Conditional simulation via entropic optimal transport: Toward non-parametric estimation of conditional Brenier maps

Ricardo Baptista[1,⋆], Aram-Alexandre Pooladian[2,⋆], Michael Brennan[3], Youssef Marzouk[3], Jonathan Niles-Weed[2,4]

[1]California Institute of Technology
[2]Center for Data Science, New York University
[3]Massachusetts Institute of Technology
[4]Courant Institute of Mathematical Sciences, New York University

November 12, 2024

Abstract

Conditional simulation is a fundamental task in statistical modeling: given a finite collection of samples from a joint distribution, it consists in generating samples from conditionals of this distribution. One promising approach is to construct conditional Brenier maps, where the components of the map push forward a reference distribution to conditionals of the target. While many estimators exist, few, if any, come with statistical or algorithmic guarantees. To this end, we propose a non-parametric estimator for conditional Brenier maps based on the computational scalability of *entropic* optimal transport. Our estimator leverages a result of Carlier et al. (2010), which shows that optimal transport maps under a rescaled quadratic cost asymptotically converge to conditional Brenier maps; our estimator is precisely the entropic analogues of these converging maps. We provide heuristic justifications for choosing the scaling parameter in the cost as a function of the number of samples by fully characterizing the Gaussian setting. We conclude by comparing the performance of the estimator to other machine learning and non-parametric approaches on benchmark datasets and Bayesian inference problems.

1 Introduction

Given access to i.i.d. samples from a joint distribution $\mu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$, our goal is to generate samples distributed according to the conditional distribution $\mu_{2|1}(\cdot|x_1) \coloneqq \mu(\cdot, x_1)/\mu_1(x_1)$ (where μ_1 is the first marginal of μ) for any $x_1 \in \mathbb{R}^{d_1}$. This sampling problem lies at the core of computational Bayesian inference, where the joint distribution

⋆Equal contribution. Correspondance to rsb@caltech.edu and ap6599@nyu.edu.

is specified by a model for observations X_1 and parameters X_2. The goal of simulation-based Bayesian inference is to sample from the posterior distribution, i.e., the conditional distribution $\mu_{2|1}(\cdot|x_1)$ corresponding to an observation x_1, given samples from the joint distribution (Cranmer et al., 2020).*

Sampling via transport. One natural framework for performing conditional simulation uses *measure transport* (Marzouk et al., 2016) by seeking a transport map that pushes forward a known source distribution to the conditional $\mu_{2|1}(\cdot|x_1)$ for any realization of the conditioning variable x_1.

Many transport approaches for conditional simulation obey the following construction: Let $\rho = \rho_1 \otimes \rho_2$ be a tensor-product source measure that is easy to sample from (e.g., the standard Gaussian) and consider a transport map T from ρ to μ of the form

$$T(x) = \begin{bmatrix} T^1(x_1) \\ T^2(x_2; x_1) \end{bmatrix}, \tag{1}$$

Theorem 2.4 of Baptista et al. (2024) shows that T^1 transports ρ_1 to μ_1, then T^2 transports ρ_2 to $\mu_{2|1}$. In particular, for ρ_1-a.e. x_1,

$$T^2(X_2; x_1) \sim \mu_{2|1}(\cdot|T^1(x_1)), \quad X_2 \sim \rho_2 . \tag{2}$$

Thus, maps of the form in (1) can perform conditional simulation. Several methods successfully learn these maps given only information from the joint distribution μ. These include conditional normalizing flows and diffusion models; see Section 6 for some examples. However, most existing approaches (i) do not provide explicit guarantees in terms of the number of samples required to obtain a good estimator, and (ii) are based on parametric models, the successful use of which requires tuning an unreasonable number of parameters. The latter is especially costly when using neural networks as the proposed estimator.

Conditional Brenier maps. Among all maps of the form in (1) that perform conditional simulation, we seek the unique transport whose component maps T^1, T^2 minimize the squared Euclidean displacement cost. We call such a transport a *conditional Brenier map*, denoted T_{CB}. A theoretical approximation scheme for finding T_{CB} was proposed by Carlier et al. (2010): For $t \in (0, 1)$, define the positive definite matrix

$$A_t \coloneqq \mathrm{diag}(\mathbf{1}_{d_1}, \sqrt{t}\mathbf{1}_{d_2}), \tag{3}$$

where $\mathbf{1}_{d_1} \coloneqq (1, \ldots, 1) \in \mathbb{R}^{d_1}$, with $d = d_1 + d_2$, and consider the weighted Euclidean cost function

$$c_t(x, y) \coloneqq \tfrac{1}{2}\|A_t(x - y)\|_2^2 . \tag{4}$$

*Joint data in this setting are cheaply obtained by sampling x_2 from its marginal (prior) distribution μ_2 and x_1 from the specified likelihood model $\mu_{1|2}(\cdot|x_2)$.

Carlier et al. (2010) show[†] the following:

Theorem 1.1 (Convergence to conditional Brenier maps). *Let $\rho, \mu \in \mathcal{P}_2(\Omega)$ with ρ having a density. Let T_t denote the corresponding optimal transport map for the cost c_t satisfying $(T_t)_\sharp \rho = \mu$. Then as $t \to 0$,*

$$\Delta(T_t, T_{\mathrm{CB}}) := \|T_t - T_{\mathrm{CB}}\|^2_{L^2(\rho)} \to 0\,.$$

Theorem 1.1 states that, in order to approximate conditional Brenier maps, it suffices to learn optimal transport maps pertaining to a rescaled quadratic cost function, namely c_t in (4). While this result is non-quantitative, it provides an optimal transport framework for conditional simulation.

Main contributions

We propose an entropic estimator for conditional Brenier maps based on the works of Pooladian and Niles-Weed (2021) and Carlier et al. (2010). In fact, we propose a general framework that allows one to use *any* estimator of the (optimal) transport map between two measures.

In addition, the risk incurred by any finite-sample estimator we propose, written $\widehat{T}_t$, has the following decomposition

$$\mathbb{E}\Delta(\widehat{T}_t, T_{\mathrm{CB}}) \lesssim \mathbb{E}\Delta(\widehat{T}_t, T_t) + \Delta(T_t, T_{\mathrm{CB}}). \tag{5}$$

We take steps toward quantifying (5) by providing a full characterization of the two error terms in the Gaussian setting. For non-Gaussian measures, we numerically evaluate the performance of the conditional entropic Brenier map relative to two baseline approaches, one based on the nearest-neighbor estimator (Manole et al., 2021) and another based on neural networks (Amos, 2022). We show that our estimator is more tractable than these other approaches and requires less tuning to maintain performance on standard conditional simulation tasks.

Notation

For $\Omega \subseteq \mathbb{R}^d$, we write $\mathcal{P}_2(\Omega)$ as the set of probability measures over Ω with finite second moment. Throughout, we write $\rho \in \mathcal{P}_2(\Omega)$ as the source measure which will always have a density with respect to Lebesgue measure over a (compact convex) set $\Omega \subseteq \mathbb{R}^d$. We denote the target measure by $\mu \in \mathcal{P}_2(\Omega)$ for the target measure. For $\Omega \subseteq \mathbb{R}^{d_1+d_2}$, we explicitly write the joint density as $\rho(x_1, x_2)$, and use $\rho_1(x_1) \in \mathcal{P}(\mathbb{R}^{d_1})$ (resp. $\rho_2(x_2) \in \mathcal{P}(\mathbb{R}^{d_2})$) to denote the first (resp. second) marginal distribution. For

[†]Their original result shows convergence of T_t to the Knothe-Rosenblatt rearrangement, a map whose Jacobian is *strictly-triangular*, by considering the weighted Euclidean cost $c_t(x, y) = \sum_{k=1}^d \frac{1}{2} t^{k-1} |x_k - y_k|^2$. The proof shows that each component of T_t converges to an optimal map between one-dimensional conditional distributions. The same argument applies block-wise to x_1, x_2 to yield the map of the form in (2) whose Jacobian is *block-triangular*.

a fixed $x \in \mathbb{R}^{d_1}$, we write $\rho_{2|1}(\cdot|x)$ to be the conditional distribution; we use the same convention for $\mu = \mu(y_1, y_2)$. We denote the push-forward constraint for a transport map T satisfying $T(X) \sim \mu$ for $X \sim \rho$ as $T_\sharp \rho = \mu$. For a positive definite matrix $C \in \mathbb{R}^{d_1+d_2}$, we write $C^{1/2}$ as the symmetric square-root of C and L_C as the block-lower Cholesky decomposition of C: the 2×2-block matrix (with blocks of size $d_1 \times d_1$ and $d_2 \times d_2$) which satisfies $\mathsf{L}_C \mathsf{L}_C^\top = C$.

2 Background

2.1 Optimal transport

We first recall some facts about optimal transport (Villani, 2009; Santambrogio, 2015) for (weighted) squared-Euclidean cost functions of the form $c(x, y) = \frac{1}{2}\|A(x - y)\|^2$ with $A \succ 0$.

We use three formulations of optimal transport (OT) between fixed measures $\rho, \mu \in \mathcal{P}_2(\Omega)$ for such costs. First, the *Monge* formulation is

$$\mathrm{OT}_0(\rho, \mu) = \inf_{T \in \mathcal{T}(\rho,\mu)} \int \tfrac{1}{2}\|A(x - T(x))\|^2 \,\mathrm{d}\rho(x)\,, \tag{6}$$

where $\mathcal{T}(\rho, \mu)$ is a family of vector-valued *transport maps* satisfying $X \sim \rho$, $T(X) \sim \mu$. For general ρ and μ, it is easy to see that such a minimizer need not exist, even for such nice cost functions. When they exist, we denote these minimizer by T_0, called a *Brenier map* (Brenier, 1991).

Next, the *primal Kantorovich* problem (Kantorovitch, 1942) is

$$\mathrm{OT}_0(\rho, \mu) = \inf_{\pi \in \Pi(\rho,\mu)} \iint \tfrac{1}{2}\|A(x - y)\|^2 \,\mathrm{d}\pi(x, y)\,, \tag{7}$$

where $\Pi(\rho, \mu)$ is the set of couplings between ρ and μ, the set of joint measures with left-marginal ρ and right-marginal μ. Since ρ and μ have finite second moment, (7) is well-posed, and a minimizer, called the *optimal plan*, always exists and is denoted by π_0.

Finally, the *dual Kantorovich* optimization problem is

$$\mathrm{OT}_0(\rho, \mu) = \sup_{(f,g) \in L^1(\rho \otimes \mu)} \int f \,\mathrm{d}\rho + \int g \,\mathrm{d}\mu\,, \tag{8}$$

which is subject to the constraint

$$f(x) + g(y) \le \tfrac{1}{2}\|A(x - y)\|^2 \quad \forall (x, y) \in \Omega \times \Omega\,. \tag{9}$$

Under the same regularity conditions on the measures, a pair of maximizers (f_0, g_0) exists, which we call optimal *Kantorovich potentials*.

These three formulations are reconciled in the following theorem, which gives an explicit formula of the optimal transport map as a function of the optimal Kantorovich

potential f_0, and says that π_0 is a deterministic function of T_0. The result for $A = I$ was proven by Brenier (1991), and generalized to a wide family of cost functions by Gangbo and McCann (1996). We present a simplified version of the statement here.

Theorem 2.1 (Brenier's theorem for rescaled quadratic costs)**.** *Suppose ρ has a density with respect to the Lebesgue measure, and μ has finite second moment. Then, for costs $c(x, y) = \frac{1}{2}\|x - y\|^2$, there exists a unique optimal transport map $T_0 \in \mathcal{T}(\rho, \mu)$ that minimizes (6) given by*

$$T_0(x) = \nabla\varphi_0(x)\,, \tag{10}$$

where $\varphi_0 = \frac{1}{2}\|\cdot\|^2 - f_0$ and f_0 is the optimal Kantorovich potential. Moreover, for costs $c(x, y) = \frac{1}{2}\|A(x - y)\|^2$ with positive definite A, the optimal transport map between ρ and μ is given by

$$T_A(x) = A^{-1}\widetilde{T}_0(Ax)\,, \tag{11}$$

where $\widetilde{T}_0$ is the Brenier map between the transformed measures $\rho_A \coloneqq (A\cdot)_\sharp\rho$ and $\mu_A \coloneqq (A\cdot)_\sharp\mu$.

2.2 Entropic optimal transport

Entropic regularization was initially introduced to improve computational tractability of the matching problem; see Cuturi (2013); Peyré and Cuturi (2019) for computational insights, and Genevay (2019) for a general overview in machine learning.

For $\varepsilon > 0$, the primal entropic optimal transport problem amounts to adding the KL divergence as a regularizer to the primal Kantorovich problem:

$$\mathrm{OT}_\varepsilon(\rho, \mu) \coloneqq \inf_{\pi\in\Pi(\rho,\mu)} \iint \tfrac{1}{2}\|A(x - y)\|^2\,\mathrm{d}\pi(x, y) + \varepsilon\mathrm{KL}(\pi\|\rho\otimes\mu)\,, \tag{12}$$

where we recall that when π has a density with respect to $\rho\otimes\mu$, we have

$$\mathrm{KL}(\pi\|\rho\otimes\mu) = \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}(\rho\otimes\mu)}\right)\mathrm{d}\pi\,,$$

and otherwise has value $+\infty$. Due to the regularizer, (12) is a strictly convex problem and admits a unique minimizer, called the *optimal entropic plan*, written π_ε.

Analogously, we have a dual formulation, written

$$\begin{aligned}\mathrm{OT}_\varepsilon(\rho, \mu) = \sup_{(f,g)\in L^1(\rho\otimes\mu)} &\int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu + \varepsilon \\ &- \varepsilon\iint e^{(f(x)+g(y)-\frac{1}{2}\|A(x-y)\|^2)/\varepsilon}\,\mathrm{d}\rho(x)\,\mathrm{d}\mu(y)\,.\end{aligned} \tag{13}$$

As $\varepsilon \to 0$, the third term in (13) converges to the hard-constraint in (9). The maximizers of (13) are called *optimal entropic Kantorovich potentials*, written $(f_\varepsilon, g_\varepsilon)$.

In addition, the entropic optimal transport problem exhibits the following *primal-dual* recovery formula (Csiszár, 1975), in which the dual variables give an explicit form for the primal solution via

$$\mathrm{d}\pi_\varepsilon(x,y) = e^{(f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|A(x-y)\|^2)/\varepsilon}\,\mathrm{d}\rho(x)\,\mathrm{d}\mu(y). \tag{14}$$

We can extend the potentials $(f_\varepsilon, g_\varepsilon)$ to lie outside the support of ρ and μ respectively, by appealing to the marginal constraints (see Mena and Niles-Weed (2019); Nutz and Wiesel (2021)). Henceforth, we write

$$\begin{aligned} f_\varepsilon(x) &= -\varepsilon \log \int e^{(g_\varepsilon(y)-\frac{1}{2}\|A(x-y)\|^2)/\varepsilon}\,\mathrm{d}\mu(y) \quad (x \in \mathbb{R}^d) \\ g_\varepsilon(y) &= -\varepsilon \log \int e^{(f_\varepsilon(x)-\frac{1}{2}\|A(x-y)\|^2)/\varepsilon}\,\mathrm{d}\rho(x) \quad (y \in \mathbb{R}^d)\,. \end{aligned}$$

2.2.1 Entropic analogues to Brenier's theorem

As before, consider the special case $c(x,y) = \frac{1}{2}\|x-y\|^2$. We can write $\varphi_\varepsilon = \frac{1}{2}\|\cdot\|^2 - f_\varepsilon$ and $\psi_\varepsilon = \frac{1}{2}\|\cdot\|^2 - g_\varepsilon$. Appealing to the expression given by the extended potentials, we have

$$\varphi_\varepsilon(x) = \varepsilon \log \int e^{(x^\top y - \psi_\varepsilon(y))/\varepsilon}\,\mathrm{d}\mu(y)\,. \tag{15}$$

Then, the *entropic Brenier map* is given by

$$T_\varepsilon(x) \coloneqq \nabla\varphi_\varepsilon(x) = \mathbb{E}_{\pi_\varepsilon}[Y|X=x]\,, \tag{16}$$

where the second equality follows from taking the gradient of (15).[‡] This explicit characterization of the entropic map (relating the gradient field to the barycentric projection of π_ε) was introduced as a means of providing a computationally tractable estimator to the Brenier map (Pooladian and Niles-Weed, 2021).

For costs of the form $\frac{1}{2}\|A(x-y)\|^2$, we arrive at a similar formula to (11) as in the unregularized case (cf. Klein et al. (2023, Lemma 2))

$$T_{A,\varepsilon}(x) = A^{-1}\widetilde{T}_\varepsilon(Ax)\,, \tag{17}$$

where $\widetilde{T}_\varepsilon$ is the entropic Brenier map between the transformed measures $\rho_A \coloneqq (A\cdot)_\sharp\rho$ and $\mu_A \coloneqq (A\cdot)_\sharp\mu$.

[‡]This is permitted by an application of the dominated convergence theorem.

2.2.2 Computational considerations with and without regularization

Given i.i.d. samples $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_n \sim \mu$, the primal Kantorovich problem can be computed by solving the linear program

$$\widehat{P} \coloneqq \operatorname*{argmin}_{P \in \mathsf{DS}_n} \langle C, P \rangle \,, \tag{18}$$

where DS_n is the set of doubly-stochastic $n \times n$ matrices (i.e., matrices with non-negative entries where the row and columns each sum to one), and $C_{ij} \coloneqq \frac{1}{2}\|A(X_i - Y_j)\|^2$. The runtime for solving this linear program is well-known to be $\mathcal{O}(n^3 \log(n))$, with the caveat that the cost matrix $C \in \mathbb{R}^{n \times n}$ must be stored in memory; see Peyré and Cuturi (2019, Chapter 3).

Incorporating entropic regularization has the benefit of improved runtime complexity. Given n samples as before, we now solve the strongly convex program

$$\widehat{P}_\varepsilon \coloneqq \operatorname*{argmin}_{P \in \mathsf{DS}_n} \langle C, P \rangle + \varepsilon H(P) \,, \tag{19}$$

where $H(P)$ is the discrete entropy of the matrix P. The most well-known algorithm for computing (19) is *Sinkhorn's algorithm* (Sinkhorn, 1964; Cuturi, 2013). Solving for $\widehat{P}_\varepsilon$ entails an iterative approach, where we solve for discrete analogues of the optimal dual variables $(\hat{f}_\varepsilon, \hat{g}_\varepsilon) \in \mathbb{R}^n \times \mathbb{R}^n$. By way of (14), this gives

$$(\widehat{P}_\varepsilon)_{ij} = e^{(\hat{f}_\varepsilon)_i/\varepsilon} e^{C_{ij}/\varepsilon} e^{(\hat{g}_\varepsilon)_j/\varepsilon} \,. \tag{20}$$

In this setting, is it known that for a given $\varepsilon > 0$, the matrix $\widehat{P}_\varepsilon$ can be computed in $\widetilde{\mathcal{O}}(n^2/(\varepsilon\delta))$ time, where $\delta > 0$ is the desired tolerance to optimality (Altschuler et al., 2017). For n sufficiently large, the quadratic runtime complexity is a significant improvement over the cubic complexity of the unregularized solver.

While the above approach is popular, one can also solve for the dual optimal vectors $(\hat{f}_\varepsilon, \hat{g}_\varepsilon) \in \mathbb{R}^n \times \mathbb{R}^n$ without storing the cost matrix $C \in \mathbb{R}^{n \times n}$. This results in a computationally feasible method for $n \gtrsim 10^5$. We refer to Peyré and Cuturi (2019, Chapter 4) for more details.

3 A family of estimators for conditional simulation

Given $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_n \sim \mu$, our three-step recipe to estimate conditional Brenier maps T_{CB} is summarized as follows: For fixed $t > 0$, let A_t be given by (3). Then

1. scale the data to obtain $\mathbb{X}_t \coloneqq (A_t X_1, \dots, A_t X_n)$ and $\mathbb{Y}_t \coloneqq (A_t Y_1, \dots, A_t Y_n)$,

2. learn an estimator $\widehat{T}_0$ of the Brenier map from the data $\mathbb{X}_t$ to $\mathbb{Y}_t$,

3. for a new sample $x \sim \rho$, unscale the pushforward map, resulting in

$$\widehat{T}_t(x) \coloneqq A_t^{-1} \widehat{T}_0(A_t x)\,. \tag{21}$$

This general approach allows practitioners to consider any estimator for the optimal transport map $\widehat{T}_0$ on the basis of data. We now discuss some estimators $\widehat{T}_t$ in detail, and comment on their scalability and practicality.

3.1 Approach 1: Conditional entropic Brenier maps

Our main estimator is based on the entropic Brenier map introduced by Pooladian and Niles-Weed (2021). Let $(\hat{f}_{\varepsilon,t}, \hat{g}_{\varepsilon,t})$ be the output of Sinkhorn's algorithm on the scaled data $\mathbb{X}_t$ and $\mathbb{Y}_t$. Following (16) and (17), the estimator is

$$\widehat{T}_{\varepsilon,t}(x) = \sum_{i=1}^n Y_i \frac{e^{((\hat{g}_{\varepsilon,t})_i - \frac{1}{2}\|A_t(x-Y_i)\|^2)/\varepsilon}}{\sum_{j=1}^n e^{((\hat{g}_{\varepsilon,t})_j - \frac{1}{2}\|A_t(x-Y_j)\|^2)/\varepsilon}}\,. \tag{22}$$

This non-parametric estimator can be evaluated in $\mathcal{O}(n)$ time, and estimated in $\mathcal{O}(n^2/\varepsilon)$ runtime. More importantly, as Sinkhorn's algorithm is GPU-friendly, this estimator is scalable to over $n \gtrsim 10^5$ sample points.

3.2 Approach 2: Nearest-Neighbor estimator

Another non-parametric estimator is due to Manole et al. (2021) which was adopted by Hosseini et al. (2023) for the purpose of conditional simulation. Let $\widehat{P}_t$ be the discrete optimal transport matching computed on the scaled data $\mathbb{X}_t$ and $\mathbb{Y}_t$. For $x \in \mathbb{R}^d$, this estimator is

$$\widehat{T}_{\mathrm{NN},t}(x) = \sum_{i=1}^n \mathbf{1}_{V_i}(A_t x) Y_{\widehat{P}_t(i)}\,, \tag{23}$$

where $(V_i)_{i=1}^n$ are Voronoi regions

$$V_i \coloneqq \{x \in \mathbb{R}^d \colon\ \|x - A_t X_i\| \le \|x - A_t X_k\| \,\forall\, k \neq i\}\,.$$

Computing the closest X_i to a new sample x has runtime $\mathcal{O}(n \log(n))$, though the overall runtime of the estimator is still $\mathcal{O}(n^3)$, since $\widehat{P}_t$ needs to be initially computed (recall the discussions in Section 2.2.2). An important caveat to this estimator is that it is only defined on the *in-sample* target points.

3.3 Approach 3: Neural networks

Neural networks are widely used to estimate optimal transport maps on the basis of data, and fall under our proposed framework as well. We follow the work of Amos (2022), which trains two multilayer perceptrons (MLPs) neural networks φ_θ and x_ϑ to define the map. In brief, the network $x_\vartheta : \mathbb{R}^d \to \mathbb{R}^d$ models the reverse transport map and $\varphi_\theta : \mathbb{R}^d \to \mathbb{R}$ models the Brenier potential. We can fit these neural networks using stochastic gradient descent on minibatches of our fixed data $\mathbb{X}_t$ and $\mathbb{Y}_t$, and express our estimator as

$$\widehat{T}_{\mathrm{MLP},t}(x) = A_t^{-1} \nabla \widehat{\varphi}_\theta(A_t x)\,. \tag{24}$$

We provide more details on the training algorithm in Appendix A.1, though we refer the interested reader to Amos (2022) for precise details.

4 Theory for conditional simulation: The Gaussian case

Our goal is to estimate the conditional Brenier map T_{CB} on the basis of samples. To explicitly quantify the error, we recall that the $L^2(\rho)$ risk incurred by all estimators from the preceding section can be decomposed as

$$\mathbb{E}\Delta(\widehat{T}_t, T_{\mathrm{CB}}) \lesssim \mathbb{E}\Delta(\widehat{T}_t, T_t) + \Delta(T_t, T_{\mathrm{CB}}) =: (\mathsf{T1}) + (\mathsf{T2})\,, \tag{25}$$

where we recall $\Delta(f,g) \coloneqq \|f-g\|^2_{L^2(\rho)}$. $(\mathsf{T1})$ denotes the statistical error, which depends on the method of choice, whereas $(\mathsf{T2})$ denotes the approximation error of T_t to T_{CB}. Ideally, our choice of t should decrease with the number of samples, resulting in a consistent estimator.

In this section, we make the estimates for (25) quantitative by considering a Gaussian-to-Gaussian transport problem, where we can leverage closed-form expressions for the estimators as a first step to understanding the error. All proofs in this section are deferred to Appendix B. We will require the following assumption:

(G) Let $\rho = \mathcal{N}(0, I_d)$ and $\mu = \mathcal{N}(m, \Sigma)$ for $\Sigma \succ 0$.

We first collect three closed-form expressions of interest.

Proposition 4.1 (Closed-form expressions)**.** *Suppose ρ and μ satisfy **(G)**. Let T_B be the optimal transport map under the squared-Euclidean cost, T_t be the optimal transport map for the c_t cost, and T_{CB} be the conditional Brenier map, all between ρ and μ. Then,*

$$\begin{aligned} T_B(x) &= m + \Sigma^{1/2} x\,, \\ T_t(x) &= m + A_t^{-2}(A_t^2 \Sigma A_t^2)^{1/2} x\,, \\ T_{\mathrm{CB}}(x) &= m + \mathsf{L}_\Sigma x\,. \end{aligned}$$

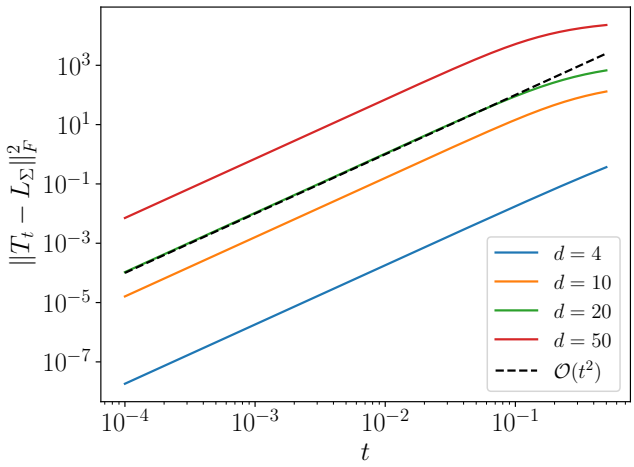

Figure 1: We observe that (T2) asymptotically converges with a rate of $\mathcal{O}(t^2)$ convergence rate with randomly generated covariance matrices of block-type.

Our main result of this section is the following theorem. The proof follows immediately from balancing the results of Proposition 4.4 (see Section 4.1) and Proposition 4.5 (see Section 4.2), and is therefore omitted.

Theorem 4.2. *Suppose ρ and μ satisfy **(G)**, and consider the following plug-in estimator*

$$\widehat{T}_t(x) = \widehat{m} + A_t^{-2}(A_t^2 \widehat{\Sigma} A_t^2)^{1/2} x \,, \tag{26}$$

where $\widehat{m}$ and $\widehat{\Sigma}$ are the empirical mean and covariance derived from the i.i.d. samples from μ. If $t(n) \asymp n^{-1/3}$, and n is sufficiently large, the plug-in estimator $\widehat{T}_t$ achieves the estimation rate

$$\mathbb{E}\|\widehat{T}_t - T_{\mathrm{CB}}\|^2_{L^2(\rho)} \lesssim n^{-2/3} \,. \tag{27}$$

Remark 4.3. It is worth mentioning that this result is far from tight. A plug-in estimator of L_Σ would directly provide parametric rates of estimation. However, this estimation approach is outside the spirit of our work, as it avoids the rescaled quadratic cost entirely.

4.1 Convergence of (T1)

Our first step is to understand how the scaling of the cost impacts the overall statistical performance of our estimator. We ultimately expect $t = t(n) \searrow 0$, which will (negatively) impact the convergence rate. Indeed, if t were ultimately fixed and not decaying with n, then the rescaling impacts the rate of estimation by at most a universal, albeit large, constant, but we would never be able to converge to the conditional Brenier map.

The following proposition tells us how the scaling of the cost impacts the statistical rates of convergence.

Proposition 4.4. *Suppose ρ and μ satisfy **(G)**, and consider the plug-in estimator from (26). The statistical rate of convergence is*

$$(\mathsf{T1}) := \mathbb{E}\|\widehat{T}_t - T_t\|^2_{L^2(\rho)} \lesssim t^{-1}n^{-1}\,. \tag{28}$$

4.2 Convergence of (T2)

We now quantify the approximation error (T2) in (25). The proof is based on a careful Taylor-expansion argument that is made tractable in the Gaussian-to-Gaussian setting.

Proposition 4.5. *Suppose ρ and μ satisfy **(G)**. Then T_t for t sufficiently small converges to the conditional Brenier map at the rate*

$$(\mathsf{T2}) := \|T_t - T_{\mathrm{CB}}\|^2_{L^2(\rho)} \lesssim_\Sigma t^2\,. \tag{29}$$

Figure 1 supports Proposition 4.5. Note that as d increases, the constant in (29) increases, though the convergence *rate* appears to always be $\mathcal{O}(t^2)$ for t sufficiently small. Details of the example are provided in Appendix A.2.

4.3 Impact of the entropic bias

Unlike the Nearest-Neighbor or the MLP estimators, the entropic Brenier map comes with an explicit bias as an artifact of the ε-regularization scheme. Here, we want to scale both parameters $\varepsilon, t \to 0$ such that the following quantity converges

$$\|T_{\varepsilon,t} - T_{\mathrm{CB}}\|^2_{L^2(\rho)} \lesssim \|T_{\varepsilon,t} - T_t\|^2_{L^2(\rho)} + \|T_t - T_{\mathrm{CB}}\|^2_{L^2(\rho)}\,. \tag{30}$$

Ultimately, we would like to provide a general rule for selecting the entropic parameter ε as a function of t. We expect $t = t(n)$ to decrease as the number of samples n increases, and similarly for $\varepsilon = \varepsilon(n)$. The following result controls the entropic bias (first term in (30)) for rescaled quadratic costs in the Gaussian setting.

Proposition 4.6. *Let $\rho = \mathcal{N}(0, A)$ and $\mu = \mathcal{N}(b, B)$, and let $c_S(x, y) := \frac{1}{2}\|S(x - y)\|^2$ for some positive-definite matrix S. Let $T_{\varepsilon,S}$ (resp. T_S) be the entropic Brenier map (resp. Brenier map) between ρ and μ. It holds that for $\varepsilon > 0$*

$$\|T_{\varepsilon,S} - T_S\|^2_{L^2(\rho)} \lesssim \mathrm{tr}(S^{-4})\varepsilon^2 + \mathcal{O}_{A,B,S}(\varepsilon^4)\,, \tag{31}$$

As a special case, we can take $S = A_t$ and, combined with Proposition 4.5, we obtain the following result for the risk of the finite-sample entropic estimator.

Theorem 4.7. *Let $T_{\varepsilon,t}$ be the entropic Brenier map under the c_t cost between ρ and μ satisfying **(G)**, and let T_{CB} be the conditional Brenier map. For $\varepsilon \asymp t^2$,*

$$\|T_{\varepsilon,t} - T_{\mathrm{CB}}\|^2_{L^2(\rho)} \lesssim_{\Sigma,d} t^2\,.$$

5 Numerical experiments

We study our proposed map estimators on several experiments already present in the literature. Our experiments fall into two broad categories: quantitative and qualitative comparisons. In the former, we consider settings where the true conditional Brenier map is known so that sampling the target conditioning or computing the mean-squared error (MSE) from (25) is possible. For the latter, such a map is not known (which is the case in practice), and so we can only visually compare the generated and target conditional samples.

Recall that two of our estimators are non-parametric: the entropic Brenier map (EOT) and the Nearest-Neighbor map (NN), while the neural network approach (MLP) is parametric. The entropic Brenier map takes the best aspects of both other estimators: it is scalable to many samples via Sinkhorn's algorithm and only requires tuning one extra variable, ε, as opposed to changing minibatch size, learning schedules, architecture, etc., which is required with neural networks.

In all of our experiments we take the reference measure to be of the form $\rho = \mu_1 \otimes \rho_2$ so that the first component of the conditional Brenier map is $T^1(x_1) = \mathrm{Id}(x_1)$. We can then sample the conditional distribution $\mu_{2|1}$ using T^2 alone, i.e., (2) becomes $T_2(X_2; x_1) \sim \mu_{2|1}(\cdot|x_1)$ for $X_2 \sim \rho_2$ and any ρ_1-a.e. x_1.

Our code is publicly available at https://github.com/aistats2025condsim/ConditionalBrenier; details of the experiments not mentioned in the main text are deferred to Appendix A.

5.1 Quantitative comparisons

5.1.1 Non-linear conditional Brenier maps

Here we consider a set of two-dimensional joint distributions $\mu(x_1, x_2)$ where $x_1 \sim \mu_1 = U[-3, 3]$ and the conditionals $\mu_{2|1}(\cdot|x_1)$ are sampled as follows:

$$\begin{aligned}
&\texttt{tanhv1}\colon\ X_2 = \tanh(x_1) + \xi \quad \xi \sim \Gamma(1, 0.3)\\
&\texttt{tanhv2}\colon\ X_2 = \tanh(x_1 + \xi), \quad \xi \sim \mathcal{N}(0, 0.05)\\
&\texttt{tanhv3}\colon\ X_2 = \xi \tanh(x_1), \quad \xi \sim \Gamma(1, 0.3).
\end{aligned}$$

For each case, we generate $n = 5000$ independent sets of samples from the source $\rho = \mu_1 \otimes \mathcal{N}(0, 1)$ and target distribution μ. For each batch of samples, we compute the three estimators and generate 2000 samples from the approximate conditionals using the estimated map.

To evaluate the error in the conditional distributions for each value of x_1, we compute a Monte-Carlo estimate of the following expected error between conditionals

$$\mathbb{E}_{X_1 \sim \mu_1}[\mathsf{D}(\mu_{2|1}(\cdot|X_1), \widehat{T}^2(\cdot; X_1)_\sharp \rho_2)]\,, \tag{32}$$

where D is some discrepancy over probability measures.

To populate Table 1, we used the same $t \in (0, 1)$ for all methods to define the rescaled data. For the approach based on entropic OT, we chose $\varepsilon = t/5$. We randomly

sample 50 i.i.d. conditional variables from μ_1 and evaluate the error in (32) with D taken to be W_2 and the Maximum-Mean-Discrepancy (MMD) metric. We repeat each experiment 10 times and report the standard deviations of the errors. For all cases, we observe that the EOT and NN estimators yield the smallest error between the true and approximate conditionals.

Dataset	Method	W_2 Error (10^{-2})	MMD Error (10^{-2})
`tanhv1`	EOT	4.26 ± 0.62	1.31 ± 0.51
	NN	5.05 ± 0.43	1.34 ± 0.43
	MLP	5.98 ± 0.88	1.97 ± 0.67
`tanhv2`	EOT	3.15 ± 0.81	2.46 ± 0.86
	NN	3.83 ± 0.68	1.91 ± 0.58
	MLP	14.39 ± 3.86	15.50 ± 5.26
`tanhv3`	EOT	2.12 ± 0.67	18.43 ± 7.78
	NN	1.77 ± 0.73	10.39 ± 7.43
	MLP	5.42 ± 0.61	27.19 ± 2.21

Table 1: Expected error between conditional distributions for three test problems; we chose $t = 6 \cdot 10^{-2}$, which seemed to provide the best performance for all methods on average.

5.1.2 Gaussian setting

Next, we examine the MSE of the estimated map on a $d = 4$-dimensional Gaussian source and target measure with $d_1 = 2$ and $d_2 = 2$. Figure 2 compares the error to the true map with increasing sample size for $t \asymp n^{-1/3}$ and $\varepsilon \asymp t^2$ (following the theoretical analysis in Section 4). While we do not believe these rates are optimal, this is a first step toward demonstrating consistency of our proposed estimator.

5.2 Qualitative comparisons

5.2.1 Two-dimensional distribution

Here we visualize the approximated conditional distributions $\mu_{2|1}(\cdot|x_1)$ for different estimators and conditioning variables x_1. We consider a two-dimensional distribution μ where the data is sampled as follows: $X_2 \sim \mathcal{N}(0,1)$ and $X_1 = x_2^2 - 1 + \xi$ with $\xi \sim \mathcal{N}(0,1)$.

Figure 3 shows the target samples from μ for learning the maps and the generated samples from the estimated maps $\widehat{T}^2(\cdot, x_1)_\sharp \rho_2$ for $x_1 \in \{-0.5, 3\}$. We observe that the maps approximate distributions with both unimodal ($x_1 = -0.5$) and bimodal structure ($x_1 = 3$). The closest is agreement found using the entropic map.

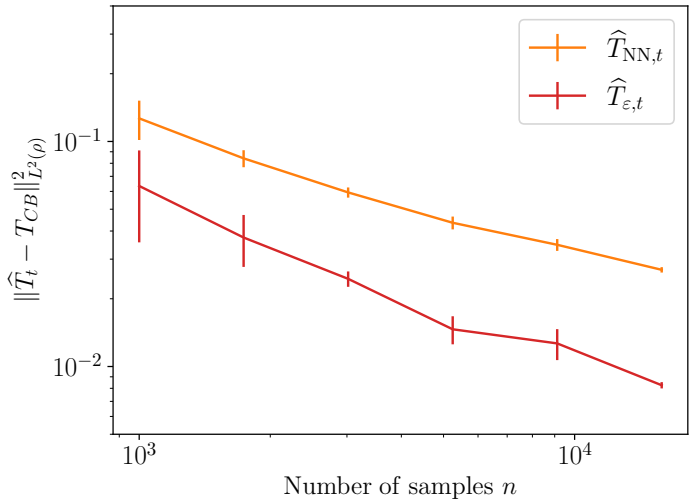

Figure 2: MSE of the estimated map for the NN and EOT estimators for a multivariate Gaussian problem with increasing sample size n.

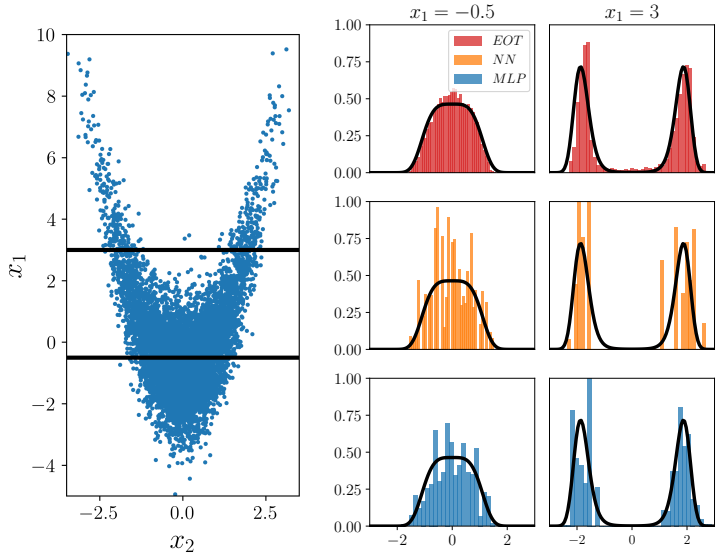

Figure 3: Left: Joint samples of $\mu(x_1, x_2)$ with slices at the conditioning values of interest $x_1 \in \{-0.5, 3\}$. Right: Generated samples from the EOT, NN, and MLP maps with the true density $\mu_{2|1}(\cdot|x_1)$ in black.

5.2.2 Posterior of Lotka–Volterra model

Lastly, we apply the entropic estimator to sample from the posterior distribution $\mu_{2|1}(\cdot|x_1)$ of a Bayesian inference problem for parameters $X_2 \in \mathbb{R}^4$ in a ordinary differential equation that models population dynamics given one observation $x_1 \in \mathbb{R}^{18}$. Figure 4 (left) presents 2.5×10^4 samples generated using the estimated entropic

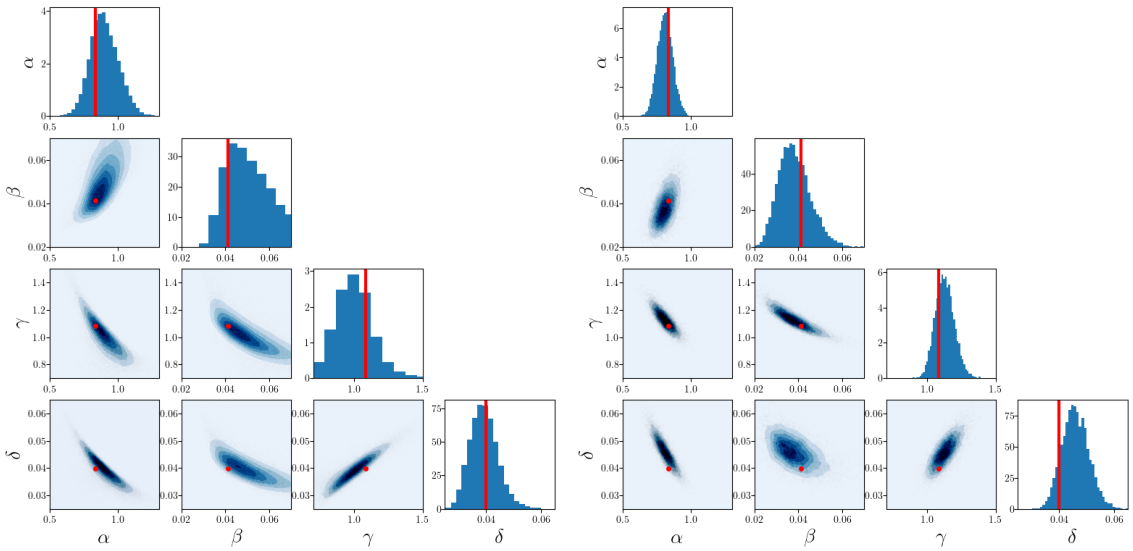

Figure 4: Comparison of samples from the posterior distribution using the entropic estimator (left) and an adaptive MCMC sampler (right).

estimator $\widehat{T}_{\varepsilon,t}$ for $\varepsilon = t = 5\times 10^{-3}$ and $n = 10^5$. The other estimators (i.e., NN and MLP) are not scalable for this high-dimensional application. We compare the generated samples to samples obtained from an adaptive Metropolis Markov chain Monte Carlo sampler, where we observe close agreement of the concentration and correlations between the conditional distributions. Moreover, the true parameter (red) that generated the observation x_1 is covered by the predicted uncertainty.

6 Related work

Estimation of (entropic) Brenier maps. The statistical estimation of optimal transport maps has been studied by e.g, Deb et al. (2021); Hütter and Rigollet (2021); Manole et al. (2021); Divol et al. (2022). To the best of our knowledge, the estimation of optimal transport maps for other costs has not appeared in the literature so far. The works of Fan et al. (2021); Klein et al. (2023); Pooladian et al. (2023b) deviate from the squared-Euclidean cost, considering task-specific costs, such as Lagrangian costs to incorporate barriers in transport, or proximal costs for sparse displacements. Statistical analyses for entropic Brenier maps (again, for the squared-Euclidean cost) have themselves been studied in numerous works; see Pooladian and Niles-Weed (2021); Goldfeld et al. (2022a,b); Rigollet and Stromme (2022); Pooladian et al. (2023a); Stromme (2023).

On the empirical side, Hosseini et al. (2023) and Alfonso et al. (2023) estimate flows or transport maps based on the notion of rescaled costs by leveraging Carlier et al. (2010). Neither work quantitatively assesses the performance of their estimator as $n \to \infty$ or provide a (heuristic) rule for $t = t(n) \to 0$. In this sense, our work takes a first step to making their estimators statistically rigorous.

Knothe–Rosenblatt map computation. On the statistical side, our work is related to that of of Irons et al. (2022) and Wang and Marzouk (2022). The authors of these works study the statistical estimation of the Knothe–Rosenblatt (KR) rearrangement (the strictly triangular map, rather than the block-triangular map considered here) via maximum likelihood estimation with the goal of obtaining rates of convergence in the KL divergence. Unlike these works, we are ultimately interested in procedures that explicitly recover the conditional Brenier map. Finally, adapted optimal transport Eckstein and Pammer (2024); Backhoff et al. (2017); Gunasingam and Wong (2024) incorporates causal constraints to recover a transport plan (in the sense of KR) that is useful for conditional simulation, but developing tractable estimators remains challenging.

Methods for conditional simulation Broadly, computational transport approaches for conditional simulation parameterize the transport map using either conditional normalizing flows Baptista et al. (2023); Wang et al. (2023) or generative adversarial networks Baptista et al. (2020). Alternatively, the transport can be defined from the flow map of a (possibly stochastic) differential equation as in Batzolis et al. (2021); Shi et al. (2022); Albergo et al. (2024). Recently, a dynamic formulation of conditional optimal transport was proposed in Kerrigan et al. (2024). These dynamic formulation use neural network-based methods, which require choosing many hyper-parameters in practice as compared to the proposed approach.

7 Conclusion

We put forth a statistically motivated approach for conditional simulation by estimating conditional Brenier maps. We propose a tractable estimator based on entropic optimal transport that approximates conditional Brenier maps from joint samples. Moreover we provide a near-complete statistical characterization of the estimator in a Gaussian-to-Gaussian setting. A natural and essential direction for future work is to extend the theoretical analyses beyond Gaussians. Similarly, it would be interesting to extend our proposed estimator to the recent Schrödinger bridge estimator developed by Pooladian and Niles-Weed (2024) for the purposes of conditional simulation.

Acknowledgements

AAP thanks NSF grant DMS-1922658 and Meta AI Research for financial support. JNW is supported by the Sloan Research Fellowship and NSF grant DMS-2339829.

References

Albergo, M. S., Goldstein, M., Boffi, N. M., Ranganath, R., and Vanden-Eijnden, E. (2024). Stochastic interpolants with data-dependent couplings. In *Forty-first*

International Conference on Machine Learning.

Alfonso, J., Baptista, R., Bhakta, A., Gal, N., Hou, A., Lyubimova, I., Pocklington, D., Sajonz, J., Trigila, G., and Tsai, R. (2023). A generative flow for conditional sampling via optimal transport. *OTML Workshop at NeurIPS*.

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30*.

Amos, B. (2022). On amortizing convex conjugates for optimal transport. *arXiv preprint arXiv:2210.12153*.

Backhoff, J., Beiglbock, M., Lin, Y., and Zalashko, A. (2017). Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562.

Baptista, R., Hosseini, B., Kovachki, N. B., and Marzouk, Y. (2020). Conditional sampling with monotone GANs: from generative models to likelihood-free inference. *arXiv e-prints*, pages arXiv–2006.

Baptista, R., Hosseini, B., Kovachki, N. B., and Marzouk, Y. M. (2024). Conditional sampling with monotone gans: from generative models to likelihood-free inference. *SIAM/ASA Journal on Uncertainty Quantification*, 12(3):868–900.

Baptista, R., Marzouk, Y., and Zahm, O. (2023). On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics*, pages 1–46.

Batzolis, G., Stanczuk, J., Schönlieb, C.-B., and Etmann, C. (2021). Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417.

Carlier, G., Galichon, A., and Santambrogio, F. (2010). From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576.

Charlier, B., Feydy, J., Glaunès, J. A., Collin, F.-D., and Durif, G. (2021). Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6.

Chewi, S. and Pooladian, A.-A. (2022). An entropic generalization of caffarelli's contraction theorem via covariance inequalities. *arXiv preprint arXiv:2203.04954*.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. (2022). Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*.

Deb, N., Ghosal, P., and Sen, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753.

Divol, V., Niles-Weed, J., and Pooladian, A.-A. (2022). Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*.

Eckstein, S. and Pammer, G. (2024). Computational methods for adapted optimal transport. *The Annals of Applied Probability*, 34(1A):675–713.

Fan, J., Liu, S., Ma, S., Chen, Y., and Zhou, H. (2021). Scalable computation of Monge maps with general costs. *arXiv preprint arXiv:2106.03812*, 4.

Flamary, R., Lounici, K., and Ferrari, A. (2019). Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*.

Gangbo, W. and McCann, R. J. (1996). The geometry of optimal transportation. *Acta Mathematica*, 177(2):113 – 161.

Gelbrich, M. (1990). On a formula for the l^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203.

Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2022a). Limit theorems for entropic optimal transport maps and the sinkhorn divergence. *arXiv preprint arXiv:2207.08683*.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2022b). Statistical inference with regularized optimal transport. *arXiv preprint arXiv:2205.04283*.

Gunasingam, M. and Wong, T.-K. L. (2024). Adapted optimal transport between gaussian processes in discrete time. *arXiv preprint arXiv:2404.06625*.

Hosseini, B., Hsu, A. W., and Taghvaei, A. (2023). Conditional optimal transport on function spaces. *arXiv preprint arXiv:2311.05672*.

Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.

Irons, N. J., Scetbon, M., Pal, S., and Harchaoui, Z. (2022). Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates. In *International Conference on Artificial Intelligence and Statistics*, pages 10161–10195. PMLR.

Kantorovitch, L. (1942). On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201.

Kerrigan, G., Migliorini, G., and Smyth, P. (2024). Dynamic conditional optimal transport through simulation-free flows. *arXiv preprint arXiv:2404.04240*.

Klein, M., Pooladian, A.-A., Ablin, P., Ndiaye, E., Niles-Weed, J., and Cuturi, M. (2023). Learning costs for structured monge displacements. *arXiv preprint arXiv:2306.11895*.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016). Sampling via measure transport: An introduction. *Handbook of uncertainty quantification*, 1:2.

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32.

Nutz, M. and Wiesel, J. (2021). Entropic optimal transport: convergence of potentials. *Probability Theory and Related Fields*, pages 1–24.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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

Pooladian, A.-A., Cuturi, M., and Niles-Weed, J. (2022). Debiaser beware: Pitfalls of centering regularized transport maps. *arXiv preprint arXiv:2202.08919*.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. (2023a). Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. *arXiv preprint arXiv:2301.11302*.

Pooladian, A.-A., Domingo-Enrich, C., Chen, R. T., and Amos, B. (2023b). Neural optimal transport with lagrangian costs. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*.

Pooladian, A.-A. and Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*.

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

Rigollet, P. and Stromme, A. J. (2022). On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*.

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

Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. (2022). Conditional simulation using diffusion schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pages 1792–1802. PMLR.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879.

Stromme, A. J. (2023). Minimum intrinsic dimension scaling for entropic optimal transport. *arXiv preprint arXiv:2306.03398*.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, S. and Marzouk, Y. (2022). On minimax density estimation via measure transport. *arXiv preprint arXiv:2207.10231*.

Wang, Z. O., Baptista, R., Marzouk, Y., Ruthotto, L., and Verma, D. (2023). Efficient neural network approaches for conditional optimal transport with applications in bayesian inference. *arXiv preprint arXiv:2310.16975*.

A Experimental details

A.1 Details of neural network estimator

For this estimator, we use a neural network to parameterize the Brenier potential φ_θ and reverse transport map x_ϑ. The neural networks have 4 hidden layers with 128 hidden units in each layer. We use an Adam optimizer to learn the parameters using a 256 batch size at each iteration and 5000 total iterations. The learning rate follows a cosine scheduler starting from the initial value of 10^{-2}, which decays over 5000 iterations by the multiplier factor of 10^{-3}. The neural networks implementation and training is performed using `OTT-JAX` (Cuturi et al., 2022) using their "Neural Dual" solver tutorial code.

A.2 Computing Figure 1

For each total dimension $d \in \{4, 10, 20, 50\}$, we randomly sample a target covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ using the following procedure. We sample matrices $A \in \mathbb{R}^{d_1\times d_1}, B \in \mathbb{R}^{d_2\times d_1}$ and $C \in \mathbb{R}^{d_2\times d_2}$ with independent standard Gaussian entries. We then construct the block covariance matrix Σ of the form

$$\Sigma = \begin{bmatrix} AA^\top & AA^\top B^\top \\ BAA^\top & BAA^\top B^\top + 0.01 I_d \end{bmatrix}.$$

Given the target $\mu = \mathcal{N}(0, \Sigma)$ and standard Gaussian source $\rho = \mathcal{N}(0, I_d)$, we compute the lower block Cholesky factor $L_\Sigma \in \mathbb{R}^{d\times d}$ satisfying $\Sigma = L_\Sigma L_\Sigma^\top$ to define the conditional Brenier map T_{CB} in Proposition 4.1. This map is compared to $T_t(x) = A_t^{-2}(A_t^2\Sigma A_t^2)1/2x$ in Proposition 4.1 to evaluate the squared errors in Figure 1. Lastly, we note that for a standard Gaussian source we have

$$\begin{aligned} \|T_t - T_{\mathrm{CB}}\|^2_{L^2(\rho)} &= \mathbb{E}_{x\sim\rho}\mathrm{Tr}\left((A_t^{-2}(A_t^2\Sigma A_t^2)^{1/2} - L_\Sigma)xx^\top(A_t^{-2}(A_t^2\Sigma A_t^2)^{1/2} - L_\Sigma)^\top\right) \\ &= \|A_t^{-2}(A_t^2\Sigma A_t^2)^{1/2} - L_\Sigma\|_F^2. \end{aligned}$$

A.3 Details for Section 5.1

For both the quantitative comparisons, we construct the entropic estimator using the `Optimal Transport Tools` (`OTT-JAX`) package (Cuturi et al., 2022). We set the maximum number of iterations for the Sinkhorn solver to 5000 and use a tolerance of 10^{-3} for checking convergence. For the nearest-neighbour estimator, we use the `Scikit-Learn` (Pedregosa et al., 2011) package to compute the nearest neighbor points in the source dataset.

For the `tanh` experiments in Section 5.1.1 we set $t = 6 \cdot 10^{-2}$ and $\varepsilon = t/5$ for all examples. For the Gaussian experiment in Section 5.1.2, we set $t = 0.1n^{-1/5}$ and $\varepsilon = t^2$ for all considered sample sizes $n \in [10^3, 10^4]$. The errors for the estimated maps in Figure 2 are computed using Monte Carlo with 10^4 independent samples from the source distribution.

A.4 Details for Section 5.2

For the experiment in Section 5.2.1, we use $n = 7500$ target samples from μ to build the estimators and set $t = 6 \cdot 10^{-2}$. We use $\varepsilon = t/5$ for the entropic estimator. The entropic and nearest-neighbor estimators are computed using `OTT` and `Scikit-Learn`, respectively, using the parameters listed in Appendix A.3. To plot the histograms in Figure 3, we generate 5000 samples from the estimated conditionals using the second component of the estimated maps $\widehat{T}^2(\cdot, x_1)$ for each x_1.

For the experiment in Section 5.2.2, we consider a target distribution specified by a prior $\mu_2(x_2)$ over parameters $x_2 = (\alpha, \beta, \delta, \gamma) \in \mathbb{R}^4$ that define the right-hand-side nonlinear ODE, and a likelihood function $\mu_{1|2}(x_1|x_2)$ given by the mapping from

parameters to noisy observations of the ODE $x_1 \in \mathbb{R}^{18}$. Given the parameters, let $P(t) = (P_1(t), P_2(t)) \in \mathbb{R}^2_+$ describe the populations of predator and prey over time in an environment, which evolve according to the coupled ODES

$$\frac{dP_1}{dt} = \alpha P_1(t) - \beta P_1(t) P_2(t)$$
$$\frac{dP_2}{dt} = -\gamma\alpha P_1(t) + \delta P_1(t) P_2(t),$$

with the initial condition $P(0) = (30, 1)$. To generate the observations, we simulate the ODEs for $t \in [0, 20]$ and sample $(\log(X_1)_{2k-1}, \log(X_1)_{2k}) \sim \mathcal{N}(\log P(k\Delta t_{obs}), \sigma^2 I_2)$ with $\Delta t_{obs} = 2$ and $\sigma = 0.01$ for $k = 1, \ldots, 9$. This sampling process defines the likelihood function $\mu_{1|2}(\cdot|x_2)$. We set the prior distribution to be log-normal, i.e., $\log X_2 \sim \mathcal{N}(m, 0.5 I_4)$ with mean $m = (-0.125, -3, -0.125, -3)$. To generate the target dataset $\{(X_1^i, X_2^i)\}_{i=1}^n \sim \mu$, we sample a set of parameters $X_2^i \sim \mu_2(\cdot)$ and paired observations $X_1^i \sim \mu_{1|2}(\cdot|x_2 = X_2^i)$.

To construct the entropic estimator for the conditional Brenier map, we consider the parameters $t = 0.01$, $\varepsilon = 0.005$, and a tolerance for the Sinkhorn solver of 10^{-4}. The estimator is computed using the `PyKeOps` package for scalability of kernel operations on the GPU with large sample sizes n (Charlier et al., 2021). For the comparison, we ran a Metropolis Hastings Markov chain Monte Carlo (MCMC) algorithm for one million steps using an initial zero-mean Gaussian proposal distribution. The proposal covariance is adapted at each step based on the Markov chain of the previous samples. We treated the first 500,000 steps of MCMC as burn-in and extracted a thinned set of 25,000 samples at uniformly spaced iterations. Figure 4 compares the MCMC samples to 25,000 samples generated using the entropic estimator $\widehat{T}^2(\cdot; x_1^*)_\sharp \rho_2$ for an observation $x_1^* \sim \mu(\cdot|x_2 = x_2^*)$ corresponding to the true parameter vector $x_2^* = (0.832, 0.041, 1.082, 0.040)$, which is displayed in red in Figure 4.

B Proofs from Section 4

Proof of Proposition 4.1. For general $\rho := \mathcal{N}(0, A)$ and $\mu := \mathcal{N}(m, B)$ the OT map has closed form (Gelbrich, 1990):

$$T_{\mathrm{B}}^{\rho\to\mu}(x) = m + A^{-1/2}(A^{1/2} B A^{1/2})^{1/2} A^{-1/2} x\,.$$

Let $\rho_t = (A_t \cdot)_\sharp \rho = \mathcal{N}(0, A_t^2)$ and similarly $\mu_t = \mathcal{N}(A_t m, A_t \Sigma A_t)$. Then it holds that

$$T_{\mathrm{B}}^{\rho_t\to\mu_t} = A_t m + A_t^{-1}(A_t^2 \Sigma A_t^2)^{1/2} A_t^{-1} x\,.$$

Using now (11), we have

$$T_t(x) = A_t^{-1}(A_t m + A_t^{-1}(A_t^2 \Sigma A_t^2)^{1/2} A_t^{-1})(A_t x) = m + A_t^{-2}(A_t^2 \Sigma A_t^2)^{1/2} x\,.$$

We now compute T_{CB}. First note that the block lower Cholesky decomposition is

$$\mathsf{L}_\Sigma = \begin{pmatrix} \Sigma_{11}^{1/2} & 0 \\ \Sigma_{21}\Sigma_{11}^{-1/2} & W^{1/2} \end{pmatrix},$$

where we recall

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

and that $W = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}^\top$. Since this is block-lower triangular, this satisfies the criteria to be a conditional Brenier map. □

Proof of Proposition 4.4. Since the error incurred from estimating the mean m is negligible, we henceforth assume that $\rho_t \coloneqq \mathcal{N}(0, A_t^2)$ and $\mu_t \coloneqq \mathcal{N}(0, A_t\Sigma A_t)$. The optimal transport map from ρ_t to μ_t is given by

$$T_t'(y) \coloneqq A_t^{-1}(A_t^2\Sigma A_t^2)^{1/2}A_t^{-1}y\,.$$

Similarly, we define $\widehat{T}_t'(y) \coloneqq A_t^{-1}(A_t^2\widehat{\Sigma}A_t^2)^{1/2}A_t^{-1}y$. The main expression we want to bound is then

$$\begin{aligned}\mathbb{E}\|\widehat{T}_t - T_t\|_{L^2(\rho)}^2 &= \mathbb{E}\|A_t^{-1}(\widehat{T}_t' - T_t')\|_{L^2(\rho_t)}^2 \\ &\le \|A_t^{-1}\|_{\mathrm{op}}^2\mathbb{E}\|\widehat{T}_t' - T_t'\|_{L^2(\rho_t)}^2 \\ &\le t^{-1}\mathbb{E}\|\widehat{T}_t' - T_t'\|_{L^2(\rho_t)}^2\,.\end{aligned}$$

To continue, we note that

$$\sqrt{\lambda_{\min}(\Sigma)}I \preceq DT_t'(y) \preceq \sqrt{\lambda_{\max}(\Sigma)}I\,, \tag{33}$$

where $\lambda_{\max}(\Sigma) > 0$ (resp. $\lambda_{\min}(\Sigma) > 0$) is the largest (resp. smallest) eigenvalue of the covariance matrix Σ. Equation (33) follows from Theorem 5 of Chewi and Pooladian (2022), where we compute $\nabla^2(-\log\rho_t) = A_t^{-2}$ and $(\lambda_{\max}(\Sigma))^{-1}A_t^{-2} \preceq \nabla^2(-\log\mu_t) \preceq (\lambda_{\min}(\Sigma))^{-1}A_t^{-2}$. By a direct application of a smoothness result by Manole et al. (2021, Theorem 6), we have

$$\mathbb{E}\|\widehat{T}_t' - T_t'\|_{L^2(\rho_t)}^2 \le \kappa(\Sigma)\mathbb{E}W_2^2(\hat{\mu}_t, \mu_t)\,,$$

where $\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, and $\hat{\mu}_t \coloneqq \mathcal{N}(0, A_t\widehat{\Sigma}A_t)$. Note that the remaining Wasserstein distance term can be upper bounded (since $\|A_t\|_{\mathrm{op}} \le 1$) as

$$\mathbb{E}W_2^2(\hat{\mu}_t, \mu_t) \le \mathbb{E}\int \|A_t(\widehat{\Sigma}^{1/2} - \Sigma^{1/2})(x)\|^2\,\mathrm{d}\rho(x) \le \mathbb{E}\|\widehat{T} - T\|_{L^2(\rho)}^2 \lesssim_\Sigma n^{-1}\,,$$

where $T(x) \coloneqq \Sigma^{1/2}x$ and similarly $\widehat{T}(x) \coloneqq \widehat{\Sigma}^{1/2}x$, and the final inequality is well-known; see Flamary et al. (2019); Divol et al. (2022).

Putting everything together, we have

$$\mathbb{E}\|\widehat{T}_t - T_t\|_{L^2(\rho)}^2 \lesssim_\Sigma t^{-1}n^{-1}\,.$$

□

Proof of Proposition 4.5. We write $Q_t \coloneqq A_t^2$. Suppose that for t small enough, M_t is a positive-definite matrix such that both

$$\|M_t^2 - (Q_t \Sigma Q_t)\|_{\mathrm{op}} \lesssim t^3\,, \quad \text{and} \quad \|Q_t^{-1} M_t - \mathsf{L}_\Sigma\|_{\mathrm{op}} \lesssim t\,. \tag{34}$$

The first inequality implies that

$$M_t^2 \preceq (Q_t \Sigma Q_t) + ct^3 I\,.$$

Since square-roots preserves the positive-definite order, an appropriate Taylor expansion gives

$$M_t \preceq (Q_t \Sigma Q_t)^{1/2} + \frac{c}{2} t^3 (Q_t \Sigma Q_t)^{-1/2} \preceq \frac{c}{2} t^2 \Sigma^{-1/2}\,.$$

Pre-multiplying by $Q_t^{-1} \preceq t^{-1} I$, we have that

$$Q_t^{-1} M_t - Q_t^{-1} (Q_t \Sigma Q_t)^{1/2} \preceq \frac{c}{2} t \Sigma^{-1/2}\,,$$

which implies $\|Q_t^{-1} M_t - Q_t^{-1}(Q_t \Sigma Q_t)^{1/2}\|_{\mathrm{op}} \leq \frac{c}{2} t (\lambda_{\min}(\Sigma))^{-1/2}$. Combined with the second inequality in (34) via triangle inequality, we arrive at

$$\|Q_t^{-1}(Q_t \Sigma Q_t)^{1/2} - \mathsf{L}_\Sigma\|_{\mathrm{op}} \lesssim_\Sigma t\,.$$

Since $\|\cdot\|_{\mathrm{F}}^2 \leq d \|\cdot\|_{\mathrm{op}}^2$, the proof is complete if we can find such an M_t. The matrix outlined in Lemma B.1 satisfies the inequalities in (34), which is the one we take to complete the proof. □

Proof of Proposition 4.6. For the linear map $x \mapsto Sx$, write $\rho_S \coloneqq (S\cdot)_\sharp \rho$ and $\mu_S \coloneqq (S\cdot)_\sharp \mu$. Specifically, $\rho_S = \mathcal{N}(0, SAS)$ and $\mu_S = \mathcal{N}(b, SBS)$. The closed-form expression for $\widetilde{T}_{\varepsilon,S}$, the entropic map from ρ_S to μ_S for $\varepsilon \geq 0$ can be computed as

$$\begin{aligned}
\widetilde{T}_{\varepsilon,S}(y) &= ((SAS)^{-1/2} M_{\varepsilon,S}^{1/2} (SAS)^{-1/2} - \tfrac{\varepsilon}{2}(SAS)^{-1}) y + Sb\,, \\
&= ((SAS)^{-1/2} [M_{\varepsilon,S}^{1/2} - \tfrac{\varepsilon}{2} I](SAS)^{-1/2}) y + Sb
\end{aligned}$$

where $M_{\varepsilon,S} = (SAS)^{1/2} SBS (SAS)^{1/2} + \frac{\varepsilon^2}{4} I$; see Pooladian et al. (2022, Proposition 3). The expression for the Brenier map between ρ_S and μ_S, $\widetilde{T}_{0,S}(y)$, follows the same formula. Using (11), the maps between ρ and μ under the cost $\frac{1}{2}\|S(x-y)\|_2^2$ are given by

$$T_{\varepsilon,S}(x) = S^{-1}(SAS)^{-1/2} [M_{\varepsilon,S}^{1/2} - \tfrac{\varepsilon}{2} I](SAS)^{-1/2} Sx + b$$

We can now directly compute

$$\begin{aligned}
\|T_{\varepsilon,S} - T_{0,S}\|_{L^2(\rho)}^2 &= \mathrm{tr}\big[\big(S^{-1}(SAS)^{1/2}[M_{\varepsilon,S}^{1/2} - M_{0,S}^{1/2} - \tfrac{\varepsilon}{2} I](SAS)^{-1/2} S\big)^2 A\big] \\
&= \mathrm{tr}\big[\big((SAS)^{-2}[M_{\varepsilon,S}^{1/2} - M_{0,S}^{1/2} - \tfrac{\varepsilon}{2} I]^2\big) A\big] \\
&= \mathrm{tr}\big[S^{-2} A^{-1} S^{-2} [M_{\varepsilon,S}^{1/2} - M_{0,S}^{1/2} - \tfrac{\varepsilon}{2} I]^2\big]\,,
\end{aligned}$$

where we use the fact that all matrices are symmetric, and thus commute under trace. A direct application of Pooladian et al. (2022, Lemma 2) tells us that

$$M_{\varepsilon,S}^{1/2} = M_{0,S}^{1/2} + \frac{\varepsilon^2}{8} M_{0,S}^{-1/2} + O(\varepsilon^4)\,.$$

Expanding the square inside the trace and applying the above Taylor expansion, we see that many terms cancel, and, dropping the remaining negative-definite terms, one arrives at

$$\|T_{\varepsilon,S} - T_{0,S}\|_{L^2(\rho)}^2 \le \frac{\varepsilon^2}{4} \operatorname{tr}(S^{-4}A^{-1}) \le \frac{\varepsilon^2}{4} \|S^{-4}\|_{\mathrm{op}} I_0(\rho)\,,$$

where recall $I_0(\rho) = \operatorname{tr}(A^{-1})$ is the Fisher information of $\rho = \mathcal{N}(0, A)$. □

Proof of Theorem 4.7. Writing the expansion (30) and using both Equation (29) and Proposition 4.6, we obtain (using that $I_0(\rho) = d$)

$$\begin{aligned}
\|T_{\varepsilon,t} - T_{\mathrm{CB}}\|_{L^2(\rho)}^2 &\le \|T_{\varepsilon,t} - T_t\|_{L^2(\rho)}^2 + \|T_t - T_{\mathrm{CB}}\|_{L^2(\rho)}^2 \\
&\lesssim_{\Sigma} 2\frac{\varepsilon^2}{4} \operatorname{tr}(A_t^{-4} I) + 2t^2 \\
&\le \frac{\varepsilon^2}{2} d t^{-2} + 2t^2 \\
&\lesssim_{\Sigma,d} t^2\,,
\end{aligned}$$

where we choose $\varepsilon^* \asymp_{\Sigma,d} t^2$ in the penultimate inequality to conclude. □

Lemma B.1 (Square-root perturbation)**.** *Consider the assumptions in Proposition 4.5. There exists a matrix M such that the following inequalities hold*

$$\|M^2 - (Q_t \Sigma Q_t)\|_{op} \lesssim t^3 \tag{35}$$

$$\|Q_t^{-1} M - \mathsf{L}_\Sigma\|_{op} \lesssim t\,, \tag{36}$$

with $Q_t = A_t^2$.

Proof. First, recall that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \mathsf{L}_\Sigma = \begin{pmatrix} \Sigma_{11}^{1/2} & 0 \\ \Sigma_{21}\Sigma_{11}^{-1/2} & S^{1/2} \end{pmatrix},$$

where $S = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}^\top$.

Let

$$M \coloneqq \begin{pmatrix} \Sigma_{11}^{1/2} & 0 \\ 0 & 0 \end{pmatrix} + t \begin{pmatrix} 0 & \Sigma_{11}^{-1/2}\Sigma_{21}^\top \\ \Sigma_{21}\Sigma_{11}^{-1/2} & S^{1/2} \end{pmatrix} + t^2 \begin{pmatrix} B_{11} & B_{21}^\top \\ B_{21} & 0 \end{pmatrix}. \tag{37}$$

We compute

$$M^2 = (Q_t \Sigma Q_t) + t^2 \begin{pmatrix} \Sigma_{11}^{1/2} B_{11} + B_{11} \Sigma_{11}^{1/2} + \Sigma_{11}^{-1/2} \Sigma_{21}^\top \Sigma_{21} \Sigma_{11}^{-1/2} & \Sigma_{11}^{1/2} B_{21}^\top + \Sigma_{11}^{-1/2} \Sigma_{21}^\top \Sigma_{11}^{-1/2} \\ B_{21} \Sigma_{11}^{1/2} + \Sigma_{11}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1/2} & 0 \end{pmatrix} + O(t^3)\,,$$

where

$$(Q_t \Sigma Q_t) = \begin{pmatrix} \Sigma_{11} & t\Sigma_{21}^\top \\ t\Sigma_{21} & t^2 \Sigma_{22} \end{pmatrix}.$$

We choose B_{11} and B_{21} to satisfy

$$\begin{aligned} &\Sigma_{11}^{1/2} B_{11} + B_{11} \Sigma_{11}^{1/2} = -\Sigma_{11}^{1/2} \Sigma_{21}^\top \Sigma_{21} \Sigma_{11}^{-1/2} \\ &B_{21} = -\Sigma_{11}^{-1} \Sigma_{21}^\top S^{1/2}\,. \end{aligned}$$

Thus, $\|M^2 - (Q_t \Sigma Q_t)\|_{\mathrm{op}} \lesssim t^3$, which is the first claim. For the second, note that

$$Q_t^{-1} M = \begin{pmatrix} \Sigma_{11} + O(t) & t\Sigma_{21}^\top + O(t) \\ \Sigma_{21} \Sigma_{11}^{-1/2} & S^{1/2} + O(t) \end{pmatrix} = \mathsf{L}_\Sigma + \begin{pmatrix} O(t) & t\Sigma_{21}^\top + O(t) \\ 0 & O(t) \end{pmatrix},$$

and thus $\|Q_t^{-1} M - \mathsf{L}_\Sigma\|_{\mathrm{op}} \lesssim t$. □