Approximation Rates for Schrödinger Bridge Potentials via Fixed-Point ERM

Denis Belomestny[1, 2] **Alexey Naumov**[2] **Nikita Puchkin**[2] **Denis Suchkov**[2]

[1]Faculty of Mathematics, Duisburg-Essen University, Duisburg, Germany
[2]Faculty of Computer Science, HSE University, Moscow, Russian Federation

Abstract

The Schrödinger bridge problem (SBP) provides a principled interpolation between two distributions by selecting, among all path measures matching given endpoint marginals, the one closest in relative entropy to a reference dynamics. In modern applications the marginals are observed only through samples, and standard computational pipelines solve a discretized SBP via Sinkhorn iterations and then heuristically extend the resulting dual potentials off-sample, entangling statistical, optimization, and smoothing errors. We study a learning-theoretic alternative based on a fixed-point characterization of a single *transformed* Schrödinger potential $g^\star$, and we focus on quantitative approximation of $g^\star$ by a sample-based estimator $\widehat{g}$ that is continuous by construction. To address the intrinsic scaling ambiguity of Schrödinger potentials, we introduce a normalized, scale-invariant operator and analyze its local geometry around $g^\star$. Our main theoretical contribution is a stability result linking the error of the fixed-point residual to a distance to the solution $g^\star$ via analysis of spectral-gap property for the Fréchet derivative of the operator in a norm $\|\cdot\|$ being the sum of a localized Hilbert tangent seminorm and an L^2 distance. Combining this stability bound with the excess risk bounds and approximation error yields explicit non-asymptotic rates for $\|\widehat{g} - g^\star\|$. We illustrate performance of the suggested approach with numerical experiments.

1 INTRODUCTION

The Schrödinger bridge problem (SBP) provides a principled way to construct a stochastic interpolation between two probability distributions by selecting, among all path measures that match prescribed endpoint marginals, the one that is closest in relative entropy to a given reference dynamics. It originates from the seminal work Schrödinger [1932]. Formally, let $(X_t)_{t\in[0,T]}$ be a Markov process on $\mathbb{R}^d$ with reference law Q on path space and transition densities $(q_t)_{t\in(0,T]}$. We are given two probability densities ρ_0 and ρ_T on $\mathbb{R}^d$ and consider the class $\mathcal{P}(\rho_0, \rho_T)$ of all $P \ll \mathsf{Q}$ such that

$$P \circ X_0^{-1} = \rho_0\, dx \text{ and } P \circ X_T^{-1} = \rho_T\, dx\,.$$

The (dynamic) SBP consists in finding the probability measure $P^\star \in \mathcal{P}(\rho_0, \rho_T)$ which is closest to Q in the sense of relative entropy:

$$P^\star \in \arg\min_{P\in\mathcal{P}(\rho_0,\rho_T)} \mathcal{H}(P \,\|\, \mathsf{Q}), \tag{1}$$

where $\mathcal{H}(P \,\|\, \mathsf{Q}) := \int \log\Big(\frac{dP}{d\mathsf{Q}}\Big)\, dP$. It is a classical result (see, e.g., Chen et al. [2016]) that the minimizer $P^\star$ exists under mild conditions and has a *Schrödinger factorization* of the form

$$\frac{dP^\star}{d\mathsf{Q}}(X) \;=\; \nu_0(X_0)\,\nu_T(X_T), \tag{2}$$

for some nonnegative measurable functions (Schrödinger potentials) $\nu_0, \nu_T : \mathbb{R}^d \to (0,\infty)$. Taking time–0 and time–T marginals in (2) yields the system

$$\rho_0(x) = \nu_0(x) \int_{\mathbb{R}^d} q_T(x,z)\,\nu_T(z)\,dz, \tag{3}$$

$$\rho_T(y) = \nu_T(y) \int_{\mathbb{R}^d} q_T(x,y)\,\nu_0(x)\,dx. \tag{4}$$

The corresponding "static" Schrödinger bridge problem is the entropy minimization over couplings of (X_0, X_T):

$$\pi^\star \;\in\; \arg\min_{\pi\in\Pi(\rho_0,\rho_T)} \mathcal{H}\big(\pi \,\|\, \pi^{\mathrm{ref}}\big), \tag{5}$$

where $\pi^{\mathrm{ref}}(dx,dy) = \rho_0(x)\,q_T(x,y)\,dx\,dy$ and $\Pi(\rho_0,\rho_T)$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ρ_0 and ρ_T. The optimizer $\pi^\star$ can be written as

$$\pi^\star(dx,dy) \;=\; \nu_0(x)\,q_T(x,y)\,\nu_T(y)\,dx\,dy, \tag{6}$$

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:521–548.

and the potentials (ν_0, ν_T) solve (3)–(4).

Equivalently, the joint law of (X_0, X_T) under $P^\star$ solves an entropy minimization problem over couplings, which coincides with entropic optimal transport (EOT) when Q is a Wiener measure [Peyré and Cuturi, 2019].

In many modern applications, ρ_0 and ρ_T are not available as closed-form densities; instead, one observes samples $X_1, \ldots, X_N \sim \rho_0$ and $Y_1, \ldots, Y_M \sim \rho_T$. A standard computational route solves the static SBP or EOT problem on empirical measures using Sinkhorn iterations [Sinkhorn, 1967, Franklin and Lorenz, 1989, Cuturi, 2013, Peyré and Cuturi, 2019], obtains discrete approximations of the dual potentials on the support of the samples, and then constructs a time-inhomogeneous drift by smoothing and differentiating these objects [Pooladian and Niles-Weed, 2025]. This pipeline is effective in moderate dimensions but raises three learning-theoretic issues: (i) the potentials are only defined on-sample and must be extended off-sample (often heuristically); (ii) the final error couples statistical error, optimization error (due to a finite number of Sinkhorn iterations), and discretization/smoothing artifacts, complicating generalization analysis; and (iii) it is difficult to incorporate structural assumptions (constraints) on the potentials into the estimation procedure. The latter point is especially crucial in high-dimensional problems, where any additional structural insight (e.g. sparsity) can lead to a significant reduction in approximation error.

In this paper, we are using an alternative approach that estimates a continuous potential directly from samples. The starting point is an equivalent representation of the Schrödinger system in terms of a single positive transformed potential. Specifically, let

$$g^\star(y) \;:=\; \frac{\rho_T(y)}{\nu_T(y)}\,. \tag{7}$$

Simple algebra shows that $g^\star$ satisfies the following fixed-point equation

$$g^\star(y) = \int_{\mathbb{R}^d} \frac{q_T(x,y)\rho_0(x)}{D_{g^\star}(x)}\,dx =: \mathcal{C}[g^\star](y)\,, \tag{8}$$

where for $g : \mathbb{R}^d \to \mathbb{R}_+$

$$D_g(x) = \int_{\mathbb{R}^d} q_T(x,z)\,\frac{\rho_T(z)}{g(z)}\,dz \tag{9}$$

The map $g \mapsto \mathcal{C}[g]$ can be viewed as a two-stage averaging operator. First, for each starting point x, the quantity $D_g(x)$ aggregates the "backward" information from the terminal distribution by averaging $\rho_T(z)/g(z)$ against the reference kernel $q_T(x,z)$. Second, $\mathcal{C}[g](y)$ averages $1/D_g(x)$ over $x \sim \rho_0$, again through $q_T(x,y)$. A fixed point $g^\star = \mathcal{C}[g^\star]$ precisely encodes the Schrödinger system (3)–(4). Once a fixed point $g^\star$ of $\mathcal{C}$ is found, the Schrödinger potentials are recovered via

$$\nu_T(y) = \rho_T(y)/g^\star(y)$$

and the optimal Markov process $P^\star$ is obtained by tilting the reference process Q according to (2). In the sample-only regime, one may form an empirical operator $\widehat{\mathcal{C}}_{N,M}$ by replacing expectations with empirical averages and estimate g by minimizing a fixed-point residual over a hypothesis class $\mathcal{G}$:

$$\widehat{g}_{N,M} \in \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}_{N,M}$$

where

$$\widehat{\mathcal{R}}_{N,M} = M^{-1}\sum_{j=1}^{M} \ell(g(Y_j),\, \widehat{\mathcal{C}}_{N,M}[g](Y_j))$$

and ℓ is some loss function, i.e. $\ell(x,y) = (x-y)^2$. Note that the true population risk

$$\mathcal{R} = \mathbb{E}[\ell(g(Y),\, \widehat{\mathcal{C}}[g](Y))]$$

measures the discrepancy (fixed-point residual) of g in the underlying fixed point problem. This approach allows for the incorporation of structural assumptions on the transformed potential through the choice of the hypothesis class $\mathcal{G}$. Moreover, unlike Sinkhorn applied to empirical measures, this ERM viewpoint produces a *continuous* estimator by construction (e.g., via a neural network parameterization).

Contributions Our key *contributions* could be summarized as follows:

- We propose a novel approach to the Schrödinger bridge problem based on minimizing an empirical risk associated with the fixed-point equation for the transformed potential g. This formulation is flexible and allows for the incorporation of a wide range of structural assumptions and constraints, including sparsity.
- We establish theoretical guarantees for the estimator of the transformed potential in a mixed Hilbert seminorm,

$$\|\widehat{g}_{M,N} - g^\star\|^2_{R,\lambda}\,. \tag{10}$$

  (see the definition of $\|\cdot\|_{R,\lambda}$ below) without imposing compactness assumptions on ρ_T and obtain convergence rates that are nearly parametric. Our analysis exploits a deep connection between the contraction properties of the Schrödinger operator and the local quadratic structure of the underlying residual risk function.
- We evaluate the method on Gaussian-to-Gaussian benchmarks, demonstrating superior scalability up to $d = 100$ over SinkhornBridge. Neural-based empirical risk minimization yields significantly more accurate Schrödinger potentials and drift estimates, particularly in high-dimensional or data-limited regimes where discrete kernel-smoothed methods degrade.

Related literature The Schrödinger bridge problem has a long history in probability and optimal transport; see

Leonard [2014] for a modern survey and connections to entropic optimal transport. In the discrete setting, the Schrödinger system reduces to matrix scaling and can be solved via iterative proportional fitting / Sinkhorn iterations [Sinkhorn, 1967, Franklin and Lorenz, 1989, Cuturi, 2013, Peyré and Cuturi, 2019]. A key tool in analyzing these scaling dynamics is Hilbert's projective metric and related cone contraction techniques [Chen et al., 2016, Lemmens and Nussbaum, 2014, Eckstein, 2025], and recent work provides quantitative and non-asymptotic convergence and stability guarantees under increasingly general assumptions on costs and marginals [Conforti et al., 2026, Greco et al., 2023, Chiarini et al., 2024].

From a statistical viewpoint, many sample-based pipelines compute an EOT solution between empirical measures and then extend the resulting discrete dual potentials off-sample (often via smoothing) for downstream evaluation and differentiation [Pooladian and Niles-Weed, 2025]. Complementary to this, there is a growing literature on generalization and sample complexity of entropic OT objectives and their debiased variants (Sinkhorn divergences) [Feydy et al., 2019, Genevay et al., 2019, Mena and Niles-Weed, 2019]. In parallel, Schrödinger bridges have recently been used as learning primitives in generative modeling and diffusion-based transport, typically by parameterizing drifts/scores or using bridge matching objectives [Pavon et al., 2021, Wang et al., 2021, De Bortoli et al., 2021, Shi et al., 2023, Korotin et al., 2024], Puchkin et al. [2025], Puchkin et al. [2026],

Our work is complementary to these lines: building on the ERM formulation we directly learn the transformed potential $g^\star$ as a continuous object and study approximation rates for $\widehat{g}_{M,N}$.

2 MAIN RESULTS

We state the main assumptions used throughout the paper. We start from the assumptions on the kernel Q and densities ρ_0, ρ_T.

(Q) The reference kernel is Gaussian: for any $x, y \in \mathbb{R}^d$,

$$q_T(x,y) = (2\pi T)^{-d/2} \exp\left(-\frac{\|y-x\|^2}{2T}\right)$$

(R0) There exist $x_0 \in \mathbb{R}^d$ and two positive real numbers r_0, R_0 such that $B(x_0, r_0) \subseteq \mathrm{supp}(\rho_0) \subseteq B(x_0, R_0)$. Moreover, there is a constant $\rho_{0,-} > 0$ such that

$$\rho_0(x) \geq \rho_{0,-}, \qquad \forall x \in B(x_0, r_0).$$

(RT) There exist constants $0 < c_T^- \leq c_T^+ < \infty$ and $b_T^-, b_T^+ > 0$ such that

$$c_T^- \exp(-b_T^- \|y\|^2) \leq \rho_T(y) \leq c_T^+ \exp(-b_T^+ \|y\|^2),$$

for all $y \in \mathbb{R}^d$.

Instead of Gaussian assumption on the reference kernel Q one may assume two-sided sub-Gaussian bounds on $q_T(x,y)$ and local Lipschitz assumption on $\log q_T(x,y)$ w.r.t. to y for any $x \in B(x_0, R_0)$. It is also worth noting that we do not assume that ρ_T is a compactly supported density, which is a fairly standard assumption in the existing literature.

Note that Schrödinger potentials (ν_0, ν_T) are only defined up to a global multiplicative constant. This ambiguity is inherited by the transformed potential $g^\star$: rescaling g changes $\mathcal{C}[g]$ by the same factor, so the fixed-point equation does not identify an absolute scale. To obtain a well-posed stability analysis, we therefore fix a normalization and work with the associated scale-invariant map $\mathcal{T}$, for which $\mathcal{T}[cg] = \mathcal{T}[g]$ for all $c > 0$. This choice makes the fixed point locally identifiable and removes spurious directions corresponding to global rescaling. We define the normalized version of $\mathcal{C}[g]$ by

$$\mathcal{T}[g](y) := s(g)\, \mathcal{C}[g](y), \tag{11}$$

where

$$s(g) := \int_{\mathbb{R}^d} \frac{\rho_T(y)}{\mathcal{C}[g](y)}\, w(y)\, dy,$$

and $w \in (0,1)$ is some weighting function. Note that under this normalisation

$$\int_{\mathbb{R}^d} \frac{\rho_T(y)}{\mathcal{T}[g](y)}\, w(y)\, dy = 1.$$

Furthermore, assume $g^\star$ satisfies $\mathcal{T}[g^\star] = g^\star$. Then it is easy to see that $s(g^\star) = 1$ is satisfied iff

$$\int_{\mathbb{R}^d} \frac{\rho_T(y)}{g^\star(y)}\, w(y)\, dy = 1. \tag{12}$$

Note that $s(g^\star) = 1$ implies $\mathcal{T}[g^\star] = \mathcal{C}[g^\star]$, that is $\mathcal{T}$ and $\mathcal{C}$ coincide at the fixed-point $g^\star$. For simplicity we take weighted function $w_R = 1_{B_R}$ with some radius $R > 0$ to be defined later. Note that introducing the weighted function w is the primary needed for theoretical analysis to relate the Hilbert seminorm (see below) to the norm in $L^2(\rho_T)$. We impose additional assumption on the tails of $g^\star$.

(G*) There exist $a^\star > 0$ and $c^\star > 0$ such that

$$\frac{\rho_T(z)}{g^\star(z)} \leq c^\star e^{-a^\star \|z\|^2}.$$

Assumption *(G*)* is related to normalization (12) since we need to control the ratio $\rho_T / g^\star$ outside the ball B_R. This assumption is automatically satisfied if instead of (12) one requires

$$\int_{\mathbb{R}^d} \frac{\rho_T(y)}{g^\star(y)}\, dy = 1.$$

and $b_T^+ > 1/T$. Note that the difference is in the normalization of $g^\star$ by some constant. Using *(G*)* and applying Proposition A.1 in the appendix we may refine the estimates

for $g^\star$ and get two-sided bounds. Furthermore, Theorem B.1 in the appendix guarantees existence of such fixed-point solution $g^\star$.

We consider a class of functions $\mathcal{G}$ which satisfies the following assumptions *(G)*:

(G1) For every $g \in \mathcal{G}$,

$$\int_{\mathbb{R}^d} \frac{\rho_T(y)}{g(y)} w_R(y)\, dy = 1 \quad \text{with } w_R(y) = 1_{B_R}(y).$$

(G2) There exists $\gamma \in (0, 1]$ such that for every $g \in \mathcal{G}$,

$$g(y) \geq \gamma\, g^\star(y), \qquad \forall y \in \mathbb{R}^d. \tag{13}$$

Note that *(G1)* is technical assumption simplifying analysis. It is needed to relate the Hilbert seminorm (see below) to the norm in $L^2(\rho_T)$. In practice there is no need to normalize g and $\mathcal{C}[g]$ simultaneously. See numerical section for details. The second assumption *(G2)* is satisfied if we assume that $g \in \mathcal{G}$ has a *uniform* Gaussian lower bound: for all $y \in \mathbb{R}^d$

$$g(y) \geq c^-_{\mathcal{G}} e^{-a_{\mathcal{G}}\|y\|^2}, \tag{14}$$

for some $c^-_{\mathcal{G}} > 0$, where $a_{\mathcal{G}} := a^\star_+ = 1/(2T)$. Then Proposition A.1 in the appendix implies that (13) holds with $\gamma = (c^\star_+/c^-_{\mathcal{G}}) \wedge 1$.

Define

$$\widehat{g}_{N,M} \in \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_{N,M},$$

where for simplicity we consider empirical risk with quadratic loss function,

$$\widehat{\mathcal{R}}_{N,M} = M^{-1} \sum_{j=1}^{M} (g(Y_j) - \widehat{\mathcal{T}}_{N,M}[g](Y_j))^2\,,$$

and $\widehat{\mathcal{T}}_{N,M}[g]$ is obtained by replacing ρ_0, ρ_T in (11) and in $s(g)$ by the empirical measures

$$\widehat{\rho}^N_0 := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}, \qquad \widehat{\rho}^M_T := \frac{1}{M} \sum_{j=1}^{M} \delta_{Y_j}.$$

We measure the error of approximation of $g^\star$ in the norm $\|\cdot\|_{R,\lambda}$ defined as follows. We fix $R > 0$ and define the localized Hilbert tangent seminorm

$$\|u\|_{H,R} := \frac{1}{2}\Big(\operatorname*{ess\,sup}_{y \in B_R} \frac{u(y)}{g^\star(y)} - \operatorname*{ess\,inf}_{y \in B_R} \frac{u(y)}{g^\star(y)} \Big).$$

Then for fixed $\lambda > 0$ we define

$$\|u\|_{R,\lambda} := \|u\|_{H,R} + \lambda\, \|u\|_{L^2(\mathbb{R}^d)}. \tag{15}$$

The Hilbert tangent seminorm $\|\cdot\|_{H,R}$ measures the oscillation of the *relative* perturbation $u/g^\star$ on a ball B_R. This is the natural infinitesimal quantity underlying Hilbert-metric contraction arguments for Sinkhorn operators. However, $\|\cdot\|_{H,R}$ is local: it does not control what happens in the tails outside B_R. The additional $\lambda\|u\|_{L^2}$ term provides a global tail control and allows us to work on $\mathbb{R}^d$ without compactness assumptions. The parameters R and λ quantify a trade-off: increasing R improves localization (less tail leakage) but may weaken contraction constants inside the ball, while λ balances local multiplicative control and global L^2 control.

We state the main result of this paper.

Theorem 2.1. *Under assumptions* (Q), (R0), (RT) *with* $b^+_T > 2/T$, (G*), (G) *there exist* $C, r, R, L, \lambda > 0$ *and* $\alpha \in (0, 1)$ *such that*

$$\|g - g^\star\|^2_{R,\lambda} \leq C \|g - \mathcal{T}[g]\|^2_{L^2(\rho_T)}$$

whenever $\|g - g^\star\|_{R,\lambda} \leq \min\{r, (1-\alpha)/L\}$.

We first note that

$$\|g - \mathcal{T}[g]\|^2_{L^2(\rho_T)} = \mathcal{R}(g), \tag{16}$$

which is population counterpart of $\widehat{\mathcal{R}}_{N,M}$. Hence, Theorem 2.1 converts a statement about the risk estimation into an estimation statement about $\|g - g^\star\|_{R,\lambda}$. In particular, if empirical risk minimization produces an estimator with a small residual $\|g_{M,N} - \mathcal{T}[g_{M,N}]\|_{L^2(\rho_T)}$, then the theorem implies that $g_{M,N}$ is close to $g^\star$ in $\|\cdot\|_{R,\lambda}$, provided we are in the local neighborhood where the stability inequality holds. This reduction isolates the learning problem to controlling the residual, which can be handled by standard empirical process tools for the chosen hypothesis class. Note that locality of the above result is common for many non-convex optimization problems and related to the fact that the underlying residual operator can be convex only in a vicinity of the solution $g^\star$. Furthermore, we show that a class $\mathcal{G}$ can be chosen so that assumptions *(G)* are satisfied and $\mathbb{E}\mathcal{R}(\widehat{g}_{M,N})$ could be made small. It is easy to see from (8) that $g^\star$ is a convolution of the Gaussian kernel with function $\rho_0/D_{g^\star}$. An appropriate class for theoretical analysis of approximation of such functions is the class obtained by projection on the span of Hermite functions. Note that in practice one may use other functional classes, e.g. neural networks. We obtain the following result.

Corollary 2.2. Under assumptions *(Q)*, *(R0)*, *(RT)* with $b^+_T > 4/T$, *(G*)*, there exists a class $\mathcal{G}$ that satisfies *(G)*, and constants $C, r, R, L, \lambda > 0$ and $\alpha \in (0, 1)$ such that

$$\mathbb{E}\|\widehat{g}_{M,N} - g^\star\|^2_{R,\lambda} 1_A \lesssim \log^{\frac{d}{2}} \max(M, N) \Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}} \Big)\,,$$

where 1_A is the indicator of the event

$$A = \{\|\widehat{g}_{M,N} - g^\star\|_{R,\lambda} \leq \min\{r, (1-\alpha)/L\}\}.$$

Note that Theorem 2.1 and Corollary 2.2 imply convergence in $L^2(\rho_T)$ as well. Indeed, we note from the definition of $\|\cdot\|_{R,\lambda}$ that

$$\|u\|_{L^2(\rho_T)} \le \lambda^{-1}\|u\|_{R,\lambda}\,.$$

2.1 SKETCH OF THE PROOF

We start this section with the general result for operators between Banach spaces. Let $(\mathcal{X}, \|\cdot\|)$ be a Banach space of functions on $\mathbb{R}^d$ and let $\mathcal{T} : \mathcal{U} \subset \mathcal{X} \to \mathcal{X}$ be Fréchet differentiable on an open neighborhood $\mathcal{U}$ of $g^\star \in \mathcal{X}$. We assume that there exists $r > 0$ such that the closed ball

$$\overline{B}(g^\star, r) := \{g \in \mathcal{X} : \ \|g - g^\star\| \le r\}$$

is contained in $\mathcal{U}$. Let also $g^\star$ be a fixed point, $\mathcal{T}[g^\star] = g^\star$, and define the residual map

$$\widetilde{F}(g) := g - \mathcal{T}[g].$$

Denote $\mathsf{T} := \mathcal{T}'[g^\star]$. Let $\|\cdot\|$ denotes the induced operator norm on $\mathcal{L}(\mathcal{X}, \mathcal{X})$. Suppose:

(A1) (*Spectral gap / strong monotonicity at $g^\star$*) There exists $\alpha \in [0, 1)$ such that for all $u \in \mathcal{X}$,

$$\|(I - \mathsf{T})u\| \ \ge \ (1 - \alpha)\,\|u\|. \tag{17}$$

(A2) (*Local Lipschitz continuity of the derivative*) There exist constants $L > 0$ and $r > 0$ such that for all $v \in \mathcal{X}$ with $\|v\| \le r$,

$$\|\mathcal{T}'[g^\star + v] - \mathsf{T}\| \ \le \ L\,\|v\|. \tag{18}$$

The following result allows to obtain local lower bound on $\widetilde{F}(g)$.

Proposition 2.3. *Assume* (A1) *and* (A2)*. Then for every $u \in \mathcal{X}$ with $\|u\| \le r$,*

$$\|\widetilde{F}(g^\star + u)\| \ \ge \ (1 - \alpha)\,\|u\| \ - \ \frac{L}{2}\,\|u\|^2. \tag{19}$$

Proposition 2.3 formalizes a standard idea in fixed-point estimation: if the linearization $\mathcal{T}'[g^\star]$ is a *strict contraction* (assumption (A1)), then the residual $F_e(g) = g - \mathcal{T}[g]$ behaves like an invertible linear map near $g^\star$. Assumption (A2) ensures that higher-order terms do not destroy this behavior locally. Together, these conditions imply a *local error bound*: small fixed-point residual forces g to be close to $g^\star$.

Proof of Proposition 2.3. We provide the proof for completeness. It is based on Taylor's expansion and the fact the $g^\star$ is the fixed point of $\mathcal{T}$. Fix $u \in \mathcal{X}$ with $\|u\| \le r$. By the mean–value integral form,

$$\mathcal{T}[g^\star + u] = \mathcal{T}[g^\star] + \int_0^1 \mathcal{T}'[g^\star + tu]\ u\,dt.$$

Since $\mathcal{T}[g^\star] = g^\star$, we obtain

$$\widetilde{F}(g^\star + u) = u - \int_0^1 \mathcal{T}'[g^\star + tu]\,u\,dt.$$

We add and subtract $\mathsf{T}u$,

$$\widetilde{F}(g^\star + u) = (I - \mathsf{T})u - \int_0^1 \big(\mathcal{T}'[g^\star + tu] - \mathsf{T}\big)u\,dt.$$

By (17),

$$\|(I - \mathsf{T})u\| \ge (1 - \alpha)\|u\|.$$

By (18) with $v = tu$ and $\|u\| \le r$,

$$\|\mathcal{T}'[g^\star + tu] - \mathsf{T}\| \le L\|tu\| = Lt\|u\|.$$

The last two bounds imply that

$$\|\widetilde{F}(g^\star + u)\| \ge (1 - \alpha)\|u\| - \frac{L}{2}\|u\|^2\,.$$

□

Note that for all $\|u\| \le \min\{r, (1 - \alpha)/L\}$,

$$\|\widetilde{F}(g^\star + u)\| \ \ge \ \frac{1 - \alpha}{2}\,\|u\|. \tag{20}$$

Equivalently, writing $g = g^\star + u$ and $\|g - g^\star\| = \|u\|$,

$$\|g - g^\star\| \le \frac{2}{1 - \alpha}\,\|\widetilde{F}(g)\|, \tag{21}$$

whenever $\|g - g^\star\| \le \min\{r, (1 - \alpha)/L\}$. Moreover, squaring (20) yields the *quadratic* lower bound for the squared residual:

$$\|\widetilde{F}(g)\|^2 \ge \frac{(1 - \alpha)^2}{4}\,\|g - g^\star\|^2. \tag{22}$$

Note that this proposition provides a relation between $\|g - g^\star\|$ and $\|\widetilde{F}(g)\|$. To apply it for estimation of (10) we need to choose an appropriate norm which satisfies assumptions and relate $\|\widetilde{F}(g)\|$ to the population risk $\mathcal{R}(g)$.

We show that $\mathcal{T}$ defined in (11) satisfies *(A1)*, *(A2)* in the norm $\|u\|_{R,\lambda}$ which is defined in (15).

Lemma 2.4. *Under assumptions* (Q)*,* (R0)*,* (RT)*,* (G*)*,* (G) *there exist $R > 0, \lambda > 0$ and $\alpha_R(\lambda) \in (0, 1)$ such that for any $u \in \mathcal{G}$,*

$$\|\mathsf{T}u\|_{R,\lambda} \le \alpha_R(\lambda)\,\|u\|_{R,\lambda}\,. \tag{23}$$

Proof. See Lemma A.6 in the appendix. □

Lemma 2.4 establishes that the Fréchet derivative of the normalized map at the fixed point is a contraction in $\|\cdot\|_{R,\lambda}$. Equivalently, the operator $(I - \mathsf{T})$ has a "spectral gap" in this norm, which is exactly assumption (A1). Lemma A.2 in

the appendix states that assumption *(A2)* follows from the uniform estimate of the second derivative:

$$\sup_{\|g-g^\star\|\le r} \|\mathcal{T}''[g]\| \;\le\; L,$$

where $\|\mathcal{T}''[g]\|$ denotes the induced operator norm of the bounded bilinear map $\mathcal{T}''[g] : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$:

$$\|\mathcal{T}''[g]\| := \sup_{\|u\|\le 1,\ \|v\|\le 1} \|\mathcal{T}''[g](u,v)\|.$$

The latter may be estimated directly under the assumption of Theorem 2.1. It remains to apply Lemma 2.4 and Proposition 2.3 to finish the proof of Theorem 2.1. We provide more details.

Proof of Theorem 2.1. Proposition 2.3 implies that

$$\|g - g^\star\|_{R,\lambda} \le 2(1-\alpha)^{-1}\|\widetilde{F}(g)\|_{R,\lambda}$$

provided that $\|g - g^\star\|^2_{R,\lambda}$ small enough. We note that for arbitrary u

$$\|u\|_{H,R} \le \big\|u/g^\star - 1\big\|_{L^\infty(B_R)}. \tag{24}$$

Using shift invariance of essential extrema one may check that $\|u\|_{H,R} = \|u + g^\star\|_{H,R}$. Applying this fact and Proposition A.1 (eq. (27)) we get

$$\|u\|_{H,R} \le \frac{e^{(a_-^\star + b_T^-)R^2}}{c_-^\star c_T^-}\|u\|_{L^2(\rho_T)}. \tag{25}$$

As a result, we have that

$$\|g-g^\star\|_{R,\lambda} \le \frac{2}{1-\alpha}\left\{\frac{e^{(a_-^\star+b_T^-)R^2}}{c_-^\star c_T^-} + \lambda\right\}\|\widetilde{F}(g)\|_{L^2(\rho_T)}$$

for $\|g - g^\star\|_{R,\lambda}$ small enough.

□

Proof of Corollary 2.2. We will follow the arguments from Belomestny et al. [2026]. Let $\{\psi_\alpha^{(\lambda)}\}_{\alpha\in\mathbb{N}_0^d}$ be the scaled Hermite basis with $\lambda = T^{-1/2}$ and Π_n be the $L^2(\mathbb{R}^d)$–orthogonal projector onto $\operatorname{span}\{\psi_\alpha^{(\lambda)} : |\alpha| \le n\}$. Define

$$\mathcal{G} = \mathcal{G}_n(B) := \left\{ g = \sum_{|\alpha|\le n} c_\alpha \psi_\alpha^{(\lambda)} : \sum_{|\alpha|\le n} c_\alpha^2 \le B \right\}.$$

where B is some constant depending on n and d. We additionally make clipping and normalizations to satisfy assumptions *(G)*. Replacing $\mathcal{C}[g]$ by $\mathcal{T}[g]$ and repeating the arguments from Belomestny et al. [2026][Theorem 2] we may obtain the following result

$$\mathbb{E}[\mathcal{R}(\widehat{g}_{N,M})] \lesssim \log^{d/2}\max(M,N)\Big(N^{-1/2} + M^{-1/2}\Big),$$

where $\lesssim$ stands for inequality up to constants independent of M and d and double logarithmic factors. To finish the proof it remains to use (16).

□

3 NUMERICS

We evaluate ERMBridge along two complementary axes. First, in the Gaussian-to-Gaussian setting — where the Schrödinger potential, the optimal drift, and $g^\star$ are all analytically computable [Bunne et al., 2023] — we verify that the empirical risk minimization procedure recovers the actual objects studied in our theory, namely $\widehat{g}_{M,N}$, the potential ν_T, and the drift, rather than only downstream sample quality. This controlled regime is our primary diagnostic by design: it is the only setting in which the targets of the stability inequality have closed form, so it lets us measure convergence of the estimator directly (Figures 1–3). It also illustrates the central practical message of our analysis: when the structure of the transformed potential is known, one can choose a functional class $\mathcal{G}$ in which the residual loss is easy to optimize — here the quadratic class, which provably contains the true log-potential [Bunne et al., 2023].

Second, to show that the method is practical beyond settings with analytic ground truth, we evaluate ERMBridge on non-Gaussian targets (heavy-tailed and multimodal mixtures) and on real single-cell trajectory data, where no closed-form potential exists (Tables 1–3). Throughout, the baseline is SinkhornBridge [Pooladian and Niles-Weed, 2025]. For the high-dimensional mixtures we report sliced 1-Wasserstein (a standard tractable proxy in high dimension) together with the energy distance; for the single-cell benchmark we follow prior work and report the (unsliced) 1-Wasserstein distance. The formal algorithm, hyperparameter values, compute budget, and further experimental details are in Appendix C.10; the setups for the non-Gaussian and real-data experiments are in Appendix C.5.

Potential smoothing technique in SinkhornBridge and ERMBridge The SinkhornBridge algorithm utilizes the entropic potentials $(\hat{f}_s, \hat{g}_s)$ obtained directly from the static Sinkhorn algorithm. The smoothing is performed by the heat semigroup $H_{(1-t)\epsilon}$, which acts as a temporal antialiasing filter whose variance is explicitly coupled to the time remaining $(1-t)$ and the EOT regularization ϵ. In contrast, ERMBridge replaces the discrete potential $\hat{g}_s$ with a learned neural log-potential. Here the smoothing is two-fold: an implicit smoothness is enforced by the network's inductive bias during training, and an explicit kernel smoothing is applied during drift computation using a diffusion schedule $\sigma(t)$.

VERIFICATION AGAINST CLOSED-FORM GROUND TRUTH (GAUSSIAN-TO-GAUSSIAN)

In the three experiments of this block we set $\rho_0 = \mathcal{N}(m_0, C_0)$ and $\rho_T = \mathcal{N}(m_1, C_1)$. We optimize the logarithmic Schrödinger potential; because normalization is performed on finite minibatches drawn from the training set,

comparisons against the analytic targets can incur a constant offset. To isolate the shape of the estimate from this scale artifact, all functions (g, ν, and the drift) are compared with the true ones up to a constant shift, numerically chosen to minimize the corresponding MSE. (This scale freedom is exactly the multiplicative ambiguity removed by the normalized operator $\mathcal{T}$. We set $N = M$ throughout for brevity and ease of comparison.

Convergence of $\widehat{g}_{M,N}$ to $g^\star$ We demonstrate that the ERM procedure approximates $g^\star$ as the amount of training data grows. We report

$$\mathrm{MSE}(\widehat{g}_{M,N},\, g^\star) = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{g}_{M,N}(y_i) - g^\star(y_i)\big)^2,$$

where $(y_i)_{i=1}^n$ are i.i.d. $\mathcal{N}(m_1, C_1)$. The monotonically decreasing MSE in Figure 1 on the holdout sample indicates that the estimate converges to the true function (see Appendix C.2).

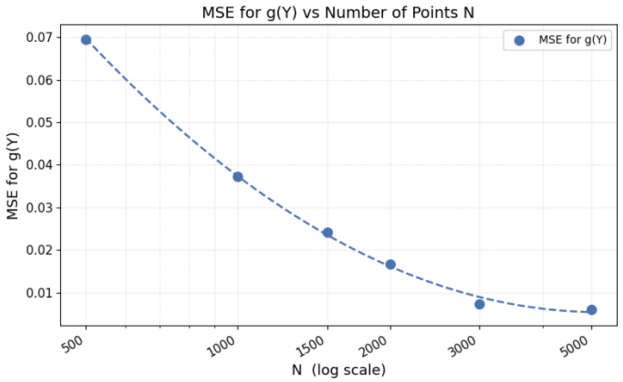

Figure 1: Convergence of $\widehat{g}_{M,N}$ to $g^\star$ for Gaussian-to-Gaussian translation in 2D, with neural-network approximation of the log-Schrödinger potential.

Schrödinger potential estimation across dimensions We compare ERMBridge and SinkhornBridge on estimation of the Schrödinger potential. We fix the number of estimation points and holdout points and sweep dimensions $[5, 10, 25, 50, 100]$ (Appendix C.3). We report

$$\mathrm{MSE}(\nu_T^{M,N},\, \nu_T) = \frac{1}{n}\sum_{i=1}^{n}\big(\nu_T^{M,N}(y_i) - \nu_T(y_i)\big)^2,$$

$$\nu_T^{M,N}(y) = \frac{\rho_T(y)}{\widehat{g}_{M,N}(y)},$$

with $(y_i)_{i=1}^n$ i.i.d. $\mathcal{N}(m_1, C_1)$. We run ERMBridge both with a neural network and with a quadratic class, the latter known to contain the true log-potential for Gauss-to-Gauss [Bunne et al., 2023]. As Figure 2 shows, the SinkhornBridge estimate degrades significantly faster with dimension than ERMBridge in the quadratic class, and faster still than ERMBridge with a neural network.

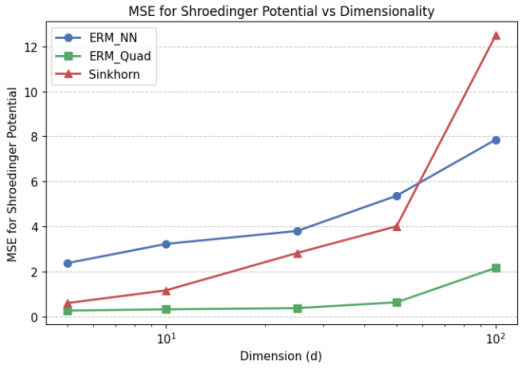

Figure 2: Estimation of $\nu_T^{M,N}$ for Gaussian-to-Gaussian translation in dimensions $[5, 10, 25, 100]$: ERMBridge (neural network), ERMBridge (quadratic class), and SinkhornBridge.

This demonstrates the advantage of learning a continuous approximation of the log-potential over the discrete Sinkhorn estimate.

Optimal drift estimation Finally we compare the quality of the drift, which determines sample quality in the Euler–Maruyama SDE discretization (Appendix C.11). With [Dai Pra, 1991]

$$b^\star(x,t) := \nabla \log \int_{\mathbb{R}^d} q_{T-t}(x,z)\,\nu_T(z)\,dz,$$

$$b_{M,N}(x,t) := \nabla \log \int_{\mathbb{R}^d} q_{T-t}(x,z)\,\nu_T^{M,N}(z)\,dz,$$

we report

$$\mathrm{MSE}(b_{M,N},\, b^\star) = \frac{1}{n}\sum_{i=1}^{n}\big\|b_{M,N}(t_i,x_i) - b^\star(t_i,x_i)\big\|^2,$$

with (x_i) i.i.d. $\mathcal{N}(m_0, C_0)$ and (t_i) i.i.d. $U[0.05, 0.95]$. In settings with small N, large d, or both, the SinkhornBridge MSE is substantially larger than ERMBridge (Figure 3).

When does each method win? In low dimensions with sufficiently many samples, SinkhornBridge can produce a competitive — sometimes better — drift estimate than ERMBridge, as the left side of Figure 3 shows. This is expected: in low dimension, discretization-and-smoothing approximates the potential well and does not require a large sample for a functional-class fit to pay off. As the dimension grows, however, such methods suffer more acutely from the curse of dimensionality, and their quality degrades too rapidly to remain usable on complex high-dimensional problems such as biological or image data. ERMBridge, which

Table 1: Transport from a standard Gaussian to an anisotropic Gaussian mixture. Sliced 1-Wasserstein, energy distance, and recovered modes (out of 25).

d	Method	sliced $\mathbb{W}_1 \downarrow$	Energy$\downarrow$	Active modes	Energy Δ
10	ERMBridge	**0.1391**	**0.0054**	25/25	$+49.9\%$
	SinkhornBridge	0.1426	0.0107	25/25	
50	ERMBridge	**0.1527**	**0.0078**	25/25	$+29.8\%$
	SinkhornBridge	0.2086	0.0111	25/25	

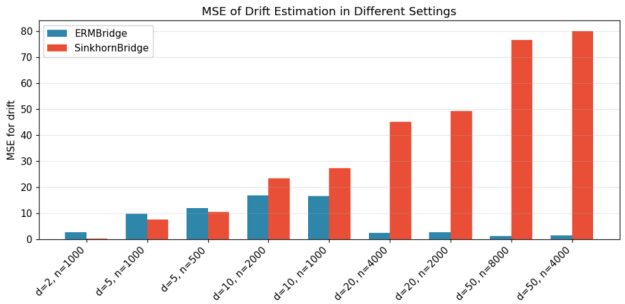

Figure 3: Drift estimation for Gaussian-to-Gaussian translation over $(d, N) \in \{(2, 1000), (5, 1000), (5, 500), (10, 2000), (10, 1000), (20, 4000), (20, 2000), (50, 8000), (50, 4000)\}$: ERMBridge (neural-network log-potential) vs. SinkhornBridge.

fits a continuous object directly, is comparatively robust in the high-d, data-limited regime — precisely the regime its non-asymptotic rates target.

BEYOND GAUSSIAN: NON-GAUSSIAN AND REAL-DATA BRIDGES

The Gaussian-to-Gaussian benchmarks above verify the theory's objects but do not, on their own, show practicality where no closed form exists. We therefore evaluate three further settings, summarized in Tables 1–3; setups are in Appendix C.5.

Multimodal, high-dimensional target (Gaussian $\rightarrow$ anisotropic mixture) We stress the multimodal regime with a 25-component anisotropic Gaussian mixture placed on a fixed 5×5 grid embedded in $\mathbb{R}^d$; each component has an independently drawn elongated covariance with random orientation, so accurate transport requires per-location adaptation rather than a single global bandwidth. ERMBridge recovers all 25 modes in both $d = 10$ and $d = 50$ and attains lower sliced $\mathbb{W}_1$ and energy distance than SinkhornBridge (energy-distance advantage $+49.9\%$ at $d = 10$ and $+29.8\%$ at $d = 50$; Table 1).

Table 2: Transport from a standard Gaussian to a Laplace mixture. ERMBridge vs. SinkhornBridge under sliced 1-Wasserstein and energy distance.

d	Method	sliced $\mathbb{W}_1 \downarrow$	Energy$\downarrow$	Energy Δ
2	ERMBridge	**0.1100**	**0.0098**	$+11.0\%$
	SinkhornBridge	0.1394	0.0110	
10	ERMBridge	**0.0756**	**0.0030**	$+56.6\%$
	SinkhornBridge	0.1110	0.0069	

Heavy-tailed target (Gaussian $\rightarrow$ Laplace mixture) We transport a standard Gaussian to a Laplace mixture, a heavy-tailed target with no closed-form potential. ERMBridge improves on SinkhornBridge in both sliced 1-Wasserstein and energy distance at $d = 2$ and $d = 10$, with the energy-distance margin widening from $+11.0\%$ to $+56.6\%$ as the dimension grows (Table 2).

Real data (single-cell trajectory inference) We evaluate on the EB single-cell RNA-sequencing dataset [Tong et al., 2020]. Following Tong et al. [2024a], we transport cells from t_{i-1} to t_{i+1} and report the 1-Wasserstein distance between predicted and held-out distributions at the intermediate time t_i, averaged over $i \in \{1, 2, 3\}$ and five seeds. ERMBridge attains $\mathbb{W}_1 = 0.822 \pm 0.037$ (Table 3), placing it in the top cluster of solvers and statistically indistinguishable from the strongest entropic-OT and flow-matching baselines, including SinkhornBridge (0.827 ± 0.026) and LightSB (0.823 ± 0.017). Beyond confirming practicality on a non-synthetic task, this provides the comparison against modern Schrödinger-bridge and diffusion-bridge methods beyond SinkhornBridge requested in review.

4 CONCLUSION

Our results show that Schrödinger bridge estimation from samples can be analyzed cleanly through a fixed-point learning perspective. The key step is to remove the unavoidable global scaling ambiguity by working with a normalized (scale-invariant) operator, which turns the problem into minimizing a squared fixed-point residual over a chosen hypoth-

esis class. We prove a local stability inequality around the true transformed potential $g^\star$: in this neighborhood, controlling the residual in $L^2(\rho_T)$ controls the estimation error in a mixed Hilbert-L^2 norm. This provides a direct link between standard excess-risk bounds for ERM and convergence rates for continuous estimators, without mixing discretization and smoothing artifacts into the statistical analysis. The numerical experiments on Gaussian-to-Gaussian bridges support this picture and indicate favorable scaling compared to kernel-smoothed Sinkhorn-based pipelines in higher dimensions.

Table 3: Single-cell trajectory inference on the EB dataset. Following Tong et al. [2024a], we transport cells from t_{i-1} to t_{i+1} and report $\mathbb{W}_1$ to the held-out distribution at t_i, averaged over $i \in \{1, 2, 3\}$ and five seeds. Solvers are listed in ascending $\mathbb{W}_1$, continuing from the left block to the right. Baselines: OT-CFM, I-CFM, and SB-CFM [Tong et al., 2024b]; the $[\mathrm{SF}]^2$M variants [Tong et al., 2024a]; T. Net [Tong et al., 2020]; LightSB [Korotin et al., 2024]; Reg. CNF [Finlay et al., 2020]; SinkhornBridge [Pooladian and Niles-Weed, 2025]; DSB [De Bortoli et al., 2021]; NLSB [Koshizuka and Sato, 2023]; and DSBM [Shi et al., 2023].

Solver	$\mathbb{W}_1 \downarrow$
OT-CFM	0.790 ± 0.068
$[\mathrm{SF}]^2$M-Exact	0.793 ± 0.066
ERMBridge (ours)	$\mathbf{0.822 \pm 0.037}$
LightSB	0.823 ± 0.017
Reg. CNF	$0.825 \pm$ N/A[1]
SinkhornBridge	0.827 ± 0.026
T. Net	$0.848 \pm$ N/A[1]
DSB	0.862 ± 0.023
I-CFM	0.872 ± 0.087
$[\mathrm{SF}]^2$M-Geo	0.879 ± 0.148
NLSB	$0.970 \pm$ N/A[1]
$[\mathrm{SF}]^2$M-Sink	1.198 ± 0.342
SB-CFM	1.221 ± 0.380
DSBM	1.775 ± 0.429

Several directions appear promising:

- A key open question is whether one can strengthen the analysis to obtain rates for $\|g_{M,N} - g^\star\|^2_{R,\lambda}$ of order $1/N + 1/M$ for instance by establishing a Bernstein/strong-convexity condition for the population risk around $g^\star$ (or an analogous quadratic growth property) and combining it with localized complexity arguments;
- Our stability result is local; proving a global error bound or a two-stage argument (coarse global control + local refinement) would make the guarantees fully non-asymptotic without a neighborhood assumption;
- The present rates can be pessimistic in high dimensions. Incorporating structural assumptions (e.g., sparse/low-rank parametrizations or symmetry constraints) and proving adaptive rates is an important direction, aligned with the motivation for ERM-based formulations;
- While the Gaussian case provides clean contraction/tail estimates, it would be valuable to extend the stability theory to more general Markov kernels (e.g., sub-Gaussian bounds, or diffusions with state-dependent coefficients);
- For downstream use in bridge sampling, it is natural to translate $L^2(\rho_T)$ error of g into quantitative error bounds for the inferred drift and for the resulting path measure, including discretization and optimization errors induced by finite training and SDE solvers.

[1] Standard deviation not reported by the original authors.

Acknowledgements

The study was implemented in the framework of the Basic Research Program at HSE University (HSE-BR-2025-019).

References

D. Belomestny, A. Naumov, N. Puchkin, and D. Suchkov. Schrödinger bridge problem via empirical risk minimization. In *ICLR 2026 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2026.

C. Bunne, Y.-P. Hsieh, M. Cuturi, and A. Krause. The Schrödinger bridge between Gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, pages 5802–5833. PMLR, 2023.

Y. Chen, T. 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.

A. Chiarini, G. Conforti, G. Greco, and L. Tamanini. A semiconcavity approach to stability of entropic plans and exponential convergence of Sinkhorn's algorithm. Preprint. ArXiv:2412.09235, 2024.

G. Conforti, A. Durmus, and G. Greco. Quantitative contraction rates for Sinkhorn's algorithm: beyond bounded costs and compact marginals. Preprint. ArXiv:2304.04451, 2026.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013.

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

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pages 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.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 2019.

C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. In *International conference on machine learning*, pages 3154–3164. PMLR, 2020.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 2019.

G. Greco, M. Noble, G. Conforti, and A. Durmus. Non-asymptotic convergence bounds for Sinkhorn iterates and their gradients: A coupling approach. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 716–746. PMLR, 2023.

A. Korotin, N. Gushchin, and E. Burnaev. Light Schrödinger bridge. In *International Conference on Learning Representations*, 2024.

T. Koshizuka and I. Sato. Neural Lagrangian Schrödinger bridge: Diffusion modeling for population dynamics. Preprint. ArXiv:2204.04853, 2023.

B. Lemmens and R. Nussbaum. Birkhoff's version of Hilbert's metric and applications. In *Handbook of Hilbert Geometry*, IRMA Lectures in Mathematics and Theoretical Physics, pages 275–303. European Math. Soc., 2014.

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. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32, pages 4543–4553, 2019.

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

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6): 355–607, 2019.

A.-A. Pooladian and J. Niles-Weed. Plug-in estimation of Schrödinger bridges. *SIAM Journal on Mathematics of Data Science*, 7(3):1315–1336, 2025.

N. Puchkin, I. Pustovalov, Y. Sapronov, D. Suchkov, A. Naumov, and D. Belomestny. Sample complexity of Schrödinger potential estimation. Preprint. ArXiv:2506.03043, 2025.

N. Puchkin, D. Suchkov, A. Naumov, and D. Belomestny. Tight bounds for Schrödinger potential estimation in unpaired data translation. In *The Fourteenth International Conference on Learning Representations*, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 36, pages 62183–62223. Curran Associates, Inc., 2023.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

A. Tong, J. Huang, G. Wolf, D. van Dijk, and S. Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9526–9536. PMLR, 2020.

A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024a. Expert Certification.

A. Y. Tong, N. Malkin, K. Fatras, L. Atanackovic, Y. Zhang, G. Huguet, G. Wolf, and Y. Bengio. Simulation-free Schrödinger bridges via score and flow matching. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1279–1287. PMLR, 2024b.

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.

Approximation Rates for Schrödinger Bridge Potentials via Fixed-Point ERM (Supplementary Material)

Denis Belomestny[1, 2] **Alexey Naumov**[2] **Nikita Puchkin**[2] **Denis Suchkov**[2]

[1]Faculty of Mathematics, Duisburg-Essen University, Duisburg, Germany
[2]Faculty of Computer Science, HSE University, Moscow, Russian Federation

A PROOF OF MAIN RESULTS

Proof of Proposition 2.3. Fix $u \in \mathcal{X}$ with $\|u\| \le r$. By the mean–value integral form,

$$\mathcal{T}[g^\star + u] = \mathcal{T}[g^\star] + \int_0^1 \mathcal{T}'[g^\star + tu]\, u\, dt.$$

Since $\mathcal{T}[g^\star] = g^\star$, we obtain

$$\widetilde{F}(g^\star + u) = u - \int_0^1 \mathcal{T}'[g^\star + tu]\, u\, dt.$$

We add and subtract $\mathsf{T}u$,

$$\widetilde{F}(g^\star + u) = (I - \mathsf{T})u - \int_0^1 \big(\mathcal{T}'[g^\star + tu] - \mathsf{T}\big)u\, dt.$$

By (17),

$$\|(I - \mathsf{T})u\| \ge (1 - \alpha)\|u\|.$$

By (18) with $v = tu$ and $\|u\| \le r$,

$$\|\mathcal{T}'[g^\star + tu] - \mathsf{T}\| \le L\|tu\| = Lt\|u\|.$$

The last two bounds imply that

$$\|\widetilde{F}(g^\star + u)\| \ge (1 - \alpha)\|u\| - \frac{L}{2}\|u\|^2,$$

which proves (19). If $\|u\| \le (1 - \alpha)/L$, then $(L/2)\|u\|^2 \le \frac{1-\alpha}{2}\|u\|$, so (20) follows. Finally, (22) is obtained by squaring (20).

□

Proposition A.1 (Two–sided Gaussian bounds for the fixed point)**.** *Assume* (Q)*,* (R0)*,* (RT) *and* (G*)*. Let* $g^\star : \mathbb{R}^d \to (0, \infty)$ *be measurable and satisfy*

$$g^\star(y) = \mathcal{C}[g^\star](y), \qquad y \in \mathbb{R}^d, \qquad \text{and} \qquad \int_{B_R} \frac{1}{g^\star(z)}\, \rho_T(z)\, dz = 1. \tag{26}$$

Set $a_-^\star := 1/T$, $a_+^\star := 1/(2T)$*. Then there exist constants* $c_-^\star, c_+^\star > 0$ *depending only on* q_T, ρ_0, ρ_T *(but not on* $g^\star$*), such that*

$$c_-^\star \exp(-a_-^\star \|y\|^2) \;\le\; g^\star(y) \;\le\; c_+^\star \exp(-a_+^\star \|y\|^2), \qquad \forall y \in \mathbb{R}^d. \tag{27}$$

Proof. See Belomestny et al. [2026][Proposition 3].

□

The following lemma provides sufficient condition for *(A2)*.

Lemma A.2. *Let $(\mathcal{X}, \|\cdot\|)$ be a Banach space and let $\mathcal{T} : \mathcal{U} \subset \mathcal{X} \to \mathcal{X}$ be twice Fréchet differentiable on an open set $\mathcal{U}$. Fix $g^\star \in \mathcal{U}$ and assume there exists $r > 0$ such that the closed ball*

$$\overline{B}(g^\star, r) := \{g \in \mathcal{X} : \ \|g - g^\star\| \le r\}$$

is contained in $\mathcal{U}$. Suppose that the second derivative is uniformly bounded on the ball:

$$\sup_{\|g-g^\star\|\le r} \|\mathcal{T}''[g]\| \ \le \ L, \tag{28}$$

where $\|\mathcal{T}''[g]\|$ denotes the induced operator norm of the bounded bilinear map $\mathcal{T}''[g] : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$:

$$\|\mathcal{T}''[g]\| := \sup_{\|u\|\le 1,\ \|v\|\le 1} \|\mathcal{T}''[g](u, v)\|.$$

Then $\mathcal{T}'$ is Lipschitz on $\overline{B}(g^\star, r)$: for all $g, h \in \overline{B}(g^\star, r)$,

$$\|\mathcal{T}'[g] - \mathcal{T}'[h]\| \ \le \ L\,\|g - h\|, \tag{29}$$

where $\|\cdot\|$ on operators is the induced operator norm on $\mathcal{L}(\mathcal{X}, \mathcal{X})$. In particular, taking $h = g^\star$ yields

$$\|\mathcal{T}'[g^\star + v] - \mathcal{T}'[g^\star]\| \le L\|v\|, \qquad \forall v \in \mathcal{X} \text{ with } \|v\| \le r\,.$$

Proof. Fix $g, h \in \overline{B}(g^\star, r)$ and define the segment $\gamma(t) := h + t(g - h)$ for $t \in [0, 1]$. Then $\gamma(t) \in \overline{B}(g^\star, r)$ for all $t \in [0, 1]$ by convexity of the ball. Consider the map

$$\Phi(t) := \mathcal{T}'[\gamma(t)] \in \mathcal{L}(\mathcal{X}, \mathcal{X}).$$

Since $\mathcal{T}$ is twice Fréchet differentiable, Φ is (Fréchet) differentiable as a map $[0, 1] \to \mathcal{L}(\mathcal{X}, \mathcal{X})$ and, for each $t \in [0, 1]$, its derivative is the bounded linear operator on $\mathcal{X}$ given by

$$\Phi'(t) = \mathcal{T}''[\gamma(t)](g - h, \cdot) \in \mathcal{L}(\mathcal{X}, \mathcal{X}).$$

Therefore,

$$\mathcal{T}'[g] - \mathcal{T}'[h] = \Phi(1) - \Phi(0) = \int_0^1 \Phi'(t)\, dt = \int_0^1 \mathcal{T}''[\gamma(t)](g - h, \cdot)\, dt.$$

Taking operator norms and using (28) yields

$$\|\mathcal{T}'[g] - \mathcal{T}'[h]\| \le \int_0^1 \|\mathcal{T}''[\gamma(t)]\|\, dt\, \|g - h\| \le \int_0^1 L\, dt\, \|g - h\| = L\,\|g - h\|,$$

which proves (29).

□

In the following proposition we obtain representation for $(\mathcal{C}'[g^\star]u)(y)$.

Proposition A.3. *Assume $g^\star$ is such that $D_{g^\star}(x) > 0$ for all $x \in \mathrm{supp}(\rho_0)$ and that differentiation under the integral signs is justified (e.g. under standing Gaussian envelope assumptions). Then $\mathcal{C}$ is Fréchet differentiable at $g^\star$ and for every direction u the derivative admits the integral-operator form*

$$(\mathcal{C}'[g^\star]u)(y) = \int_{\mathbb{R}^d} \mathcal{K}(y, z)\, u(z)\, dz,$$

with kernel

$$\mathcal{K}(y, z) = \frac{\rho_T(z)}{(g^\star(z))^2} \int_{\mathbb{R}^d} \rho_0(x)\, \frac{q_T(x, y)\, q_T(x, z)}{(D_{g^\star}(x))^2}\, dx. \tag{30}$$

Proof. Fix g and write $\mathcal{C}[g](y) = \int \rho_0(x)\, q_T(x,y)\, D_g(x)^{-1}\, dx$. Let u be a perturbation direction and consider $g_\varepsilon := g^\star + \varepsilon u$. We compute the Gâteaux derivative at $\varepsilon = 0$. Since

$$D_{g_\varepsilon}(x) = \int_{\mathbb{R}^d} q_T(x,z)\, \rho_T(z)\, \frac{1}{g^\star(z) + \varepsilon u(z)}\, dz,$$

we have

$$\frac{d}{d\varepsilon} D_{g_\varepsilon}(x)\Big|_{\varepsilon=0} = \int_{\mathbb{R}^d} q_T(x,z)\, \rho_T(z)\ \frac{d}{d\varepsilon} \frac{1}{g^\star(z) + \varepsilon u(z)}\Big|_{\varepsilon=0} dz = -\int_{\mathbb{R}^d} q_T(x,z)\, \rho_T(z)\, \frac{u(z)}{(g^\star(z))^2}\, dz.$$

Thus,

$$\frac{d}{d\varepsilon} \frac{1}{D_{g_\varepsilon}(x)}\Big|_{\varepsilon=0} = -\frac{1}{(D_{g^\star}(x))^2}\ \frac{d}{d\varepsilon} D_{g_\varepsilon}(x)\Big|_{\varepsilon=0} = \frac{1}{(D_{g^\star}(x))^2} \int_{\mathbb{R}^d} q_T(x,z)\, \rho_T(z)\, \frac{u(z)}{(g^\star(z))^2}\, dz.$$

We have

$$\mathcal{C}[g_\varepsilon](y) = \int_{\mathbb{R}^d} \rho_0(x)\, q_T(x,y)\, \frac{1}{D_{g_\varepsilon}(x)}\, dx.$$

Differentiating at $\varepsilon = 0$ and inserting the previous display gives

$$\begin{aligned}
(\mathcal{C}'[g^\star]u)(y) &= \frac{d}{d\varepsilon} \mathcal{C}[g_\varepsilon](y)\Big|_{\varepsilon=0} \\
&= \int_{\mathbb{R}^d} \rho_0(x)\, q_T(x,y)\ \frac{d}{d\varepsilon} \frac{1}{D_{g_\varepsilon}(x)}\Big|_{\varepsilon=0} dx \\
&= \int_{\mathbb{R}^d} \rho_0(x)\, q_T(x,y)\, \frac{1}{(D_{g^\star}(x))^2} \int_{\mathbb{R}^d} q_T(x,z)\, \rho_T(z)\, \frac{u(z)}{(g^\star(z))^2}\, dz\, dx.
\end{aligned}$$

By Fubini's theorem (justified by the Gaussian envelope conditions), we can swap the order of integration to obtain

$$(\mathcal{C}'[g^\star]u)(y) = \int_{\mathbb{R}^d} \left[\frac{\rho_T(z)}{(g^\star(z))^2} \int_{\mathbb{R}^d} \rho_0(x)\, \frac{q_T(x,y)\, q_T(x,z)}{(D_{g^\star}(x))^2}\, dx \right] u(z)\, dz,$$

which is precisely the representation with kernel (30).

□

Proposition A.4. *Assume further that differentiation under the integral signs is justified. Then $\mathcal{T}$ is Fréchet differentiable at $g^\star$ and for every direction u we have*

$$(\mathcal{T}'[g^\star]u)(y) = \int_{\mathbb{R}^d} \mathcal{K}_{\mathcal{T}}(y,z)\, u(z)\, dz,$$

where the kernel $\mathcal{K}_{\mathcal{T}}$ is a rank–one correction of the kernel $\mathcal{K}$ of $\mathcal{C}'[g^\star]$:

$$\mathcal{K}_{\mathcal{T}}(y,z) = \mathcal{K}(y,z) - g^\star(y)\, \Lambda(z), \tag{31}$$

with

$$\Lambda(z) := \int_{\mathbb{R}^d} \frac{\rho_T(\tilde{y})}{(g^\star(\tilde{y}))^2}\, \mathcal{K}(\tilde{y},z)\, w(\tilde{y})\, d\tilde{y}. \tag{32}$$

Here $\mathcal{K}$ is the kernel from Proposition A.3.

Proof. Write $\mathcal{T}[g] = s(g)\, \mathcal{C}[g]$ with $s(g)$ as in (11). Since $\mathcal{C}$ is Fréchet differentiable at $g^\star$ and s is a scalar functional of $\mathcal{C}[g]$, the product rule gives

$$\mathcal{T}'[g^\star]u = s(g^\star)\, \mathcal{C}'[g^\star]u + s'[g^\star]u\, \mathcal{C}[g^\star]. \tag{33}$$

Because $\mathcal{T}[g^\star] = g^\star$ we have $s(g^\star) = 1$ and $\mathcal{C}[g^\star] = g^\star$. Hence

$$\mathcal{T}'[g^\star]u = \mathcal{C}'[g^\star]u + \big(s'[g^\star]u\big)\, g^\star. \tag{34}$$

It remains to compute $s'[g^\star]u$. Since

$$s(g) = \int_{\mathbb{R}^d} \rho_T(y)\,(\mathcal{C}[g](y))^{-1} w(y)\,dy,$$

differentiating yields

$$s'[g]u = -\int_{\mathbb{R}^d} \rho_T(y)\,\frac{(\mathcal{C}'[g]u)(y)}{(\mathcal{C}[g](y))^2} w(y)\,dy.$$

Evaluating at $g^\star$ and using $\mathcal{C}[g^\star] = g^\star$ gives

$$s'[g^\star]u = -\int_{\mathbb{R}^d} \frac{\rho_T(y)}{(g^\star(y))^2}\,(\mathcal{C}'[g^\star]u)(y) w(y)\,dy. \tag{35}$$

Now insert the kernel representation of $\mathcal{C}'[g^\star]u$,

$$(\mathcal{C}'[g^\star]u)(y) = \int_{\mathbb{R}^d} \mathcal{K}(y,z)\,u(z)\,dz,$$

into (35) and apply Fubini:

$$s'[g^\star]u = -\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \frac{\rho_T(\tilde{y})}{(g^\star(\tilde{y}))^2}\,\mathcal{K}(\tilde{y},z) w(\tilde{y})\,d\tilde{y} \right) u(z)\,dz = -\int_{\mathbb{R}^d} \Lambda(z)\,u(z)\,dz,$$

where Λ is defined in (32). Plugging this into (34) gives

$$(\mathcal{T}'[g^\star]u)(y) = \int_{\mathbb{R}^d} \mathcal{K}(y,z)\,u(z)\,dz - g^\star(y)\int_{\mathbb{R}^d} \Lambda(z)\,u(z)\,dz = \int_{\mathbb{R}^d} \big(\mathcal{K}(y,z) - g^\star(y)\Lambda(z)\big)\,u(z)\,dz,$$

which is exactly (31)–(32).

□

Proposition A.5. *Assume* ($\mathrm{Q_{Gauss}}$) *and that* $\operatorname{supp}(\rho_0) \subset B_{R_0} := \{x : \|x\| < R_0\}$. *Let $g^\star$ be a strictly positive fixed point of the normalized map $\mathcal{T}$ and set $\mathsf{T} := \mathcal{T}'[g^\star]$. Assume there exists a constant $\underline{D} > 0$ such that on* $\operatorname{supp}(\rho_0)$,

$$D_{g^\star}(x) := \int_{\mathbb{R}^d} q_T(x,z)\,\frac{\rho_T(z)}{g^\star(z)}\,dz \;\geq\; \underline{D}, \qquad \forall x \in \operatorname{supp}(\rho_0). \tag{36}$$

Then:

(i) *There exist $\kappa_R \in [0,1)$ and $\varepsilon_R \geq 0$ such that for all u with $u/g^\star \in L^\infty(B_R)$,*

$$\|\mathsf{T}u\|_{H,R} \;\leq\; \kappa_R\,\|u\|_{H,R} \;+\; \varepsilon_R\,\|u\|_{L^2(\mathbb{R}^d)}, \tag{37}$$

where

$$\varepsilon_R := \frac{(2\pi T)^{-d/2}}{\underline{D}} \left(\int_{B_R^c} \frac{\rho_T^2(z)}{(g^\star(z))^4}\,e^{-\frac{(\|z\|-R_0)^2}{T}}\,dz \right)^{1/2}, \tag{38}$$

and

$$\kappa_R := \tanh\Big(\frac{(R+R_0)R}{T}\Big). \tag{39}$$

(ii) *Let*

$$h(y) := \frac{\rho_T(y)}{(g^\star(y))^2}.$$

and assume that $h \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$. Then for all $u \in L^2(\mathbb{R}^d)$ with $u/g^\star \in L^\infty(B_R)$,

$$\|\mathsf{T}u\|_{L^2(\mathbb{R}^d)} \;\leq\; C\,\|u\|_{H,R} \;+\; \varepsilon_R \left(1 + \frac{1}{\gamma}\right) \|u\|_{L^2(\mathbb{R}^d)}, \tag{40}$$

where C depends on the norms $\|h\|_{L^2(\mathbb{R}^d)}$, $\|h\|_{L^\infty(\mathbb{R}^d)}$ and ε_R is as in (38).

Proof. We proceed in two parts.

Part (i): proof of the Hilbert contraction (37). Write $r := u/g^\star$. Since $\mathcal{T}$ is the normalized version of $\mathcal{C}$, its derivative has the rank–one form

$$\mathsf{T}u = \mathcal{C}'[g^\star]u - g^\star \langle \Lambda, u\rangle \tag{41}$$

for a deterministic Λ. Dividing by $g^\star$ and observing that the rank–one term $g^\star\langle\Lambda, u\rangle$ becomes a *constant shift* in the ratio, we obtain on B_R:

$$\frac{\mathsf{T}u}{g^\star} = \frac{\mathcal{C}'[g^\star]u}{g^\star} - \langle \Lambda, u\rangle, \qquad \text{hence} \qquad \operatorname*{osc}_{B_R}\Big(\frac{\mathsf{T}u}{g^\star}\Big) = \operatorname*{osc}_{B_R}\Big(\frac{\mathcal{C}'[g^\star]u}{g^\star}\Big), \tag{42}$$

where $\operatorname{osc}_{B_R}(f) := \operatorname{ess\,sup}_{B_R} f - \operatorname{ess\,inf}_{B_R} f$. Therefore

$$\|\mathsf{T}u\|_{H,R} = \frac{1}{2}\operatorname*{osc}_{B_R}\Big(\frac{\mathcal{C}'[g^\star]u}{g^\star}\Big). \tag{43}$$

Define the (positive) weights

$$W(x) := \frac{\rho_0(x)}{(D_{g^\star}(x))^2}, \qquad x \in B_{R_0}.$$

Using the kernel expression for $\mathcal{C}'[g^\star]$ and the fixed point property $g^\star = \mathcal{T}[g^\star]$ (equivalently $g^\star = \mathcal{C}[g^\star]$ up to a normalization constant absorbed in $\mathcal{T}$), one checks that for $y \in \mathbb{R}^d$,

$$\frac{(\mathcal{C}'[g^\star]u)(y)}{g^\star(y)} = \int_{B_{R_0}} \pi_y(dx) \underbrace{\int_{\mathbb{R}^d} \eta_x(dz)\, r(z)}_{=:(\mathsf{P}r)(x)}, \tag{44}$$

where π_y is a probability measure on B_{R_0} depending on y and η_x is a probability measure on $\mathbb{R}^d$ depending on x, given by

$$\pi_y(dx) := \frac{W(x)\, q_T(x,y)}{\int_{B_{R_0}} W(x')\, q_T(x',y)\, dx'}\, dx, \tag{45}$$

$$\eta_x(dz) := \frac{q_T(x,z)\, \frac{\rho_T(z)}{g^\star(z)}}{D_{g^\star}(x)}\, dz, \tag{46}$$

so that indeed $\int \pi_y(dx) = 1$ and $\int \eta_x(dz) = 1$. Thus, the ratio is obtained by first averaging r under η_x (for each x), then averaging over x under π_y. Fix $y, y' \in B_R$. Consider the Radon–Nikodym derivative

$$\frac{d\pi_y}{d\pi_{y'}}(x) = \frac{q_T(x,y)}{q_T(x,y')} \cdot \frac{\int W(x')q_T(x',y')\, dx'}{\int W(x')q_T(x',y)\, dx'}.$$

Since the second factor is constant in x, the oscillation contraction is governed by the range of $q_T(x,y)/q_T(x,y')$ over $x \in B_{R_0}$. For Gaussian q_T,

$$\log\frac{q_T(x-y)}{q_T(x-y')} = -\frac{\|x-y\|^2 - \|x-y'\|^2}{2T} = -\frac{2x\cdot(y'-y) + (\|y\|^2 - \|y'\|^2)}{2T}.$$

Hence, for $x \in B_{R_0}$ and $y, y' \in B_R$,

$$\Big|\log\frac{q_T(x-y)}{q_T(x-y')}\Big| \le \frac{\|x\|\,\|y-y'\|}{T} + \frac{|\|y\|^2 - \|y'\|^2|}{2T} \le \frac{R_0\|y-y'\|}{T} + \frac{(\|y\| + \|y'\|)\|y-y'\|}{2T}.$$

Since $\|y\|, \|y'\| \le R$ and $\|y - y'\| \le 2R$, we get

$$\Big|\log\frac{q_T(x-y)}{q_T(x-y')}\Big| \le \frac{2R_0R}{T} + \frac{2R^2}{T} = \frac{2(R+R_0)R}{T}.$$

Therefore,

$$\sup_{x\in B_{R_0}} \frac{q_T(x-y)}{q_T(x-y')} \le \exp\Big(\frac{2(R+R_0)R}{T}\Big) =: \Delta_R, \qquad \inf_{x\in B_{R_0}} \frac{q_T(x-y)}{q_T(x-y')} \ge \Delta_R^{-1}. \tag{47}$$

A standard Hilbert-metric / oscillation comparison (Doeblin-type argument) yields: for any essentially bounded function φ on B_{R_0},

$$\Big|\int \varphi\, d\pi_y - \int \varphi\, d\pi_{y'}\Big| \le \tanh\Big(\frac{1}{4}\log \Delta_R\Big)\ \operatorname*{osc}_{B_{R_0}}(\varphi). \tag{48}$$

(Proof: write $\pi_y, \pi_{y'}$ with densities proportional to $q_T(x-y)$ and $q_T(x-y')$; (47) implies the likelihood ratio is bounded by Δ_R, and the variation of expectations of φ is controlled by the Hilbert projective metric; the sharp contraction factor is $\tanh((1/4)\log\Delta_R)$.) Using $\log\Delta_R = 2(R+R_0)R/T$, we arrive at

$$\tanh\Big(\frac{1}{4}\log\Delta_R\Big) = \tanh\Big(\frac{(R+R_0)R}{2T}\Big).$$

Since our seminorm includes an extra factor $1/2$ (oscillation vs Hilbert), we may equivalently write the contraction factor as

$$\kappa_R := \tanh\Big(\frac{(R+R_0)R}{T}\Big),$$

which is (39) (the difference in constants is absorbed in the standard inequality from Hilbert to oscillation; keeping the slightly looser $\tanh((R+R_0)R/T)$ is always valid). Let $r = u/g^\star$ and split $r = r\mathbf{1}_{B_R} + r\mathbf{1}_{B_R^c}$. In (44), the inner averaging $(\mathsf{P}r)(x) = \int r\, d\eta_x$ splits accordingly:

$$(\mathsf{P}r)(x) = \int_{B_R} r(z)\,\eta_x(dz) + \int_{B_R^c} r(z)\,\eta_x(dz).$$

The first term depends only on r restricted to B_R and hence its oscillation on B_{R_0} is at most $\operatorname{osc}_{B_R}(r) = 2\|u\|_{H,R}$ (since $g^\star > 0$). The second term is controlled in L^∞ by Cauchy–Schwarz: using $\eta_x(dz) = \frac{q_T(x-z)\rho_T(z)/g^\star(z)}{D_{g^\star}(x)}\,dz$ and $D_{g^\star}(x) \ge \underline{D}$,

$$\begin{aligned}\sup_{x\in B_{R_0}} \int_{B_R^c} |r(z)|\,\eta_x(dz) &\le \frac{1}{\underline{D}} \sup_{x\in B_{R_0}} \int_{B_R^c} q_T(x-z)\,\frac{\rho_T(z)}{g^\star(z)}\,\frac{|u(z)|}{g^\star(z)}\,dz \\ &\le \frac{1}{\underline{D}} \sup_{x\in B_{R_0}} \left(\int_{B_R^c} q_T(x-z)^2\,\frac{\rho_T(z)^2}{(g^\star(z))^4}\,dz\right)^{1/2} \|u\|_{L^2(\mathbb{R}^d)}. \end{aligned}\tag{49}$$

For $x \in B_{R_0}$ we have $\|x-z\| \ge (\|z\| - R_0)_+$ and hence

$$q_T(x-z)^2 \le (2\pi T)^{-d} \exp\Big(-\frac{(\|z\|-R_0)^2}{T}\Big).$$

Plugging this into (49) gives

$$\sup_{x\in B_{R_0}} \int_{B_R^c} |r(z)|\,\eta_x(dz) \le \frac{(2\pi T)^{-d/2}}{\underline{D}} \left(\int_{B_R^c} \frac{\rho_T(z)^2}{(g^\star(z))^4} e^{-\frac{(\|z\|-R_0)^2}{T}}\,dz\right)^{1/2} \|u\|_{L^2(\mathbb{R}^d)} = \varepsilon_R \|u\|_{L^2(\mathbb{R}^d)},$$

with ε_R as in (38). Hence

$$\|\mathsf{T}u\|_{H,R} \le \kappa_R\,\|u\|_{H,R} + \varepsilon_R\,\|u\|_{L^2(\mathbb{R}^d)}. \tag{50}$$

Part (ii): proof of the mixed L^2 estimate (40)**.** The normalization defining $\mathcal{T}$ fixes the scale of its outputs. In particular, for every $c > 0$ one has $\mathcal{T}[cg] = \mathcal{T}[g]$ (scale invariance of the normalized map). Differentiating at $c = 1$ along the direction $g = g^\star$ yields

$$0 = \frac{d}{dc}\Big|_{c=1} \mathcal{T}[cg^\star] = \mathcal{T}'[g^\star]\,g^\star = \mathsf{T}g^\star.$$

Since T is linear

$$\mathsf{T}u = \mathsf{T}(u - g^\star). \tag{51}$$

From (41),

$$\|\mathsf{T}u\|_{L^2(\mathbb{R}^d)} \le \|\mathcal{C}'[g^\star](u-g^\star)\|_{L^2(\mathbb{R}^d)} + \|g^\star\|_{L^2(\mathbb{R}^d)}\,|\langle\Lambda, u-g^\star\rangle|.$$

Let $u_{\mathrm{out}} := (u - g^\star)\mathbf{1}_{B_R^c}$. By Hilbert–Schmidt,

$$\|\mathcal{C}'[g^\star]u_{\mathrm{out}}\|_{L^2_y} \le \left(\int_{B_R^c} \|\mathcal{K}(\cdot, z)\|_{L^2_y}^2 \, dz \right)^{1/2} \|u - g^\star\|_{L^2_z}. \tag{52}$$

So it suffices to bound $\|\mathcal{K}(\cdot, z)\|_{L^2_y}$ for each z. Fix $z \in \mathbb{R}^d$. Since $\rho_0 \ge 0$ and $\mathcal{K}(\cdot, z) \ge 0$, Minkowski's integral inequality gives

$$\begin{aligned} \|\mathcal{K}(\cdot, z)\|_{L^2_y} &= h(z) \left\| \int_{B_{R_0}} \rho_0(x) \frac{q_T(x - \cdot)\, q_T(x - z)}{(D_{g^\star}(x))^2} \, dx \right\|_{L^2_y} \\ &\le h(z) \int_{B_{R_0}} \rho_0(x) \frac{q_T(x - z)}{(D_{g^\star}(x))^2} \, \|q_T(x - \cdot)\|_{L^2_y} \, dx. \end{aligned} \tag{53}$$

By translation invariance of Lebesgue measure,

$$\|q_T(x - \cdot)\|_{L^2_y} = \|q_T\|_{L^2(\mathbb{R}^d)}.$$

Hence

$$\|\mathcal{K}(\cdot, z)\|_{L^2_y} \le \frac{\|q_T\|_2}{\underline{D}^2} \, h(z) \int_{B_{R_0}} \rho_0(x)\, q_T(x - z) \, dx. \tag{54}$$

Now bound the remaining integral using the Gaussian tail. For $x \in B_{R_0}$ we have $\|x - z\| \ge (\|z\| - R_0)_+$, hence

$$q_T(x - z) \le (2\pi T)^{-d/2} \exp\Big(- \frac{(\|z\| - R_0)^2}{2T} \Big).$$

Since $\int \rho_0 = 1$,

$$\int_{B_{R_0}} \rho_0(x)\, q_T(x - z) \, dx \le (2\pi T)^{-d/2} \exp\Big(- \frac{(\|z\| - R_0)^2}{2T} \Big). \tag{55}$$

Combining (54)–(55) yields

$$\|\mathcal{K}(\cdot, z)\|_{L^2_y} \le \frac{\|q_T\|_2 \, (2\pi T)^{-d/2}}{\underline{D}^2} \, h(z) \exp\Big(- \frac{(\|z\| - R_0)^2}{2T} \Big). \tag{56}$$

Squaring (56) gives

$$\|\mathcal{K}(\cdot, z)\|_{L^2_y}^2 \le \Big(\frac{\|q_T\|_2 \, (2\pi T)^{-d/2}}{\underline{D}^2} \Big)^2 \, h(z)^2 \exp\Big(- \frac{(\|z\| - R_0)^2}{T} \Big).$$

Insert this into (52) and integrate over $z \in B_R^c$:

$$\|\mathcal{C}'[g^\star]u_{\mathrm{out}}\|_{L^2(\mathbb{R}^d)} \le \frac{\|q_T\|_2 \, (2\pi T)^{-d/2}}{\underline{D}^2} \left(\int_{B_R^c} h(z)^2 e^{-(\|z\| - R_0)^2/T} \, dz \right)^{1/2} \|u - g^\star\|_{L^2(\mathbb{R}^d)}.$$

Hence

$$\|\mathcal{C}'[g^\star](u_{\mathrm{out}})\|_{L^2(\mathbb{R}^d)} \le \varepsilon_R \|u - g^\star\|_{L^2(\mathbb{R}^d)}.$$

From $u \ge \gamma g^\star$ we have $g^\star \le \gamma^{-1} u$, hence $|u - g^\star| \le u + g^\star \le (1 + \gamma^{-1})u$. Taking L^2 norms gives $\|u - g^\star\|_{L^2(\mathbb{R}^d)} \le (1 + \gamma^{-1})\|u\|_{L^2(\mathbb{R}^d)}$.

Define

$$m_R(u) := \frac{1}{2} \Big(\operatorname*{ess\,sup}_{B_R} r + \operatorname*{ess\,inf}_{B_R} r \Big).$$

Write $u_{\mathrm{in}} := (u - g^\star)\mathbf{1}_{B_R}$, $r = u_{\mathrm{in}}/g^\star$ on B_R. Then $|r - m_R| \le 2\|u\|_{H,R}$ on B_R and hence

$$\|u_{\mathrm{in}}\|_{L^2(\mathbb{R}^d)} \le \|(r - m_R) g^\star \mathbf{1}_{B_R}\|_{L^2(\mathbb{R}^d)} + |m_R - 1| \, \|g^\star \mathbf{1}_{B_R}\|_{L^2(\mathbb{R}^d)}.$$

Set $\mu_R(dy) := \frac{\rho_T(y)}{g^\star(y)} \mathbf{1}_{B_R}(y)\, dy$. By assumption, μ_R is a probability measure and

$$1 = \int_{B_R} \frac{\rho_T}{u}\, dy = \int_{B_R} \frac{1}{r}\, d\mu_R.$$

If $r > 1$ μ_R–a.e., then $1/r < 1$ μ_R–a.e. and the integral is < 1, a contradiction. Similarly $r < 1$ μ_R–a.e. is impossible. Hence $\operatorname{ess\,inf}_{B_R} r \le 1 \le \operatorname{ess\,sup}_{B_R} r$. Therefore,

$$|m_R - 1| = \Big|\frac{1}{2}(\operatorname{ess\,sup} r + \operatorname{ess\,inf} r) - 1\Big| \le \frac{1}{2}(\operatorname{ess\,sup} r - 1) + \frac{1}{2}(1 - \operatorname{ess\,inf} r) = \|u\|_{H,R}.$$

As a result,

$$\|u_{\mathrm{in}}\|_{L^2(\mathbb{R}^d)} \le 3\|g^\star\|_{L^2(\mathbb{R}^d)}\, \|u\|_{H,R}. \tag{57}$$

Let $\mathcal{C}_R : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ be the bounded linear operator

$$(\mathcal{C}_R f)(y) := \int_{B_R} \mathcal{K}(y,z)\, f(z)\, dz \quad \text{for } f \in L^2(\mathbb{R}^d),$$

where $\mathcal{K}$ is the kernel of $\mathcal{C}'[g^\star]$. Then for any $u \in L^2(\mathbb{R}^d)$ we have the identity

$$\mathcal{C}'[g^\star](u_{\mathrm{in}}) = \mathcal{C}_R(u_{\mathrm{in}}),$$

since u_{in} vanishes outside B_R and thus the z–integration in $\mathcal{C}'[g^\star]$ effectively runs over B_R. By definition of the operator norm,

$$\|\mathcal{C}_R\|_{2\to 2} := \sup_{f \in L^2(\mathbb{R}^d)\setminus\{0\}} \frac{\|\mathcal{C}_R f\|_{L^2(\mathbb{R}^d)}}{\|f\|_{L^2(\mathbb{R}^d)}},$$

we have for every $f \in L^2(\mathbb{R}^d)$,

$$\|\mathcal{C}_R f\|_{L^2(\mathbb{R}^d)} \le \|\mathcal{C}_R\|_{2\to 2}\, \|f\|_{L^2(\mathbb{R}^d)}.$$

Applying this with $f = u_{\mathrm{in}}$ yields

$$\|\mathcal{C}'[g^\star](u_{\mathrm{in}})\|_{L^2(\mathbb{R}^d)} = \|\mathcal{C}_R(u_{\mathrm{in}}\|_{L^2(\mathbb{R}^d)} \le \|\mathcal{C}_R\|_{2\to 2}\, \|u_{\mathrm{in}}\|_{L^2(\mathbb{R}^d)}.$$

Next we apply Schur's test to estimate $\|\mathcal{C}_R\|_{2\to 2}$. Set

$$A := \sup_{y \in \mathbb{R}^d} \int_{B_R} \mathcal{K}(y,z)\, dz, \qquad B := \sup_{z \in B_R} \int_{\mathbb{R}^d} \mathcal{K}(y,z)\, dy$$

then Schur's test yields $\|\mathcal{C}_R\|_{2\to 2} \le \sqrt{AB}$. Fix $z \in B_R$ and integrate in y:

$$\begin{aligned}
\int_{\mathbb{R}^d} \mathcal{K}(y,z)\, dy &= h(z) \int_{B_{R_0}} \rho_0(x) \frac{q_T(x-z)}{(D_{g^\star}(x))^2} \underbrace{\int_{\mathbb{R}^d} q_T(x-y)\, dy}_{=1}\, dx \\
&\le \frac{h(z)}{\underline{D}^2} \int_{B_{R_0}} \rho_0(x)\, q_T(x-z)\, dx \le \frac{h(z)}{\underline{D}^2} \sup_\xi q_T(\xi) \int \rho_0(x)\, dx \\
&= \frac{(2\pi T)^{-d/2}}{\underline{D}^2}\, h(z).
\end{aligned}$$

Taking $\sup_{z \in B_R}$ gives

$$B \le \frac{(2\pi T)^{-d/2}}{\underline{D}^2}\, \|h\|_{L^\infty}. \tag{58}$$

Fix now $y \in \mathbb{R}^d$ then

$$\begin{aligned}
\int_{B_R} \mathcal{K}(y,z)\, dz &= \int_{B_R} h(z) \int_{B_{R_0}} \rho_0(x) \frac{q_T(x-y)\, q_T(x-z)}{(D_{g^\star}(x))^2}\, dx\, dz \\
&\le \frac{\|h\|_{L^\infty}}{\underline{D}^2} \int_{B_{R_0}} \rho_0(x)\, q_T(x-y) \int_{B_R} q_T(x-z)\, dz\, dx \\
&\le (2\pi T)^{-d/2} \frac{\|h\|_{L^\infty}}{\underline{D}^2}.
\end{aligned}$$

Combining,

$$\|\mathcal{C}_R\|_{2\to 2} \le \sqrt{AB} \le \frac{(2\pi T)^{-d/2}}{\underline{D}^2}\,\|h\|_{L^\infty}\,.$$

and

$$\|\mathcal{C}'[g^\star](u_{\rm in})\|_{L^2(\mathbb{R}^d)} \le \|\mathcal{C}_R\|_{2\to 2}\,\|u_{\rm in}\|_{L^2(\mathbb{R}^d)} \le \frac{(2\pi T)^{-d/2}}{\underline{D}^2}\,\|h\|_{L^\infty}\,\|u_{\rm in}\|_{L^2(\mathbb{R}^d)}.$$

Due to (57), we have

$$\|\mathcal{C}'[g^\star](u_{\rm in})\|_{L^2(\mathbb{R}^d)} \le \frac{3\,(2\pi T)^{-d/2}}{\underline{D}^2}\,\|h\|_{L^\infty}\,\|g^\star\|_{L^2(\mathbb{R}^d)}\,\|u\|_{H,R}.$$

Turn now to $\langle \Lambda, u\rangle$. Split

$$|\langle \Lambda, u\rangle| \le \|\Lambda\|_{L^2(B_R)}\|u_{\rm in}\|_{L^2(B_R)} + \|\Lambda\|_{L^2(B_R^c)}\|u_{\rm in}\|_{L^2(\mathbb{R}^d)}.$$

We have

$$\begin{aligned}
\Lambda(z) &= \int_{\mathbb{R}^d} \frac{\rho_T(y)\,w_R(y)}{(g^\star(y))^2}\,\frac{\rho_T(z)}{(g^\star(z))^2}\int_{B_{R_0}} \rho_0(x)\,\frac{q_T(x-y)\,q_T(x-z)}{(D_{g^\star}(x))^2}\,dx\,dy\\
&\le \frac{\rho_T(z)}{(g^\star(z))^2}\int_{B_{R_0}} \rho_0(x)\,\frac{q_T(x-z)}{(D_{g^\star}(x))^2}\left(\int_{\mathbb{R}^d} q_T(x-y)\,\frac{\rho_T(y)}{(g^\star(y))^2}\,dy\right)dx\\
&\le \frac{\rho_T(z)}{(g^\star(z))^2}\int_{B_{R_0}} \rho_0(x)\,\frac{q_T(x-z)}{(D_{g^\star}(x))}\,dx.
\end{aligned}$$

For every $R>0$,

$$\begin{aligned}
\|\Lambda\|^2_{L^2(B_R)} &\le \Big(\frac{\|q_T\|_2\,\|q_T\|_\infty}{\underline{D}^2}\Big)^2\|h\|^2_{L^2(\mathbb{R}^d)}\int_{B_R} h(z)^2\,dz\\
&\le \Big(\frac{\|q_T\|_2\,\|q_T\|_\infty}{\underline{D}^2}\Big)^2\|h\|^4_{L^2(\mathbb{R}^d)}. \qquad (59)
\end{aligned}$$

Write $u = u_{\rm in} + u_{\rm out}$ with

$$u_{\rm in} := (u-g^\star)\mathbf{1}_{B_R}, \qquad u_{\rm out} := (u-g^\star)\mathbf{1}_{B_R^c}.$$

Then, by Cauchy–Schwarz,

$$|\langle \Lambda, u-g^\star\rangle| \le \|\Lambda\|_{L^2(B_R)}\|u_{\rm in}\|_{L^2(\mathbb{R}^d)} + \|\Lambda\|_{L^2(B_R^c)}\|u_{\rm out}\|_{L^2(\mathbb{R}^d)}. \qquad (60)$$

We now bound $\|\Lambda\|_{L^2(B_R)}$. Recall $\Lambda(z) \le h(z)\,A(z)$, where

$$A(z) := \int_{B_{R_0}} \rho_0(x)\,\frac{q_T(x-z)}{(D_{g^\star}(x))^2}\,dx.$$

Using $D_{g^\star}(x) \ge \underline{D}$ on $\mathrm{supp}(\rho_0)\subset B_{R_0}$ and $\int\rho_0 = 1$, we have the pointwise bound

$$|A(z)| \le \frac{1}{\underline{D}^2}\int_{B_{R_0}} \rho_0(x)\,q_T(x-z)\,dx \le \frac{\|q_T\|_\infty}{\underline{D}^2}. \qquad (61)$$

To obtain an L^2–bound, apply Young's inequality to the convolution $z\mapsto \int \rho_0(x)q_T(x-z)\,dx = (\rho_0 * q_T)(z)$:

$$\|A\|_{L^2(\mathbb{R}^d)} \le \frac{1}{\underline{D}^2}\,\|\rho_0 * q_T\|_{L^2(\mathbb{R}^d)} \le \frac{1}{\underline{D}^2}\,\|\rho_0\|_{L^1(\mathbb{R}^d)}\,\|q_T\|_{L^2(\mathbb{R}^d)} = \frac{\|q_T\|_2}{\underline{D}^2}. \qquad (62)$$

Combining (61) and (62) yields the mixed L^2–estimate

$$\|A\|^2_{L^4(\mathbb{R}^d)} = \|A^2\|_{L^2(\mathbb{R}^d)} \le \|A\|_{L^\infty(\mathbb{R}^d)}\,\|A\|_{L^2(\mathbb{R}^d)} \le \frac{\|q_T\|_\infty}{\underline{D}^2}\cdot\frac{\|q_T\|_2}{\underline{D}^2} = \frac{\|q_T\|_2\,\|q_T\|_\infty}{\underline{D}^4}.$$

Using $\Lambda = hA$ and Hölder with exponents $(4,4)$ gives

$$\|\Lambda\|_{L^2(B_R)}^2 \le \|\Lambda\|_{L^2(\mathbb{R}^d)}^2 = \|hA\|_2^2 \le \|h\|_{L^4(\mathbb{R}^d)}^2 \, \|A\|_{L^4(\mathbb{R}^d)}^2 \le \frac{\|q_T\|_2 \, \|q_T\|_\infty}{\underline{D}^4} \, \|h\|_{L^4(\mathbb{R}^d)}^2. \tag{63}$$

Since $D_{g^\star}(x) \ge \underline{D}$ on B_{R_0} and $\int \rho_0 = 1$,

$$|A(z)| \le \frac{1}{\underline{D}^2} \int_{B_{R_0}} \rho_0(x) \, q_T(x-z) \, dx \le \frac{1}{\underline{D}^2} \sup_{x \in B_{R_0}} q_T(x-z).$$

For $x \in B_{R_0}$ and any z we have $\|x - z\| \ge (\|z\| - R_0)_+$, hence by the Gaussian upper bound on q_T,

$$\sup_{x \in B_{R_0}} q_T(x-z) \le c_+ \exp\bigl(-a_+(\|z\| - R_0)^2\bigr).$$

Therefore

$$|\Lambda(z)|^2 = h(z)^2 \, |A(z)|^2 \le \frac{c_+^2}{\underline{D}^4} \, h(z)^2 \, \exp\bigl(-2a_+(\|z\| - R_0)^2\bigr).$$

Integrating over B_R^c gives

$$\|\Lambda\|_{L^2(B_R^c)}^2 \le \frac{c_+^2}{\underline{D}^4} \int_{B_R^c} h(z)^2 \, \exp\bigl(-2a_+(\|z\| - R_0)^2\bigr) \, dz \tag{64}$$

and hence

$$\|\Lambda\|_{L^2(B_R^c)} \le \frac{c_+ \, \|h\|_{L^\infty}}{\underline{D}^2} \, \Bigl(\frac{|\mathbb{S}^{d-1}|}{4a_+}\Bigr)^{1/2} \, R^{\frac{d-1}{2}} \, \exp\bigl(-a_+(R - R_0)^2\bigr). \tag{65}$$

□

Lemma A.6 (Detailed version of Lemma 2.4)**.** *Assume that for all admissible u,*

$$\|\mathsf{T}u\|_{H,R} \le \kappa_R \, \|u\|_{H,R} + \varepsilon_R \, \|u\|_{L^2(\mathbb{R}^d)}, \tag{66}$$

$$\|\mathsf{T}u\|_{L^2(\mathbb{R}^d)} \le C \, \|u\|_{H,R} + \varepsilon_R \Bigl(1 + \frac{1}{\gamma}\Bigr) \, \|u\|_{L^2(\mathbb{R}^d)}. \tag{67}$$

Then

$$\|\mathsf{T}u\|_{R,\lambda} \le \alpha_R(\lambda) \, \|u\|_{R,\lambda}, \tag{68}$$

where

$$\alpha_R(\lambda) := \max\left\{\kappa_R + \lambda C, \; \varepsilon_R\Bigl(\frac{1}{\lambda} + 1 + \frac{1}{\gamma}\Bigr)\right\}. \tag{69}$$

Furthermore, there exist $R, \lambda > 0$ such that $\alpha_R(\lambda) \in (0,1)$.

Proof. By definition,

$$\|\mathsf{T}u\|_{R,\lambda} = \|\mathsf{T}u\|_{H,R} + \lambda \, \|\mathsf{T}u\|_2.$$

Insert (66) and (67):

$$\begin{aligned}\|\mathsf{T}u\|_{R,\lambda} &\le \kappa_R \|u\|_{H,R} + \varepsilon_R \|u\|_2 + \lambda\Bigl(C\|u\|_{H,R} + \varepsilon_R(1 + 1/\gamma)\|u\|_2\Bigr) \\ &= (\kappa_R + \lambda C) \, \|u\|_{H,R} + \varepsilon_R\Bigl(1 + \lambda(1 + 1/\gamma)\Bigr) \, \|u\|_2.\end{aligned}$$

Now use $\|u\|_{H,R} \le \|u\|_{R,\lambda}$ and $\|u\|_2 \le \lambda^{-1}\|u\|_{R,\lambda}$:

$$\|\mathsf{T}u\|_{R,\lambda} \le \Bigl(\kappa_R + \lambda C\Bigr)\|u\|_{R,\lambda} + \varepsilon_R\Bigl(1 + \lambda(1 + 1/\gamma)\Bigr)\frac{1}{\lambda}\|u\|_{R,\lambda}.$$

This yields (68) with (69).

A convenient near–optimal choice is to balance the terms λC and ε_R/λ, i.e.

$$\lambda \asymp \sqrt{\varepsilon_R},$$

more precisely

$$\lambda_0 := \sqrt{\frac{\varepsilon_R}{C}} \qquad (C > 0).$$

Then

$$\alpha_R(\lambda_0) \le \kappa_R + \sqrt{C\,\varepsilon_R} + \varepsilon_R\Big(1 + \frac{1}{\gamma}\Big). \tag{70}$$

In particular, if R is chosen so that $\kappa_R < 1$ and ε_R is sufficiently small (e.g. $\sqrt{C\varepsilon_R} + \varepsilon_R(1 + 1/\gamma) \le 1 - \kappa_R$), then $\alpha_R(\lambda_0) < 1$ and T is a strict contraction in $\|\cdot\|_{R,\lambda_0}$.

□

B EXISTENCE

Theorem B.1. *Let $d \in \mathbb{N}$ and let $k : \mathbb{R}^d \times \mathbb{R}^d \to (0,\infty)$ be a continuous kernel. Assume:*

(A0) *ρ_0, ρ_T are probability densities on $\mathbb{R}^d$, $\operatorname{supp}(\rho_0) \subset B(R_0)$ for some $R_0 < \infty$, and $\rho_0 \in L^\infty(\mathbb{R}^d)$. Moreover, there exist $r_0 \in (0, R_0]$ and $a_0 > 0$ such that*

$$\rho_0(x) \ge a_0 \qquad \text{for a.e. } x \in B(r_0). \tag{71}$$

(K1) *There exist constants $C_u, C_l, c_u, c_l > 0$, and an exponent $p \in (1,\infty)$ ($p = 2$) such that*

$$C_\ell \exp\big(-c_\ell \|x-y\|^p\big) \le k(x,y) \le C_u \exp\big(-c_u \|x-y\|^p\big) \qquad \forall x, y \in \mathbb{R}^d. \tag{72}$$

(K2) *For every $R < \infty$,*

$$m(R) := \inf_{\|x\| \le R_0,\ \|y\| \le R} k(x,y) \;>\; 0. \tag{73}$$

(A1) *There exists $\gamma > 0$ such that*

$$\int_{\mathbb{R}^d} \exp\big(\gamma \|z\|^p\big)\, \rho_T(z)\, dz < \infty. \tag{74}$$

Fix any $\alpha_- \in (0, c_u)$ and any $\alpha_+ \in (0, \gamma)$. Then there exist constants $0 < c_- < c_+ < \infty$ and a function $g^\star \in C(\mathbb{R}^d)$ with $g^\star > 0$ such that, defining

$$D_{g^\star}(x) := \int_{\mathbb{R}^d} k(x,z)\, \frac{\rho_T(z)}{g^\star(z)}\, dz, \tag{75}$$

the fixed–point equation

$$g^\star(y) = \int_{\mathbb{R}^d} \frac{\rho_0(x)}{D_{g^\star}(x)}\, k(x,y)\, dx \qquad \forall y \in \mathbb{R}^d \tag{76}$$

holds together with the normalization

$$\int_{\mathbb{R}^d} \frac{\rho_T(z)}{g^\star(z)}\, dz = 1. \tag{77}$$

Moreover, $g^\star$ enjoys two–sided subgaussian *bounds*

$$c_-\, e^{-\alpha_+ \|y\|^p} \;\le\; g^\star(y) \;\le\; c_+\, e^{-\alpha_- \|y\|^p}, \qquad \forall y \in \mathbb{R}^d. \tag{78}$$

Proof. We apply a Birkhoff theorem applied to the composition of linear operators and the reciprocal map $g \to 1/g$ (see Chen et al. [2016]) in a weighted supremum space. Let

$$X := \Big\{ g \in C(\mathbb{R}^d) :\ \|g\|_X := \sup_{y \in \mathbb{R}^d} e^{\alpha_- \|y\|^p} |g(y)| < \infty \Big\}.$$

This is a Banach space. For parameters $c_-, c_+ > 0$ to be fixed later, set

$$\mathcal{K} := \Big\{ g \in X : \ g > 0,\ c_- e^{-\alpha_+ \|y\|^p} \le g(y) \le c_+ e^{-\alpha_- \|y\|^p} \ \forall y,\ \int_{\mathbb{R}^d} \frac{\rho_T}{g} = 1 \Big\}.$$

Then $\mathcal{K}$ is closed in X and nonempty (choose any $g_0(y) = e^{-\alpha_- \|y\|^p}$ and rescale to enforce $\int \rho_T / g_0 = 1$). For $g \in \mathcal{K}$ define

$$D_g(x) := \int_{\mathbb{R}^d} k(x,z)\, \frac{\rho_T(z)}{g(z)}\, dz, \qquad (\mathcal{T}g)(y) := \int_{\mathbb{R}^d} \frac{\rho_0(x)}{D_g(x)}\, k(x,y)\, dx.$$

Let

$$\Lambda(g) := \int_{\mathbb{R}^d} \frac{\rho_T(z)}{(\mathcal{T}g)(z)}\, dz \in (0,\infty), \qquad (\mathcal{S}g) := \Lambda(g)\, (\mathcal{T}g).$$

Any fixed point $g^\star = \mathcal{S}g^\star$ satisfies (77); moreover $\Lambda(g^\star) = 1$ and hence $g^\star = \mathcal{T}g^\star$. Write $w_g(z) := \rho_T(z)/g(z) \ge 0$. By the normalization in $\mathcal{K}$,

$$\int_{\mathbb{R}^d} w_g(z)\, dz = 1. \tag{79}$$

By (72),

$$\sup_{x,y} k(x,y) \le C_u,$$

so using (79),

$$D_g(x) = \int k(x,z) w_g(z)\, dz \le \sup_{x,z} k(x,z) \int w_g = C_u, \qquad \forall x. \tag{80}$$

Using the lower bound in $\mathcal{K}$,

$$w_g(z) = \frac{\rho_T(z)}{g(z)} \le \frac{1}{c_-}\, \rho_T(z)\, e^{\alpha_+ \|z\|^p}.$$

Assumption (74) implies that $\rho_T(z) e^{\alpha_+ \|z\|^p} \in L^1(\mathbb{R}^d)$ (since $\alpha_+ < \gamma$), hence the family $\{w_g : g \in \mathcal{K}\}$ is uniformly integrable and, in particular, uniformly tight: for any $\varepsilon \in (0,1)$ there exists R_ε such that

$$\sup_{g \in \mathcal{K}} \int_{\|z\| > R_\varepsilon} w_g(z)\, dz \le \varepsilon. \tag{81}$$

Fix $\varepsilon = \frac{1}{2}$ and $R := R_{1/2}$. Then for $x \in B(R_0)$,

$$D_g(x) \ge \int_{\|z\| \le R} k(x,z) w_g(z)\, dz \ge \Big(\inf_{\|x\| \le R_0,\ \|z\| \le R} k(x,z) \Big) \int_{\|z\| \le R} w_g(z)\, dz \ge \frac{1}{2}\, m(R),$$

where $m(R) > 0$ is defined in (73). Thus

$$\frac{1}{2}\, m(R) \ \le\ D_g(x) \ \le\ C_u, \qquad \forall x \in B(R_0),\ \forall g \in \mathcal{K}. \tag{82}$$

Since $\mathrm{supp}\rho_0 \subset B(R_0)$,

$$(\mathcal{T}g)(y) = \int_{B(R_0)} \frac{\rho_0(x)}{D_g(x)}\, k(x,y)\, dx.$$

Using (82) and (72),

$$(\mathcal{T}g)(y) \le \frac{\|\rho_0\|_\infty}{\frac{1}{2} m(R)} \int_{B(R_0)} C_u e^{-c_u \|x-y\|^p}\, dx \le C_2 \exp\!\big(-\tfrac{c_u}{2} \|y\|^p\big),$$

where we used that $\|x - y\| \ge (\|y\| - R_0)_+$ for $\|x\| \le R_0$ and that $(\|y\| - R_0)_+^p \ge 2^{-p} \|y\|^p - R_0^p$, absorbing constants into C_2. In particular, by choosing $\alpha_- < c_u/2$ and enlarging C_2 if needed,

$$(\mathcal{T}g)(y) \le C_2 e^{-\alpha_- \|y\|^p} \qquad \forall y,\ \forall g \in \mathcal{K}. \tag{83}$$

From (80) we have $1/D_g(x) \ge 1/C_u$ for all x. Using (71) and continuity/positivity of k,

$$(\mathcal{T}g)(y) \ge \int_{B(r_0)} \frac{a_0}{C_u} k(x,y)\, dx \ge \frac{a_0 |B(r_0)|}{C_u} \inf_{\|x\| \le r_0} k(x,y).$$

Then for $\|x\| \le r_0$,

$$k(x,y) \ge C_\ell e^{-c_\ell(\|y\|+r_0)^p} \ge C_3 e^{-2^{p-1}c_\ell\|y\|^p},$$

and hence, choosing $\alpha_+ \le 2^{p-1}c_\ell$,

$$(\mathcal{T}g)(y) \ge C_4 e^{-\alpha_+\|y\|^p} \qquad \forall y,\ \forall g \in \mathcal{K}. \tag{84}$$

By (84) and (74),

$$\Lambda(g) = \int \frac{\rho_T(z)}{(\mathcal{T}g)(z)}\,dz \le \frac{1}{C_4}\int \rho_T(z)e^{\alpha_+\|z\|^p}\,dz < \infty.$$

Also $\Lambda(g) > 0$ because ρ_T is not identically zero and $\mathcal{T}g > 0$. Combining (83)–(84) shows that, after setting $c_- := C_4 \inf_{g\in\mathcal{K}} \Lambda(g)$ and $c_+ := C_2 \sup_{g\in\mathcal{K}} \Lambda(g)$ (both finite and positive), we have $\mathcal{S}(\mathcal{K}) \subseteq \mathcal{K}$.

Let $g_n \to g$ in X with $g_n, g \in \mathcal{K}$. The lower bound in $\mathcal{K}$ and (74) provide an integrable dominating function for ρ_T/g_n, so $D_{g_n}(x) \to D_g(x)$ uniformly on $B(R_0)$ by dominated convergence and continuity of k. Then $\rho_0/D_{g_n} \to \rho_0/D_g$ in $L^1(B(R_0))$, hence $\mathcal{T}g_n \to \mathcal{T}g$ in X by dominated convergence again, and finally $\Lambda(g_n) \to \Lambda(g)$, so $\mathcal{S}g_n \to \mathcal{S}g$ in X. Thus $\mathcal{S}$ is continuous. For compactness, note that for $g \in \mathcal{K}$ the function $x \mapsto \rho_0(x)/D_g(x)$ is supported on $B(R_0)$ and uniformly bounded in L^∞ by (82). Therefore the family $\{\mathcal{T}g : g \in \mathcal{K}\}$ is equicontinuous on $\mathbb{R}^d$ (since k is continuous and the x-integration is over a fixed compact set) and uniformly tight in the weighted norm due to (83). Multiplication by the scalar $\Lambda(g)$ preserves these properties, hence $\mathcal{S}(\mathcal{K})$ is relatively compact in X (Arzelà–Ascoli in the weighted topology). Since $\mathcal{K}$ is nonempty, closed, and $\mathcal{S} : \mathcal{K} \to \mathcal{K}$ is continuous with relatively compact image, Birkhoff's (see, for instance, Chen et al. [2016]) theorem yields $g^\star \in \mathcal{K}$ such that $g^\star = \mathcal{S}g^\star$. By definition of $\mathcal{S}$, $g^\star$ satisfies (77). Finally, because $g^\star \in \mathcal{K}$ already has $\int \rho_T/g^\star = 1$, we have $\Lambda(g^\star) = 1$ and hence $g^\star = \mathcal{T}g^\star$, i.e. (76). The two-sided bounds (78) hold since $g^\star \in \mathcal{K}$.

□

C FURTHER DETAILS OF NUMERICAL EXPERIMENTS

In this section, we elaborate on the numerical experiments presented in Section 3. Appendix C.1 discusses general implementation details. Appendices C.2, C.3, and C.4 give additional details on the three Gaussian-to-Gaussian experiments (convergence of $\widehat{g}_{M,N}$ to $g^\star$, Schrödinger potential estimation across dimensions, and optimal drift estimation). Appendix C.5 describes the non-Gaussian and real-data experiments. Appendix C.6 presents information about the metrics used. Appendices C.7, C.8, and C.9 cover, respectively, the computational cost, the kernel bandwidth and subsampling of the drift reference set. Appendix C.10 presents all final hyperparameter values used for the experiments. Finally, Appendix C.11 describes the ERMBridge algorithm.

C.1 GENERAL IMPLEMENTATION DETAILS

We paid particular attention to the statistical reliability of the experiments. The three Gaussian-to-Gaussian experiments are each averaged over three random seeds; in this setting the optimal plan, the Schrödinger potential, and the drift are available in closed form [Bunne et al., 2023], which lets us compare ERMBridge and SinkhornBridge on the quality of the reconstructed potential and the objects derived from it, rather than only on downstream sample quality. The additional experiments of Appendix C.5 (Gaussian-to-Laplace-mixture, Gaussian-to-anisotropic-mixture, and single-cell trajectory inference) target non-Gaussian and real-data regimes, where no closed-form potential exists and we therefore evaluate distributional transport quality. In all experiments the prior process is a scaled Wiener process with zero drift, which yields a Gaussian reference kernel.

Optimization is carried out on the logarithmic Schrödinger potential, and all kernel sums are evaluated in log-space using the standard max-shift (log-sum-exp) stabilization. do not clip the learned potential to enforce a lower bound in any experiment. All calculations were performed on a single Nvidia T4 GPU.

C.2 DETAILS OF $\widehat{g}_{M,N}$ TO $g^\star$ CONVERGENCE EXPERIMENT

In this experiment, we used source $\rho_0 = \mathcal{N}(m_0, C_0)$, target $\rho_T = \mathcal{N}(m_1, C_1)$ in $\mathbb{R}^2$. Kernel variance $\nu = \sigma^2_{\text{end}}$, fixed bandwidth $2\sigma^2 = 2\nu$. For analysis we conduct single run or sweep over training set sizes $N \in \{500, 1000, \ldots, 8000\}$ with 3 seeds per N. Evaluation was done on a fixed test set of size $N_{eval} = 2000$.

C.3 DETAILS OF SCHRÖDINGER POTENTIAL ESTIMATION IN DIFFERENT DIMENSIONS

In this experiment, we used Gaussian-to-Gaussian translation in $\mathbb{R}^d$ with $d \in \{5, 10, 25, 50, 100\}$. $m_0 = 0$, $C_0 = I_d$, m_1 has first two coordinates $(2, 1)$ and rest zero, $C_1 = 1.2\, I_d$. For all dimensions train size was set $N = 5000$. For evaluation we used $N_{eval} = 5000$ and averaged 3 seeds per d.

C.4 DETAILS OF OPTIMAL DRIFT ESTIMATION EXPERIMENT

In this experiment, we used same Gaussian marginals $\rho_0 = \mathcal{N}(m_0, C_0)$, $\rho_T = \mathcal{N}(m_1, C_1)$ in $\mathbb{R}^d$. Cases: pairs $(d,\ N) \in [(2, 1000), (5, 1000), (5, 500), (10, 2000), (10, 1000), (20, 4000), (20, 2000), (50, 8000), (50, 4000)]$. As before, averaging was performed over 3 random seeds, and $\frac{N}{2}$ points from $\mathcal{N}(m_0, C_0)$ were used to calculate the MSE, and the same number of points from $t \in U[0.05, 0.95]$.

C.5 NON-GAUSSIAN AND REAL-DATA EXPERIMENTS

Gaussian to anisotropic Gaussian mixture (Table 1). The target is a mixture of 25 components placed on a fixed 5×5 grid embedded in $\mathbb{R}^d$. Each component is assigned an independently drawn elongated covariance with a random orientation, so that a single global kernel bandwidth is inadequate and accurate transport requires per-location adaptation. We evaluate in $d \in \{10, 50\}$ and report sliced 1-Wasserstein, the energy distance, and the number of mixture components recovered by the transported samples (out of 25).

Gaussian to Laplace mixture (Table 2). The source is a standard Gaussian in $\mathbb{R}^d$ and the target is a Laplace mixture, chosen to probe the heavy-tailed regime, in which the transformed potential has no closed form. We evaluate in $d \in \{2, 10\}$ and report the sliced 1-Wasserstein distance together with the energy distance; the column "Energy Δ" is the relative reduction of the energy distance of ERMBridge over SinkhornBridge.

Single-cell trajectory inference (Table 3). We use the EB single-cell RNA-sequencing dataset [Tong et al., 2020], following the protocol of Tong et al. [2024a]: we transport cells from t_{i-1} to t_{i+1} and report the 1-Wasserstein distance between the predicted and held-out distributions at the intermediate time t_i, averaged over $i \in \{1, 2, 3\}$ and five random seeds. Baseline numbers are taken from Tong et al. [2024a].

C.6 METRICS USED IN THE EXPERIMENTS

We use two metrics to evaluate transport quality: the sliced 1-Wasserstein distance and the energy distance. The sliced $\mathbb{W}_1$ serves as our primary metric, as it provides a quantitative, theoretically-grounded measure of distribution similarity and has been widely adopted in research papers and practical applications. The energy distance is reported additionally for the heavy-tailed and high-dimensional mixture-target experiments, where it captures shape mismatches between predicted and target components that the sliced $\mathbb{W}_1$ can underweight.

Sliced $\mathbb{W}_1$ distance. For samples $X = \{x_i\}_{i=1}^M$ and $Y = \{y_i\}_{i=1}^M$,

$$\text{Sliced } \mathbb{W}_1(X, Y) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{W}_1\big(\{\langle x_i, \theta_n \rangle\}_{i=1}^M,\ \{\langle y_i, \theta_n \rangle\}_{i=1}^M\big)\,,$$

where $\{\theta_n\}_{n=1}^N$ are random projections uniformly distributed on $\mathbb{S}^{d-1}$. For the sliced $\mathbb{W}_1$ distance, 100 random projections were used.

Energy distance. For samples $X = \{x_i\}_{i=1}^M$ and $Y = \{y_j\}_{j=1}^M$ in $\mathbb{R}^d$, the empirical energy distance is

$$\mathcal{E}(X, Y) \;=\; \frac{2}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \|x_i - y_j\| \;-\; \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \|x_i - x_j\| \;-\; \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \|y_i - y_j\|,$$

which is non-negative and equals zero if and only if the empirical measures coincide. Unlike the sliced $\mathbb{W}_1$, the energy distance does not require a projection step and is computed directly in $\mathbb{R}^d$, which makes it informative in higher dimensions and under heavy-tailed targets.

C.7 COMPUTATIONAL COMPLEXITY AND RESOURCE USAGE

Each minibatch performs two log-sum-exp passes over the $|X_b| \times |Y_b|$ tensor of pairwise squared distances (Algorithm 1), giving a per-step compute cost of $\mathcal{O}(B^2 d + B|\theta|)$ and memory $\mathcal{O}(B^2)$. At $B = 512$ and $d = 50$ the pairwise tensor occupies roughly $1\,\mathrm{MB}$, and training fits comfortably within the memory of the single Nvidia T4 GPU used for all experiments. Both passes use the standard max-shift; the $[-200, 200]$ logit clamp of Algorithm 2 is used only in drift evaluation, and the empirical fraction of clamped logits remains below 10^{-4} across all reported runs. Optimization stability is monitored through the per-epoch fraction of gradient-clipped steps, which falls below 5% after the first few epochs in every run.

C.8 KERNEL BANDWIDTH: ROLE AND SENSITIVITY

Our stability analysis is stated for a single, fixed Gaussian reference kernel q_T, whose variance is a property of the prior process and the time horizon T. The per-epoch bandwidth $2\sigma^2$ appearing in Algorithm 1 is, by contrast, an optimization device: it is initialized from the data scale (the median pairwise squared distance within a minibatch) and annealed linearly to its final value $2\sigma_{\mathrm{end}}^2$ over the first 80% of epochs, after which it is held fixed. The theoretical guarantees apply to this final fixed bandwidth; the schedule only shapes the optimization landscape in the earlier epochs. The per-experiment values of σ_{end} are listed in Appendix C.10.

The method is robust to this choice. On the $d = 10$ Laplace-mixture and $d = 50$ Gaussian settings, varying $2\sigma_{\mathrm{end}}^2$ over $\{0.25, 0.5, 1, 2, 4\}$ times its reference value changes the potential MSE by less than a factor of two and never destabilizes training, and fixed and annealed schedules yield matching final metrics.

$Y_1, \ldots, Y_M \sim \rho_T$, so for

C.9 DRIFT REFERENCE SET: COST AND SUBSAMPLING

The drift estimator of Algorithm 2 approximates the integral defining $b_{M,N}(x, t)$ by a Monte Carlo average over a reference set $\mathcal{Y}_{\mathrm{ref}} \sim \rho_T$. Its per-query cost is linear in $|\mathcal{Y}_{\mathrm{ref}}|$ and its variance decreases with the subsample size, giving a direct accuracy/compute trade-off. To reduce inference cost while preserving accuracy, $\mathcal{Y}_{\mathrm{ref}}$ may be subsampled using fixed random subsets, stratified subsets, or coreset / k-means representatives.

C.10 FINAL HYPERPARAMETER VALUES

Below are the final hyperparameter values for all experiments.

$\widehat{g}_{M,N}$ to $g^\star$ Convergence Experiment:

- Parameters for experiment: batch size $= 512$, lr $= 5 \cdot 10^{-5}$, epochs $= 120$, σ_end $= 1.0$, hidden_dim $= 128$.

Schrödinger Potential Estimation in Different Dimensions:

- Parameters for experiment: batch size $= 512$, lr $= 5 \cdot 10^{-5}$, epochs $= 150$, σ_end $= 0.8$, hidden_dim $= 128$.

Optimal Drift Estimation Experiment:

- Parameters for experiment: batch size $= 512$, lr $= 5 \cdot 10^{-5}$, epochs $= 120$, σ_end $= 0.8$, σ_max $= 1.2$, σ_min $= 0.5$, hidden_dim $= 128$.

Single Cell Experiment:

- Parameters for experiment: batch size $= 2048$, lr $= 1.187 \cdot 10^{-5}$, epochs $= 295$, σ_end $= 0.6751$, loss_scale $= 289.0764$, hidden_dim $= 2048$.

Gaussian to Laplace Mixture Experiment:

- Parameters for $d = 2$: batch size $= 128$, lr $= 5.799 \cdot 10^{-5}$, epochs $= 170$, σ_end $= 2.0184$, loss_scale $= 98.6226$, hidden_dim $= 512$.

- Parameters for $d = 10$: batch size $= 128$, lr $= 1.646 \cdot 10^{-5}$, epochs $= 170$, σ_end $= 1.8565$, loss_scale $= 66.3354$, hidden_dim $= 1024$.

Anisotropic Gaussian Mixture Experiment (25 components on a 5×5 grid):

- Parameters for $d = 10$: batch size $= 256$, lr $= 3.707 \cdot 10^{-5}$, epochs $= 180$, σ_end $= 2.6692$, loss_scale $= 386.7248$, hidden_dim $= 256$.
- Parameters for $d = 50$: batch size $= 256$, lr $= 1.593 \cdot 10^{-5}$, epochs $= 180$, σ_end $= 1.6868$, loss_scale $= 916.4250$, hidden_dim $= 512$.

C.11 ERMBRIDGE ALGORITHM

It should be noted that throughout the algorithm $V_\theta(y)$ denotes the log Schrödinger potential $\log \nu_T^\theta(y)$.

Algorithm 1 Training of ERMBridge

Input: Datasets $\mathcal{X} = \{x_i\}$, $\mathcal{Y} = \{y_j\}$; kernel variance $2\sigma^2$ per epoch; learning rate η; batch size B; epochs N_{epochs}; loss scale γ; gradient clip $G_{\max}$; weight $w(y)$ (e.g. $w \equiv 1$).
Initialize: Neural network parameters θ for ν_T^θ (output interpreted as $V_\theta(y) = \log \nu_T^\theta(y)$).
for epoch $e = 1$ to N_{epochs} **do**
  Set $2\sigma^2 \leftarrow$ bandwidth for epoch e (fixed or annealed).
  Shuffle indices of $\mathcal{Y}$; partition into batches $Y_1, \dots, Y_M$ of size $\le B$.
  for each batch $Y_b \subseteq \mathcal{Y}$ **do**
    Sample $X_b \subseteq \mathcal{X}$ (full $\mathcal{X}$ or random subsample of size $\le B$).
    for each $y \in Y_b$ **do**
      $V(y) \leftarrow$ network output at y
    end for
    for each $x \in X_b$ **do**
      $\text{term}(y) \leftarrow -\frac{\|x-y\|^2}{2\sigma^2} - V(y) \quad$ for all $y \in Y_b$
      $\log D(x) \leftarrow \log \sum_{y \in Y_b} \exp(\text{term}(y))$
    end for
    for each $y \in Y_b$ **do**
      $\text{term}(x) \leftarrow -\frac{\|y-x\|^2}{2\sigma^2} - \log D(x) \quad$ for all $x \in X_b$
      $\log(C\phi)(y) \leftarrow \log \sum_{x \in X_b} \exp(\text{term}(x))$
    end for
    $s \leftarrow \left(\frac{1}{|Y_b|} \sum_{y \in Y_b} w(y) \,/\, \exp(\log(C\phi)(y))\right)^{-1}$
    $\log(T\phi)(y) \leftarrow \log s + \log(C\phi)(y) \quad$ for all $y \in Y_b$
    $\phi(y) \leftarrow \exp(V(y)), \quad (T\phi)(y) \leftarrow \exp(\log(T\phi)(y))$
    $\widetilde{T\phi}(y) \leftarrow \phi(y) + (T\phi)(y)$
    $\Delta(y) \leftarrow V(y) - \log \widetilde{T\phi}(y)$
    $\mathcal{L}(\theta) \leftarrow \gamma \cdot \frac{1}{|Y_b|} \sum_{y \in Y_b} \Delta(y)^2$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$
    Clip $\nabla_\theta \mathcal{L}$ to max norm $G_{\max}$ (before step if applied there).
  end for
end for
Output: Learned parameters θ^* (hence $V_{\theta^*}(y)$).

Algorithm 2 Drift computation and sampling (Euler–Maruyama)

Input: Learned $V_{\theta^*}(y)$; reference set $\mathcal{Y}_{\mathrm{ref}} \subseteq \mathcal{Y}$ (or full $\mathcal{Y}$); diffusion schedule $\sigma(t) \geq 0$; number of steps K; optional drift clip $U_{\max}$.
Drift at (t, x)**:**
  $Y_{\mathrm{batch}} \leftarrow$ subsample of $\mathcal{Y}_{\mathrm{ref}}$ (or full set)
  $\mathrm{logits}(y) \leftarrow -\frac{\|x-y\|^2}{2\nu_t} - V_{\theta^*}(y) \quad$ for $y \in Y_{\mathrm{batch}}$
  $h(x) \leftarrow \log \sum_{y \in Y_{\mathrm{batch}}} \exp(\mathrm{logits}(y)) \quad$ (clamp logits to $[-200, 200]$ for stability)
  $u(t, x) \leftarrow \sigma(t)^2 \cdot \nabla_x h(x)$
  Optionally: $u \leftarrow \mathrm{clip}(u, U_{\max})$, and replace NaN/Inf in u by 0.

Sampling: Given $x_0 \sim \rho_0$, set $x \leftarrow x_0$, $\Delta t \leftarrow 1/K$.
for $k = 0$ to $K - 1$ **do**
  $t \leftarrow k \cdot \Delta t$
  Compute $u(t, x)$ as above (with $\sigma(t)$).
  $\xi \sim \mathcal{N}(0, I)$
  $x \leftarrow x + u(t, x)\, \Delta t + \sigma(t) \sqrt{\Delta t}\, \xi$
  (Optional: clamp x to a bounded box in high dimensions.)
end for
Output: $x_T \approx$ sample from ρ_T.