# Discrete Schrödinger Bridges with Applications to Two-Sample Homogeneity Testing

**Zaid Harchaoui**
Department of Statistics
University of Washington
zaid@uw.edu

**Lang Liu**
Department of Statistics
University of Washington
liu16@uw.edu

**Soumik Pal***
Department of Mathematics
University of Washington
soumikpal@gmail.com

## Abstract

We introduce an entropy-regularized statistic that defines a divergence between probability distributions. The statistic is the transport cost of a coupling which admits an expression as a weighted average of Monge couplings with respect to a Gibbs measure. This coupling is related to the static Schrödinger bridge given a finite number of particles. We establish the asymptotic consistency of the statistic as the sample size goes to infinity and show that the population limit is the solution of Föllmer's entropy-regularized optimal transport. The proof technique relies on a chaos decomposition for paired samples. We illustrate the interest of the approach on the two-sample homogeneity testing problem.

## 1 Introduction

In 1932, Schrödinger [23] considered the following lazy gas experiment; see, e.g., [5] for a review. Image $n$ indistinguishable particles in $\mathbb{R}^d$ moving independently as Brownian motion at temperature $\epsilon$. At time $t = 0$, we observe that the empirical distribution of their initial locations approximately equals some density $\rho_0$. At time $t = 1$, we observe that the empirical distribution of their terminal locations approximately equals another density $\rho_1$, which differs significantly from what it should be by the law of large numbers, i.e.,

$$\rho_1(y) \neq \int \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|x-y\|^2}{2\epsilon}\right) \rho_0(x)dx.$$

It is clear that this situation is unlikely to happen. Schrödinger then inquires for, among all unlikely ways in which this could happen, the most likely path for each particle. As Föllmer shows in [9], the paths are determined by first solving for the (static) Schrödinger bridge (which is introduced in Section 2) and then connecting the two end points by a Brownian bridge with diffusion $\epsilon$.

Although Schrödinger's lazy gas experiment is typically defined in the dynamic setting for Brownian motion, its static counterpart, Schrödinger bridges [9, 15], can be defined more generally. In

*Authors listed in alphabetical order. Z.H. acknowledges support from NSF grant DMS-1810975 and CCF-1740551. L.L. acknowledges support from NSF grant DMS-1612483 and CCF-1740551. S.P. acknowledges support from NSF grant DMS-1612483.

Presented at the NeurIPS 2021 Optimal Transport and Machine Learning workshop on December 13. 

continuum, the Schrödinger bridge can be made precise as the solution to the entropy-regularized optimal transport (EOT) between two densities $\rho_0$ and $\rho_1$ [16], where the entropy is given by the negative differential entropy. Recently, Schrödinger bridges have been used in score-based generative modeling [4] and Markov chain Monte Carlo [1].

In its entropy-regularized form, the Schrödinger bridge problem is closely related to the EOT between two discrete distributions [7, 8], where the entropy is given by the negative Shannon entropy. This discrete EOT is particularly attractive both from a computational viewpoint [7] and from a statistical viewpoint [20]. When we only have access to i.i.d. samples from $\rho_0$ and $\rho_1$, one may use the solution to the discrete EOT between the empirical distributions to estimate the Schrödinger bridge. However, it remains largely unclear if this estimation is consistent. Existing works either focus on the case when both $\rho_0$ and $\rho_1$ are discrete [3, 13] or is limited to the regularized cost rather than the solution [10, 17].

In this work, we introduce the so-called discrete Schrödinger bridge which recovers Schrödinger's original discrete set-up as the Schrödinger bridge connecting two empirical distributions. We show that it is the solution to a modified discrete EOT problem. We prove its convergence towards the Schrödinger bridge in continuum. Finally, we design a novel Schrödinger bridge statistic and apply it to two-sample homogeneity testing on synthetic and real data.

## 2 Background

**Schrödinger bridges in continuum.** Let $\nu_0$ and $\nu_1$ be two probability measures on $\mathbb{R}^d$. Given $\epsilon \in \mathbb{R}_+$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, we assume that the following Markov transition kernel density is well-defined:

$$p_\epsilon(x, y) = \frac{1}{Z_\epsilon(x)} \exp\left[-\frac{1}{\epsilon} c(x, y)\right],$$

where $Z_\epsilon(x)$ is the normalizing constant. For instance, when $c$ is the quadratic cost, this is the transition density of Brownian motion with diffusion $\epsilon$ considered in Schrödinger's lazy gas experiment. Suppose that $(W_0, W_1)$ is distributed according to this Markov transition kernel. Informally, the Schrödinger bridge connecting $\nu_0$ and $\nu_1$ at temperature $\epsilon$ is the joint law of $(W_0, W_1)$ conditioned to have $W_0 \sim \nu_0$ and $W_1 \sim \nu_1$. In continuum, when $\nu_0 = \rho_0$ and $\nu_1 = \rho_1$ are densities, it can be made precise as the solution to the EOT problem:

$$\min_{\nu \in \Pi(\rho_0, \rho_1)} \left[\int c(x, y) d\nu(x, y) + \epsilon H(\nu)\right], \tag{1}$$

where $\Pi(\rho_0, \rho_1)$ is the set of joint distributions (couplings) with marginals $\rho_0$ and $\rho_1$, and $H(\nu)$ is the entropy of $\nu$ defined as $H(\nu) = \int \nu(x, y) \log \nu(x, y) dx dy$ if $\nu$ is a density and infinity otherwise.

**Characterization of Schrödinger bridges.** Since the entropy $H$ is strongly convex, the EOT problem has a unique solution $\mu_\epsilon$. Even though $\mu_\epsilon$ is usually not explicit, it admits the following expression due to [6, 22]. There exists two measurable functions $a_\epsilon$ and $b_\epsilon$, to be called the Schrödinger potentials, such that, for $\xi(x, y) = \exp\left\{-\frac{1}{\epsilon}[c(x, y) - a_\epsilon(x) - b_\epsilon(y)]\right\}$,

$$\mu_\epsilon(x, y) = \xi(x, y)\rho_0(x)\rho_1(y). \tag{2}$$

Note that $\mu_\epsilon \in \Pi(\rho_0, \rho_1)$. This implies

$$\int \xi(x, y)\rho_0(x) dx \overset{\text{a.s.}}{=} \int \xi(x, y)\rho_1(y) dy \overset{\text{a.s.}}{=} 1. \tag{3}$$

## 3 Discrete Schrödinger bridges

Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two independent i.i.d. samples from densities $\rho_0$ and $\rho_1$, respectively. Let $\mathcal{S}_n$ be the set of permutations on $[n] := \{1, \ldots, n\}$. Every permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$ can be viewed as a matching between these two sets of random variables, and it induces a coupling $\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$ with marginals given by the two empirical distributions $\hat{\rho}_0^n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and

$\hat{\rho}_1^n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$. Its associated cost is then $c(X, Y_\sigma) := \sum_{i=1}^n c(X_i, Y_{\sigma_i})$. If we weigh each permutation $\sigma$ by the (random) weight $w(\sigma) := \exp(-\frac{1}{\epsilon}c(X, Y_\sigma))$, then we obtain a Gibbs probability distribution on $\mathcal{S}_n$, i.e., $q_\epsilon^*(\sigma) := w(\sigma)/\sum_{\tau\in\mathcal{S}_n} w(\tau)$. Now we mix all permutations by defining

$$\hat{\mu}_\epsilon^n := \sum_{\sigma\in\mathcal{S}_n} q_\epsilon^*(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}. \tag{4}$$

**Interpretation as discrete Schrödinger bridges.** It can be shown that $\hat{\mu}_\epsilon^n$ recovers Schrödinger's original discrete set-up as the Schrödinger bridge connecting $\hat{\rho}_0^n$ and $\hat{\rho}_1^n$ at temperature $\epsilon$. To see this, consider a realization $X_i = x_i$ and $Y_i = y_i$ for $i \in [n]$. Then $\hat{\rho}_0^n$ and $\hat{\rho}_1^n$ are (nonrandom) discrete distributions supported on $n$ categories. Imagine $n$ independent particles $\{W^i\}_{i=1}^n$, starting from positions $W_0^i = x_i, i \in [n]$, make jumps according to the Markov transition kernel $p_\epsilon(x_i, \cdot)$, respectively. Let $L_n(1) := \frac{1}{n}\sum_{i=1}^n \delta_{W_1^i}$ be the empirical distribution of their terminal locations and $L_n(0,1) := \frac{1}{n}\sum_{i=1}^n \delta_{(W_0^i, W_1^i)}$ be the joint empirical distribution at two time points. It can be shown [18] that the law of $L_n(0,1)$ given $L_n(1) = \hat{\rho}_1^n$ is exactly given by $\hat{\mu}_\epsilon^n$ (with $X_i = x_i$ and $Y_i = y_i$, $i \in [n]$). In other words, for each matching $\sigma \in \mathcal{S}_n$,

$$\mathbb{P}\left(L_n(0,1) = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, y_{\sigma_i})} \;\middle|\; L_n(1) = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\right) = q_\epsilon^*(\sigma). \tag{5}$$

We will refer to $\hat{\mu}_\epsilon^n$ the discrete Schrödinger bridge.

**Connection to discrete entropy-regularized optimal transport.** The discrete EOT problem is a discrete counterpart of the EOT problem in continuum. To be more specific, it reads

$$\min_{\nu\in\Pi(\hat{\rho}_0^n, \hat{\rho}_1^n)} \left[\sum_{i=1}^n\sum_{j=1}^n c(X_i, Y_j)\nu(X_i, Y_j) + \epsilon\mathrm{Ent}(\nu)\right], \tag{6}$$

where $\mathrm{Ent}(\nu) = \sum_{i=1}^n\sum_{j=1}^n \nu(X_i, Y_j)\log\nu(X_i, Y_j)$. The solution to (4) is a smoothed version of the optimal Monge coupling between $\hat{\rho}_0^n$ and $\hat{\rho}_1^n$. In the same spirit, the discrete Schrödinger bridge can be viewed as another smoothed version of the optimal Monge coupling. Recall that $\hat{\mu}_\epsilon^n$ is a convex combination of all Monge couplings with weights $\{\exp(-\frac{1}{\epsilon}c(X, Y_\sigma))\}_{\sigma\in\mathcal{S}_n}$. That is, a permutation gets exponentially small weight if its associated cost is relatively large. In fact, $\hat{\mu}_\epsilon^n$ is the solution to a modified EOT problem.

**Lemma 1.** *Let $\mathcal{P}(\mathcal{S}_n)$ be the set of probabilities on $\mathcal{S}_n$. Then we have*

$$q_\epsilon^* = \operatorname*{arg\,min}_{q\in\mathcal{P}(\mathcal{S}_n)} \left[\sum_{i=1}^n\sum_{j=1}^n c(X_i, Y_j)\nu_q(X_i, Y_j) + \frac{\epsilon}{n}\mathrm{Ent}(q)\right], \tag{7}$$

*where $\nu_q := \sum_{\sigma\in\mathcal{S}_n} q(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$. In particular, $\hat{\mu}_\epsilon^n = \nu_{q_\epsilon^*}$.*

*Proof.* For any probability $q$ on $\mathcal{S}_N$, consider the relative entropy $H(q \mid q_\epsilon^*)$ of $q$ with respect to $q_\epsilon^*$.

$$\begin{aligned}
H(q \mid q_\epsilon^*) &= \sum_{\sigma\in\mathcal{S}_n} q(\sigma)\log\frac{q(\sigma)}{q_\epsilon^*(\sigma)} = \sum_{\sigma\in\mathcal{S}_n} q(\sigma)\log\left(\frac{q(\sigma)\sum_{\tau\in\mathcal{S}_N} w(\tau)}{w(\sigma)}\right) \\
&= \mathrm{Ent}(q) + \log\left[\sum_{\tau\in\mathcal{S}_n} w(\tau)\right]\sum_{\sigma\in\mathcal{S}_n} q(\sigma) + \frac{1}{\epsilon}\sum_{\sigma\in\mathcal{S}_n} c(X, Y_\sigma)q(\sigma) \\
&= \frac{n}{\epsilon}\sum_{i=1}^n\sum_{j=1}^n c(X_i, Y_j)\nu_q(X_i, Y_j) + \mathrm{Ent}(q) + \log\sum_{\tau\in\mathcal{S}_n} w(\tau).
\end{aligned}$$

Hence, (7) is equivalent to minimizing $H(q \mid q_\epsilon^*)$, which is uniquely minimized at $q = q_\epsilon^*$. □

**Schrödinger bridge statistic.** Since $\hat{\mu}_\epsilon^n$ is a smoothed version of the optimal Monge coupling, its cost of transport

$$T_n := \int c(x,y) d\hat{\mu}_\epsilon^n(x,y) = \sum_{\sigma \in \mathcal{S}_n} q_\epsilon^*(\sigma) \frac{1}{n} \sum_{i=1}^n c(X_i, Y_{\sigma_i}), \tag{8}$$

can be used to measure the difference between $\hat{\rho}_0^n$ and $\hat{\rho}_1^n$. In particular, it can be used as a statistic to test for homogeneity between two samples. We will discuss it in detail in Section 5.

## 4 Asymptotic properties

In this section, we summarize asymptotic properties of the discrete Schrödinger bridge $\hat{\mu}_\epsilon^n$ and the Schrödinger bridge statistic $T_n$. Due to the length constraint, we give the exact statements in Appendix and will present the full proofs in a different venue.

The next theorem shows that $\hat{\mu}_\epsilon^n$ is a consistent estimator of the Schrödinger bridge in continuum $\mu_\epsilon$.

**Theorem 1** (Consistency)**.** *Under appropriate assumptions, as $n \to \infty$, $T_n$ converges in probability to $\theta := \int c d\mu_\epsilon$. In particular, $\hat{\mu}_\epsilon^n$ converges weakly to $\mu_\epsilon$, in probability.*

*Proof sketch.* We use a change of measure argument. Firstly, we show that, under the measure $\mu_\epsilon^n$ (i.e., $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mu_\epsilon$), $T_n$ can be rewritten as the conditional expectation of $c(X_1, Y_1)$ given the $\sigma$-algebra generated by the pair of random measures $(\hat{\rho}_0^n, \hat{\rho}_1^n)$. Concretely, $T_n = \mu_\epsilon^n[c(X_1, Y_1) \mid \sigma(\hat{\rho}_0^n, \hat{\rho}_1^n)]$. Secondly, under the measure $\mu_\epsilon^n$, $T_n \to_{a.s.} \mu_\epsilon^n[T_n] = \theta$ by the reverse martingale convergence theorem. Thirdly, we prove a contiguity result in the sense of Le Cam [24, Chapter 6] and transfer the convergence result to the desired one under the measure $(\rho_0 \otimes \rho_1)^n$ from which our data are generated. Finally, repeating the above argument for each continuous bounded function $\eta$ yields $\int \eta d\hat{\mu}_\epsilon^n \to_p \int \eta d\mu_\epsilon$. The weak convergence then follows. □

The Schrödinger bridge statistic $T_n$ is a rather complicated function of the two empirical distributions $\hat{\rho}_0^n$ and $\hat{\rho}_1^n$. Our next result shows that it can be well approximated by linear functions of the two distributions. Given a probability measure $\nu$ and $p \geq 1$, we define $\mathbf{L}^p(\nu)$ the space of functions that have finite $p$-th moment under $\nu$ and $\mathbf{L}_0^p(\nu)$ be the subset of $\mathbf{L}^p(\nu)$ whose expectation under $\nu$ is zero.

**Theorem 2** (First order chaos decomposition)**.** *Under appropriate assumptions. There exists two functions $f \in \mathbf{L}_0^2(\rho_0)$ and $g \in \mathbf{L}_0^2(\rho_1)$ such that*

$$T_n - \theta = \frac{1}{n} \sum_{i=1}^n [f(X_i) + g(Y_i)] + o_p(n^{-1/2}). \tag{9}$$

*In particular, as $n \to \infty$, $\sqrt{n}(T_n - \theta) \to_d \mathcal{N}(0, \varsigma^2)$ where $\varsigma^2 := \rho_0[f^2] + \rho_1[g^2]$.*

*Proof sketch.* Thanks to the contiguity result in Theorem 1, it suffices to prove (9) under the measure $\mu_\epsilon^n$. Recall from (2) that $\mu_\epsilon(x,y) = \xi(x,y)\rho_0(x)\rho_1(y)$.

*Step 1.* Find functions $f \in \mathbf{L}_0^2(\rho_0)$ and $g \in \mathbf{L}_0^2(\rho_1)$ such that

$$\begin{aligned} \mu_\epsilon^n[n(T_n - \theta) \mid X_i] &= \mu_\epsilon^n \left[ \sum_{k=1}^n [f(X_k) + g(Y_k)] \,\middle|\, X_i \right] = f(X_i) + \mu_\epsilon[g(Y_i) \mid X_i] \\ \mu_\epsilon^n[n(T_n - \theta) \mid Y_i] &= \mu_\epsilon^n \left[ \sum_{k=1}^n [f(X_k) + g(Y_k)] \,\middle|\, Y_i \right] = \mu_\epsilon[f(X_i) \mid Y_i] + g(Y_i) \end{aligned} \tag{10}$$

for all $i \in [n]$. By symmetry and exchangeability, we obtain

$$\begin{aligned} \kappa_{1,0}(x) &:= \mu_\epsilon^n[n(T_n - \theta) \mid X_i](x) = \int [c(x,y) - \theta] \xi(x,y) \rho_1(y) dy \\ \kappa_{0,1}(y) &:= \mu_\epsilon^n[n(T_n - \theta) \mid Y_i](y) = \int [c(x,y) - \theta] \xi(x,y) \rho_0(x) dx. \end{aligned}$$

**Algorithm 1** Gibbs sampling for discrete Schrödinger bridges

**Input:** samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, functions $c$ and $\xi$, burn-in $B$ and number of iterations $L$.
**Initialization:** $\sigma^{(0)} \leftarrow \mathrm{id}$.
**for** $t = 1, \ldots, L$ **do**
  Randomly generate $i \neq j \in [n]$.
  Compute $r \leftarrow \xi\big(X_i, Y_{\sigma_j^{(t-1)}}\big)\xi\big(X_j, Y_{\sigma_i^{(t-1)}}\big)/\xi\big(X_i, Y_{\sigma_i^{(t-1)}}\big)\xi\big(X_j, Y_{\sigma_j^{(t-1)}}\big)$.
  Generate $a \sim \mathrm{Bern}(r)$.
  **if** a = 1 **then**
    Obtain $\sigma^{(t)}$ from $\sigma^{(t-1)}$ by swapping the entries $\sigma_i^{(t-1)}$ and $\sigma_j^{(t-1)}$.
  **else**
    Set $\sigma^{(t)} \leftarrow \sigma^{(t-1)}$.
  **end if**
**end for**
**Output:** $T \leftarrow \frac{1}{L-B}\sum_{t=B+1}^{L} \frac{1}{n} c(X, Y_{\sigma^{(t)}})$.

If we define two linear operators $\mathcal{A} : \mathbf{L}^2(\rho_0) \to \mathbf{L}^2(\rho_1)$ and its adjoint $\mathcal{A}^* : \mathbf{L}^2(\rho_1) \to \mathbf{L}^2(\rho_0)$ by

$$(\mathcal{A}f)(y) = \int f(x)\xi(x,y)\rho_0(x)dx \quad \text{and} \quad (\mathcal{A}^* g)(x) = \int g(y)\xi(x,y)\rho_1(y)dy. \tag{11}$$

Then (10) becomes

$$\kappa_{1,0}(X_i) = f(X_i) + \mathcal{A}^* g(X_i) \quad \text{and} \quad \kappa_{0,1}(Y_i) = g(Y_i) + \mathcal{A}f(Y_i). \tag{12}$$

Hence, we can derive $f$ and $g$ by solving this linear system, i.e.,

$$f = (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1}) \quad \text{and} \quad g = (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0}). \tag{13}$$

*Step 2.* Control the remainder $R_1 := T_n - \theta - \frac{1}{n}\sum_{i=1}^n [f(X_i) + g(Y_i)]$ under the measure $\mu_\epsilon^n$. Some algebra shows that $R_1 = U_n / D_n$, where

$$U_n = \frac{1}{n \cdot n!} \sum_{\sigma \in \mathcal{S}_n} \widetilde{c}(X, Y_\sigma)\xi^{\otimes}(X, Y_\sigma) \quad \text{and} \quad D_n = \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \xi^{\otimes}(X, Y_\sigma)$$

with $\widetilde{c}(x,y) = c(x,y) - \theta - f(x) - g(y)$ and $\xi^{\otimes}(X, Y_\sigma) = \prod_{i=1}^n \xi(X_i, Y_{\sigma_i})$. By a change of measure argument, we obtain

$$\mu_\epsilon^n[|R_1|] = \mathbb{E}[|U_n|] \leq \sqrt{\mathbb{E}[U_n^2]},$$

where $\mathbb{E}$ is the expectation under $(\rho_0 \otimes \rho_1)^n$. Note that $U_n$ is a U-statistic of order $n$. To prove $\mathbb{E}[U_n^2] = o(n^{-1})$, we derive a novel Hoeffding decomposition [24, Chapter 11] of $U_n$ and upper bound its second moment. □

*Remark* 1. Our results actually hold for a large class of functions beyond the cost function: given an arbitrary function $\eta \in \mathbf{L}^1(\mu_\epsilon)$, the same results hold for $T_n(\eta) := \int \eta d\hat{\mu}_\epsilon^n$ and $\theta(\eta) := \int \eta d\mu_\epsilon$.

## 5 Two-sample testing with the Schrödinger bridge statistic

Given the two independent i.i.d. samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, suppose that we are interested in determining whether they come from the same distribution. This can be formalized as a two-sample hypothesis testing problem:

$$\mathbf{H}_0 : \rho_0 = \rho_1 \leftrightarrow \mathbf{H}_1 : \rho_0 \neq \rho_1. \tag{14}$$

That is, we test the null hypothesis that they come from the same distribution against the alternative hypothesis that they do not. To proceed, we define a test statistic, i.e., a real-valued function of the data, such that the larger it is the less likely $\mathbf{H}_0$ is true. Then we choose a threshold $h_n$ and adopt the decision rule (or test) $\mathbb{1}\{S_n > h_n\}$, that is, we reject the null if the test statistic exceeds the threshold. The performance of a test can be measured by two quantities: the type I error rate, i.e., the

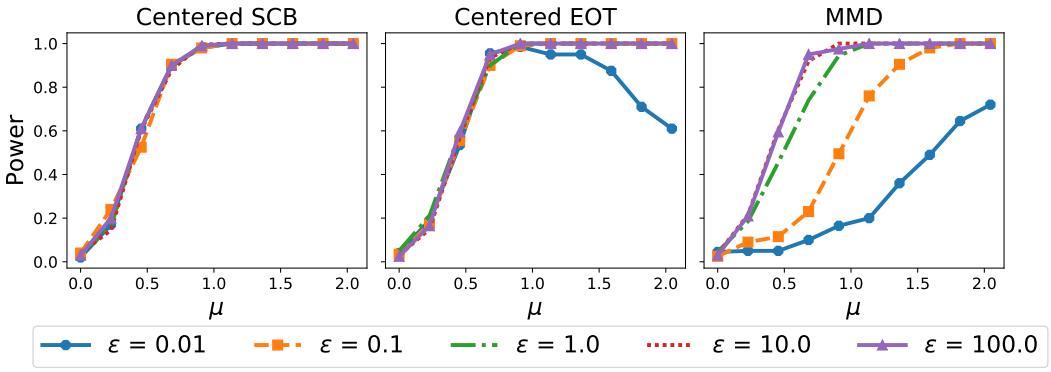


Figure 1: Statistical power versus $\mu$ for the pair $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$.

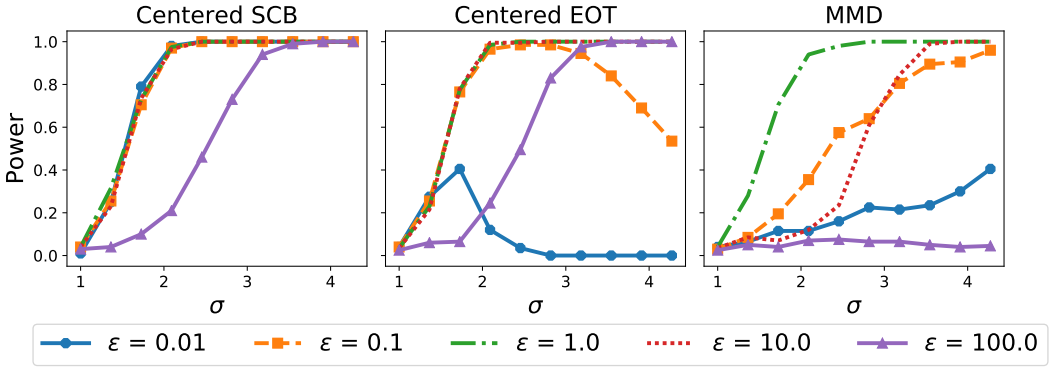


Figure 2: Statistical power versus $\sigma$ for the pair $\mathcal{N}(0,1)$ and $\mathcal{N}(0,\sigma^2)$.

probability of rejecting the null given that the null is true, and the statistical power, i.e., the probability of rejecting the null given that the null is not true.

Since the discrete Schrödinger bridge can be used to measure the difference between the two empirical measures $\hat{\rho}_0^n$ and $\hat{\rho}_1^n$, it can be used as a test statistic in two-sample testing. However, there are two caveats. First, the Schrödinger bridge statistic is biased in the sense that its limit $\theta \neq 0$ when $\rho_0 = \rho_1$. We instead use the centered Schrödinger bridge statistic

$$\bar{T}_n := T_n(\hat{\rho}_0^n, \hat{\rho}_1^n) - \frac{1}{2}T_n(\hat{\rho}_0^n, \hat{\rho}_0^n) - \frac{1}{2}T_n(\hat{\rho}_1^n, \hat{\rho}_1^n), \tag{15}$$

where $T_n(\hat{\rho}_i^n, \hat{\rho}_j^n)$ is the statistic associated with the discrete Schrödinger bridge connecting $\hat{\rho}_i^n$ and $\hat{\rho}_j^n$ for $i, j \in \{0,1\}$. Note that this centering procedure also appears in [19] where the discrete EOT is used for two-sample testing. Second, since the space of permutations $\mathcal{S}_n$ is prohibitively large, it is infeasible to compute the Schrödinger bridge statistic exactly. We adopt here a Gibbs sampling approach and summarize the procedure in Algorithm 1.

## 6 Experiments

In this section, we apply the (centered) Schrödinger bridge (SCB) statistic to two-sample testing on both synthetic and real data. We compare its type I error rate and statistical power with the discrete EOT [19] and the maximum mean discrepancy (MMD) [12]. For comparison purposes, all the thresholds are determined by permutation test: we 1) randomly permute the pooled sample $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$, 2) compute the test statistic for the permuted sample, and 3) repeat previous two steps for 500 times and choose the threshold as the upper 5-percentile of these statistics. This procedure guarantees that the type I error rates of the three tests are all close to 0.05.

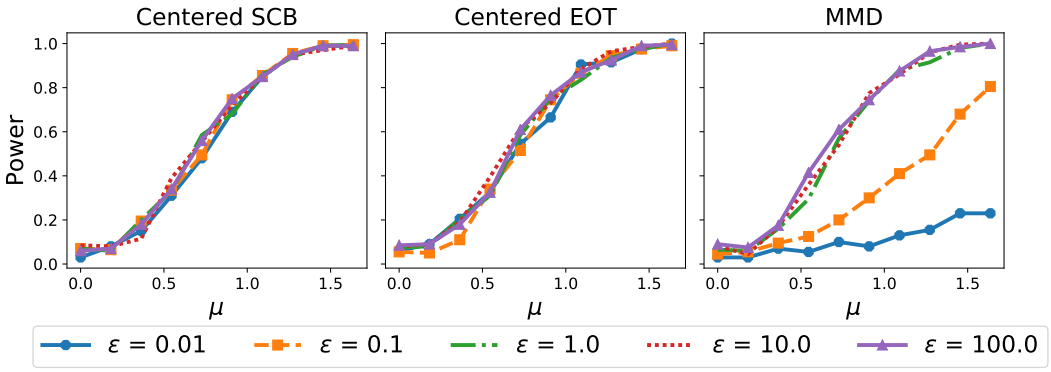


Figure 3: Statistical power versus $\mu$ for the pair $\mathrm{VM}(0,1)$ and $\mathrm{VM}(\mu,1)$.

### 6.1 Synthetic data

**Settings.** We consider 4 different pairs of distributions: 1) $\mathcal{N}(0,1)$ v.s. $\mathcal{N}(\mu,1)$, 2) $\mathcal{N}(0,1)$ v.s. $\mathcal{N}(0,\sigma^2)$, 3) $\mathrm{VM}(0,1)$ v.s. $\mathrm{VM}(\mu,1)$, and 4) $\mathrm{VM}(0,1)$ v.s. $\mathrm{VM}(0,\kappa)$, where $\mathrm{VM}(\mu,\kappa)$ is the von Mises distribution with location $\mu$ and concentration $\kappa$. For each pair of distributions $(\rho_0,\rho_1)$, we independently generate $n=50$ i.i.d. observations from each of the distributions. Then we perform the three tests and record their decisions. For the Schrödinger bridge test and EOT test, we use the quadratic cost and set $\epsilon \in \{0.01,0.1,1,10,100\}$. For the MMD test, we use the RBF kernel $k(x,x') = \exp(-\|x-x'\|^2/\epsilon)$ and set $\epsilon \in \{0.01,0.1,1,10,100\}$. We repeat the whole procedure 200 times and compute the rejection frequency. We plot the rejection frequency as we vary the parameter (e.g., $\mu$ in the first pair). When $\rho_0=\rho_1$, the rejection frequency is an estimate of the type I error rate; when $\rho_0\neq\rho_1$, it is an estimate of the statistical power.

**Normal distribution.** The results for normal distributions are in Figure 1 and Figure 2. When the two distributions differ in mean, the Schrödinger bridge test demonstrates similar performance across different values of $\epsilon$. The EOT test shows a similar behavior except for $\epsilon=0.01$: the statistical power increases in the beginning and then decreases as $\mu$ increases. This decline is due to the computational instability of the Sinkhorn algorithm used to compute the EOT when $\epsilon$ is relatively small. As for the MMD test, its performance largely depends on the parameter in the RBF kernel. The three tests perform analogously with their best parameter. When the two distributions differ in variance, most of the findings are the same. The parameter $\epsilon=100$ gives significantly worse performance and the instability issue in the EOT test is more prominent.

**Von Mises distribution.** The results for von Mises distributions are in Figure 3 and Figure 4. The Schrödinger bridge test and EOT test performs similarly without the instability issue. The performance of the MMD test heavily depends on the parameter $\epsilon$, and its statistical power with the best parameter is close to the ones of the Schrödinger bridge test and the EOT test.

### 6.2 Real data

**Settings.** We compare the three tests on the MNIST dataset [14]. Given two digits $m_1$ and $m_2$, we randomly sample $n \in \{5,10,15,20,25,30\}$ images from each of the two classes. For the Schrödinger bridge test and EOT test, we define the cost on images as follows: for each image $M$, we normalize it and view it as a discrete distribution; the cost between two images is then chosen as the Wasserstein-2 distance between the corresponding discrete distributions. Again, we set the regularization parameter $\epsilon \in \{0.01,0.1,1,10,100\}$. For the MMD test, we use $k(M_1,M_2) := \exp(-c(M_1,M_2)/\varepsilon)$ as the kernel on images, where $c$ is the cost defined above. We repeat the whole procedure 200 times and plot the rejection frequency as we vary the sample size.

**Results.** The results for $m_1=m_2=3$ is shown in Figure 5. The type I error rate of all the tests are close to 0.05 with different parameters. The results for $m_1=3$ and $m_2=5$ is presented in Figure 6. All the tests performs similarly with the MMD test with $\varepsilon=0.01$ slightly better. All the

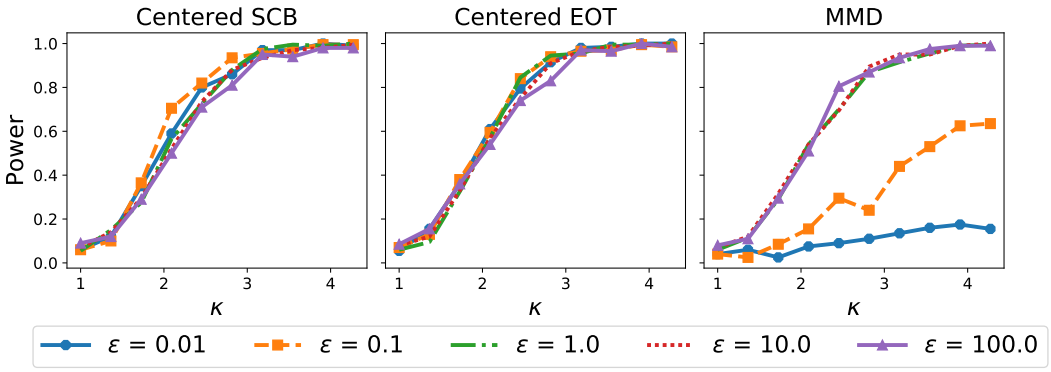


Figure 4: Statistical power versus $\kappa$ for the pair $\mathrm{VM}(0, 1)$ and $\mathrm{VM}(0, \kappa)$.

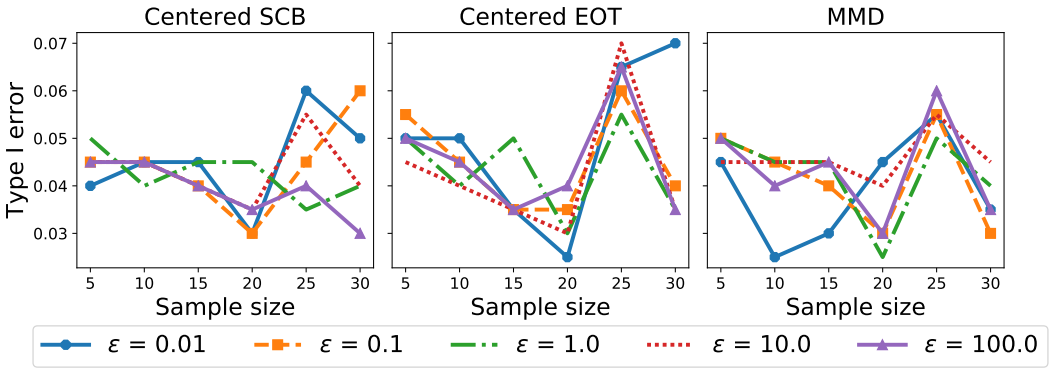


Figure 5: Type I error rate versus sample size for digits 3 and 3.

tests achieve power 1 with a relatively small sample size, and their performance are robust to the value of parameters considered in the experiments.

## References

[1] E. Bernton, J. Heng, A. Doucet, and P. E. Jacob. Schrödinger bridge samplers. *arXiv preprint*, 2019.

[2] P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Number 1. Springer-Verlag New York, 1998.

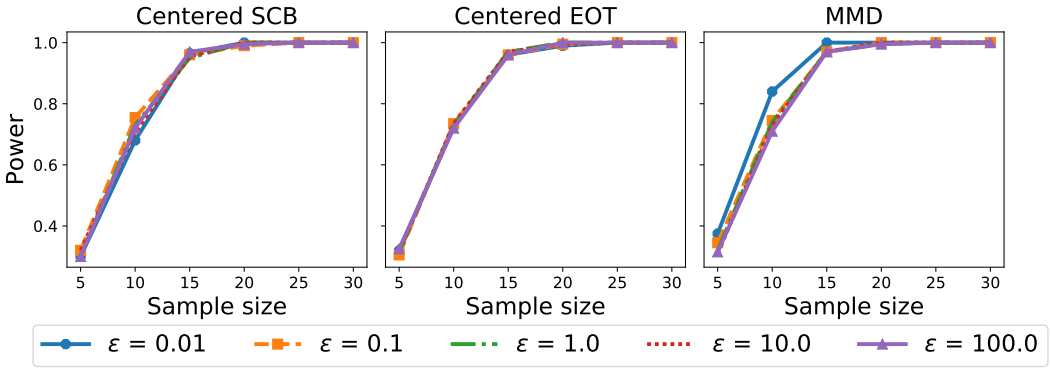


Figure 6: Statistical power versus sample size for digits 3 and 5.

[3] J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electron. J. Statist.*, 13(2):5120–5150, 2019.

[4] V. D. Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *arXiv preprint*, 2021.

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

[6] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

[7] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[8] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[9] H. Föllmer. Random fields and diffusion processes. In *École d'été de probabilités de Saint-Flour XV-XVII-1985-87*, volume 1362 of *Lecture Notes in Mathematics*. Springer, Berlin, 1988.

[10] A. Genevay, L. Chizat, F. R. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *AISTATS*, pages 1574–1583, 2019.

[11] I. Gohberg, S. Goldberg, and M. Kaashoek. *Classes of Linear Operators Vol. 1*. Birkhäuser, Basel, 1990.

[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[13] M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.*, 2(2):419–443, 2020.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] C. Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[16] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.*, 34(4):1533–1574, 2014.

[17] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*, pages 4543–4553, 2019.

[18] S. Pal and T.-K. L. Wong. Multiplicative Schrödinger problem and the Dirichlet transport. *Probability Theory and Related Fields*, 178(1):613–654, 2020.

[19] A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

[20] P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11):1228 – 1235, 2018.

[21] H. H. Rugh. Cones and gauges in complex spaces: Spectral gaps and complex Perron-Frobenius theory. *Annals of Mathematics*, 171(3):1707–1752, 2010.

[22] L. Rüschendorf and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statistics & Probability Letters*, 17:369–375, 1993.

[23] E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Ann. Inst. H. Poincaré*, 2:269–310, 1932.

[24] A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

# A Appendix

Given a probability measure $\nu$ and $p \geq 1$, let $\mathbf{L}^p(\nu)$ be the space of functions that have finite $p$-th norm under $\nu$. We follow the standard abuse of keeping the same notation for an absolutely continuous measure and its density.

**Assumption 1.** All the results stated below hold under the following assumptions.

1. $c$ is a nonnegative continuous cost function such that $c(x, y) = 0$ if and only if $x = y$, and satisfies the following asymptotic growth bound: for some $a, b > 0$ and for some $p \geq 1$,
$$c(x, y) \leq a + b\left(|x|^p + |y|^p\right), \text{ as } |x|, |y| \to \infty. \tag{16}$$
2. $\rho_0$ and $\rho_1$ have finite $p$th moment. Consequently, $\int c(x, y)d\mu_\epsilon < \infty$.
3. The Schrödinger potentials are integrable, i.e., $a_\epsilon \in \mathbf{L}^1(\rho_0)$ and $b_\epsilon \in \mathbf{L}^1(\rho_1)$. See [22] for sufficient conditions.

Let $\eta$ be any function on $\mathbb{R}^d \times \mathbb{R}^d$ integrable under $\mu_\epsilon$. Let $T_n = T_n(\eta) := \int \eta(x, y)d\hat{\mu}_\epsilon^n$ and let $\theta = \int \eta(x, y)d\mu_\epsilon$. Explicitly, $T_n = \sum_{\sigma\in\mathcal{S}_n} q_\epsilon^*(\sigma)\frac{1}{N}\sum_{i=1}^n \eta(X_i, Y_{\sigma_i})$.

**Theorem 2.** ***(Consistency.)*** *As $n \to \infty$, $T_n$ converges in probability to $\theta$ for all $\eta \in \mathbf{L}^1(\mu_\epsilon)$. In particular, $\hat{\nu}_\epsilon^n$ converges weakly to $\mu_\epsilon$, in probability.*

$T_n$ is a rather complicated function of the two empirical distributions $(\widehat{\rho}_0^n, \widehat{\rho}_1^n)$. Our next result shows that it can, in fact, be well approximated by linear functions of the two measures in a way that is similar to the first term of a Taylor expansion of smooth functions. Let us start by defining some operators on various $\mathbf{L}^2$ spaces.

**Definition 1.** Define linear operators $\mathcal{A} : \mathbf{L}^2(\rho_0) \to \mathbf{L}^2(\rho_1)$ and its adjoint $\mathcal{A}^* : \mathbf{L}^2(\rho_1) \to \mathbf{L}^2(\rho_0)$ by
$$(\mathcal{A}f)(y) = \int f(x)\xi(x, y)\rho_0(x)dx, \quad (\mathcal{A}^*g)(x) = \int g(y)\xi(x, y)\rho_1(y)dy. \tag{17}$$

It can be shown that $\mathcal{A}$ is a well-defined linear operator, and $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ are two Markov operators defined on $\mathbf{L}^2(\rho_0)$ and $\mathbf{L}^2(\rho_1)$, respectively. We denote by $I_\nu : \mathbf{L}^2(\nu) \to \mathbf{L}^2(\nu)$ the identity operator on $\mathbf{L}^2(\nu)$. When the context is clear, we will write $I$ for short. We further make the following assumptions.

**Assumption 2.** All the results stated below hold under the following additional assumptions.

1. $\xi \in \mathbf{L}^2(\rho_0 \otimes \rho_1)$, $\eta \in \mathbf{L}^2(\mu_\epsilon)$ and $\eta\xi \in \mathbf{L}^2(\rho_0 \otimes \rho_1)$.
2. As a consequence [2, Appendix A.4], the operator $\mathcal{A}$ is compact. Then the operators $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ admit eigenvalue decomposition $\mathcal{A}^*\mathcal{A}\alpha_k = s_k^2\alpha_k$ and $\mathcal{A}\mathcal{A}^*\beta_k = s_k^2\beta_k$ for all $k \geq 0$ with $s_0 = 1$, $\alpha_0 = \beta_0 = \mathbf{1}$ and $0 \leq s_k \leq 1$ for all $k \geq 0$. Moreover, it holds that $\mathcal{A}\alpha_k = s_k\beta_k$ and $\mathcal{A}^*\beta_k = s_k\alpha_k$; see [11, Chapter 6.1]. We call $\{s_k\}_{k\geq 0}$ the singular values of $\mathcal{A}$ and $\mathcal{A}^*$, and call $\{\alpha_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq 0}$ the singular functions.
3. The operators $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ have positive eigenvalue gap, *i.e.*, $s_k \leq s_1 < 1$ for all $k \geq 1$. By Jentzsch's Theorem [21, Theorem 7.2], a sufficient condition is that $\xi$ is bounded.

**Theorem 3.** ***(First order chaos decomposition)*** *Recall $\theta := \int \eta(x, y)d\mu_\epsilon$. Define*
$$\kappa_{1,0}(x) := \int [\eta(x, y) - \theta]\xi(x, y)\rho_1(y)dy, \tag{18}$$
$$\kappa_{0,1}(y) := \int [\eta(x, y) - \theta]\xi(x, y)\rho_0(x)dx. \tag{19}$$
*Then, $T_n - \theta = \mathcal{L}_1 + o_p\big(1/\sqrt{n}\big)$, where*
$$\mathcal{L}_1 := \frac{1}{n}\sum_{i=1}^n [(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_i)].$$
*In particular, we have $\sqrt{n}(T_n - \theta) \to_d \mathcal{N}(0, \varsigma^2)$, where $\varsigma^2 = \varsigma^2(\eta)$, as a function of $\eta$, is given by*
$$\varsigma^2 = \mathbb{E}\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_1)^2\right] + \mathbb{E}\left[(I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_1)^2\right].$$