ, m, b, and v only.

In Theorem 1, we assume that ρ_0 is the density of $\mathcal{N}(0, I_d)$. Though it is a standard choice of initial distribution in practice, we would like to emphasize that unbounded support of ρ_0 significantly complicates the proof and makes the problem even more challenging.

The problem of Schrödinger potential estimation was also studied in [Korotin et al., 2024] and [Pooladian and Niles-Weed, 2024]. In [Korotin et al., 2024], the authors suggest an algorithm called Light Schrödinger Bridge, which is based on minimization of the empirical KL-divergence between entropic optimal transport plans. This slightly differs from our setup, since we aim to minimize empirical KL-divergence between marginal endpoint distributions. The reason is that Korotin, Gushchin, and Burnaev [2024] are motivated by the style transfer task, where the initial distribution is also unknown. In contrast, we focus on generative modelling where the initial distribution ρ_0 is available to learner. In [Korotin et al., 2024, Theorem A.1], the authors consider the case when admissible potentials are Gaussian mixtures with K components. Assuming that both initial and finite distibutions have a compact support, they prove a $\mathcal{O}(n^{-1/2})$ upper bound on the Rademacher complexity of such class. On the other hand, we allow the support of ρ_0 and ρ_T^* to be unbounded. Besides, the rate of convergence presented in Theorem 1 may be much faster than $\mathcal{O}(n^{-1/2})$ if the target distribution is close to $\{\rho_T^\psi : \psi \in \Psi\}$. In the realizable case (that is, $\rho_T^* \in \{\rho_T^\psi : \psi \in \Psi\}$) the right-hand side in Theorem 1 becomes $\mathcal{O}(\log^2 n/n)$. Finally Theorem 1 provides a high-probability upper bound on the excess risk while the result of Korotin et al. [2024] holds in expectation. In [Pooladian and Niles-Weed, 2024] the authors study properties of a plug-in Sinkhorn-based estimator. Similarly to Korotin et al. [2024], they consider the case of compactly supported initial and target measures. However, they assume that these measures are supported on smooth k-dimensional submanifolds. They derive a $\mathcal{O}(n^{-1/2} + (T-\tau)^{-k-2} n^{-1})$ bound on the squared total variation distance between *path measures* up to moment $\tau < T$. Unfortunately, the second term grows very fast when τ approaches T, and there are no guarantees whether the marginal endpoint distributions will be close to each other.

In Theorem 1, we focus on the statistical error leaving study of the approximation out of the scope of the present paper. The reason is that there are few results on properties of the true log-potential $\psi^*(x) = \log\big(\nu_T(\mathrm{d}x)/\mathrm{d}x\big)$. However, we would like to note that, according to our findings (see Lemma B.2 and (5)), if ψ^* fulfils Assumption 3, then for any $\psi \in \Psi$ and $y \in \mathbb{R}^d$

$$\log \frac{\rho_T^*(y)}{\rho_T^\psi(y)} \lesssim |\psi(y) - \psi^*(y)| \\ + \left(\mathcal{T}_\infty|\psi - \psi^*|\right)^{1/\mathcal{K}(T)} \|\Sigma^{-1/2}(y-m)\|^{2-2/\mathcal{K}(T)} e^{\mathcal{O}(e^{-bT}\|\Sigma^{-1/2}(y-m)\|^2)},$$

where $1 \leqslant \mathcal{K}(T) \leqslant 1 + \mathcal{O}(\sqrt{d}e^{-bT})$. In the proof of Theorem 1 (see Step 4), we show that the expectation

$$\mathbb{E}_{Y\sim\rho_T^*} \left\|\Sigma^{-1/2}(Y-m)\right\|^{2-2/\mathcal{K}(T)} e^{\mathcal{O}(e^{-bT}\|\Sigma^{-1/2}(Y-m)\|^2)}$$

is finite, provided that $bT \geqslant (5+\log d) \vee \log\left(160b\left(\mathrm{v}^2 \vee 1\right)\left\|\Sigma^{-1}\right\|\right)$. This allows us to relate the KL-divergence between ρ_T^* and ρ_T^ψ with the distances between the corresponding log-potentials:

$$\mathsf{KL}\left(\rho_T^*, \rho_T^\psi\right) \lesssim \|\psi - \psi^*\|_{L_1(\rho_T^*)} + \left(\mathcal{T}_\infty|\psi - \psi^*|\right)^{1/\mathcal{K}(T)}.$$

5 Proof sketch of Theorem 1

In this section, we discuss main ideas used in the proof of Theorem 1. Rigorous derivations are deferred to Appendix A. Since the proof is quite long, we split it into several steps.

Step 1: log-density properties. Let us note that Assumptions 3 and 4 concern properties of log-potentials $\psi \in \Psi$ while empirical risks include marginal densities ρ_T^ψ. For this reason, before we consider the empirical process

$$\frac{1}{n}\sum_{i=1}^n \log\frac{\rho_T^*(Y_i)}{\rho_T^\psi(Y_i)} - \mathsf{KL}\left(\rho_T^*, \rho_T^\psi\right), \quad \psi \in \Psi,$$

we have to study the random variables $\log\left(\rho_T^*(Y_i)/\rho_T^\psi(Y_i)\right)$, $1 \leqslant i \leqslant n$. Using basic properties of the Ornstein-Uhlenbeck operator, we show that

$$-\log \rho_T^\psi(y) \lesssim -\psi(y) + \left\|\Sigma^{-1/2}(y-m)\right\|^2.$$

In view of Assumption 3, this means that $-\log \rho_T^\psi(y)$ grows as fast as a quadratic function. Since the target distribution is sub-Gaussian and has a bounded density, this yields that the random variables $\log\left(\rho_T^*(Y_i)/\rho_T^\psi(Y_i)\right)$, $1 \leqslant i \leqslant n$, are sub-exponential. More specifically, applying Lemma C.3 we obtain the following upper bound on their Orlicz norm:

$$\left\|\log \frac{\rho_T^*(Y_i)}{\rho_T^\psi(Y_i)}\right\|_{\psi_1} \lesssim \Lambda d + M + d \quad \text{for all } i \in \{1, \dots, n\}.$$

Step 2: ε-net argument and Bernstein's inequality. The result obtained on the first step allows us to use concentration inequalities for sub-exponential random variables. Let us fix $\varepsilon \in (0, R)$ and let Θ_ε stand for the minimal ε-net of Θ with respect to the ℓ_∞-norm. We denote the set of corresponding log-potentials by Ψ_ε:

$$\Psi_\varepsilon = \{\psi_\theta : \theta \in \Theta_\varepsilon\}.$$

Using Bernstein's inequality for unbounded random variables (see, for instance, [Lecué and Mitchell, 2012, Proposition 5.2]) and the union bound, we obtain that

$$\left|\mathsf{KL}\left(\rho_T^*, \rho_T^\psi\right) - \frac{1}{n}\sum_{i=1}^n \log \frac{\rho_T^*(Y_i)}{\rho_T^\psi(Y_i)}\right| \lesssim \sqrt{\operatorname{Var}\left(\log \frac{\rho_T^*(Y_1)}{\rho_T^\psi(Y_1)}\right) \frac{\log(2|\Psi_\varepsilon|/\delta)}{n}} + \frac{(\Lambda d + M + d)\log n \log(2|\Psi_\varepsilon|/\delta)}{n}$$

with probability at least $(1-\delta)$ simultaneously for all $\psi \in \Psi_\varepsilon$.

Step 3: bounding the loss variance. One of the key ingredients in the proof of Theorem 1, which allows us to hope for faster rates of convergence than $\mathcal{O}(n^{-1/2})$, is analysis of the variance of $\log\left(\rho_T^*(Y_1)/\rho_T^\psi(Y_1)\right)$, $\psi \in \Psi$. Despite the fact that the admissible log-potentials may be unbounded, we are still able to show that the class Ψ satisfies a Bernstein-type condition

$$\operatorname{Var}\left(\log \frac{\rho_T^*(Y_1)}{\rho_T^\psi(Y_1)}\right) \lesssim (\Lambda d + M + d)\log n \left(\mathsf{KL}\left(\rho_T^*, \rho_T^\psi\right) + \frac{1}{n}\right).$$

Steps 4 and 5: from ε-net to a uniform Bernstein-type bound. The hardest and technically involved part of the proof is to show that the losses $\log\left(\rho_T^*(y)/\rho_T^\psi(y)\right)$ and $\log\left(\rho_T^*(y)/\rho_T^\phi(y)\right)$ do not differ too much, once the corresponding log-potentials ψ and ϕ are close to each other. This follows from Lemma B.2, which relies on properties of the Ornstein-Uhlenbeck operator established in and Lemma B.3. We would like to note that the unbounded support of the initial density ρ_0 significantly complicates the proof of Lemma B.2. Nevertheless, we prove that

$$\log \frac{\rho_T^\psi(y)}{\rho_T^\phi(y)} \lesssim |\psi(y)-\phi(y)| + (\mathcal{T}_\infty|\psi-\phi|)^{1/\mathcal{K}(T)} \|\Sigma^{-1/2}(y-m)\|^{2-2/\mathcal{K}(T)} e^{\mathcal{O}(e^{-bT}\|\Sigma^{-1/2}(y-m)\|^2)},$$

where $1 \leqslant \mathcal{K}(T) \leqslant 1 + \mathcal{O}(\sqrt{d}e^{-bT})$. Though the right-hand side depends exponentially on the squared norm of $\Sigma^{-1/2}(y-m)$, the coefficient $\mathcal{O}(e^{-bT})$ is quite small, which is enough for our purposes.

Steps 6 and 7: choice of ε and the final bound. The rest of the proof is quite standard. On Step 6, we choose an appropriate ε and obtain a uniform Berstein-type inequality

$$\mathsf{KL}\left(\rho_T^*, \rho_T^\psi\right) - \frac{1}{n}\sum_{i=1}^n \log \frac{\rho_T^*(Y_i)}{\rho_T^\psi(Y_i)} \lesssim \sqrt{\Upsilon(n,\delta)\,\mathsf{KL}\left(\rho_T^*, \rho_T^\psi\right)} + \Upsilon(n,\delta),$$

where

$$\Upsilon(n,\delta) = (\Lambda d + M + d)\left(d + \log \frac{RLn}{\delta} + (M \vee \log \Lambda)\sqrt{d}e^{-bT}\right)\frac{D\log n}{n},$$

which holds simultaneously for all $\psi \in \Psi$ with probability at least $(1-2\delta)$. After that, we transform it into the desired excess risk bound and finish the proof.

References

R. Baptista, A.-A. Pooladian, M. Brennan, Y. Marzouk, and J. Niles-Weed. Conditional simulation via entropic optimal transport: Toward non-parametric estimation of conditional brenier maps. *arXiv preprint arXiv:2411.07154*, 2024.

A. Beurling. An automorphism of product measures. *Ann. Math. (2)*, 72:189–200, 1960. ISSN 0003-486X. doi: 10.2307/1970151.

Y. Chen, T. Georgiou, and M. Pavon. Entropic and displacement interpolation: a computational approach using the hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

Y. Chen, T. T. Georgiou, and M. Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *Siam Review*, 63(2):249–313, 2021.

A. Chiarini, G. Conforti, G. Greco, and L. Tamanini. A semiconcavity approach to stability of entropic plans and exponential convergence of Sinkhorn's algorithm, 2024. URL `https://arxiv.org/abs/2412.09235`.

G. Conforti. Weak semiconvexity estimates for schrödinger potentials and logarithmic sobolev inequality for schrödinger bridges, 2024. URL `https://arxiv.org/abs/2301.00083`.

G. Conforti, A. Durmus, and G. Greco. Quantitative contraction rates for sinkhorn algorithm: beyond bounded costs and compact marginals, 2024. URL `https://arxiv.org/abs/2304.04451`.

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied Mathematics and Optimization*, 23(1):313–329, 1991. doi: 10.1007/BF01445134.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34: 17695–17709, 2021.

S. Eckstein. Hilbert's projective metric for functions of bounded growth and exponential convergence of Sinkhorn's algorithm. *Probability Theory and Related Fields*, pages 1–37, 2025.

R. Fortet. Résolution d'un système d'équations de M. Schrödinger. *J. Math. Pures Appl. (9)*, 19: 83–105, 1940. ISSN 0021-7824.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

N. Gushchin, S. Kholkin, E. Burnaev, and A. Korotin. Light and optimal Schrödinger bridge matching. In *Forty-first International Conference on Machine Learning*, 2024a. URL `https://doi.org/10.48550/arXiv.2402.03207`.

N. Gushchin, D. Selikhanovych, S. Kholkin, E. Burnaev, and A. Korotin. Adversarial Schrödinger bridge matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL `https://arxiv.org/abs/2405.14449`.

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

A. Korotin, N. Gushchin, and E. Burnaev. Light Schrödinger bridge. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://doi.org/10.48550/arXiv.2310.01174`.

G. Lecué and C. Mitchell. Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, 6(none):1803 – 1837, 2012. doi: 10.1214/12-EJS730. URL `https://doi.org/10.1214/12-EJS730`.

C. Leonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar. I^{2}sb: Image-to-image Schrödinger bridge. In *Fortieth International Conference on Machine Learning*, 2023. URL `https://doi.org/10.48550/arXiv.2302.05872`.

H. D. March and P. Henry-Labordere. Building arbitrage-free implied volatility: Sinkhorn's algorithm and variants, 2023. URL https://arxiv.org/abs/1902.04456.

M. Pavon, G. Trigila, and E. G. Tabak. The data-driven Schrödinger bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573, 2021.

S. Peluchetti. Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling. *Journal of Machine Learning Research*, 24(374):1–51, 2023.

A.-A. Pooladian and J. Niles-Weed. Plug-in estimation of Schrödinger bridges. *arXiv preprint arXiv:2408.11686*, 2024.

G. Rapakoulias, A. R. Pedram, and P. Tsiotras. Go with the flow: Fast diffusion for gaussian mixture models. Preprint. ArXiv:2412.09059v3, 2024.

P. Rigollet and J.-C. Hütter. High-dimensional statistics. Preprint. ArXiv:2310.19244, 2023.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

E. Schrödinger. über die umkehrung der naturgesetze. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse*, pages 144–153, 1932.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion Schrödinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023.

M. G. Silveri, A. O. Durmus, and G. Conforti. Theoretical guarantees in KL for diffusion flow matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ia4WUCwHA9.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2314570.

A. Stromme. Sampling from a Schrödinger bridge. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4058–4067. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/stromme23a.html.

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

F. Vargas, P. Thodoroff, A. Lamacraft, and N. Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

G. Wang, Y. Jiao, Q. Xu, Y. Wang, and C. Yang. Deep generative learning via Schrödinger bridge. In *International conference on machine learning*, pages 10794–10804. PMLR, 2021.

Q. Zhang and Y. Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022. URL https://doi.org/10.48550/arXiv.2111.15141.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main result Theorem 1 fully corresponds to that stated in the Abstract and Introduction, confirming the specified Contribution of the article.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions we impose are discussed in Section 4 of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For the main result and auxiliary results, the assumptions are explicitly stated in Section 4. For Theorem 1, a sketch of the proof is given in Section 5, and references are given to auxiliary results in the Appendix, which are also completely proved.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper does not include numerical experiments and presents a theoretical derivation of a statistical bound, therefore the question about reproducibility does not apply to it.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper does not include numerical experiments and presents a theoretical derivation of a statistical bound.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include numerical experiments and presents a theoretical derivation of a statistical bound.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include numerical experiments and presents a theoretical derivation of a statistical bound.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include numerical experiments and presents a theoretical derivation of a statistical bound.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper is of purely theoretical nature, and the proposed methods do not deal with sensitive attributes that could induce unfairness or privacy issues.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This article does not have a direct social impact on society, as it is of theoretical nature. We are not aware of any cases where high-probability upper bound on the KL-divergence has a strong social impact on society.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This article does not contain anything that would require this kind of protection.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This article does not contain any existing assets that need to be referenced.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This article does not contain any new assets that would fit this question.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects, as it is theoretical in nature.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects, as it is of theoretical nature.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used in core method development in this research. We used LLM writing and editing purposes only.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.