SCHRÖDINGER BRIDGE PROBLEM VIA EMPIRICAL RISK MINIMIZATION

Denis Belomestny
Duisburg-Essen University and HSE University
denis.belomestny@uni-due.de

Alexey Naumov, Nikita Puchkin & Denis Suchkov
HSE University
{anaumov,npuchkin,d.suchkov}@hse.ru

ABSTRACT

We study the Schrödinger bridge problem when the endpoint distributions are available only through samples. Classical computational approaches estimate Schrödinger potentials via Sinkhorn iterations on empirical measures and then construct a time-inhomogeneous drift by differentiating a kernel-smoothed dual solution. In contrast, we propose a learning-theoretic route: we rewrite the Schrödinger system in terms of a single positive *transformed* potential that satisfies a nonlinear fixed-point equation and estimate this potential by *empirical risk minimization* over a function class. We establish uniform concentration of the empirical risk around its population counterpart under sub-Gaussian assumptions on the reference kernel and terminal density. We plug the learned potential into a stochastic control representation of the bridge to generate samples. We illustrate performance of the suggested approach with numerical experiments.

1 INTRODUCTION

The Schrödinger bridge problem (SBP) provides a principled way to interpolate between two probability distributions by selecting, among all stochastic processes matching prescribed endpoint marginals, the one that is closest to a reference dynamics in relative entropy. 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 $(q_t)_{t\in(0,T]}$ with respect to the Lebesgue measure, so that $\mathsf{Q}\big(X_T \in [y, y+dy) \,\big|\, X_0 = x\big) = q_T(x,y)\,dy$. We are given two probability densities ρ_0 and ρ_T on $\mathbb{R}^d$ and consider the class

$$\mathcal{P}(\rho_0,\rho_T) \;:=\; \Big\{ P \ll \mathsf{Q} \,\Big|\, P\circ X_0^{-1} = \rho_0\,dx,\; P\circ X_T^{-1} = \rho_T\,dx \Big\}.$$

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:

$$P^\star \;\in\; \arg\min_{P\in\mathcal{P}(\rho_0,\rho_T)} \mathcal{H}(P\,\|\,\mathsf{Q}), \qquad \mathcal{H}(P\,\|\,\mathsf{Q}) := \int \log\!\left(\frac{dP}{d\mathsf{Q}}\right) dP. \tag{1}$$

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):

$$\pi^\star \;\in\; \arg\min_{\pi\in\Pi(\rho_0,\rho_T)} \mathcal{H}\big(\pi \,\|\, \pi^{\mathrm{ref}}\big), \qquad \pi^{\mathrm{ref}}(dx,dy) = \rho_0(x)\, q_T(x,y)\, dx\, dy, \tag{5}$$

where $\Pi(\rho_0,\rho_T)$ 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). In the case $q_T(x)$ is the transition density of the SDE $dX_t = \sigma dW_t$ with some $\sigma > 0$, (6) coincides with the solution of the entropy-regularized optimal transport (EOT). From a computational standpoint, EOT is often solved by Sinkhorn iterations on empirical measures, yielding discrete approximations of the Schrödinger potentials. Building on this viewpoint, recent works plug these (discrete) potentials into a stochastic control representation of the bridge, producing a drift field after additional smoothing and differentiation. While effective in moderate dimensions, this pipeline raises two conceptual challenges from a learning perspective. First, the potential is estimated only on the support of the empirical samples, and must be extended off-sample (e.g. via kernel smoothing) to yield a continuous drift. Second, the overall error blends optimization error (finite Sinkhorn iterations), statistical error (finite samples), and discretization/smoothing error, which complicates generalization analysis.

In this paper, we propose to estimate Schrödinger potentials directly in a functional form as a learning problem. We rewrite the Schrödinger system in terms of a single positive transformed potential $g^\star$ that satisfies a nonlinear fixed-point equation

$$g^\star = \mathcal{C}[g^\star], \tag{7}$$

for a nonlinear integral operator $\mathcal{C}$ depending on ρ_0, ρ_T, and q_T (see Section 2 for the details). When only samples are available,

$$X_1, \ldots, X_N \sim \rho_0, \qquad Y_1, \ldots, Y_M \sim \rho_T,$$

we form an empirical operator $\widehat{\mathcal{C}}_{N,M}$ by replacing expectations with empirical averages. Rather than enforcing the fixed point through iterative proportional fitting (Sinkhorn), we estimate $g^\star$ by minimizing an empirical residual loss over a hypothesis class $\mathcal{G}$ of positive functions (e.g. neural networks):

$$\widehat{g}_{N,M} \in \arg\min_{g\in\mathcal{G}} \frac{1}{M}\sum_{j=1}^{M} \ell\Big(g(Y_j),\, \widehat{\mathcal{C}}_{N,M}[g](Y_j)\Big), \tag{8}$$

where ℓ is minimized at equality (e.g. squared loss). The resulting objective can be optimized using stochastic-gradient methods. Crucially, in contrast to Sinkhorn-type algorithms, which implicitly produce potentials only at sampled locations and require interpolation or smoothing to obtain continuous objects, the learned estimator is continuous by construction. This is essential in downstream tasks such as: computing drift fields for Schrödinger bridges, simulating controlled diffusions and sensitivity analysis and gradient-based control. Moreover, a key benefit of the functional-learning viewpoint is that it naturally accommodates sparse representations of the Schrödinger potential g. Depending on the choice of hypothesis class, one may obtain: sparse basis expansions (e.g. wavelets, Fourier features, or kernel dictionaries), low-rank representations induced by bottleneck neural networks and implicit sparsity through regularization. Such sparsity can lead to faster evaluation of the potential and its gradients. In translation-invariant settings, sparsity can be particularly effective, as the potential often exhibits low-frequency structure or localized features that can be captured with a small number of active components.

Once a potential (or log-potential) is learned, we use the stochastic control representation of the Schrödinger bridge; see (Dai Pra, 1991). Let $h(t,x)$ denote the time-evolved Schrödinger potential given by

$$h(t,x) = \int \nu_T(y)\, q_{T-t}(x,y)\, dy.$$

Then the transition density of the process $(X_t^\star)_{t\in(0,T]}$ with the law $P^\star$ is given by

$$q^h(y,T|x,t) = \frac{q_{T-t}(x,y)h(T,y)}{h(x,t)}$$

In the case of diffusion processes $dX_t = bdt + \sigma dW_t$ this corresponds to the change of drift of basic process $(X_t)_{t\in(0,T]}$ by $a\nabla \log h$ where $a = \sigma\sigma^\top$. This yields a practical sampler: starting from $x_0 \sim \rho_0$, we simulate an SDE with the learned drift (e.g. via Euler–Maruyama) to obtain approximate bridge samples at intermediate times and at time T.

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

- We reformulate the Schrödinger system as a single nonlinear fixed-point equation for a transformed potential $g^\star$, and propose an ERM estimator based on minimizing the empirical fixed-point residual over some class of transformed potentials $\mathcal{G}$. This gives a flexible framework for the study of the empirical Schrödinger problem.
- When the reference kernel Q is Gaussian, we show that the population fixed point $g^\star$ admits a rapidly converging Hermite function expansion. We derive an explicit L^2 approximation bound for the degree-n Hermite function projector and combine it with the uniform concentration bound for the empirical risk to obtain an end-to-end risk guarantee with near-parametric dependence on sample size up to polylog factors.
- We numerically illustrate the performance of the method on (i) two-dimensional Swiss roll to S-curve example, (ii) Gaussian mixture transport under train–test shift, and (iii) single-cell population interpolation. We demonstrate performance that are comparable or improve on existing baselines.

1.1 Related work

The SBP originates in Schrödinger's 1932 work Schrödinger (1932) on the most likely stochastic evolution between two prescribed marginals under a reference dynamics. Modern treatments emphasize its connections to reciprocal processes, large deviations, and optimal transport; see, e.g., the survey of Leonard (2014) for background and further references. A stochastic control viewpoint on SBP (and related reciprocal diffusions) appears in Dai Pra (1991), and computational perspectives exploiting projective/Hilbert-metric structure were developed in, e.g., Chen et al. (2016).

Entropy-regularized optimal transport and Sinkhorn The static SBP coincides with EOT, which has become a central tool in computational OT due to its stability and algorithmic efficiency. EOT can be solved in the dual by the Sinkhorn algorithm (iterative proportional fitting / matrix scaling), popularized in ML by Cuturi (2013) and rooted in earlier matrix-scaling results such as Franklin & Lorenz (1989); see also Peyré & Cuturi (2019) for a comprehensive overview. Recent theoretical work has sharpened our understanding of Sinkhorn's contraction and convergence properties beyond classical bounded/compact settings and has provided non-asymptotic bounds for the iterates and their gradients; see, e.g., Conforti et al. (2023); Greco et al. (2023).

Estimating Schrödinger bridges from samples In the statistical setting where ρ_0 and ρ_T are only accessible through samples, a standard approach is to solve EOT between empirical measures (via Sinkhorn) to obtain discrete approximations of the Schrödinger potentials and then construct a sampler for the dynamic bridge. A representative recent instance is SinkhornBridge of Pooladian & Niles-Weed (2024), which plugs (approximate) dual solutions into a stochastic control representation to produce a time-inhomogeneous drift. Our approach differs at the estimation stage: instead of computing discrete potentials by Sinkhorn iterations on empirical measures and subsequently extending/smoothing them, we estimate a *continuous* potential by ERM over a function class (e.g. neural networks). This viewpoint is tailored to learning-theoretic analysis (uniform concentration, approximation error) and yields a potential that generalizes off-sample by construction.

SBP in generative modeling and data-to-data translation A growing body of work connects SBP to modern generative modeling via controlled diffusions and score-based methods. Examples include diffusion Schrödinger bridges and their applications to generative modeling (De Bortoli et al., 2021), learning-based SB variants such as neural Lagrangian Schrödinger bridges (Koshizuka & Sato, 2022), and Schrödinger bridge matching objectives (Shi et al., 2023). Related computationally efficient alternatives include LightSB (Korotin et al., 2024), LightSB-OU (Puchkin et al., 2026). In applications to biological time interpolation and dynamical modeling, SB/OT ideas also appear in, e.g., TrajectoryNet (Tong et al., 2020) and subsequent simulation-free or matching-based formulations (Tong et al., 2023b), as well as minibatch OT-based training objectives (Tong et al.,

2023a). Continuous normalizing-flow baselines are often trained with stabilizing regularization and architectural constraints; see, e.g., Finlay et al. (2020).

Notations Given $l, u : \mathbb{R}^d \to \mathbb{R}$ the bracket $[l, u]$ is defined as the collection of all functions $f : \mathbb{R}^d \to \mathbb{R}$ for which $l(x) \le f(x) \le u(x)$ for all $x \in \mathbb{R}^d$. For a class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ define its bracketing number as follows:

$$\mathcal{N}_{[]}(\mathcal{F}, \|\cdot\|_\infty, \delta) := \inf\Big\{ m \in \mathbb{N} : \exists (l_j, u_j)_{j=1}^m \text{ s.t. } \|u_j - l_j\|_\infty \le \delta,\ \mathcal{G} \subseteq \bigcup_{j=1}^m [l_j, u_j] \Big\}.$$

For $f : \mathbb{R}^d \to \mathbb{R}$ define its Fourier transform by $\widehat{f}(\omega) := \int_{\mathbb{R}^d} f(x) e^{-i\omega\cdot x}\, dx$. For $K \subset \mathbb{R}^d$ we use notation $\|f\|_{L^\infty(K)} := \sup_{x\in K} |f(x)|$. For a probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ we write $\|f\|_{L^p(\mu)} := \{\int_{\mathbb{R}^d} |f(x)|^p \mu(dx)\|^{1/p}$. We also define clipping by $\mathrm{clip}(f, a, b) = f$ if $a \le f \le b$, a if $f < a$ and b if $f > b$.

2 Estimator based on the empirical risk minimization

First we recast (3), (4) in terms of a single transformed potential and obtain a nonlinear fixed–point problem. Define

$$g^\star(y) \ := \ \frac{\rho_T(y)}{\nu_T(y)}, \tag{9}$$

that is, $\nu_T(y) = \rho_T(y)/g^\star(y)$. From (3) we obtain

$$\nu_0(x) \ = \ \frac{\rho_0(x)}{\displaystyle\int q_T(x,z)\, \nu_T(z)\, dz} \ = \ \frac{\rho_0(x)}{\displaystyle\int q_T(x,z)\, \frac{\rho_T(z)}{g^\star(z)}\, dz}. \tag{10}$$

For each $x \in \mathbb{R}^d$ and $g : \mathbb{R}^d \to (0, \infty)$ define

$$D_g(x) := \int_{\mathbb{R}^d} q_T(x,z)\, \frac{\rho_T(z)}{g(z)}\, dz. \tag{11}$$

Substituting (10) and (9) into (4) gives

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

Cancelling the common factor $\rho_T(y)$ on both sides and multiplying by $g^\star(y)$, we obtain the nonlinear fixed–point equation

$$g^\star(y) = \mathcal{C}[g^\star](y) \ \text{ with } \mathcal{C}[g](y) := \int_{\mathbb{R}^d} \frac{q_T(x,y)\rho_0(x)}{D_g(x)}\, dx\,. \tag{12}$$

So the (static) Schrödinger bridge problem can be equivalently formulated as the problem of finding a positive function $g^\star$ solving the nonlinear fixed–point equation $g^\star = \mathcal{C}[g^\star]$. Once a fixed point $g^\star$ is found, the Schrödinger potentials are recovered via $\nu_T(y) = \rho_T(y)/g^\star(y)$ and (10), and the optimal Markov process $P^\star$ is obtained by tilting the reference process Q according to (2).

In many applications the marginal densities ρ_0 and ρ_T are not available explicitly. Instead, we observe independent samples

$$X_1, \ldots, X_N \sim \rho_0, \qquad Y_1, \ldots, Y_M \sim \rho_T,$$

and seek to recover the fixed point $g^\star$ solving $g^\star = \mathcal{C}[g^\star]$. This gives rise to a statistical version of the Schrödinger bridge problem. Define the empirical measures

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

Replacing ρ_T by its empirical counterparts in (11) we obtain

$$\widehat{D}(x) := \frac{1}{M}\sum_{k=1}^{M} q_T(x, Y_k)\,\frac{1}{g(Y_k)}\,.$$

In Lemma 4 we show that $\overline{D} \geq D_g(x) \geq \underline{D}$ for all $x \in \mathrm{supp}(\rho_0)$. But we can't prove the same lower bound for $\widehat{D}(x)$. We introduce a clipping operator $\mathcal{T}_{[a,b]}$ of the form $\mathcal{T}_{[a,b]}[f](x) = \mathrm{clip}(f, a, b)$. Replacing ρ_0 and ρ_T by their empirical counterparts in (12) yields the empirical nonlinear operator

$$\widehat{\mathcal{C}}_{N,M}[g](y) := \frac{1}{N}\sum_{i=1}^{N} \frac{q_T(X_i, y)}{\mathcal{T}_{[\underline{D},\overline{D}]}[\widehat{D}](X_i)}. \tag{13}$$

The fixed–point condition $g^\star = \mathcal{C}[g^\star]$ is approximated by enforcing

$$g(Y_j) \;\approx\; \widehat{\mathcal{C}}_{N,M}[g](Y_j), \qquad j = 1, \ldots, M.$$

To estimate g, we recast this system as an ERM problem. Let $\ell : (0,\infty)\times(0,\infty) \to [0,\infty)$ be a loss function minimized at equality, e.g. $\ell(u,v) = \frac{1}{2}(u-v)^2$. The empirical risk is defined by

$$\widehat{\mathcal{R}}_{N,M}(g) := \frac{1}{M}\sum_{j=1}^{M} \ell\Big(g(Y_j),\, \widehat{\mathcal{C}}_{N,M}[g](Y_j)\Big)\,. \tag{14}$$

The statistical Schrödinger bridge estimator is then given by

$$\widehat{g}_{N,M} \;\in\; \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}_{N,M}(g), \tag{15}$$

where $\mathcal{G}$ is an admissible class of positive functions or parametric models (e.g. neural networks). In the next section, we study the behavior of $\widehat{\mathcal{R}}_{N,M}(g)$ under a set of realistic assumptions that allow us to apply tools from empirical process theory.

3 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) There exist constants $c_-, c_+ > 0$ and $a_-, a_+ > 0$ such that

$$c_- \exp(-a_-\|x-y\|^2) \;\leq\; q_T(x,y) \;\leq\; c_+ \exp(-a_+\|x-y\|^2), \qquad \forall x, y \in \mathbb{R}^d.$$

Moreover, q_T is globally Lipschitz with constant L_q.

(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

$$c_T^- \exp(-b_T^-\|y\|^2) \;\leq\; \rho_T(y) \;\leq\; c_T^+ \exp(-b_T^+\|y\|^2), \qquad \forall y \in \mathbb{R}^d.$$

The nondegeneracy condition in *(R0)* prevents the normalization factors from becoming arbitrarily small on $\mathrm{supp}(\rho_0)$ and is used to obtain uniform bounds for $D_g(x)$ and two-sided bound for the fixed point $g^\star$. Note that we don't assume that ρ_T is compactly supported. Instead, we assume two-sided sub-Gaussian behavior. We now impose additional assumptions on the loss.

(L) The loss $\ell : (0,\infty)\times(0,\infty) \to \mathbb{R}$ is locally bounded and jointly Lipschitz, i.e. for any $K > 0$ there exist constants $B_\ell = B_\ell(K), L_\ell = L_\ell(K) > 0$ such that for all $(u,v), (u',v') \in (0,K)^2$,

$$|\ell(u,v)| \leq B_\ell, \qquad |\ell(u,v) - \ell(u',v')| \leq L_\ell\big(|u-u'| + |v-v'|\big).$$

Finally, we impose assumptions on the hypothesis class.

(G) Assume that the class $\mathcal{G}$ satisfies the following assumptions: there exist constants $c_{\mathcal{G}}^-, c_{\mathcal{G}}^+ > 0$ and $a_{\mathcal{G}} > 0$ such that for all $g \in \mathcal{G}$,

$$c_{\mathcal{G}}^- e^{-a_{\mathcal{G}}\|y\|^2} \le g(y) \le c_{\mathcal{G}}^+, \qquad \forall y \in \mathbb{R}^d\,. \tag{16}$$

Note that *(G)* enforces that all candidates $g \in \mathcal{G}$ are uniformly bounded above and, critically, bounded below by a Gaussian. The lower bound prevents instabilities caused by the ratio $q_T(x,z)/g(z)$ in the normalizer D_g and yields a manageable envelope for empirical-process arguments. Below we additionally assume that $b_T^+ > 4a_G$ which ensures that these envelopes are square-integrable under ρ_T.

Theorem 1. *Suppose that the assumptions* (Q)*,* (R0)*,* (RT)*,* (L)*,* (G) *hold with* $b_T^+ > 4a_{\mathcal{G}}$ *and* $K = c_{\mathcal{G}}^+(1 + (c_+/\underline{D})$*. Then we have for all* $N, M \ge 1$*,*

$$\mathbb{E}[\mathcal{R}(\widehat{g}_{N,M})] \le \inf_{g\in\mathcal{G}} \mathcal{R}(g) + 2\mathbb{E}\Big[\sup_{g\in\mathcal{G}} \big|\widehat{\mathcal{R}}_{N,M}(g) - \mathcal{R}(g)\big|\Big] \tag{17}$$

and

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}} \big|\widehat{\mathcal{R}}_{N,M}(g) - \mathcal{R}(g)\big|\Big] \le C_1 \left(\sqrt{\frac{\log\log(N)}{N}} + \frac{1}{\sqrt{M}}\right) \int_0^{C_2} \sqrt{\log \mathcal{N}_{[]}\Big(\mathcal{G}, \|\cdot\|_\infty, \frac{\varepsilon}{C_3}\Big)}\, d\varepsilon\,,$$

where C_1, C_2 *and* C_3 *are positive constants depending on constants from the assumptions* (Q)*,* (R0)*,* (RT)*,* (L) *and* (G)*.*

Proof. The detailed proof is contained in Section B.1. □

3.1 Approximation error

In this section, we show that by choosing an appropriate function class $\mathcal{G}$, we can achieve a small value of $\mathbb{E}[\mathcal{R}(\widehat{g}_{N,M})]$. For simplicity, we assume that the transition kernel is Gaussian. Recall (12) that the function $g^\star$ has the following form

$$g^\star(y) = \int_{\mathbb{R}^d} w(x)\, q(y-x)\, dx$$

with $w(x) = \rho_0(x)/D_{g^\star}(x)$, that is, it represents a convolution of compactly supported function with the Gaussian kernel. In this case, the class of Hermite polynomials is a natural candidate for the class of functions G. We have brought to the appendix the main results concerning Hermite polynomials, in particular estimates of Hermite coefficients of $g^\star$. We also note that $g^\star$ is not uniquely determined since the potentials ν_0 and ν_T are determined up to a multiplicative constant. We assume that

$$\int \frac{1}{g^\star(y)} \rho_T(y) dy = \int \nu_T(y) dy = 1. \tag{18}$$

More precisely, assume the following condition.

($\mathrm{Q_{Gauss}}$) The reference kernel is Gaussian:

$$q_T(z) = (2\pi T)^{-d/2} \exp\left(-\frac{\|z\|^2}{2T}\right), \qquad z \in \mathbb{R}^d.$$

The following theorem provides a bound for $\mathbb{E}[\mathcal{R}(\widehat{g}_{N,M})]$.

Theorem 2. *Assume conditions* ($\mathrm{Q_{Gauss}}$)*,* (R0)*,* (RT) *with* $b_T^+ > 4/T$*,* (L) *and* (18)*. Then there exists a class* $\mathcal{G}$ *such that* (G) *is satisfied and*

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

where $\lesssim$ *stands for inequality up to constants independent of* M *and* d *and double logarithmic factors.*

Proof. The detailed proof is contained in Section C.3. □

4 NUMERICS

In this section, we present an experimental analysis of the proposed algorithm and its comparison with SinkhornBridge (Pooladian & Niles-Weed, 2024). This algorithm was chosen for comparison because it is closest to ours in terms of its training procedure. However, because our algorithm learns a continuous Schrödinger log-potential function rather than a discrete solution to the optimal transport problem, we were able to demonstrate significant superiority across several generative modelling and data-to-data translation tasks. In the following, the algorithm we proposed will be referred to as ERM-Bridge for brevity. The formal description of the algorithm, the hyperparameter values, and additional experimental details are provided in Appendix D.6.

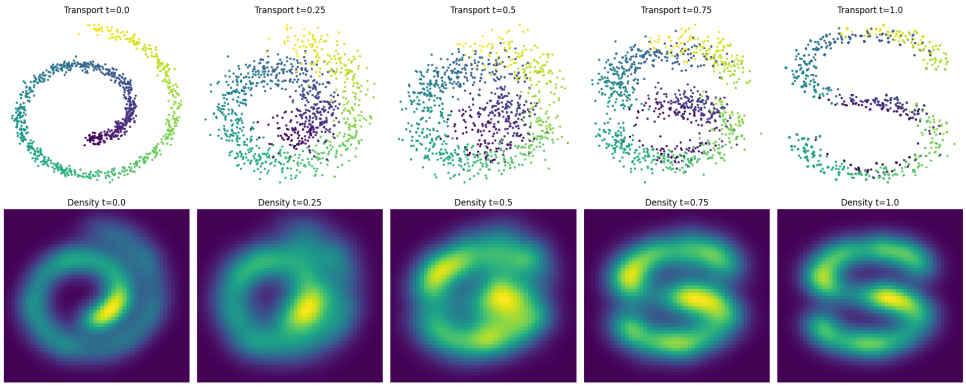

Figure 1: Sample translation from Swiss-Roll to S-Curve and density map for ERM-Bridge at time $t \in [0, 0.25, 0.5, 0.75, 1]$.

Swiss-Roll to S-Curve Experiment We evaluate the validity of our algorithm on a classic two-dimensional example. In this experiment, we examined the translation of the Swiss-Roll distribution into an S-Curve and present the distribution at time $t \in [0, 0.25, 0.5, 0.75, 1]$ for clarity, see Figure 1. The figure clearly shows that the algorithm reliably learns the distribution and the correct density map (the density was approximated using KDE on a two-dimensional surface). It is also worth noting that our algorithm demonstrated the best training and sampling time on this problem. The results and hyperparameters are reported in Appendix D.6.

Figure 2: Plot of sliced Wasserstein distance as a function of KL between the distribution on the train and sampling for ERM-Bridge and SinkhornBridge.

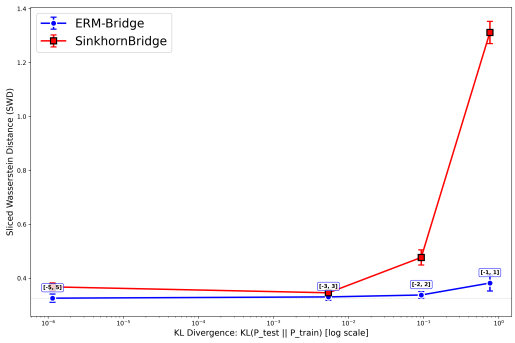

Table 1: The quality of intermediate distribution restoration for single-cell data for various algorithms, including ERM-Bridge and SinkhornBridge.

Solver	$\mathbb{W}_1$ metric
OT-CFM (Tong et al., 2023a)	0.790 ± 0.068
$[SF]^2$M-Exact (Tong et al., 2023b)	0.793 ± 0.066
ERM-Bridge	0.809 ± 0.030
LightSB-OU (Puchkin et al., 2026)	0.815 ± 0.016
SinkhornBridge (Pooladian & Niles-Weed, 2024)	0.818 ± 0.028
LightSB (Korotin et al., 2024)	0.823 ± 0.017
Reg. CNF (Finlay et al., 2020)	$0.825 \pm$ N/A[a]
T. Net (Tong et al., 2020)	$0.848 \pm$ N/A[a]
DSB (De Bortoli et al., 2021)	0.862 ± 0.023
I-CFM (Tong et al., 2023a)	0.872 ± 0.087
$[SF]^2$M-Geo (Tong et al., 2023b)	0.879 ± 0.148
NLSB (Koshizuka & Sato, 2022)	$0.970 \pm$ N/A[a]
$[SF]^2$M-Sink (Tong et al., 2023b)	1.198 ± 0.342
SB-CFM (Tong et al., 2023a)	1.221 ± 0.380
DSBM (Shi et al., 2023)	1.775 ± 0.429

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

Evaluation on a Gaussian Mixture To further illustrate the importance of the network approximated log-potential, we evaluated our algorithm on a Gaussian mixture problem. For this experiment, we used mixtures of 25 Gaussians from a uniform grid with standard, identical covariance

matrices. Transport quality was quantified with the sliced approximation of the Wasserstein distance $\mathbb{W}_1$ as the metric, see Appendix D.2 for the definition of metrics.

In this experiment, both algorithms were trained on 3000 samples from a truncated normal distribution on $[-10, 10]$, and sampled on starting points from a truncated normal distribution on $[-1, 1], [-2, 2], [-5, 5]$. This experiment demonstrates how the algorithms considered behave when the starting distributions in the training set and the sampling set may differ, see Figure 2. This situation is quite typical for data-to-data transport, where both the starting and target distributions are available only from samples, and it is impossible to ensure complete identity between the training and test sets. It is clearly seen that the continuous potential-like function learned by the neural network in our algorithm behaves better than the discrete optimal transport SinkhornBridge. By performing an equal amount of time and grid point iteration of hyperparameter search for both our algorithm and SinkhornBridge, the baseline achieved $\mathbb{W}_1 = 1.3115$ for $[-1, 1]$ case, while our algorithm achieved $\mathbb{W}_1 = 0.3818$. The hyperparameter values are reported in Appendix D.6.

Evaluation on the 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 in (Tong et al., 2020). We then predicted the cell distribution at the intermediate time t_i and computed the Wasserstein distance $\mathbb{W}_1$ between the predicted distribution and the ground truth. The results were averaged across all three setups ($i = 1, 2, 3$), see Table 1.

To ensure statistical robustness, we repeated the experiment five times. By performing an equal amount of time and grid point iteration of hyperparameters for both our algorithm and SinkhornBridge, the baseline achieved $\mathbb{W}_1 = 0.818$, while our algorithm achieved $\mathbb{W}_1 = 0.809$. The hyperparameter values are reported in Appendix D.6.

5 Conclusion

We studied the *statistical* problem of estimating Schrödinger bridge potentials from finite samples. We rewrote the Schrödinger system as a single nonlinear fixed-point equation $g = C[g]$ for a transformed potential, and proposed an ERM estimator obtained by minimizing an empirical fixed-point residual over a hypothesis class. This yields a continuous potential estimator by construction, and naturally pairs with the stochastic-control representation of the Schrödinger bridge. Several directions remain for future work, including extensions beyond sub-Gaussian tail assumptions and bounds on the error between $g_{M,N}$ and $g^\star$. The latter will require to study local behavior of derivative of $\mathcal{C}[g]$ near $g^\star$ in the Hilbert metric (spectral gap condition).

Acknowledgments

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University №139-15-2025-009.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. 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.

Giovanni Conforti, Alain Durmus, and Giacomo Greco. 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.

Marco Cuturi. 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.

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

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709, 2021.

Chris Finlay, Joern-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. 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.

Joel Franklin and Jens Lorenz. 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.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2016. ISBN 9781107043169. doi: 10.1017/CBO9781107337312.

Giacomo Greco, Maxence Noble, Giovanni Conforti, and Alain Durmus. Non-asymptotic convergence bounds for sinkhorn iterates and their gradients: A coupling approach. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 716–746. PMLR, 2023. URL https://proceedings.mlr.press/v195/greco23a.html.

Alexander Korotin, Nikita Gushchin, and Evgeny Burnaev. 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.

Takeshi Koshizuka and Issei Sato. 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.

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

Gabriel Peyré and Marco Cuturi. 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.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. 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.

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

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

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. 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.

Alexander Tong, Jessie Huang, Guy Wolf, David van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9526–9536. PMLR, 2020. URL https://proceedings.mlr.press/v119/tong20a.html.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. 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.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. 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.

A Properties of the empirical operator under sub-Gaussian assumptions

Proposition 3 (Two–sided Gaussian bounds for the fixed point)**.** *Assume* (Q)*,* (R0)*,* (RT)*. Let* $g^\star : \mathbb{R}^d \to (0,\infty)$ *be measurable and satisfy*

$$g^\star(y) = \mathcal{C}[g^\star](y), \qquad y \in \mathbb{R}^d, \qquad \text{and} \qquad \int_{\mathbb{R}^d} \frac{1}{g^\star(z)}\, \rho_T(z)\, dz = 1. \tag{19}$$

Set $a_-^\star := 2a_-$, $a_+^\star := \frac{a_+}{2}$*. 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*

$$c_-^\star \exp(-a_-^\star \|y\|^2) \;\le\; g^\star(y) \;\le\; c_+^\star \exp(-a_+^\star \|y\|^2), \qquad \forall y \in \mathbb{R}^d. \tag{20}$$

Remark 1. *It is important to note that the behavior of the function* $g^\star$ *is determined only by the conditions on the transition kernel* q *and densities* ρ_0, ρ_T*.*

Proof. For each $x \in \mathbb{R}^d$ define

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

Then the fixed–point equation can be written as

$$g^\star(y) = \int_{\mathbb{R}^d} \frac{\rho_0(x)}{D_{g^\star}(x)}\, q_T(x,y)\, dx.$$

We claim that there exist constants $0 < \underline{D}^\star \le \overline{D}^\star < \infty$, depending only on q_T, ρ_0, ρ_T, such that

$$\underline{D}^\star \;\le\; D_{g^\star}(x) \;\le\; \overline{D}^\star, \qquad \forall x \in \operatorname{supp}(\rho_0). \tag{21}$$

By *(Q)*,

$$q_T(x,z) \le c_+ \exp(-a_+\|x-z\|^2) \le c_+,$$

hence, using the normalization in (19),

$$D_{g^\star}(x) = \int q_T(x,z)\, \frac{\rho_T(z)}{g^\star(z)}\, dz \le c_+ \int \frac{\rho_T(z)}{g^\star(z)}\, dz = c_+. \tag{22}$$

Thus we may take

$$\overline{D}^\star := c_+. \tag{23}$$

Define

$$m(x) := \int_{\mathbb{R}^d} q_T(x,z)\, \rho_T(z)\, dz.$$

Using *(Q)*, *(RT)* and the compactness of $B(x_0, R_0)$, one checks that

$$m_- \;\le\; m(x) \;\le\; m_+, \qquad \forall x \in \operatorname{supp}(\rho_0)$$

where $m_+ := c_+$, and

$$m_- := c_- c_T^- \exp\big(-2a_- R_\star^2\big)\Big(\frac{\pi}{2a_- + b_T^-}\Big)^{d/2}$$

with $R_\star := \|x_0\| + R_0$. Where we used that fact for any $\gamma > 0$,

$$\int_{\mathbb{R}^d} \exp(-\gamma\|z\|^2)\, dz = \left(\frac{\pi}{\gamma}\right)^{d/2} < \infty, \tag{24}$$

and

$$\|x - z\|^2 \le 2\|x\|^2 + 2\|z\|^2\,. \tag{25}$$

For each such x, define a probability measure ν_x by

$$\nu_x(dz) := \frac{q_T(x,z)\,\rho_T(z)}{m(x)}\, dz.$$

Then

$$D_{g^\star}(x) = m(x)\,\mathbb{E}_{\nu_x}\Big[\frac{1}{g^\star(Z)}\Big].$$

We now compare ν_x and ρ_T. The Radon–Nikodym derivative is

$$\frac{d\nu_x}{d\rho_T}(z) = \frac{q_T(x,z)}{m(x)}.$$

By the two–sided Gaussian bounds in *(Q)* and the bounds on m, there exist constants $0 < k_- \le k_+ < \infty$ such that

$$k_- \;\le\; \frac{d\nu_x}{d\rho_T}(z) \;\le\; k_+, \qquad \forall x \in \mathrm{supp}(\rho_0),\ z \in \mathbb{R}^d.$$

For any nonnegative measurable h we therefore have

$$k_- \int h(z)\,\rho_T(z)\,dz \;\le\; \int h(z)\,\frac{d\nu_x}{d\rho_T}(z)\,\rho_T(z)\,dz \;\le\; k_+ \int h(z)\,\rho_T(z)\,dz,$$

that is,

$$k_-\,\mathbb{E}_{\rho_T}[h(Z)] \;\le\; \mathbb{E}_{\nu_x}[h(Z)] \;\le\; k_+\,\mathbb{E}_{\rho_T}[h(Z)].$$

Apply this with $h(z) = \frac{1}{g^\star(z)} \ge 0$. Using the normalization in (19), we obtain

$$k_- \le \mathbb{E}_{\nu_x}\Big[\frac{1}{g^\star(Z)}\Big] \le k_+.$$

Consequently,

$$D_{g^\star}(x) = m(x)\,\mathbb{E}_{\nu_x}\Big[\frac{1}{g^\star(Z)}\Big] \ge m_-\,k_- =: \underline{D}^\star > 0 \tag{26}$$

for all $x \in \mathrm{supp}(\rho_0)$. Together with the upper bound, this proves (21). Using the fixed–point equation and (21),

$$g^\star(y) = \int_{\mathrm{supp}(\rho_0)} \frac{\rho_0(x)}{D_{g^\star}(x)}\, q_T(x,y)\,dx \;\le\; \frac{1}{\underline{D}} \int_{B(x_0,R_0)} \rho_0(x)\, q_T(x,y)\,dx.$$

By the upper Gaussian bound in *(Q)*,

$$q_T(x,y) \le c_+ \exp\big(-a_+\|x-y\|^2\big).$$

Using

$$\|x-y\|^2 \ge \frac{1}{2}\|y\|^2 - \|x\|^2 \ge \frac{1}{2}\|y\|^2 - R_\star^2,$$

we obtain

$$q_T(x,y) \le c_+ \exp(a_+ R_\star^2)\, \exp\Big(-\frac{a_+}{2}\|y\|^2\Big), \qquad x \in B(x_0,R_0).$$

Therefore

$$g^\star(y) \le \frac{c_+ \exp(a_+ R_\star^2)}{\underline{D}^\star} \int_{\mathrm{supp}(\rho_0)} \rho_0(x)\,dx\, \exp\Big(-\frac{a_+}{2}\|y\|^2\Big).$$

Since ρ_0 is a probability density, $\int_{\mathrm{supp}(\rho_0)} \rho_0(x)\,dx = 1$, and we may set

$$c_+^\star := \frac{c_+ \exp(a_+ R_\star^2)}{\underline{D}^\star}, \qquad a_+^\star := \frac{a_+}{2},$$

to obtain

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

Using again the fixed–point equation and (21),

$$g^\star(y) = \int_{\mathrm{supp}(\rho_0)} \frac{\rho_0(x)}{D_{g^\star}(x)}\, q_T(x,y)\,dx \;\ge\; \frac{1}{\overline{D}^\star} \int_{\mathrm{supp}(\rho_0)} \rho_0(x)\, q_T(x,y)\,dx.$$

By *(R0)*,

$$g^\star(y) \ge \frac{\rho_{0,-}}{\overline{D}^\star} \int_{B(x_0,r_0)} q_T(x,y)\,dx.$$

Using the lower Gaussian bound in *(Q)*,

$$q_T(x,y) \ge c_- \exp\big(-a_- \|x-y\|^2\big),$$

and $\|x\| \le \|x_0\| + r_0$ for $x \in B(x_0, r_0)$, we have

$$\|x-y\| \le \|x\| + \|y\| \le \|x_0\| + r_0 + \|y\|$$

and therefore

$$\|x-y\|^2 \le (\|x_0\| + r_0 + \|y\|)^2 \le 2r_\star^2 + 2\|y\|^2,$$

where $r_\star = \|x_0\| + r_0$. Thus,

$$q_T(x,y) \ge c_- \exp\big(-a_-\,(2r_\star^2 + 2\|y\|^2)\big), \qquad x \in B(x_0,r_0).$$

It follows that

$$\int_{B(x_0,r_0)} q_T(x,y)\,dx \ge c_- \exp(-2a_- r_\star^2)\, \exp(-2a_-\|y\|^2)\, \lambda^d(B(x_0,r_0)),$$

where $\lambda^d(B(x_0,r_0))$ is the Lebesgue measure of $B(x_0,r_0)$. Therefore

$$g^\star(y) \ge \frac{\rho_{0,-}\, c_-\, \lambda^d(B(x_0,r_0))\, \exp(-2a_- r_\star^2)}{\overline{D}^\star} \exp(-2a_-\|y\|^2).$$

Setting

$$c_-^\star := \frac{\rho_{0,-}\, c_-\, \lambda^d(B(x_0,r_0))\, \exp(-2a_- r_\star^2)}{\overline{D}^\star}, \qquad a_-^\star := 2a_-,$$

we obtain

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

Combining the upper and lower bounds completes the proof of (20). □

B UNIFORM CONCENTRATION OF THE EMPIRICAL RISK

Lemma 4. *Suppose that the assumptions* (Q), (R0), (RT) *and* (G) *hold. Assume additionally that* $b_T^+ > 2a_{\mathcal{G}}$. *Then there exist constants* $\underline{D}, \overline{D}, L_{D,\infty}, L_{\mathcal{C},\infty} > 0$ *such that*

$$\underline{D} \;\le\; D_g(x) \;\le\; \overline{D}, \qquad x \in B(x_0,R_0),\ g \in \mathcal{G}; \tag{27}$$

$$\|D_g - D_h\|_{L^\infty(B(x_0,R_0))} \le L_{D,\infty} \|g-h\|_\infty; \tag{28}$$

$$\|\mathcal{C}[g] - \mathcal{C}[h]\|_\infty \le L_{\mathcal{C},\infty}\, \|g-h\|_{L^\infty(\mathbb{R}^d)}; \tag{29}$$

$$\|\mathcal{C}[g]\|_\infty \le (c_+/\underline{D})\, \|g\|_{L^\infty(\mathbb{R}^d)}\,. \tag{30}$$

Note that $\underline{D}, \overline{D}$ *do not depend on the class* $\mathcal{G}$.

Proof. The proof of (27) repeats the proof of Proposition 3. We make necessary changes. First, we replace (22) by

$$D_g(x) = \int q_T(x,z)\,\frac{\rho_T(z)}{g(z)}\,dz \le c_+ c_T^+ (c_{\mathcal{G}}^-)^{-1} \int \exp\Big(-\Big(b_T^+ - a_{\mathcal{G}}\Big)\|z\|^2\Big)\,dz$$
$$= c_+ c_T^+ (c_{\mathcal{G}}^-)^{-1} \left(\frac{\pi}{b_T^+ - a_{\mathcal{G}}}\right)^{d/2} =: \overline{D}\,.$$

Recall the proof of (26). Note that

$$\mathbb{E}_{\rho_T}[1/g] \ge (c_{\mathcal{G}}^+)^{-1}\,,$$

and we can take $\underline{D} := m_-\, k_-(c_{\mathcal{G}}^+)^{-1}$. We prove (28). By definition,

$$D_g(x) - D_h(x) = \int_{\mathbb{R}^d} q_T(x,z)\,\rho_T(z)\,\Big(\frac{1}{g(z)} - \frac{1}{h(z)}\Big)\,dz.$$

Using

$$\left|\frac{1}{g(z)} - \frac{1}{h(z)}\right| = \frac{|g(z)-h(z)|}{g(z)h(z)} \le \|g-h\|_\infty \cdot \frac{1}{g(z)h(z)}$$

and the lower bound in *(G)* (applied to both g and h), we obtain for all z,

$$\frac{1}{g(z)h(z)} \le (c_{\mathcal{G}}^-)^{-2}\exp(2a_{\mathcal{G}}\|z\|^2).$$

Hence

$$|D_g(x) - D_h(x)| \le \|g-h\|_\infty\,(c_{\mathcal{G}}^-)^{-2} \int_{\mathbb{R}^d} q_T(x,z)\,\rho_T(z)\,\exp(2a_{\mathcal{G}}\|z\|^2)dz. \tag{31}$$

We now bound the integral uniformly in $x \in B(x_0,R_0)$. By *(Q)* and the inequality

$$\|x-z\|^2 \ge \frac{1}{2}\|z\|^2 - \|x\|^2,$$

we have, for all $x \in B(x_0,R_0)$,

$$q_T(x,z) \le c_+\exp(-a_+\|x-z\|^2) \le c_+\exp(a_+R_\star^2)\exp\Big(-\frac{a_+}{2}\|z\|^2\Big)$$

with $R_\star := \|x_0\| + R_0$. Combining this with *(RT)* yields

$$q_T(x,z)\,\rho_T(z)\,\exp(2a_{\mathcal{G}}\|z\|^2) \le c_+c_T^+\exp(a_+R_\star^2)\,\exp\Big(-\Big(b_T^+ - 2a_{\mathcal{G}} + \frac{a_+}{2}\Big)\|z\|^2\Big)\,.$$

By (24), uniformly for $x \in B(x_0,R_0)$,

$$|D_g(x) - D_h(x)| \le L_{D,\infty}\|g-h\|_\infty\,,$$

where

$$L_{D,\infty} := (c_{\mathcal{G}}^-)^{-2}c_+c_T^+\exp(a_+R_\star^2)\left(\frac{\pi}{b_T^+ - 2a_{\mathcal{G}} + a_+/2}\right)^{d/2}. \tag{32}$$

This proves (28). For each $y \in \mathbb{R}^d$,

$$\mathcal{C}[g](y) - \mathcal{C}[h](y) = \int_{\mathrm{supp}(\rho_0)} \rho_0(x)\,q_T(x,y)\,\left(\frac{1}{D_g(x)} - \frac{1}{D_h(x)}\right)\,dx.$$

By (27) and (28) we get

$$|\mathcal{C}[g](y) - \mathcal{C}[h](y)| \le \frac{L_{D,\infty}\|g-h\|_\infty}{\underline{D}^2}\int_{\mathrm{supp}(\rho_0)}\rho_0(x)\,q_T(x,y)\,dx$$
$$\le L_{\mathcal{C},\infty}\|g-h\|_\infty,$$

where

$$L_{\mathcal{C},\infty} = L_{D,\infty}c_+\,. \tag{33}$$

This proves (29).

□

B.1 Proof of Theorem 1

Define an intermediate risk that uses the empirical average in Y_j but the population operator $\mathcal{C}[g]$:

$$\widetilde{\mathcal{R}}_M(g) := \frac{1}{M}\sum_{j=1}^{M} \ell\big(g(Y_j), \mathcal{C}[g](Y_j)\big).$$

Then

$$\widehat{\mathcal{R}}_{N,M}(g) - \mathcal{R}(g) = \underbrace{\big(\widehat{\mathcal{R}}_{N,M}(g) - \widetilde{\mathcal{R}}_M(g)\big)}_{\text{(I)}} + \underbrace{\big(\widetilde{\mathcal{R}}_M(g) - \mathcal{R}(g)\big)}_{\text{(II)}}.$$

Taking supremum over $g \in \mathcal{G}$ and expectations yields

$$\mathbb{E}\sup_{g\in\mathcal{G}}\big|\widehat{\mathcal{R}}_{N,M}(g) - \mathcal{R}(g)\big| \;\le\; T_1 + T_2,$$

where

$$T_1 := \mathbb{E}\sup_{g\in\mathcal{G}}\big|\widehat{\mathcal{R}}_{N,M}(g) - \widetilde{\mathcal{R}}_M(g)\big|, \qquad T_2 := \mathbb{E}\sup_{g\in\mathcal{G}}\big|\widetilde{\mathcal{R}}_M(g) - \mathcal{R}(g)\big|.$$

For fixed g, $\widetilde{\mathcal{R}}_M(g)$ is the empirical average of the i.i.d. variables

$$f_g(Y_j) := \ell\big(g(Y_j), \mathcal{C}[g](Y_j)\big), \qquad j = 1, \dots, M.$$

Let $\mathcal{F} := \big\{f_g : g \in \mathcal{G}\big\}$. By Lemma 4,

$$|\mathcal{C}[g](y) - \mathcal{C}[g'](y)| \le L_{\mathcal{C},\infty}\|g - g'\|_{L^\infty(\mathbb{R}^d)}, \quad |\mathcal{C}[g](y)| \le (c_+/\underline{D})\|g\|_{L^\infty(\mathbb{R}^d)}.$$

By *(L)* and boundedness of any $g \in \mathcal{G}$,

$$|f_g(y) - f_{g'}(y)| \le L_\ell\left(|g(y) - g'(y)| + |\mathcal{C}[g](y) - \mathcal{C}[g'](y)|\right)$$

with $L_\ell = L_\ell\big(c_{\mathcal{G}}^+(1 + (c_+/\underline{D})\big)$. Combining with *(G)* and *(RT)*, we obtain that $\mathcal{F}$ has envelope F with $\|F\|_\infty \le B_\ell = B_\ell((1 + (c_+/\underline{D}))\, c_{\mathcal{G}}^+))$. Theorem 3.5.13 in Giné & Nickl (2016) implies

$$\begin{aligned} T_2 &\lesssim \frac{1}{\sqrt{M}}\int_0^{8\|F\|_\infty} \sqrt{\log(2N_{[]}(\mathcal{F}, L^2(\rho_T), \varepsilon))}\, d\varepsilon \\ &\lesssim \frac{1}{\sqrt{M}}\int_0^{8B_\ell} \sqrt{\log(2N_{[]}(\mathcal{G}, L^2(\rho_T), \varepsilon/L_\ell))}\, d\varepsilon. \end{aligned} \tag{34}$$

We now bound

$$T_1 = \mathbb{E}\sup_{g\in\mathcal{G}}\big|\widehat{\mathcal{R}}_{N,M}(g) - \widetilde{\mathcal{R}}_M(g)\big|.$$

By the Lipschitz property *(L)*, for each g,

$$\begin{aligned} \big|\widehat{\mathcal{R}}_{N,M}(g) - \widetilde{\mathcal{R}}_M(g)\big| &= \left|\frac{1}{M}\sum_{j=1}^{M}\Big(\ell\big(g(Y_j), \widehat{\mathcal{C}}_{N,M}[g](Y_j)\big) - \ell\big(g(Y_j), \mathcal{C}[g](Y_j)\big)\Big)\right| \\ &\le \frac{L_\ell}{M}\sum_{j=1}^{M}\big|\widehat{\mathcal{C}}_{N,M}[g](Y_j) - \mathcal{C}[g](Y_j)\big|. \end{aligned}$$

Taking supremum over $g \in \mathcal{G}$ and expectations,

$$T_1 \le L_\ell\,\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{M}\sum_{j=1}^{M}\big|\widehat{\mathcal{C}}_{N,M}[g](Y_j) - \mathcal{C}[g](Y_j)\big|\right].$$

Fix a large compact set $K \subset \mathbb{R}^d$ (to be defined later) and consider

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sup_{y\in K}\big|\widehat{\mathcal{C}}_{N,M}[g](y) - \mathcal{C}[g](y)\big|\right].$$

By Lemma 4,

$$0 < \underline{D} \;\le\; D_g(x) \;\le\; \overline{D} < \infty, \quad x \in B(x_0, R_0), \quad g \in \mathcal{G}. \tag{35}$$

Define

$$\widetilde{\mathcal{C}}_N[g](y) := \frac{1}{N} \sum_{i=1}^{N} \frac{q_T(X_i, y)}{D_g(X_i)}.$$

Then, for each (g, y),

$$\widehat{\mathcal{C}}_{N,M}[g](y) - \mathcal{C}[g](y) = \underbrace{\left(\widehat{\mathcal{C}}_{N,M}[g](y) - \widetilde{\mathcal{C}}_N[g](y)\right)}_{\text{(A)}} + \underbrace{\left(\widetilde{\mathcal{C}}_N[g](y) - \mathcal{C}[g](y)\right)}_{\text{(B)}}.$$

Thus,

$$\sup_{g\in\mathcal{G}} \sup_{y\in\mathbb{R}^d} \left|\widehat{\mathcal{C}}_{N,M}[g](y) - \mathcal{C}[g](y)\right| \le \sup_{g\in\mathcal{G}} \sup_{y\in K} |(\mathrm{A})| + \sup_{g\in\mathcal{G}} \sup_{y\in K} |(\mathrm{B})|.$$

For fixed (g, y) define

$$\phi_{g,y}(x) := \frac{q_T(x, y)}{D_g(x)}.$$

Then

$$\widetilde{\mathcal{C}}_N[g](y) = \frac{1}{N} \sum_{i=1}^{N} \phi_{g,y}(X_i), \qquad \mathcal{C}[g](y) = \mathbb{E}_{X\sim\rho_0}[\phi_{g,y}(X)].$$

By (35) and the upper Gaussian bound on q_T in *(Q)*, together with compactness of $B(x_0, R_0)$, there exists $F_\infty < \infty$ (take $F_\infty = c_+/\underline{D}$) such that

$$|\phi_{g,y}(x)| \le F_\infty, \quad x \in B(x_0, R_0),\ y \in \mathbb{R}^d,\ g \in \mathcal{G}.$$

Fix $g, h \in \mathcal{G}$ and $x \in B(x_0, R_0)$. By Lemma 4,

$$|D_g(x) - D_h(x)| \le L_{D,\infty} \|g - h\|_\infty .$$

Hence, the map $(g, y) \mapsto \phi_{g,y}(x)$ is Lipschitz on $\mathcal{G} \times K$ with respect to $\|\cdot\|_\infty$ in g and the Euclidean norm in y. For the y–dependence this follows from smoothness (or Lipschitz continuity) of q_T and boundedness of $1/D_g(x)$ on $B(x_0, R_0)$. Thus, for all $x \in B(x_0, R_0)$, $g, h \in \mathcal{G}$, and $y, y' \in K$,

$$|\phi_{g,y}(x) - \phi_{h,y'}(x)| \le L_g \, \|g - h\|_{L^\infty(\mathbb{R}^d)} + L_y \, \|y - y'\|, \tag{36}$$

where

$$L_g := \frac{c_+ \, L_{D,\infty}}{\underline{D}^2}, \qquad L_y := \frac{L_q}{\underline{D}}. \tag{37}$$

Let

$$\Phi_K := \left\{x \mapsto \phi_{g,y}(x) : g \in \mathcal{G}, y \in K\right\}.$$

Note that

$$\log \mathcal{N}_{[]}\big(\Phi_K, L^2(\rho_0), \varepsilon\big) \;\le\; \log \mathcal{N}_{[]}\Big(\mathcal{G}, \|\cdot\|_\infty, \frac{\varepsilon}{2L_g}\Big) \;+\; C_d \, \log\Big(\frac{L_y \, \mathrm{diam}(K)}{\varepsilon}\Big), \tag{38}$$

where $C_d \le d$ is a dimensional constant (coming from covering K in Euclidean norm). Then it follows from Theorem 3.5.13 in Giné & Nickl (2016),

$$\sup_{g\in\mathcal{G}} \sup_{y\in\mathbb{R}^d} |(\mathrm{B})| \lesssim \frac{1}{\sqrt{N}} \int_0^{8F_\infty} \sqrt{\log \mathcal{N}_{[]}\Big(\mathcal{G}, \|\cdot\|_\infty, \frac{\varepsilon}{2L_g}\Big)}\, d\varepsilon + \frac{1}{\sqrt{N}} \sqrt{\log\Big(\frac{L_y \, \mathrm{diam}(K)}{F_\infty}\Big)}. \tag{39}$$

Furthermore, for a fixed (g, y),

$$(\mathrm{A}) = \frac{1}{N} \sum_{i=1}^{N} q_T(X_i, y) \left(\frac{1}{\widehat{D}_g(X_i)} - \frac{1}{D_g(X_i)}\right).$$

By (35), $u \mapsto 1/u$ is Lipschitz on $[\underline{D}, \overline{D}]$ with constant $L_D := \overline{D}^2/\underline{D}^2$, so

$$\left| \frac{1}{\mathcal{T}_{[\underline{D},\overline{D}]}[\widehat{D}_g](X_i)} - \frac{1}{D_g(X_i)} \right| \le L_D \big|\widehat{D}_g(X_i) - D_g(X_i)\big|.$$

By *(Q)* we have

$$q_T(x,y) \le c_+ \quad \text{for all } x, y \in \mathbb{R}^d.$$

Hence, for all $X_i \in B(x_0, R_0)$ and all $y \in \mathbb{R}^d$,

$$q_T(X_i, y) \le c_+.$$

Thus,

$$\sup_{y\in K} |(\mathrm{A})| \le \frac{c_+ L_D}{N} \sum_{i=1}^{N} \big|\widehat{D}_g(X_i) - D_g(X_i)\big|.$$

Taking supremum over $g \in \mathcal{G}$ and expectations,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \sup_{y\in K} |(\mathrm{A})|\right] \le c_+ L_D\, \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \big|\widehat{D}_g(X_i) - D_g(X_i)\big|\right].$$

Now for each fixed x and g,

$$D_g(x) = \mathbb{E}_{Z\sim\rho_T}\big[\psi_g(x,Z)\big], \qquad \widehat{D}_g(x) = \frac{1}{M}\sum_{k=1}^{M} \psi_g(x, Y_k),$$

with

$$\psi_g(x,z) := q_T(x,z)\, \frac{1}{g(z)}.$$

Let

$$\Psi := \big\{(x,z) \mapsto \psi_g(x,z) : x \in B(x_0,R_0),\ g \in \mathcal{G}\big\}.$$

Due to *(G)*,

$$g(z) \ge c_{\mathcal{G}}^{-} e^{-a_{\mathcal{G}}\|z\|^2}, \qquad \forall z \in \mathbb{R}^d,$$

and we have for all $x \in B(x_0, R_0)$ and all $g \in \mathcal{G}$,

$$\psi_g(x,z) \le F(z) := \frac{c_+}{c_{\mathcal{G}}^{-}}\, e^{a_{\mathcal{G}}\|z\|^2}.$$

Since ρ_T satisfies the Gaussian upper bound (RT) with $b_T^+ > 2a_{\mathcal{G}}$, we get $F \in L^2(\rho_T)$ and

$$\|F\|_{L^2(\rho_T)} \le \frac{c_+}{c_{\mathcal{G}}^{-}} \sqrt{c_T^+} \left(\frac{\pi}{b_T - 2a_{\mathcal{G}}}\right)^{d/4}.$$

Theorem 3.5.13 in Giné & Nickl (2016) gives

$$\mathbb{E}\Big[\sup_{x\in B(x_0,R_0)} \sup_{g\in\mathcal{G}} \big|\widehat{D}_g(x) - D_g(x)\big|\Big] \lesssim \frac{1}{\sqrt{M}} \int_0^{8\|F\|_{L^2(\rho_T)}} \sqrt{\log(2N_{[]}(\Psi, L^2(\rho_T), \varepsilon))}\, d\varepsilon. \quad (40)$$

Fix $x \in B(x_0, R_0)$ and $z \in \mathbb{R}^d$. Write

$$\psi_g(x,z) - \psi_h(x,z) = q_T(x,z)\Big(\frac{1}{g(z)} - \frac{1}{h(z)}\Big).$$

We have

$$\frac{1}{g(z)} \le \frac{1}{c_{\mathcal{G}}^{-}} e^{a_{\mathcal{G}}\|z\|^2}, \qquad \frac{1}{h(z)} \le \frac{1}{c_{\mathcal{G}}^{-}} e^{a_{\mathcal{G}}\|z\|^2}.$$

Hence, using $|1/u - 1/v| = |u - v|/(uv)$,

$$\Big|\frac{1}{g(z)} - \frac{1}{h(z)}\Big| \le \frac{|g(z) - h(z)|}{(c_{\mathcal{G}}^{-})^2}\, e^{2a_{\mathcal{G}}\|z\|^2}.$$

Multiplying by $|q_T(x,z)| \le c_+$ gives

$$|\psi_g(x,z) - \psi_h(x,z)| \le \frac{c_+}{(c_{\mathcal{G}}^-)^2}\, e^{2a_{\mathcal{G}}\|z\|^2}\, \|g-h\|_\infty.$$

Taking $L^2(\rho_T)$ norms yields

$$\|\psi_g(x,\cdot) - \psi_h(x,\cdot)\|_{L^2(\rho_T)} \le \frac{c_+}{(c_{\mathcal{G}}^-)^2}\Big(\int e^{4a_{\mathcal{G}}\|z\|^2}\rho_T(z)\,dz\Big)^{1/2}\|g-h\|_\infty.$$

Hence,

$$\|\psi_g(x,\cdot) - \psi_h(x,\cdot)\|_{L^2(\rho_T)} \;\le\; \frac{c_+}{(c_{\mathcal{G}}^-)^2}\, M_\rho\, \|g-h\|_\infty$$

and as a result we have ,

$$\log \mathcal{N}_{[]}\big(\Psi, L^2(\rho_T), \varepsilon\big) \;\le\; \log \mathcal{N}_{[]}\Big(\mathcal{G}, \|\cdot\|_\infty, \frac{\varepsilon}{C_g}\Big) \;+\; d\,\log\Big(\frac{C\,R_0}{\varepsilon}\Big),$$

where $C_g := \frac{c_+}{(c_{\mathcal{G}}^-)^2}\, M_\rho$. We choose $K = B(0,R)$ with $R > R_0$. Fix $g \in \mathcal{G}$ and $y \in \mathbb{R}^d$ with $\|y\| \ge R$. For any $x \in B(0,R_0)$, we have

$$\|x-y\| \;\ge\; \|y\| - \|x\| \;\ge\; R - R_0,$$

hence by the kernel upper bound

$$q_T(x,y) \le c_+ \exp\big(-a_+\|x-y\|^2\big) \le c_+ \exp\big(-a_+(R-R_0)^2\big).$$

Using $D_g(x) \ge \underline{D}$ and that ρ_0 is a probability density,

$$\begin{aligned}|\mathcal{C}[g](y)| &= \Big|\int_{B(0,R_0)} \rho_0(x)\,\frac{q_T(x,y)}{D_g(x)}\,dx\Big| \le \frac{1}{\underline{D}}\int_{B(0,R_0)}\rho_0(x)\,q_T(x,y)\,dx \\ &\le \frac{1}{\underline{D}}\sup_{x\in B(0,R_0)} q_T(x,y) \le \frac{c_+}{\underline{D}}\exp\big(-a_+(R-R_0)^2\big).\end{aligned}$$

Furthermore, using $\mathcal{T}_{[\underline{D},\overline{D}]}[\widehat{D}_g](X_i) \ge \underline{D}$ and $X_i \in B(0,R_0)$,

$$|\widehat{\mathcal{C}}_{N,M}[g](y)| \le \frac{1}{N}\sum_{i=1}^N \frac{q_T(X_i,y)}{\mathcal{T}_{[\underline{D},\overline{D}]}[\widehat{D}_g](X_i)} \le \frac{1}{\underline{D}}\sup_{x\in B(0,R_0)} q_T(x,y) \le \frac{c_+}{\underline{D}}\exp\big(-a_+(R-R_0)^2\big).$$

Therefore, for all $g \in \mathcal{G}$ and all $\|y\| \ge R$,

$$\big|\widehat{\mathcal{C}}_{N,M}[g](y) - \mathcal{C}[g](y)\big| \le |\widehat{\mathcal{C}}_{N,M}[g](y)| + |\mathcal{C}[g](y)| \le \frac{2c_+}{\underline{D}}\exp\big(-a_+(R-R_0)^2\big).$$

Then for every $R > R_0$,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sup_{y\notin B(0,R)}\big|\widehat{\mathcal{C}}_{N,M}[g](y) - \mathcal{C}[g](y)\big|\right] \le \frac{2c_+}{\underline{D}}\,\exp\big(-a_+(R-R_0)^2\big). \tag{41}$$

C Approximation error

Lemma 5. *Suppose that the assumptions* (Q)*,* (R0)*,* (RT) *and* (G) *hold. Assume additionally that* $b_T^+ > 4a_{\mathcal{G}}$*. Then there exists constant* $L_{\mathcal{C},2} > 0$ *such that*

$$\|\mathcal{C}[g] - \mathcal{C}[h]\|_{L^2(\rho_T)} \le L_{\mathcal{C},2}\,\|g-h\|_{L^2(\rho_T)}. \tag{42}$$

Proof. Similarly to (31),

$$|D_g(x) - D_h(x)| \le (c_{\mathcal{G}}^-)^{-2}\int q_T(x,z)\,\rho_T(z)\,\exp(2a_{\mathcal{G}}\|z\|^2)\,|g(z)-h(z)|\,dz.$$

Apply Cauchy–Schwarz w.r.t. $\rho_T(z)\,dz$:

$$|D_g(x) - D_h(x)| \le (c_{\mathcal{G}}^-)^{-2}\Big(\int q_T^2(x,z)\exp(4a_{\mathcal{G}}\|z\|^2)\,\rho_T(z)\,dz\Big)^{1/2}\|g-h\|_{L^2(\rho_T)}.$$

Taking the supremum over $x \in B(x_0, R_0)$ yields

$$\|D_g - D_h\|_{L^\infty(B(x_0,R_0))} \le L_{D,2}\,\|g-h\|_{L^2(\rho_T)}, \tag{43}$$

for some constant $L_{D,2} < \infty$ provided

$$\sup_{x\in B(x_0,R_0)} \int q_T^2(x,z)e^{4a_{\mathcal{G}}\|z\|^2}\,\rho_T(z)\,dz$$

is finite. We now check this finiteness under *(Q)* and the upper tail in *(RT)*. By *(Q)*,

$$q_T^2(x,z) \le c_+^2\exp(-2a_+\|x-z\|^2) \le c_+^2\exp(2a_+R_\star^2)\exp(-a_+\|z\|^2),$$

We obtain

$$q_T^2(x,z)e^{4a_{\mathcal{G}}\|z\|^2} \le c_+^2\exp(2a_+R_\star^2)\,\exp\Big(-\big(a_+ - 4a_{\mathcal{G}} + b_T^+\big)\|z\|^2\Big).$$

This is integrable if $a_+ + b_T^+ - 4a_{\mathcal{G}} > 0$. Hence $L_{D,2} < \infty$ and (43) holds. By (43) it holds that

$$|\mathcal{C}[g](y) - \mathcal{C}[h](y)| \le \frac{L_{D,2}}{\underline{D}^2}\|g-h\|_{L^2(\rho_T)}\underbrace{\int_{K_0}\rho_0(x)\,q_T(x,y)\,dx}_{=:m_0(y)}.$$

Now take $L^2(\rho_T)$ norms in y:

$$\|\mathcal{C}[g] - \mathcal{C}[h]\|_{L^2(\rho_T)} \le \frac{L_{D,2}}{\underline{D}^2}\,\|m_0\|_{L^2(\rho_T)}\,\|g-h\|_{L^2(\rho_T)}.$$

Finally, $\|m_0\|_{L^2(\rho_T)} < \infty$ under *(Q),(R0),(RT)* since $m_0(y)$ is sub-Gaussian in y (a compactly supported mixture of sub-Gaussian kernels) and ρ_T has sub-Gaussian tails. Therefore (42) holds with

$$L_{\mathcal{C},2} := \frac{L_{D,2}}{\underline{D}^2}\,\|m_0\|_{L^2(\rho_T)} < \infty.$$

□

Corollary 5.1. *Let $g^\star \in \mathcal{G}$ be a fixed point of $\mathcal{C}$, i.e. $g^\star = \mathcal{C}[g^\star]$. Then under assumptions of Lemma 5, it holds for all $g \in \mathcal{G}$,*

$$\big|\mathcal{R}(g) - \mathcal{R}(g^\star)\big| \le L_\ell\big(1 + L_{\mathcal{C},2}\big)\,\|g-g^\star\|_{L^2(\rho_T)}. \tag{44}$$

Proof. Using *(L)* and the fixed point property $\mathcal{C}[g^\star] = g^\star$, we have almost surely (with $Y \sim \rho_T$)

$$\Big|\ell\big(g(Y),\mathcal{C}[g](Y)\big) - \ell\big(g^\star(Y),g^\star(Y)\big)\Big| \le L_\ell\Big(|g(Y)-g^\star(Y)| + |\mathcal{C}[g](Y)-g^\star(Y)|\Big).$$

Taking expectations and applying Cauchy–Schwarz gives

$$\big|\mathcal{R}(g) - \mathcal{R}(g^\star)\big| \le L_\ell\Big(\|g-g^\star\|_{L^2(\rho_T)} + \|\mathcal{C}[g]-g^\star\|_{L^2(\rho_T)}\Big).$$

Since $g^\star = \mathcal{C}[g^\star]$,

$$\|\mathcal{C}[g]-g^\star\|_{L^2(\rho_T)} = \|\mathcal{C}[g]-\mathcal{C}[g^\star]\|_{L^2(\rho_T)} \le L_{\mathcal{C},2}\,\|g-g^\star\|_{L^2(\rho_T)}$$

by Lemma 5. Combining these inequalities yields (44). □

C.1 BARGMANN TRANSFORM

For $f \in L^2(\mathbb{R}^d)$, the (Segal-)Bargmann transform is defined by

$$(\mathcal{B}f)(z) := \pi^{-d/4} \int_{\mathbb{R}^d} \exp\Big(-\tfrac{1}{2}|y|^2 + \sqrt{2}\, z \cdot y - \tfrac{1}{2}\, z \cdot z \Big)\, f(y)\, dy, \qquad z \in \mathbb{C}^d,$$

where $z \cdot y = \sum_{j=1}^d z_j y_j$ and $z \cdot z = \sum_{j=1}^d z_j^2$. The map $\mathcal{B}$ is unitary from $L^2(\mathbb{R}^d)$ onto the Fock space $\mathcal{F}^2(\mathbb{C}^d)$ of entire functions. Define the Hermite polynomials by

$$\mathrm{He}_n(x) := (-1)^n e^{x^2} \frac{d^n}{dx^n}\Big(e^{-x^2}\Big), \qquad n \in \mathbb{N}_0.$$

Then the corresponding $L^2(\mathbb{R})$–normalized Hermite functions are

$$\psi_n(x) := \frac{1}{2^{n/2}\pi^{1/4}\sqrt{n!}}\, \mathrm{He}_n(x)\, e^{-x^2/2}, \qquad x \in \mathbb{R},\ \ n \in \mathbb{N}_0. \tag{45}$$

A direct computation (using the generating function of Hermite polynomials) shows that for all $n \in \mathbb{N}_0$,

$$(\mathcal{B}\psi_n)(z) = \frac{z^n}{\sqrt{n!}}, \qquad z \in \mathbb{C}.$$

Indeed, by definition,

$$(\mathcal{B}\psi_0)(z) = \pi^{-1/4} \int_{\mathbb{R}} e^{-\frac{1}{2}y^2 + \sqrt{2}zy - \frac{1}{2}z^2}\ \pi^{-1/4} e^{-y^2/2}\, dy = \pi^{-1/2} \int_{\mathbb{R}} e^{-y^2 + \sqrt{2}zy - \frac{1}{2}z^2}\, dy.$$

Complete the square:

$$-y^2 + \sqrt{2}zy = -\big(y - \tfrac{z}{\sqrt{2}}\big)^2 + \tfrac{1}{2}z^2,$$

hence

$$-y^2 + \sqrt{2}zy - \tfrac{1}{2}z^2 = -\big(y - \tfrac{z}{\sqrt{2}}\big)^2.$$

Therefore,

$$(\mathcal{B}\psi_0)(z) = \pi^{-1/2} \int_{\mathbb{R}} e^{-(y - \frac{z}{\sqrt{2}})^2}\, dy = \pi^{-1/2} \int_{\mathbb{R}} e^{-u^2}\, du = 1.$$

Let $f \in \mathcal{S}(\mathbb{R})$ (Schwartz) so that all integrations by parts are justified. Write

$$K(y, z) := \exp\Big(-\tfrac{1}{2}y^2 + \sqrt{2}zy - \tfrac{1}{2}z^2 \Big).$$

Then

$$(\mathcal{B}(a^* f))(z) = \pi^{-1/4} \int_{\mathbb{R}} K(y, z)\, \frac{1}{\sqrt{2}}\Big(y - \frac{d}{dy}\Big) f(y)\, dy = \frac{1}{\sqrt{2}} \pi^{-1/4} \Big(I_1 - I_2 \Big),$$

where

$$I_1 := \int_{\mathbb{R}} K(y, z)\, y\, f(y)\, dy, \qquad I_2 := \int_{\mathbb{R}} K(y, z)\, f'(y)\, dy$$

and

$$a^* := \frac{1}{\sqrt{2}}\Big(y - \frac{d}{dy}\Big).$$

Integrate by parts in I_2 (boundary terms vanish since $K(\cdot, z)$ has Gaussian decay and f is Schwartz):

$$I_2 = -\int_{\mathbb{R}} \partial_y K(y, z)\, f(y)\, dy.$$

Compute $\partial_y K$:

$$\partial_y K(y, z) = (-y + \sqrt{2}z)\, K(y, z).$$

Hence

$$I_2 = -\int_{\mathbb{R}} (-y + \sqrt{2}z)\, K(y, z)\, f(y)\, dy = \int_{\mathbb{R}} y\, K(y, z)\, f(y)\, dy - \sqrt{2}z \int_{\mathbb{R}} K(y, z)\, f(y)\, dy.$$

Therefore,

$$I_1 - I_2 = \sqrt{2}z \int_{\mathbb{R}} K(y,z)\, f(y)\, dy,$$

and so

$$(\mathcal{B}(a^* f))(z) = \frac{1}{\sqrt{2}}\pi^{-1/4} \cdot \sqrt{2}z \int_{\mathbb{R}} K(y,z)\, f(y)\, dy = z\,(\mathcal{B}f)(z).$$

Thus we have shown

$$\mathcal{B} \circ a^* = (\text{multiplication by } z) \circ \mathcal{B} \qquad \text{on } \mathcal{S}(\mathbb{R}). \tag{46}$$

By the definition of ψ_n we have $\psi_n := \frac{1}{\sqrt{n}}\, a^* \psi_{n-1}$, $n \geq 1$ and (46) implies

$$(\mathcal{B}\psi_n)(z) = \frac{1}{\sqrt{n}}\,(\mathcal{B}(a^*\psi_{n-1}))(z) = \frac{1}{\sqrt{n}}\, z\,(\mathcal{B}\psi_{n-1})(z).$$

Starting from $(\mathcal{B}\psi_0)(z) = 1$, we obtain recursively

$$(\mathcal{B}\psi_n)(z) = \frac{z^n}{\sqrt{n!}}, \qquad n \in \mathbb{N}_0.$$

This identity reflects the fact that the Bargmann transform diagonalizes the harmonic oscillator: Hermite functions are mapped to monomials. The d–dimensional Hermite functions factorize as

$$\psi_\alpha(x) = \prod_{j=1}^{d} \psi_{\alpha_j}(x_j), \qquad \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d.$$

Likewise, the Bargmann kernel factorizes coordinatewise:

$$e^{-\frac{1}{2}|y|^2 + \sqrt{2}\, z\cdot y - \frac{1}{2} z\cdot z} = \prod_{j=1}^{d} e^{-\frac{1}{2}y_j^2 + \sqrt{2}\, z_j y_j - \frac{1}{2} z_j^2}.$$

Applying Fubini's theorem and the one–dimensional identity yields

$$(\mathcal{B}\psi_\alpha)(z) = \prod_{j=1}^{d} (\mathcal{B}\psi_{\alpha_j})(z_j) = \prod_{j=1}^{d} \frac{z_j^{\alpha_j}}{\sqrt{\alpha_j!}} = \frac{z^\alpha}{\sqrt{\alpha!}}, \tag{47}$$

where $z^\alpha = \prod_{j=1}^d z_j^{\alpha_j}$ and $\alpha! = \prod_{j=1}^d \alpha_j!$. Since $\{\psi_\alpha\}_{\alpha \in \mathbb{N}_0^d}$ is an orthonormal basis of $L^2(\mathbb{R}^d)$, every $f \in L^2(\mathbb{R}^d)$ admits the Hermite expansion

$$f = \sum_{\alpha \in \mathbb{N}_0^d} \langle f, \psi_\alpha \rangle\, \psi_\alpha \quad (\text{in } L^2(\mathbb{R}^d)).$$

Because $\mathcal{B}$ is unitary, we may apply it termwise:

$$\mathcal{B}f = \sum_{\alpha \in \mathbb{N}_0^d} \langle f, \psi_\alpha \rangle\, \mathcal{B}\psi_\alpha.$$

Using (47), we obtain the power–series representation

$$(\mathcal{B}f)(z) = \sum_{\alpha \in \mathbb{N}_0^d} \langle f, \psi_\alpha \rangle\, \frac{z^\alpha}{\sqrt{\alpha!}}, \qquad z \in \mathbb{C}^d.$$

Thus, the Bargmann transform converts the Hermite expansion of f into the Taylor expansion of the entire function $\mathcal{B}f$. In this correspondence,

$$\langle f, \psi_\alpha \rangle = \frac{\partial^\alpha (\mathcal{B}f)(0)}{\sqrt{\alpha!}},$$

which explains why bounds on the growth of $\mathcal{B}f$ immediately yield decay estimates for Hermite coefficients.

C.2 Approximation by Hermite polynomials

Since we are interested in the case $T = 1$ we omit T from the notation of q. Then

$$q(z) := (2\pi)^{-d/2} \exp\Big(- \frac{\|z\|^2}{2} \Big).$$

Proposition 6. *Let $d \geq 1$ let $w : \mathbb{R}^d \to [0, \infty)$ be compactly supported with*

$$\mathrm{supp}(w) \subseteq \{x \in \mathbb{R}^d : \ \|x\| \leq R\}, \qquad M_0 := \int_{\mathbb{R}^d} w(x)\, dx < \infty.$$

Define

$$g^\star(y) := (w * q)(y) = \int_{\mathbb{R}^d} w(x)\, q(x - y)\, dx.$$

Let $\{\psi_\alpha\}_{\alpha \in \mathbb{N}_0^d}$ be the d-dimensional tensor-product Hermite basis orthonormal in $L^2(\mathbb{R}^d)$ and define the Hermite coefficients

$$c_\alpha := \langle g^\star, \psi_\alpha \rangle_{L^2(\mathbb{R}^d)}, \qquad \alpha \in \mathbb{N}_0^d.$$

Then for every multi-index $\alpha \in \mathbb{N}_0^d$ with $m := |\alpha| \geq 1$,

$$|c_\alpha| \ \leq \ C_d\, m^{1/4} \Big(\frac{e^{1/2}\beta}{\sqrt{m}} \Big)^m, \qquad C_d = \pi^{-d/4} 2^{-d/2} M_0, \qquad \beta := R\sqrt{\frac{d}{2}}. \tag{48}$$

Equivalently,

$$|c_\alpha| \ \leq \ C_d\, m^{1/4} \exp\Big(- \frac{m}{2} \log m + m \log(e^{1/2}\beta) \Big). \tag{49}$$

In particular, if $m \geq (e^{1/2}\beta)^2 = e\beta^2$, then $\log(e^{1/2}\beta) \leq \frac{1}{2} \log m$ and

$$|c_\alpha| \ \leq \ C_d\, m^{1/4} \exp\Big(- \frac{m}{4} \log m \Big). \tag{50}$$

Remark 2. *Consider the Gaussian kernel with variance $T > 1$,*

$$q_T(x - y) = (2\pi T)^{-d/2} \exp\Big(- \frac{|x - y|^2}{2T} \Big), \qquad g^\star = w * q_T.$$

Although the unscaled Bargmann transform of $g^\star$ exhibits Gaussian (quadratic) growth in the complex variable z, this can be compensated by a suitable scaling of the Hermite basis. More precisely, let $\lambda \in (0, 1)$ be chosen such that

$$\lambda^2 = \frac{1}{T}, \qquad \textit{equivalently} \qquad \lambda = T^{-1/2}.$$

If one expands $g^\star$ in the scaled Hermite basis

$$\psi_\alpha^{(1/\lambda)}(x) = \lambda^{d/2} \prod_{j=1}^d \psi_{\alpha_j}(\lambda x_j), \qquad \alpha \in \mathbb{N}_0^d,$$

and considers the associated scaled Bargmann transform, then the quadratic term in the exponential prefactor of the Bargmann representation is exactly neutralized by the scaling. As a result, the Bargmann transform of $g^\star$ in the scaled variables obeys a linear–exponential *growth bound of the form*

$$\sup_{\max_j |z_j| \leq r} \big| (\mathcal{B} g^\star)(z) \big| \ \leq \ C_{d,T} \, \exp\big(b_{d,T}\, r \big), \qquad r > 0,$$

with constants $C_{d,T}, b_{d,T} > 0$ depending only on d, T, and the support radius of w, but not *on r. In particular, under this proper choice of the scaling parameter λ, the Bargmann transform exhibits the same linear-exponential growth as in the critical case $T = 1$, and the resulting scaled Hermite coefficients decay super-geometrically in the total degree $|\alpha|$. This observation explains why the scaling $\lambda = T^{-1/2}$ is natural and optimal for extending the $T = 1$ coefficient estimates to general variances $T > 1$.*

Proof. The Bargmann transform maps Hermite functions to normalized monomials:

$$(\mathcal{B}\psi_\alpha)(z) = \frac{z^\alpha}{\sqrt{\alpha!}}, \qquad \alpha \in \mathbb{N}_0^d,$$

and therefore, for $g^\star \in L^2(\mathbb{R}^d)$,

$$(\mathcal{B}g^\star)(z) = \sum_{\alpha\in\mathbb{N}_0^d} c_\alpha \, \frac{z^\alpha}{\sqrt{\alpha!}}, \qquad z \in \mathbb{C}^d.$$

In particular,

$$c_\alpha = \frac{\partial^\alpha(\mathcal{B}g^\star)(0)}{\sqrt{\alpha!}}. \tag{51}$$

For $r > 0$ set the closed polydisk $D_r := \{z \in \mathbb{C}^d : \max_j |z_j| \le r\}$. By the multivariate Cauchy estimate,

$$|\partial^\alpha F(0)| \le \alpha!\, r^{-|\alpha|} \sup_{z\in D_r} |F(z)|, \qquad \alpha \in \mathbb{N}_0^d.$$

Apply this with $F = \mathcal{B}g^\star$ and combine with (51):

$$|c_\alpha| \le \sqrt{\alpha!}\, r^{-m} \sup_{z\in D_r} |(\mathcal{B}g^\star)(z)|, \qquad m := |\alpha|. \tag{52}$$

Insert (60) (see Proposition 8) into (52):

$$|c_\alpha| \le C_0 \sqrt{\alpha!}\, r^{-m} e^{\beta r}, \qquad r > 0.$$

Minimize $\phi(r) := \beta r - m \log r$ over $r > 0$. Since $\phi'(r) = \beta - \frac{m}{r}$, the unique minimizer is

$$r_* = \frac{m}{\beta}.$$

Hence

$$|c_\alpha| \le C_0 \sqrt{\alpha!} \left(\frac{\beta}{m}\right)^m e^{\beta(m/\beta)} = C_0 \sqrt{\alpha!} \left(\frac{e\beta}{m}\right)^m. \tag{53}$$

Since $\alpha! \le m!$ for $m = |\alpha|$, we have $\sqrt{\alpha!} \le \sqrt{m!}$. By Stirling's estimate there exists an absolute constant $C > 0$ such that

$$\sqrt{m!} \;\le\; C\, m^{1/4} \Big(\frac{m}{e}\Big)^{m/2}, \qquad m \ge 1. \tag{54}$$

Combining (53) and (54) yields

$$|c_\alpha| \le C\, C_0\, m^{1/4} \Big(\frac{m}{e}\Big)^{m/2} \Big(\frac{e\beta}{m}\Big)^m.$$

Simplify:

$$\Big(\frac{m}{e}\Big)^{m/2} \Big(\frac{e\beta}{m}\Big)^m = m^{m/2} e^{-m/2} \cdot e^m \beta^m m^{-m} = e^{m/2} \beta^m m^{-m/2} = \Big(\frac{e^{1/2}\beta}{\sqrt{m}}\Big)^m.$$

Therefore

$$|c_\alpha| \le C\, C_0\, m^{1/4} \left(\frac{e^{1/2}\beta}{\sqrt{m}}\right)^m,$$

which is (48). □

Proposition 7. *Let $d \ge 1$ and let $\{\psi_\alpha\}_{\alpha\in\mathbb{N}_0^d}$ be the d–dimensional Hermite basis in $L^2(\mathbb{R}^d)$. Let Π_n^F denote the $L^2(\mathbb{R}^d)$–orthogonal projector onto*

$$\operatorname{span}\{\psi_\alpha : \ |\alpha| \le n\}.$$

Under assumptions of Proposition 6 for every $n \in \mathbb{N}_0$ with $n + 1 > 2K^2$, one has

$$\|g^\star - \Pi_n^F g^\star\|_{L^2(\mathbb{R}^d)} \;\le\; \widetilde{C}_d\, C_d\, (n+1)^{\frac{d}{2}-\frac{1}{4}} \left(\frac{K}{\sqrt{n+1}}\right)^{n+1} \tag{55}$$

where C_d is given in (48) *and*

$$\widetilde{C}_d = \left(1 + 2^{d-1/2}\, \frac{\Gamma(d+1/2)}{(\ln 2)^{d+1/2}}\right)^{1/2}.$$

Proof. By orthonormality of $\{\psi_\alpha\}$ and the definition of Π_n^F,

$$\|g^\star - \Pi_n^F g^\star\|_{L^2(\mathbb{R}^d)}^2 = \sum_{|\alpha|>n} |c_\alpha|^2. \tag{56}$$

Using (48) with $m := |\alpha|$ gives

$$\sum_{|\alpha|>n} |c_\alpha|^2 \le C_d^2 \sum_{|\alpha|>n} m^{1/2} \Big(\frac{K^2}{m}\Big)^m.$$

Group by total degree $m = |\alpha|$ and let

$$N_d(m) := \#\{\alpha \in \mathbb{N}_0^d : \ |\alpha| = m\} = \binom{m+d-1}{d-1}.$$

Then

$$\sum_{|\alpha|>n} |c_\alpha|^2 \le C_d^2 \sum_{m=n+1}^{\infty} N_d(m)\, m^{1/2} \Big(\frac{K^2}{m}\Big)^m. \tag{57}$$

Using the standard estimate $N_d(m) \le \widetilde{C}_d\, m^{d-1}$, we obtain

$$\sum_{|\alpha|>n} |c_\alpha|^2 \le \widetilde{C}_d\, C_d^2 \sum_{m=n+1}^{\infty} m^{d-\frac{1}{2}} \Big(\frac{K^2}{m}\Big)^m. \tag{58}$$

Set

$$q := \frac{K^2}{n+1} \in (0,1).$$

For $m \ge n+1$ we have $\frac{K^2}{m} \le q$, hence

$$\Big(\frac{K^2}{m}\Big)^m \le q^m.$$

Therefore,

$$\begin{aligned} \sum_{m=n+1}^{\infty} m^{d-\frac{1}{2}} \Big(\frac{K^2}{m}\Big)^m &\le q^{n+1}(n+1)^{d-1/2} \sum_{k=0}^{\infty} \Big(1 + \frac{k}{n+1}\Big)^{d-1/2} q^k \\ &\le q^{n+1}(n+1)^{d-1/2} \left(1 + 2^{d-1/2}\, \frac{\Gamma(d+1/2)}{(-\ln q)^{d+1/2}}\right) \end{aligned}$$

Insert this into (58) and take square roots:

$$\|g^\star - \Pi_n^F g^\star\|_2 \le \widetilde{C}_d\, C_d\, (n+1)^{\frac{d}{2}-\frac{1}{4}}\, q^{(n+1)/2}.$$

Finally, $q^{(n+1)/2} = \Big(\frac{K}{\sqrt{n+1}}\Big)^{n+1}$ and $1 - q = 1 - \frac{K^2}{n+1}$, which gives (55). □

Proposition 8. *Let assumptions of Proposition 6 hold. For every $z \in \mathbb{C}^d$,*

$$(\mathcal{B}g^\star)(z) = \pi^{-d/4}\, 2^{-d/2} \int_{\mathbb{R}^d} w(x)\, \exp\Big(-\frac{|x|^2}{4} + \frac{1}{\sqrt{2}}\, x \cdot z\Big)\, dx. \tag{59}$$

Moreover for every $r > 0$,

$$\sup_{\max_j |z_j| \le r} |(\mathcal{B}g^\star)(z)| \ \le\ \pi^{-d/4}\, 2^{-d/2}\, M_0\, \exp\Big(\frac{R}{\sqrt{2}}\, \sqrt{d}\, r\Big). \tag{60}$$

Proof. By definition and Fubini,

$$(\mathcal{B}g^\star)(z) = \pi^{-d/4} e^{-\frac{1}{2} z\cdot z} \int_{\mathbb{R}^d} w(x)\, (2\pi)^{-d/2} \int_{\mathbb{R}^d} \exp\Big(-\tfrac{1}{2}|y|^2 + \sqrt{2}\, z \cdot y - \tfrac{1}{2}|x-y|^2\Big)\, dy\, dx.$$

Expand $|x-y|^2 = |x|^2 - 2x\cdot y + |y|^2$ to obtain

$$-\tfrac{1}{2}|y|^2 - \tfrac{1}{2}|x-y|^2 + \sqrt{2}\, z\cdot y = -|y|^2 + (x+\sqrt{2}\, z)\cdot y - \tfrac{1}{2}|x|^2.$$

Complete the square:

$$-|y|^2 + (x+\sqrt{2}z)\cdot y = -\Big|y - \frac{x+\sqrt{2}z}{2}\Big|^2 + \frac{|x+\sqrt{2}z|^2}{4}.$$

Therefore,

$$\int_{\mathbb{R}^d} \exp\Big(-|y|^2 + (x+\sqrt{2}z)\cdot y\Big)\, dy = \pi^{d/2}\exp\Big(\frac{|x+\sqrt{2}z|^2}{4}\Big).$$

Hence

$$\begin{aligned}(2\pi)^{-d/2}&\int_{\mathbb{R}^d} \exp\Big(-\tfrac{1}{2}|y|^2 + \sqrt{2}\, z\cdot y - \tfrac{1}{2}|x-y|^2\Big)\, dy \\ &= (2\pi)^{-d/2}\,\pi^{d/2}\,\exp\Big(-\frac{|x|^2}{2} + \frac{|x+\sqrt{2}z|^2}{4}\Big) = 2^{-d/2}\exp\Big(-\frac{|x|^2}{4} + \frac{1}{\sqrt{2}}x\cdot z + \frac{1}{2}z\cdot z\Big),\end{aligned}$$

where we used $|x+\sqrt{2}z|^2 = |x|^2 + 2\sqrt{2}\, x\cdot z + 2\, z\cdot z$. Multiplying by the prefactor $e^{-\frac{1}{2}z\cdot z}$ from the Bargmann kernel cancels the quadratic term in $z\cdot z$, yielding (59). From (59) and $e^{-|x|^2/4}\le 1$,

$$|(\mathcal{B}g^\star)(z)| \le \pi^{-d/4}\,2^{-d/2}\int_{\mathbb{R}^d}|w(x)|\,\Big|\exp\Big(\frac{1}{\sqrt{2}}x\cdot z\Big)\Big|\,dx = \pi^{-d/4}\,2^{-d/2}\int_{\mathbb{R}^d}|w(x)|\,\exp\Big(\frac{1}{\sqrt{2}}x\cdot \mathrm{Re}\, z\Big)\,dx.$$

If $\mathrm{supp}(w)\subseteq\{|x|\le R\}$, then $x\cdot \mathrm{Re}\, z \le |x|\,|\,\mathrm{Re}\, z| \le R\,|\,\mathrm{Re}\, z|$, and thus

$$|(\mathcal{B}g^\star)(z)| \le \pi^{-d/4}\,2^{-d/2}\,M_0\,\exp\Big(\frac{R}{\sqrt{2}}|\,\mathrm{Re}\, z|\Big).$$

On the polydisk $\max_j |z_j|\le r$ we have $|\,\mathrm{Re}\, z| \le |z| \le \sqrt{d}\, r$, hence

$$\sup_{\max_j|z_j|\le r}|(\mathcal{B}g^\star)(z)| \le \pi^{-d/4}\,2^{-d/2}\,M_0\,\exp\Big(\frac{R}{\sqrt{2}}\sqrt{d}\, r\Big),$$

which is (60). □

Lemma 9. *Let $\mathcal{F}_B = \{f_c(x) := \sum_{|\alpha|\le n} c_\alpha\psi_\alpha(x), \sum_{|\alpha|\le n}|c_\alpha|^2 \le B\}$. Then there exists some absolute constant $C>0$ such that*

$$\log \mathcal{N}_{[]}(\mathcal{F}_B, \|\cdot\|_\infty, \varepsilon) \le p\log\left(1 + \frac{BC\sqrt{p}}{\varepsilon}\right),$$

where $p = \binom{n+d}{d}$.

Proof. Denote the set of coefficients by $\mathcal{C} := \{c = (c_\alpha, |\alpha|\le n)\in\mathbb{R}^p$ with $\sum_{|\alpha|\le n}|c_\alpha|^2 \le B\}$. Note that

$$\mathcal{N}(\mathcal{C}, \|\cdot\|_2, \varepsilon) \le \left(1 + \frac{2B}{\varepsilon}\right)^p,$$

Let $f_c, f_{c'} \in \mathcal{F}_B$. Then

$$\|f_c - f_{c'}\|_\infty \le \sqrt{p}\max_{|\alpha|\le n}\|\psi_\alpha\|_\infty \|c-c'\|_2 \le C^d\sqrt{p}\|c-c'\|_2,$$

where C is some absolute constant. Hence,

$$\mathcal{N}_{[]}(\mathcal{F}_B, \|\cdot\|_\infty, \varepsilon) \le \left(1 + \frac{4BC\sqrt{p}}{\varepsilon}\right)^p$$

□

Proposition 10. *Let H_n be the* physicists' Hermite polynomials *defined by the generating function*

$$\sum_{n=0}^{\infty} H_n(x)\,\frac{t^n}{n!} = e^{2xt-t^2}, \qquad x,t\in\mathbb{R}.$$

Define the (normalized) Hermite functions

$$\psi_n(x) := \frac{1}{(2^n n!\sqrt{\pi})^{1/2}}\,H_n(x)\,e^{-x^2/2}, \qquad x\in\mathbb{R},\ n\in\mathbb{N}_0, \tag{61}$$

and the creation operator

$$a^* := \frac{1}{\sqrt{2}}\Big(x-\frac{d}{dx}\Big).$$

Then for every $n\geq 1$,

$$a^*\psi_{n-1} = \sqrt{n}\,\psi_n, \qquad \textit{equivalently} \qquad \psi_n = \frac{1}{\sqrt{n}}\,a^*\psi_{n-1}. \tag{62}$$

Proof. From the generating function, differentiate with respect to x:

$$\sum_{n=0}^{\infty} H_n'(x)\,\frac{t^n}{n!} = \partial_x\big(e^{2xt-t^2}\big) = 2t\,e^{2xt-t^2} = 2t\sum_{n=0}^{\infty} H_n(x)\,\frac{t^n}{n!} = \sum_{n=1}^{\infty} 2n\,H_{n-1}(x)\,\frac{t^n}{n!}.$$

Comparing coefficients of t^n gives

$$H_n'(x) = 2n\,H_{n-1}(x), \qquad n\geq 1. \tag{63}$$

Next, differentiate the generating function with respect to t:

$$\sum_{n=0}^{\infty} H_{n+1}(x)\,\frac{t^n}{n!} = \partial_t\big(e^{2xt-t^2}\big) = (2x-2t)e^{2xt-t^2} = 2x\sum_{n=0}^{\infty} H_n(x)\,\frac{t^n}{n!} - 2\sum_{n=0}^{\infty} H_n(x)\,\frac{t^{n+1}}{n!}.$$

Rewrite the last term as $\sum_{n=0}^{\infty} 2n\,H_{n-1}(x)\,\frac{t^n}{n!}$ and compare coefficients:

$$H_{n+1}(x) = 2x\,H_n(x) - 2n\,H_{n-1}(x), \qquad n\geq 1. \tag{64}$$

Using the product rule,

$$\frac{d}{dx}\big(H_{n-1}(x)e^{-x^2/2}\big) = H_{n-1}'(x)e^{-x^2/2} - xH_{n-1}(x)e^{-x^2/2}.$$

Hence

$$\begin{aligned}\Big(x-\frac{d}{dx}\Big)\big(H_{n-1}e^{-x^2/2}\big) &= xH_{n-1}e^{-x^2/2} - \Big(H_{n-1}'e^{-x^2/2} - xH_{n-1}e^{-x^2/2}\Big)\\ &= \big(2xH_{n-1} - H_{n-1}'\big)e^{-x^2/2}.\end{aligned} \tag{65}$$

Now use (63) with $n-1$:

$$H_{n-1}'(x) = 2(n-1)H_{n-2}(x).$$

Insert this into (65) and apply (64) with index $n-1$:

$$2xH_{n-1} - H_{n-1}' = 2xH_{n-1} - 2(n-1)H_{n-2} = H_n.$$

Therefore we have shown

$$\Big(x-\frac{d}{dx}\Big)\big(H_{n-1}(x)e^{-x^2/2}\big) = H_n(x)e^{-x^2/2}. \tag{66}$$

Let

$$N_n := (2^n n!\sqrt{\pi})^{-1/2}, \quad \text{so that} \quad \psi_n = N_n\,H_n\,e^{-x^2/2}.$$

Using (66),

$$a^*\psi_{n-1} = \frac{1}{\sqrt{2}}\Big(x-\frac{d}{dx}\Big)\big(N_{n-1}H_{n-1}e^{-x^2/2}\big) = \frac{N_{n-1}}{\sqrt{2}}\,H_n\,e^{-x^2/2}.$$

It remains to compare $\frac{N_{n-1}}{\sqrt{2}}$ with $\sqrt{n}\,N_n$:

$$\frac{N_{n-1}}{\sqrt{2}} = \frac{1}{\sqrt{2}}\,(2^{n-1}(n-1)!\sqrt{\pi})^{-1/2} = (2^n(n-1)!\sqrt{\pi})^{-1/2},$$

and

$$\sqrt{n}\,N_n = \sqrt{n}\,(2^n n!\sqrt{\pi})^{-1/2} = (2^n(n-1)!\sqrt{\pi})^{-1/2}.$$

Thus $\frac{N_{n-1}}{\sqrt{2}} = \sqrt{n}\,N_n$, and consequently

$$a^*\psi_{n-1} = \sqrt{n}\,N_n\,H_n\,e^{-x^2/2} = \sqrt{n}\,\psi_n,$$

which is (62). □

C.3 Proof of Theorem 2

Using $\mathcal{R}(g^\star) = 0$ and Corollary 5.1 we get

$$\inf_{g\in\mathcal{G}} \mathcal{R}(g) \le L_\ell\big(1 + L_{\mathcal{C},2}\big) \inf_{g\in\mathcal{G}} \|g - g^\star\|_{L^2(\rho_T)}\,. \tag{67}$$

Let $\{\psi_\alpha^{(\lambda)}\}_{\alpha\in\mathbb{N}_0^d}$ be the scaled Hermite basis with $\lambda = T^{-1/2}$. Let Π_n be the $L^2(\mathbb{R}^d)$–orthogonal projector onto $\operatorname{span}\{\psi_\alpha^{(\lambda)} : |\alpha| \le n\}$. Define

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

We choose $B \asymp nd^d$. By Proposition 7, there exist some constant $C > 0$ such that for $n \ge C\log\max(M,N)$, $\Pi_n g^\star \in \mathcal{G}$, and

$$\|g^\star - \Pi_n g^\star\|_{L^2(\mathbb{R}^d)} \lesssim \left(N^{-1/2} + M^{-1/2}\right) \log^{d/2}\left(\max\{M,N\}\right)\,. \tag{68}$$

Note that by Proposition 3, $g^\star(y) \in \left[c_-^\star e^{-a_-^\star\|y\|^2}, c_+^\star\right] := I$ for all y by (20). We need to ensure condition (16). Define the clipping operator $\Pi^{\mathrm{clip}} : \mathbb{R} \to \mathbb{R}$ by

$$\Pi^{\mathrm{clip}}(t) := \min\left\{c_+^\star, \max\left\{t, c_-^\star e^{-a_-^\star\|y\|^2}\right\}\right\}.$$

Then for any $f \in \mathcal{G}$, $\Pi^{\mathrm{clip}}(f)(y)$ satisfies (16). For each fixed y, the map $t \mapsto \Pi(t)$ is the metric projection onto the interval I, hence it is 1–Lipschitz. Applying this gives the pointwise contraction for any $y \in \mathbb{R}^d$,

$$|g^\star(y) - \Pi^{\mathrm{clip}}(\Pi_n g^\star(y))| = \left|\Pi^{\mathrm{clip}}(g^\star(y)) - \Pi^{\mathrm{clip}}(\Pi_n g^\star(y))\right| \le |g^\star(y) - \Pi_n g^\star(y)|,$$

In particular, clipping can only *decrease* any $L^2(\mathbb{R}^d)$–error:

$$\|g^\star(y) - \Pi^{\mathrm{clip}}(\Pi_n g^\star(y))\|_{L^2(\mathbb{R}^d)} \le \|g^\star(y) - \Pi_n g^\star(y)\|_{L^2(\mathbb{R}^d)}\,.$$

It follows from the last inequality, (67) and (68) that

$$\inf_{g\in\mathcal{G}} \mathcal{R}(g) \lesssim \left(N^{-1/2} + M^{-1/2}\right) \log^{d/2}\left(\max\{M,N\}\right)\,.$$

Note that for chosen n, Lemma 9 and Proposition 6 imply that there exists some constant $C' > 0$ such that

$$\log \mathcal{N}_{[]}\Big(\mathcal{G}, \|\cdot\|_\infty, \varepsilon\Big) \lesssim \log^{d/2}\left(\max\{M,N\}\right)\log(C'/\varepsilon)\,.$$

Hence, we may conclude the statement by applying Theorem 1.

D Further Details of Numerical Experiments

In this section, we elaborate on details of numerical experiments presented in Section 4. The section is organized as follows. In Appendix D.1, we discuss general implementation details. Appendix D.2 presents information about the metrics used and describes the algorithm. Appendix D.3 provides additional information about the Swiss-Roll to S-Curve two dimensional problem experiment. Appendix D.4 provides additional details about the 25 Gaussian mixture Extrapolation experiment. Appendix D.5 provides additional details about the biological data experiment. Finally, Appendix D.6 presents all final hyperparameter values used for the experiments.

D.1 General Implementation Details

In our experiments, we paid close attention to hyperparameter selection. Due to the algorithm's specifics, careful hyperparameter tuning is essential for stable training and generation. This was accomplished by using Optuna (Akiba et al., 2019). We used 50 trial optimization for Optuna over the sliced $\mathbb{W}_1$ metric in the 25 Gaussian experiment and 100 trial optimization for Optuna over the sliced $\mathbb{W}_1$ metric for Single Cell data.

The calculations were fully performed on the Nvidia T4 GPU. For the SinkhornBridge algorithm, we used the hyperparameter values provided in the official repository, as we were unable to conduct extensive testing of the algorithm and select its hyperparameters ourselves.

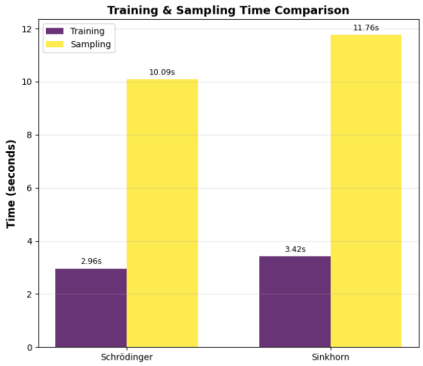

Figure 3: Comparison of training and sampling times for ERM-Bridge and SinkhornBridge on the Swiss-Roll to S-Curve translation task. We used 2000 training points for both algorithms.

D.2 Metrics Used in the Experiments

Sliced Wasserstein Distance was used as main metric since it provides quantitative, theoretically-grounded measures of distribution similarity and has been widely adopted in research papers and practical applications.

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

where $\{\theta_n\}_{n=1}^{N}$ are random projections uniformly distributed on $\mathbb{S}^{d-1}$.

For the Sliced Wasserstein Distance, 100 random projections were used.

We also provide a description of the training algorithm for the proposed algorithm training 1 and sampling 2. The pseudocode provided allows a reader to visually evaluate how our approach differs from SinkhornBridge.

D.3 Details of Evaluation on the Swiss-Roll to S-Curve Experiment

In this experiment, we validated the algorithm's performance and sampling results at various intermediate time points on a common two dimensional problem in generative modelling. Optuna did not perform hyperparameter optimization for this problem, as it is not complex and both algorithms solve it well for any reasonable choice of training and sampling parameters.

We also compared the running time of our algorithm and SinkhornBridge on this problem, demonstrating significant sampling speedup.

D.4 Details of Evaluation on the Gaussian Mixture

In this experiment, we aimed to clearly demonstrate the behavior of our proposed algorithm and SinkhornBridge with explicit data bias in the sampling dataset relative to the train dataset. To this end, we created a synthetic example using a uniform grid of 25 Gaussians in two-dimensional space and a truncated Normal distribution of $[-10, 10]$ in the train dataset and $[-1, 1], [-3, 3], [-5, 5], [-10, 10]$ in test.

The problems arising from the discrete nature of SinkhornBridge become apparent upon visual inspection of the plots. The value of the experiment is that similar problems demonstrated in the synthetic example also arise with real data. For example, our algorithm demonstrated higher quality on Single Cell data D.5, where both the initial and final distributions are only available as samples.

D.5 Details of Evaluation on the Single Cell data

In this experiment, we aimed to demonstrate the performance of our algorithm on the data-to-data translation task. The continuous log potential not only yielded the best metric value but was also easier to optimize, requiring less sampling time.

The results for the LightSB model were taken from the paper (Korotin et al., 2024).

D.6 Final Hyperparameter Values

Below are the final hyperparameter values for all experiments.

Swiss-Roll to S-Curve Experiment:

- Parameters for experiment: batch size $= 1000$, lr $= 2 \cdot 10^{-3}$, epochs $= 1500$, σ_end $= 0.5$, loss_scale $= 1.0$, hidden_dim $= 64$.

Gaussian Mixture Experiment:

- Parameters for experiment: batch size $= 64$, lr $= 5 \cdot 10^{-4}$, epochs $= 140$, σ_end $= 0.9$, loss_scale $= 0.11$ hidden_dim $= 128$.

Single Cell Experiment:

- Parameters for experiment: batch size $= 2048$, lr $= 10^{-4}$, epochs $= 141$, σ_end $= 0.4216$, loss_scale $= 196.5431$, hidden_dim $= 2048$.

Algorithm 1 Training of ERM-Bridge

Input: Datasets $\mathcal{X}, \mathcal{Y}$, kernel bandwidth σ, learning rate η, batch size B, the number of iterations N_{steps}
Initialize: Neural network parameters θ for ϕ_θ

for i from 1 to N_{steps} **do**
  Sample batches $X_b \sim \mathcal{X}$ and $Y_b \sim \mathcal{Y}$

  for $y \in Y_b$ **do**
    $V(y) \leftarrow \log \phi_\theta(y)$
  end for

  for $x \in X_b$ **do**
    $\text{term}(y) \leftarrow -\|x - y\|^2/(2\sigma^2) - V(y) \quad$ for all $y \in Y_b$
    $\log D(x) \leftarrow \log \sum_{y \in Y_b} \exp(\text{term}(y))$
  end for

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

  $\Delta(y) \leftarrow V(y) - \log(C\phi)(y)$
  $\mathcal{L}(\theta) \leftarrow \frac{1}{|Y_b|} \sum_{y \in Y_b} (\Delta(y) - \text{mean}(\Delta))^2$

  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$
end for
Output: Learned potential parameters θ^*

Algorithm 2 Sampling via Learned Continuous Drift

Input: Trained model ϕ_{θ^*}, Initial sample x_0, Reference target samples $\mathcal{Y}_{ref}$, Total time T, Steps K
Initialize: $x \leftarrow x_0$, $t \leftarrow 0$, $\Delta t \leftarrow T/K$

for $k = 0$ to $K - 1$ **do**
  Set noise level σ_t (e.g., via cosine schedule)
  $\nu_t \leftarrow \sigma_t^2(T - t)$

  for $j = 1 \dots |\mathcal{Y}_{ref}|$ **do**
    $L_j \leftarrow -\frac{\|x - y_j\|^2}{2\nu_t} - \log \phi_{\theta^*}(y_j)$
  end for
  $h(x) \leftarrow \log \sum_j \exp(L_j)$
  $g \leftarrow \nabla_x h(x)$
  $u(x, t) \leftarrow \sigma_t^2 \, g$

  Sample noise $\xi \sim \mathcal{N}(0, I)$
  $x \leftarrow x + u(x, t)\Delta t + \sigma_t \sqrt{\Delta t}\, \xi$
  $t \leftarrow t + \Delta t$
end for
Return: Transported sample x_T