On the approximation of Schrödinger bridge potentials

Anonymous Authors[1]

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 $\mathcal{H}(P\,\|\,\mathsf{Q}) := \int \log\left(\frac{dP}{d\mathsf{Q}}\right) dP$. 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

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

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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

coincides with entropic optimal transport (EOT) when Q is a Wiener measure (Peyré & Cuturi, 2019).

In many modern applications, ρ_0 and ρ_T are not available as closed-form densities; instead, one observes samples $X_1, \dots, X_N \sim \rho_0$ and $Y_1, \dots, Y_M \sim \rho_T$. A standard computational route solves the static SBP or EOT problem on empirical measures using Sinkhorn iterations (Sinkhorn, 1967; Franklin & Lorenz, 1989; Cuturi, 2013; Peyré & 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 & Niles-Weed, 2024). 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 from the recent paper (Belomestny et al., 2026) 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). (Belomestny et al., 2026) studies behavior of $\mathbb{E}[\mathcal{R}(\widehat{g}_{N,M})]$ under a set of assumptions on ρ_0, ρ_T, transition kernel and a class $\mathcal{G}$. It also states that low residual risk shows that $\widehat{g}_{N,M}$ is close to the *fixed-point* of $g \mapsto C[g]$, but translating this into a bound on $||\widehat{g}_{N,M} - g^\star||$ for some norm is non-trivial.

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. In this paper to obtain a well-posed stability analysis, we therefore fix a normalization for $\mathcal{C}[g]$ and work with the associated scale-invariant map $\mathcal{T}$, for which $\mathcal{T}[cg] = \mathcal{T}[g]$ for all $c > 0$. We define $g^\star$ to be a fixed point of $\mathcal{T}[g]$ and replace $\mathcal{C}[g]$ by its normalized version in the definition of $\widehat{\mathcal{R}}_{N,M}$ and $\mathcal{R}$.

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

- 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. We also provide additional experiments for non-Gaussian targets.

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 & Lorenz, 1989; Cuturi, 2013; Peyré & 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 & 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., 2023; 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 & Niles-Weed, 2024). 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 & 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., 2025a), (Puchkin et al., 2025b),

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.

As stated in the introduction, Schrödinger potentials (ν_0, ν_T) are only defined up to a global multiplicative constant. 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 $g \in \mathcal{G}$,

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

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

$$\widehat{\rho}_0^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}, \qquad \widehat{\rho}_T^M := \frac{1}{M}\sum_{j=1}^{M} \delta_{Y_j}.$$

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*

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

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

We first note that

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

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 nonconvex 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*

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

where 1_A *is the indicator of the event*

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

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 $T'[g^\star]$ is a *strict contraction* (assumption (A1)), then the residual $F_e(g) = g - T[g]$ 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:

$$\sup_{\|g-g^\star\|\le r} \|\mathcal{T}''[g]\| \;\le\; L, \tag{24}$$

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

$$\|u\|_{H,R} \le \left\|u/g^\star - 1\right\|_{L^\infty(B_R)}. \tag{25}$$

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. (28)) we get

$$\|u\|_{H,R} \le \frac{e^{(a_-^\star + b_T^-)R^2}}{c_-^\star c_T^-}\|u\|_{L^2(\rho_T)}. \tag{26}$$

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 Colorally 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 $\mathrm{span}\{\psi_\alpha^{(\lambda)} : |\alpha| \le n\}$. 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. 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

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

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

In this section, we would like to demonstrate experimentally the correctness of our empirical risk minimization procedure and compare its properties with SinkhornBridge (Pooladian & Niles-Weed, 2024) on some important examples.

The first three experiments described below investigate the Gaussian-to-Gaussian translation, i.e., $\rho_0 = \mathcal{N}(m_0, C_0)$ and $\rho_T = \mathcal{N}(m_1, C_1)$. In this case, the Schrödinger potential corresponding to the optimal translation, the optimal drift, and $g^\star$ are all analytically computable (see (Bunne et al., 2023)), allowing us to more accurately evaluate the performance of the proposed algorithm and SinkhornBridge. In all experiments we assumed $N = M$ for brevity and ease of comparison.

In following experiments, optimization was performed for the logarithmic Schrödinger potential. When compared with the true potentials, this can lead to estimation inaccuracies due to the fact that normalization was performed on batches from the training set, which inevitably has a finite number of elements. To avoid this, all functions (i.e., g, ν, and drift) were compared with the true ones up to a constant shift, numerically selected to minimize the corresponding MSE deviation.

Below, the algorithm we proposed in the article will be referred to as ERMBridge for brevity. The formal algorithm, hyperparameter values, and further experimental details are provided in C.9.

Potential smoothing technique in SinkhornBridge and ERMBridge We need to note that 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. The variance of this filter is explicitly coupled to the time remaining $(1-t)$ and the EOT regularization ϵ. In contrast, the proposed 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 neural network's inductive bias during training, and an explicit kernel smoothing is applied during drift computation using a diffusion schedule $\sigma(t)$.

Analysis of $\widehat{g}_{M,N}$ convergence to $g^\star$ In this experiment, our goal was to demonstrate that the empirical risk minimization procedure we are investigating leads to a true approximation of the function $g^\star$ as the number of data points available to the model during training increases. We evaluate the mean-squared error for $\widehat{g}_{M,N}$, defined as

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

where $(y_i)_{i=1}^n$ are i.i.d. $\mathcal{N}(m_1, C_1)$ random vectors.

The monotonically decreasing mean-squared error on Figure 1 between $\widehat{g}_{M,N}$ and $g^\star$ estimated from the holdout sample indicates that the estimate converges to the true function (see C.3 for details).

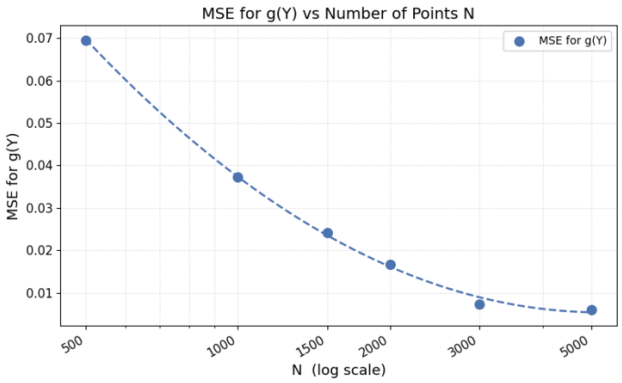

Figure 1. Results of $\widehat{g}_{M,N}$ convergence to $g^\star$ in the case of Gaussian-to-Gaussian translation in 2D for ERMBridge with neural network approximation of the log-Schrödinger potential.

Analysis of Schrödinger potential estimation in different dimensions The experiment below allows us to directly compare ERMBridge with SinkhornBridge in terms of Schrödinger potential estimation. We fixed the number of points for estimating the Schrödinger potential and the number of points for calculating the mean-squared error on the holdout sample, and then ran the algorithms for different dimensions $[5, 10, 25, 50, 100]$ (see C.4 for details). We evaluate the mean-squared error for $\nu_T^{M,N}$, defined as

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

where

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

and $(y_i)_{i=1}^n$ are i.i.d $\mathcal{N}(m_1, C_1)$ random vectors.

It should be noted that we performed the analysis for ERMBridge not only in the class of neural networks for approximating the log-Schrödinger potential but also in the class of quadratic functions, which is known to contain the true log-Schrödinger potential for the Gauss-to-Gauss translation problem (Bunne et al., 2023). As Figure 2 shows, the quality of the potential estimate using SinkhornBridge degrades significantly faster than that using ERMBridge in the class of quadratic functions and even faster than that using ERMBridge in the class of neural networks.

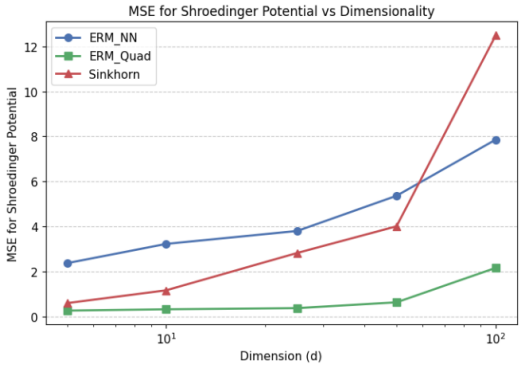

Figure 2. Estimation results of $\nu_T^{M,N}$ in the case of Gaussian-to-Gaussian translation in dimensions $[5, 10, 25, 100]$ for ERMBridge with approximation of the log-Schrödinger potential by a neural network, ERMBridge with approximation of the log-Schrödinger potential in the class of quadratic functions, and for SinkhornBridge, respectively.

This demonstrates the advantage of using the empirical risk minimization procedure for estimating the continuous approximation of the logarithmic Schrödinger potential as a function over the discrete estimate offered by the Sinkhorn procedure.

Comparative analysis of optimal drift estimation In our third experiment, the goal was to demonstrate the advantages of ERMBridge over SinkhornBridge in terms of the quality of the approximation of the true drift, since it is this drift that determines the final quality of the generated samples in the Euler-Maruyama SDE discretization (see C.8 for details).

We define the optimal drift (see (Dai Pra, 1991))

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

and the corresponding estimated drift as

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

Similarly as before, we evaluate the mean-squared error of

$b_{M,N}$, defined as

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

where $(x_i)_{i=1}^n$ are i.i.d $\mathcal{N}(m_0, C_0)$ random vectors and $(t_i)_{i=1}^n$ are i.i.d $U[0.05, 0.95]$ random variables.

It should be noted that in cases with insufficient N, high d, or both, the mean-squared error for SinkhornBridge is significantly larger than that of ERMBridge (see Figure 3). The experiment clearly demonstrates the advantage of the neural network-estimated log-potential over the discrete, kernel-smoothed Sinkhorn potential when the data is limited or the dimensionality is high (see C.5 for details).

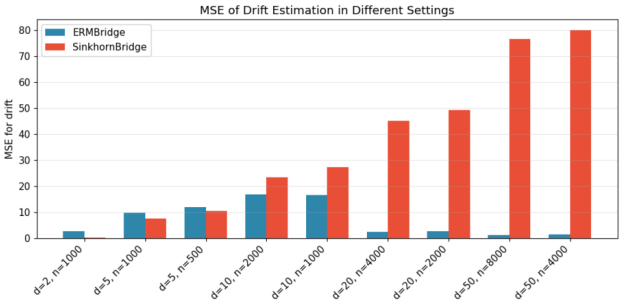

Figure 3. Results of $b_{M,N}$ estimation in the case of Gaussian-to-Gaussian translation in the $(d, N) \in [(2, 1000), (5, 1000), (5, 500), (10, 2000), (10, 1000), (20, 4000), (20, 2000), (50, 8000), (50, 4000)]$ settings for ERMBridge with neural network approximation of the log-Schrödinger potential and SinkhornBridge, respectively.

Although in low dimensions and with a sufficient number of points, SinkhornBridge can produce a reasonably good estimate of the optimal drift, surpassing that of ERMBridge, its quality degrades too rapidly with increasing dimensions, making it effectively unsuitable for complex problems such as working with high-dimensional biological or image data.

Extension to non-Gaussian mixture targets To verify that the advantages observed in the analytical Gaussian setting carry over to regimes without closed-form potentials, we additionally compare ERMBridge against SinkhornBridge on two non-Gaussian targets, with ρ_0 a standard Gaussian in both cases: (i) a Laplace mixture, probing the heavy-tailed regime, and (ii) an anisotropic Gaussian mixture of 25 components arranged on a fixed 5×5 grid embedded in $\mathbb{R}^d$, where each component has an independently drawn elongated covariance with random orientation, so that per-location kernel adaptation matters. Since ground-truth potentials are unavailable, we evaluate sample transport quality using the Energy distance, the standard metric in this setting, and additionally report the sliced 1-Wasserstein distance.

Table 1. Sample transport quality on non-Gaussian mixture targets. Both methods recover all 25 modes in the anisotropic mixture setting.

d	Method	sliced $\mathbb{W}_1 \downarrow$	Energy$\downarrow$	ΔEnergy
Gaussian to Laplace mixture				
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	
Gaussian to anisotropic Gaussian mixture				
10	ERMBridge	**0.1391**	**0.0054**	+49.9%
	SinkhornBridge	0.1426	0.0107	
50	ERMBridge	**0.1527**	**0.0078**	+29.8%
	SinkhornBridge	0.2086	0.0111	

As reported in Table 1, ERMBridge dominates SinkhornBridge on both metrics across all settings, with the relative gap in Energy distance growing as the dimension increases, mirroring the trend already observed in the analytical Gaussian-to-Gaussian potential and drift estimation experiments. Since both methods recover all modes in the multimodal setting, the improvement reflects transport accuracy rather than mode coverage.

Evaluation on real biological data To demonstrate that the conclusions above carry over to real-world problems, we additionally evaluated ERMBridge on the single-cell trajectory inference task on the EB dataset (Tong et al., 2020). ERMBridge attains $\mathbb{W}_1 = 0.822$, statistically indistinguishable from SinkhornBridge ($\mathbb{W}_1 = 0.827$) and on par with other state-of-the-art entropic-OT and flow-matching baselines. The full experimental protocol and a comparison against a broader set of methods are provided in Appendix C.7.

References

Belomestny, D., Naumov, A., Puchkin, N., and Suchkov, D. Schrödinger bridge problem via empirical risk minimization. *arXiv e-prints*, art. arXiv:2602.08374, February 2026. doi: 10.48550/arXiv.2602.08374.

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

Chen, Y., Georgiou, T. T., and Pavon, M. Entropic and displacement interpolation: A computational approach using the Hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016. doi: 10.1137/16M1061382. URL https://doi.org/10.1137/16M1061382.

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

Conforti, G., Durmus, A., and Greco, G. Quantitative contraction rates for sinkhorn algorithm: Beyond bounded costs and compact marginals. *arXiv e-prints*, April 2023. doi: 10.48550/arXiv.2304.04451. URL https://doi.org/10.48550/arXiv.2304.04451.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013. URL https://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf.

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

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

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

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. 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*, pp. 2681–2690. PMLR, 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. How to train your neural ODE: the world of Jacobian and kinetic regularization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3154–3164. PMLR, 2020. URL https://proceedings.mlr.press/v119/finlay20a.html.

Franklin, J. and Lorenz, J. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989. doi: 10.1016/0024-3795(89)90490-4. URL https://doi.org/10.1016/0024-3795(89)90490-4.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. 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*, pp. 1574–1583. PMLR, 2019. URL https://proceedings.mlr.press/v89/genevay19a.html.

Greco, G., Noble, M., Conforti, G., and Durmus, A. 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*, pp. 716–746. PMLR, 2023. URL https://proceedings.mlr.press/v195/greco23a.html.

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

Koshizuka, T. and Sato, I. Neural lagrangian schrödinger bridge: Diffusion modeling for population dynamics. *arXiv preprint arXiv:2204.04853*, 2022. doi: 10.48550/arXiv.2204.04853. URL https://doi.org/10.48550/arXiv.2204.04853.

Lemmens, B. and Nussbaum, R. Birkhoff's version of Hilbert's metric and applications. In Papadopoulos, A. and Troyanov, M. (eds.), *Handbook of Hilbert Geometry*, IRMA Lectures in Mathematics and Theoretical Physics, pp. 275–303. European Math. Soc., December 2014. URL https://kar.kent.ac.uk/41206/.

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

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4543–4553, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5acdc9ca5d99ae66afdfe1eea0e3b26b-Abstract.html.

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

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, February 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

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

2024. doi: 10.48550/arXiv.2408.11686. URL https://doi.org/10.48550/arXiv.2408.11686.

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

Puchkin, N., Suchkov, D., Naumov, A., and Belomestny, D. Tight bounds for schrödinger potential estimation in unpaired data translation, 2025b. URL https://arxiv.org/abs/2508.07392.

Schrödinger, E. Über die Umkehrung der Naturgesetze. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse*, pp. 144–153, 1932.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. Diffusion schrödinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023. doi: 10.48550/arXiv.2303.16852. URL https://doi.org/10.48550/arXiv.2303.16852.

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

Tong, A., Huang, J., Wolf, G., van Dijk, D., and Krishnaswamy, S. 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*, pp. 9526–9536. PMLR, 2020. URL https://proceedings.mlr.press/v119/tong20a.html.

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023a. doi: 10.48550/arXiv.2302.00482. URL https://doi.org/10.48550/arXiv.2302.00482. 2023a.

Tong, A., Malkin, N., Fatras, K., Atanackovic, L., Zhang, Y., Huguet, G., Wolf, G., and Bengio, Y. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023b. doi: 10.48550/arXiv.2307.03672. URL https://doi.org/10.48550/arXiv.2307.03672. 2023b.

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

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*

$$g^\star(y) = \mathcal{C}[g^\star](y), \qquad y \in \mathbb{R}^d, \qquad \text{and} \qquad \int_{B_R} \frac{1}{g^\star(z)}\, \rho_T(z)\, dz = 1. \tag{27}$$

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

$$c_-^\star \exp(-a_-^\star \|y\|^2) \;\le\; g^\star(y) \;\le\; c_+^\star \exp(-a_+^\star \|y\|^2), \qquad \forall y \in \mathbb{R}^d. \tag{28}$$

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 this ball:

$$\sup_{\|g - g^\star\| \le r} \|\mathcal{T}''[g]\| \;\le\; L, \tag{29}$$

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 $g, h \in \overline{B}(g^\star, r)$,

$$\|\mathcal{T}'[g] - \mathcal{T}'[h]\| \;\le\; L\, \|g - h\|, \tag{30}$$

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 (29) 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 (30). □

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

$$\mathcal{K}(y, z) = \frac{\rho_T(z)}{(g^\star(z))^2} \int_{\mathbb{R}^d} \rho_0(x)\, \frac{q_T(x, y)\, q_T(x, z)}{(D_{g^\star}(x))^2}\, dx. \tag{31}$$

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 (31). □

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 $\mathcal{C}'[g^\star]$:

$$\mathcal{K}_{\mathcal{T}}(y,z) = \mathcal{K}(y,z) - g^\star(y)\, \Lambda(z), \tag{32}$$

with

$$\Lambda(z) := \int_{\mathbb{R}^d} \frac{\rho_T(\tilde{y})}{(g^\star(\tilde{y}))^2}\, \mathcal{K}(\tilde{y},z)\, w(\tilde{y})\, d\tilde{y}. \tag{33}$$

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

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

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

$$\mathcal{T}'[g^\star]u = \mathcal{C}'[g^\star]u + \big(s'[g^\star]u\big)\, g^\star. \tag{35}$$

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

$$s'[g^\star]u = -\int_{\mathbb{R}^d} \frac{\rho_T(y)}{(g^\star(y))^2}\, (\mathcal{C}'[g^\star]u)(y) w(y)\, dy. \tag{36}$$

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 (36) 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 (33). Plugging this into (35) 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 (32)–(33). □

Proposition A.5. *Assume* ($\mathrm{Q}_{\mathrm{Gauss}}$) *and that* $\mathrm{supp}(\rho_0) \subset B_{R_0} := \{x : \|x\| < R_0\}$. *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* $\mathrm{supp}(\rho_0)$,

$$D_{g^\star}(x) := \int_{\mathbb{R}^d} q_T(x,z)\,\frac{\rho_T(z)}{g^\star(z)}\,dz \;\geq\; \underline{D}, \qquad \forall x \in \mathrm{supp}(\rho_0). \tag{37}$$

Then:

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

$$\|\mathsf{T}u\|_{H,R} \;\leq\; \kappa_R\,\|u\|_{H,R} \;+\; \varepsilon_R\,\|u\|_{L^2(\mathbb{R}^d)}, \tag{38}$$

where

$$\varepsilon_R := \frac{(2\pi T)^{-d/2}}{\underline{D}}\left(\int_{B_R^c} \frac{\rho_T^2(z)}{(g^\star(z))^4}\,e^{-\frac{(\|z\|-R_0)^2}{T}}\,dz\right)^{1/2}, \tag{39}$$

and

$$\kappa_R := \tanh\Big(\frac{(R+R_0)R}{T}\Big). \tag{40}$$

(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* $u/g^\star \in L^\infty(B_R)$,

$$\|\mathsf{T}u\|_{L^2(\mathbb{R}^d)} \;\leq\; C\,\|u\|_{H,R} \;+\; \varepsilon_R\left(1+\frac{1}{\gamma}\right)\|u\|_{L^2(\mathbb{R}^d)}, \tag{41}$$

where C *depends on the norms* $\|h\|_{L^2(\mathbb{R}^d)}$, $\|h\|_{L^\infty(\mathbb{R}^d)}$ *and* ε_R *is as in* (39).

Proof. We proceed in two parts.

Part (i): proof of the Hilbert contraction (38)**.** Write $r := u/g^\star$. Since $\mathcal{T}$ is the normalized version of $\mathcal{C}$, its derivative has the rank–one form

$$\mathsf{T}u = \mathcal{C}'[g^\star]u - g^\star\,\langle \Lambda, u\rangle \tag{42}$$

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:

$$\frac{\mathsf{T}u}{g^\star} = \frac{\mathcal{C}'[g^\star]u}{g^\star} - \langle\Lambda,u\rangle, \qquad \text{hence} \qquad \operatorname*{osc}_{B_R}\Big(\frac{\mathsf{T}u}{g^\star}\Big) = \operatorname*{osc}_{B_R}\Big(\frac{\mathcal{C}'[g^\star]u}{g^\star}\Big), \tag{43}$$

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

$$\|\mathsf{T}u\|_{H,R} = \frac{1}{2}\operatorname*{osc}_{B_R}\Big(\frac{\mathcal{C}'[g^\star]u}{g^\star}\Big). \tag{44}$$

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 $y \in \mathbb{R}^d$,

$$\frac{(\mathcal{C}'[g^\star]u)(y)}{g^\star(y)} = \int_{B_{R_0}} \pi_y(dx)\;\underbrace{\int_{\mathbb{R}^d} \eta_x(dz)\,r(z)}_{=:(\mathsf{P}r)(x)}, \tag{45}$$

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

$$\pi_y(dx) := \frac{W(x)\, q_T(x,y)}{\int_{B_{R_0}} W(x')\, q_T(x',y)\, dx'}\, dx, \tag{46}$$

$$\eta_x(dz) := \frac{q_T(x,z)\, \frac{\rho_T(z)}{g^\star(z)}}{D_{g^\star}(x)}\, dz, \tag{47}$$

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

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

Therefore,

$$\sup_{x\in B_{R_0}} \frac{q_T(x-y)}{q_T(x-y')} \le \exp\Big(\frac{2(R+R_0)R}{T}\Big) =: \Delta_R, \qquad \inf_{x\in B_{R_0}} \frac{q_T(x-y)}{q_T(x-y')} \ge \Delta_R^{-1}. \tag{48}$$

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

$$\Big|\int \varphi\, d\pi_y - \int \varphi\, d\pi_{y'}\Big| \le \tanh\Big(\frac{1}{4}\log \Delta_R\Big)\ \operatorname*{osc}_{B_{R_0}}(\varphi). \tag{49}$$

(Proof: write $\pi_y, \pi_{y'}$ with densities proportional to $q_T(x-y)$ and $q_T(x-y')$; (48) 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 (40) (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 (45), 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 $\mathrm{osc}_{B_R}(r) = 2\|u\|_{H,R}$ (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 $D_{g^\star}(x) \geq \underline{D}$,

$$\begin{aligned}\sup_{x\in B_{R_0}} \int_{B_R^c} |r(z)|\,\eta_x(dz) &\leq \frac{1}{\underline{D}} \sup_{x\in B_{R_0}} \int_{B_R^c} q_T(x-z)\,\frac{\rho_T(z)}{g^\star(z)}\,\frac{|u(z)|}{g^\star(z)}\,dz \\ &\leq \frac{1}{\underline{D}} \sup_{x\in B_{R_0}} \left(\int_{B_R^c} q_T(x-z)^2\,\frac{\rho_T(z)^2}{(g^\star(z))^4}\,dz\right)^{1/2} \|u\|_{L^2(\mathbb{R}^d)}. \end{aligned} \tag{50}$$

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

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

Plugging this into (50) gives

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

with ε_R as in (39). Hence

$$\|\mathsf{T}u\|_{H,R} \leq \kappa_R\,\|u\|_{H,R} + \varepsilon_R\,\|u\|_{L^2(\mathbb{R}^d)}. \tag{51}$$

Part (ii): proof of the mixed L^2 estimate (41). 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

$$\mathsf{T}u = \mathsf{T}(u - g^\star). \tag{52}$$

From (42),

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

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

$$\|\mathcal{C}'[g^\star]u_{\mathrm{out}}\|_{L^2_y} \leq \left(\int_{B_R^c} \|\mathcal{K}(\cdot,z)\|_{L^2_y}^2\,dz\right)^{1/2} \|u - g^\star\|_{L^2_z}. \tag{53}$$

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

$$\begin{aligned}\|\mathcal{K}(\cdot,z)\|_{L^2_y} &= h(z) \left\| \int_{B_{R_0}} \rho_0(x)\,\frac{q_T(x-\cdot)\,q_T(x-z)}{(D_{g^\star}(x))^2}\,dx \right\|_{L^2_y} \\ &\leq h(z) \int_{B_{R_0}} \rho_0(x)\,\frac{q_T(x-z)}{(D_{g^\star}(x))^2}\,\|q_T(x-\cdot)\|_{L^2_y}\,dx. \end{aligned} \tag{54}$$

By translation invariance of Lebesgue measure,

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

Hence

$$\|\mathcal{K}(\cdot,z)\|_{L^2_y} \leq \frac{\|q_T\|_2}{\underline{D}^2}\,h(z) \int_{B_{R_0}} \rho_0(x)\,q_T(x-z)\,dx. \tag{55}$$

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

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

Since $\int \rho_0 = 1$,

$$\int_{B_{R_0}} \rho_0(x)\, q_T(x - z)\, dx \leq (2\pi T)^{-d/2} \exp\Big(- \frac{(\|z\| - R_0)^2}{2T} \Big). \tag{56}$$

Combining (55)–(56) yields

$$\|\mathcal{K}(\cdot, z)\|_{L^2_y} \leq \frac{\|q_T\|_2\, (2\pi T)^{-d/2}}{\underline{D}^2}\, h(z) \exp\Big(- \frac{(\|z\| - R_0)^2}{2T} \Big). \tag{57}$$

Squaring (57) gives

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

Insert this into (53) and integrate over $z \in B_R^c$:

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

Hence

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

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

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 $|r - m_R| \leq 2 \|u\|_{H,R}$ on B_R and hence

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

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 $\operatorname{ess\,inf}_{B_R} r \leq 1 \leq \operatorname{ess\,sup}_{B_R} r$. Therefore,

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

As a result,

$$\|u_{\mathrm{in}}\|_{L^2(\mathbb{R}^d)} \leq 3 \|g^\star\|_{L^2(\mathbb{R}^d)}\, \|u\|_{H,R}. \tag{58}$$

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 $f = u_{\text{in}}$ yields

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

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

$$B \le \frac{(2\pi T)^{-d/2}}{\underline{D}^2}\, \|h\|_{L^\infty}. \tag{59}$$

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

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

Due to (58), we have

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

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

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

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 $R > 0$,

$$\begin{aligned}\|\Lambda\|^2_{L^2(B_R)} &\le \Big(\frac{\|q_T\|_2\, \|q_T\|_\infty}{\underline{D}^2} \Big)^2 \|h\|^2_{L^2(\mathbb{R}^d)} \int_{B_R} h(z)^2\, dz \\ &\le \Big(\frac{\|q_T\|_2\, \|q_T\|_\infty}{\underline{D}^2} \Big)^2 \|h\|^4_{L^2(\mathbb{R}^d)}. \end{aligned} \tag{60}$$

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

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

Then, by Cauchy–Schwarz,

$$|\langle \Lambda, u - g^\star \rangle| \le \|\Lambda\|_{L^2(B_R)} \|u_{\mathrm{in}}\|_{L^2(\mathbb{R}^d)} + \|\Lambda\|_{L^2(B_R^c)} \|u_{\mathrm{out}}\|_{L^2(\mathbb{R}^d)}. \tag{61}$$

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

$$|A(z)| \le \frac{1}{\underline{D}^2} \int_{B_{R_0}} \rho_0(x)\, q_T(x-z)\, dx \le \frac{\|q_T\|_\infty}{\underline{D}^2}. \tag{62}$$

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

$$\|A\|_{L^2(\mathbb{R}^d)} \le \frac{1}{\underline{D}^2} \|\rho_0 * q_T\|_{L^2(\mathbb{R}^d)} \le \frac{1}{\underline{D}^2} \|\rho_0\|_{L^1(\mathbb{R}^d)}\, \|q_T\|_{L^2(\mathbb{R}^d)} = \frac{\|q_T\|_2}{\underline{D}^2}. \tag{63}$$

Combining (62) and (63) 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

$$\|\Lambda\|^2_{L^2(B_R)} \le \|\Lambda\|^2_{L^2(\mathbb{R}^d)} = \|hA\|_2^2 \le \|h\|^2_{L^4(\mathbb{R}^d)}\, \|A\|^2_{L^4(\mathbb{R}^d)} \le \frac{\|q_T\|_2\, \|q_T\|_\infty}{\underline{D}^4}\, \|h\|^2_{L^4(\mathbb{R}^d)}. \tag{64}$$

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,

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

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

$$\|\Lambda\|^2_{L^2(B_R^c)} \le \frac{c_+^2}{\underline{D}^4}\int_{B_R^c} h(z)^2\,\exp\bigl(-2a_+(\|z\|-R_0)^2\bigr)\,dz \tag{65}$$

and hence

$$\|\Lambda\|_{L^2(B_R^c)} \le \frac{c_+\,\|h\|_{L^\infty}}{\underline{D}^2}\,\Bigl(\frac{|\mathbb{S}^{d-1}|}{4a_+}\Bigr)^{1/2}\,R^{\frac{d-1}{2}}\,\exp\bigl(-a_+(R-R_0)^2\bigr). \tag{66}$$

□

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

$$\|\mathsf{T}u\|_{H,R} \le \kappa_R\,\|u\|_{H,R} + \varepsilon_R\,\|u\|_{L^2(\mathbb{R}^d)}, \tag{67}$$

$$\|\mathsf{T}u\|_{L^2(\mathbb{R}^d)} \le C\,\|u\|_{H,R} + \varepsilon_R\Bigl(1+\frac{1}{\gamma}\Bigr)\,\|u\|_{L^2(\mathbb{R}^d)}. \tag{68}$$

Then

$$\|\mathsf{T}u\|_{R,\lambda} \le \alpha_R(\lambda)\,\|u\|_{R,\lambda}, \tag{69}$$

where

$$\alpha_R(\lambda) := \max\left\{\kappa_R + \lambda C,\ \varepsilon_R\Bigl(\frac{1}{\lambda}+1+\frac{1}{\gamma}\Bigr)\right\}. \tag{70}$$

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 (67) and (68):

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

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 (69) with (70).

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

$$\alpha_R(\lambda_0) \le \kappa_R + \sqrt{C\,\varepsilon_R} + \varepsilon_R\Bigl(1+\frac{1}{\gamma}\Bigr). \tag{71}$$

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*

$$\rho_0(x) \geq a_0 \qquad \text{for a.e. } x \in B(r_0). \tag{72}$$

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

$$C_\ell \exp\big(-c_\ell \|x - y\|^p\big) \leq k(x, y) \leq C_u \exp\big(-c_u \|x - y\|^p\big) \qquad \forall x, y \in \mathbb{R}^d. \tag{73}$$

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

$$m(R) := \inf_{\|x\| \leq R_0,\ \|y\| \leq R} k(x, y) \ > \ 0. \tag{74}$$

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

$$\int_{\mathbb{R}^d} \exp\big(\gamma \|z\|^p\big)\, \rho_T(z)\, dz < \infty. \tag{75}$$

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

$$D_{g^\star}(x) := \int_{\mathbb{R}^d} k(x, z)\, \frac{\rho_T(z)}{g^\star(z)}\, dz, \tag{76}$$

the fixed–point equation

$$g^\star(y) = \int_{\mathbb{R}^d} \frac{\rho_0(x)}{D_{g^\star}(x)}\, k(x, y)\, dx \qquad \forall y \in \mathbb{R}^d \tag{77}$$

holds together with the normalization

$$\int_{\mathbb{R}^d} \frac{\rho_T(z)}{g^\star(z)}\, dz = 1. \tag{78}$$

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

$$c_-\, e^{-\alpha_+ \|y\|^p} \ \leq \ g^\star(y) \ \leq \ c_+\, e^{-\alpha_- \|y\|^p}, \qquad \forall y \in \mathbb{R}^d. \tag{79}$$

Proof. We apply a Schauder fixed–point argument 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

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

Then $\mathcal{K}$ is convex and 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 (78); moreover $\Lambda(g^\star) = 1$ and hence $g^\star = \mathcal{T}g^\star$. Write $w_g(z) := \rho_T(z)/g(z) \geq 0$. By the normalization in $\mathcal{K}$,

$$\int_{\mathbb{R}^d} w_g(z)\, dz = 1. \tag{80}$$

By (73),

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

so using (80),

$$D_g(x) = \int k(x,z) w_g(z)\, dz \leq \sup_{x,z} k(x,z) \int w_g = C_u, \qquad \forall x. \tag{81}$$

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

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

Assumption (75) 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

$$\sup_{g\in\mathcal{K}} \int_{\|z\|>R_\varepsilon} w_g(z)\, dz \leq \varepsilon. \tag{82}$$

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

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

where $m(R) > 0$ is defined in (74). Thus

$$\frac{1}{2}\, m(R) \ \leq\ D_g(x)\ \leq\ C_u, \qquad \forall x \in B(R_0),\ \forall g \in \mathcal{K}. \tag{83}$$

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 (83) and (73),

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

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

$$(\mathcal{T}g)(y) \leq C_2 e^{-\alpha_-\|y\|^p} \qquad \forall y,\ \forall g \in \mathcal{K}. \tag{84}$$

From (81) we have $1/D_g(x) \geq 1/C_u$ for all x. Using (72) and continuity/positivity of k,

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

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

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

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

$$(\mathcal{T}g)(y) \geq C_4 e^{-\alpha_+\|y\|^p} \qquad \forall y,\ \forall g \in \mathcal{K}. \tag{85}$$

By (85) and (75),

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

Also $\Lambda(g) > 0$ because ρ_T is not identically zero and $\mathcal{T}g > 0$. Combining (84)–(85) 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 (75) 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 (83). 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 (84). 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, convex, closed, and $\mathcal{S} : \mathcal{K} \to \mathcal{K}$ is continuous with relatively compact image, Schauder's fixed point theorem yields $g^\star \in \mathcal{K}$ such that $g^\star = \mathcal{S}g^\star$. By definition of $\mathcal{S}$, $g^\star$ satisfies (78). 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. (77). The two–sided bounds (79) hold since $g^\star \in \mathcal{K}$. □

C. Further Details of Numerical Experiments

In this section, we elaborate on details of the numerical experiments presented in Section 3. The section is organized as follows. In Appendix C.1, we discuss general implementation details. Appendix C.2 presents information about the metrics used. Appendix C.3 provides additional information about the $\widehat{g}_{M,N}$ to $g^\star$ convergence experiment. Appendix C.4 provides additional details about Schrödinger potential estimation in different dimensions. Appendix C.5 provides additional details about the optimal drift estimation experiment. Appendix C.6 describes the non-Gaussian mixture experiments. Appendix C.7 describes the single-cell experiment. Appendix C.8 describes the ERMBridge algorithm. Finally, Appendix C.9 presents all final hyperparameter values used for the experiments.

C.1. General Implementation Details

In our experiments, we paid special attention to statistical reliability. For the analytical Gaussian-to-Gaussian experiments (Appendices C.3–C.5), the resulting metrics were averaged across runs on three different random seeds to ensure a fair comparison; for the mixture experiments (Appendix C.6) we similarly averaged across three seeds, and for the single-cell experiment (Appendix C.7) we followed the protocol of Tong et al. (2020) and repeated each evaluation five times.

The first three experiments were performed in the Gauss-to-Gauss setting, in which the Schrödinger potential, optimal drift and $g^\star$ are available analytically, allowing us to compare ERMBridge and SinkhornBridge in terms of the quality of reconstruction of the optimal plan and the functions resulting from it rather than through downstream sample quality. In the mixture and single-cell experiments, ground-truth potentials are unavailable, so we instead evaluate the quality of the generated samples using distributional metrics defined in Appendix C.2. In all experiments the prior process was assumed to be a scaled Wiener process with zero drift.

The calculations were fully performed on an Nvidia T4 GPU.

C.2. Metrics Used in the Experiments

For the analytical Gaussian-to-Gaussian experiments, the evaluation metrics (MSE against the closed-form $g^\star$, ν_T and $b^\star$) are defined directly in the main text (Section 3). For the mixture experiments and the single-cell experiment, ground-truth potentials are unavailable and we instead evaluate transport quality through distributional metrics: the sliced 1-Wasserstein distance and the energy distance for the mixture experiments, and the standard 1-Wasserstein distance $\mathbb{W}_1$ for the single-cell experiment.

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 (squared) 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.3. Details of $\widehat{g}_{M,N}$ to $g^\star$ Convergence Experiment

In this experiment, we used source $\rho_0 = \mathcal{N}(m_0, C_0)$ and target $\rho_T = \mathcal{N}(m_1, C_1)$ in $\mathbb{R}^2$. Kernel variance $\nu = \sigma_{\text{end}}^2$, fixed bandwidth $2\sigma^2 = 2\nu$. For analysis we conducted a 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 $N_{\text{eval}} = 2000$.

C.4. 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, the train size was set to $N = 5000$. For evaluation we used $N_{\text{eval}} = 5000$ and averaged over 3 seeds per d.

C.5. Details of Optimal Drift Estimation Experiment

In this experiment, we used the 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 $N/2$ points from $\mathcal{N}(m_0, C_0)$ were used to calculate the MSE, with the same number of points sampled from $t \in U[0.05, 0.95]$.

C.6. Details of Non-Gaussian Mixture Experiments

In this experiment we evaluate ERMBridge and SinkhornBridge on two non-Gaussian targets, with the source ρ_0 chosen to be a standard Gaussian $\mathcal{N}(0, I_d)$ in both cases.

Gaussian to Laplace mixture. The target is a uniform mixture of $K = 5$ isotropic Laplace components in $\mathbb{R}^d$ with mode locations sampled uniformly on the sphere of radius $r = 4$ and shared scale parameter $b = 1$, evaluated for $d \in \{2, 10\}$. We used a training set of $N = 4000$ source-target pairs and a held-out evaluation set of $N_{\text{eval}} = 4000$ samples per side. The sliced $\mathbb{W}_1$ distance was computed with 100 random projections.

Gaussian to anisotropic Gaussian mixture. The target is a uniform mixture of 25 Gaussian components arranged on a fixed 5×5 grid embedded in the first two coordinates of $\mathbb{R}^d$, with grid spacing $\Delta = 4$ and the remaining coordinates centered at the origin. Each component has an independently drawn elongated covariance: anisotropy ratio $10:1$ in two random orthogonal directions, with random orientation, so that per-location kernel adaptation matters. The marginal variance along the remaining $d - 2$ directions is set to $\sigma_\perp^2 = 0.25$. We evaluated $d \in \{10, 50\}$, used $N = 8000$ training pairs and $N_{\text{eval}} = 8000$ evaluation samples per side, and reported the number of recovered modes (a target component is counted as recovered if at least one generated sample lies within Mahalanobis radius 3 of its mean).

C.7. Evaluation on single-cell data

For a more comprehensive comparison with SinkhornBridge, we conducted experiments on biological data (Tong et al., 2020). Following the original paper, we formulated the problem as transporting the cell distribution at time t_{i-1} to time t_{i+1} for $i \in \{1, 2, 3\}$. We use results for other methods from (Tong et al., 2023b), whose authors were the first to consider this setup. We then predicted the cell distribution at the intermediate time t_i and computed the Wasserstein distance $\mathbb{W}_1$

Table 2. Quality of intermediate distribution restoration on single-cell data for various solvers, including ERMBridge and SinkhornBridge. Methods are sorted by $\mathbb{W}_1$ (lower is better).

Solver	$\mathbb{W}_1 \downarrow$	Solver	$\mathbb{W}_1 \downarrow$
OT-CFM (Tong et al., 2023a)	0.790 ± 0.068	DSB (De Bortoli et al., 2021)	0.862 ± 0.023
$[\mathrm{SF}]^2$M-Exact (Tong et al., 2023b)	0.793 ± 0.066	I-CFM (Tong et al., 2023a)	0.872 ± 0.087
LightSB-OU (Puchkin et al., 2025b)	0.815 ± 0.016	$[\mathrm{SF}]^2$M-Geo (Tong et al., 2023b)	0.879 ± 0.148
ERMBridge	0.822 ± 0.037	NLSB (Koshizuka & Sato, 2022)	$0.970 \pm$ N/A[a]
LightSB (Korotin et al., 2024)	0.823 ± 0.017	$[\mathrm{SF}]^2$M-Sink (Tong et al., 2023b)	1.198 ± 0.342
Reg. CNF (Finlay et al., 2020)	$0.825 \pm$ N/A[a]	SB-CFM (Tong et al., 2023a)	1.221 ± 0.380
SinkhornBridge (Pooladian & Niles-Weed, 2024)	0.827 ± 0.026	DSBM (Shi et al., 2023)	1.775 ± 0.429
T. Net (Tong et al., 2020)	$0.848 \pm$ N/A[a]		

[a] The authors did not report the standard deviation.

between the predicted distribution and the ground truth. The results were averaged across all three setups ($i = 1, 2, 3$); see Table 2. To ensure statistical robustness, we repeated the experiment five times. ERMBridge achieved $\mathbb{W}_1 = 0.822$, while SinkhornBridge achieved $\mathbb{W}_1 = 0.827$. The corresponding hyperparameter values are reported in Appendix C.9.

C.8. 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**
      $\mathrm{term}(y) \leftarrow -\frac{\|x-y\|^2}{2\sigma^2} - V(y) \quad$ for all $y \in Y_b$
      $\log D(x) \leftarrow \log \sum_{y \in Y_b} \exp(\mathrm{term}(y))$
    end for
    for each $y \in Y_b$ **do**
      $\mathrm{term}(x) \leftarrow -\frac{\|y-x\|^2}{2\sigma^2} - \log D(x) \quad$ for all $x \in X_b$
      $\log(C\phi)(y) \leftarrow \log \sum_{x \in X_b} \exp(\mathrm{term}(x))$
    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.

C.9. 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 $= 5e-5,$ epochs $= 120,$ σ_end $= 1.0,$ hidden_dim $= 128.$

Schrödinger Potential Estimation in Different Dimensions:

- Parameters for experiment: batch size $= 512,$ lr $= 5e-5,$ epochs $= 150,$ σ_end $= 0.8,$ hidden_dim $= 128.$

Optimal Drift Estimation Experiment:

- Parameters for experiment: batch size $= 512,$ lr $= 5e-5,$ epochs $= 120,$ σ_end $= 0.8,$ σ_max $= 1.2,$ σ_min $= 0.5,$ hidden_dim $= 128.$

Gaussian to Laplace Mixture Experiment:

- Parameters for $d = 2$: batch size $= 128,$ lr $= 5.799\text{e-}5,$ epochs $= 170,$ σ_end $= 2.0184,$ loss_scale $= 98.6226,$ hidden_dim $= 512.$
- Parameters for $d = 10$: batch size $= 128,$ lr $= 1.646\text{e-}5,$ 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 $= 3.707\text{e-}4,$ epochs $= 180,$ σ_end $= 2.6692,$ loss_scale $= 386.7248,$ hidden_dim $= 256.$
- Parameters for $d = 50$: batch size $= 256,$ lr $= 1.593\text{e-}4,$ epochs $= 180,$ σ_end $= 1.6868,$ loss_scale $= 916.4250,$ hidden_dim $= 512.$

Single Cell Experiment:

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