

---

# Synthesis and Analysis of Data as Probability Measures With Entropy-Regularized Optimal Transport

---

Brendan Mallery<sup>(†)</sup>, James M. Murphy<sup>(†,\*)</sup>, Shuchin Aeron<sup>(□,\*)</sup>  
Tufts University

## Abstract

We consider synthesis and analysis of probability measures using the entropy-regularized Wasserstein-2 cost and its unbiased version, the Sinkhorn divergence. The synthesis problem consists of computing the barycenter, with respect to these costs, of reference measures given a set of coefficients belonging to the simplex. The analysis problem consists of finding the coefficients for the closest barycenter in the Wasserstein-2 distance to a given measure. Under the weakest assumptions on the measures thus far in the literature, we compute the derivative of the entropy-regularized Wasserstein-2 cost. We leverage this to establish a characterization of barycenters with respect to the entropy-regularized Wasserstein-2 cost as solutions that correspond to a fixed point of an average of the entropy-regularized displacement maps. This characterization yields a finite-dimensional, convex, quadratic program for solving the analysis problem when the measure being analyzed is a barycenter with respect to the entropy-regularized Wasserstein-2 cost. We show that these coefficients, as well as the value of the barycenter functional, can be estimated from samples with dimension-independent rates of convergence, and that barycentric coefficients are stable with respect to perturbations in the Wasserstein-2 metric. We employ the barycentric coefficients as features for classification of corrupted point cloud data, and show that compared to neural network baselines, our approach is more efficient in small training data regimes.

## 1 INTRODUCTION

Modeling data as probability measures is an emerging theme with applications in signal analysis (Cazelles et al., 2020; Cheng et al., 2023), image processing (Rabin et al., 2012; Nadeem et al., 2020; Simon and Aberdam, 2020), natural language processing (NLP) (Xu et al., 2018) and beyond. In this setting, each data point is treated as a probability distribution over an appropriate domain. We seek to develop useful data processing models and methods—akin to those widely used in statistical signal processing, such as principal component analysis (PCA) (Pearson, 1901), archetypal analysis (Cutler and Breiman, 1994), and non-negative matrix factorization (NMF) (Lee and Seung, 2000)—that allow for *synthesis* of new distributions from a canonical reference set of distributions, as well as *analysis* of a given distribution in terms of these references.

Wasserstein-2 barycenters provide a natural method to perform both synthesis and analysis of data using the methods of computational optimal transport (OT) (Peyré and Cuturi, 2019). One can synthesize a new distribution from a given set of reference measures with their barycenter (Dognin et al., 2019; Rabin et al., 2012; Chzhen et al., 2020), and analyze a measure by assigning it barycentric coefficients with respect to a set of reference measures (Bonneel et al., 2016; Schmitz et al., 2018; Werenski et al., 2022; Mueller et al., 2023). Wasserstein-2 barycenters can geometrically interpolate between the reference measures, which allows one to model data outside of their support that is useful for image processing (Simon and Aberdam, 2020) and time-series modeling (Cheng et al., 2021, 2023). This is in contrast to barycenters defined via maximum mean discrepancy distances, which are linear mixtures of the reference measures (Cohen et al., 2020), or via information theoretic divergences such as relative entropy (Cover, 1999), which require overlapping support of the underlying measures.

However, computation of optimal transport costs is not parallelizable and can incur computational complex-

---

<sup>(†)</sup> Department of Mathematics. <sup>(□)</sup> Department of Electrical and Computer Engineering. \* co-last authors. Correspondence to: brendan.mallery@tufts.edu. Proceedings of the 28<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

ity that is cubic in the number of samples. Furthermore, synthesizing even an approximate Wasserstein-2 barycenter to a fixed degree of accuracy in arbitrary dimensions cannot be done in polynomial time (Altschuler and Boix-Adsera, 2022), and known rates of estimation of both synthesis and analysis scale exponentially in the ambient dimension  $d$  (Werenski et al., 2022) due to the curse of dimensionality that is inherent to estimating the Wasserstein-2 costs and maps (Fournier and Guillin, 2015; Hütter and Rigollet, 2021; Pooladian and Niles-Weed, 2021; Deb et al., 2021).

On the other hand, entropy-regularized optimal transport does not suffer from these computational and statistical issues. Fast computation via parallelization is enabled via the Sinkhorn-Knopp algorithm (Sinkhorn and Knopp, 1967; Cuturi, 2013) and the population entropy-regularized optimal transport costs and maps can be estimated from samples with dimension-free rates of convergence (Mena and Niles-Weed, 2019; Bigot et al., 2019; Pooladian and Niles-Weed, 2021; Rigollet and Stromme, 2022; Werenski et al., 2023; Masud et al., 2023; Stromme, 2023).

Motivated by this, in this paper we consider entropy-regularized variants of Wasserstein-2 barycenters, which are defined in (5) and (6). We are particularly concerned with analysis of data via (regularized) barycentric coefficients, which has until recently been relatively under-studied in comparison to synthesis.

The rest of the paper is organized as follows. In Section 1.1 we present the necessary notation and background. We then detail our main contributions and related work in Section 1.2. The main technical results are established in Sections 2, 3, and 4 along with numerical results. Applications to point cloud classification are presented in Section 5<sup>1</sup>. We defer our proofs and additional experiments to the Appendix.

## 1.1 Notation and Background

**Notation:**  $\Omega$  is a closed subset of  $\mathbb{R}^d$  and  $\mathcal{P}(\Omega)$  is the set of probability measures with support contained in  $\Omega$ .  $|\Omega|$  is the diameter of  $\Omega$  with respect to the Euclidean distance. For  $\mu \in \mathcal{P}(\Omega)$ , its expectation is denoted by  $\mathbb{E}(\mu)$ , its variance is denoted by  $\text{Var}(\mu)$ , and its second moment is denoted by  $M_2(\mu)$ .  $\mathcal{P}_2(\Omega)$  denotes the subset of  $\mu \in \mathcal{P}(\Omega)$  with finite  $M_2(\mu)$ .  $\mathcal{G}(\Omega)$  denotes the set of subgaussian probability measures with support contained in  $\Omega$ , and  $\mathcal{G}_\sigma(\Omega)$  is the set of subgaussian probability measures with support in  $\Omega$  with subgaussian constant  $\leq \sigma$ .  $\mathcal{P}^n(\Omega)$  denotes the subset of  $\mathcal{P}(\Omega)$  supported on  $n$  or fewer points. We let

$\delta_z$  denote the Dirac probability measure supported at  $z \in \mathbb{R}^d$ . For a probability measure  $\mu$ ,  $\hat{\mu}^n$  is the random (empirical) measure defined by  $\hat{\mu}^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ , where  $X_1, X_2, \dots, X_n$  are i.i.d. samples from  $\mu$ . The ball of radius (with respect to the Euclidean distance)  $t$  centered at  $x \in \mathbb{R}^d$  is denoted by  $B_t(x)$ . For a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  and a finite, signed measure  $\mu$  on  $\mathbb{R}^d$ ,  $\langle f, \mu \rangle := \int f d\mu$ . For any  $\mu \ll \nu$ ,  $\frac{d\mu}{d\nu}$  denotes the Radon-Nikodym derivative of  $\mu$  with respect to  $\nu$ , i.e., a measurable function such that  $\int_A \frac{d\mu}{d\nu} d\nu = \mu(A)$ , for any measurable  $A$ .

For  $\mu, \nu \in \mathcal{P}_2(\Omega)$ , the *Wasserstein-2 distance* (Villani, 2009) between  $\mu$  and  $\nu$  is:

$$OT_2(\mu, \nu) := \inf_{\zeta \in \Pi(\mu, \nu)} \sqrt{\int \frac{1}{2} \|x - y\|^2 d\zeta(x, y)},$$

where  $\Pi(\mu, \nu) \subset \mathcal{P}_2(\Omega \times \Omega)$  is the set of couplings with marginals  $\mu$  and  $\nu$ . For any  $\epsilon > 0$ , the *entropy-regularized, or entropy-regularized Wasserstein-2 cost* is (Nutz, 2021):

$$OT_2^\epsilon(\mu, \nu) := \inf_{\zeta \in \Pi(\mu, \nu)} \int \frac{1}{2} \|x - y\|^2 d\zeta(x, y) + \epsilon KL(\zeta \| \mu \otimes \nu), \quad (1)$$

where  $KL(\zeta \| \mu \otimes \nu)$  is the relative entropy between  $\zeta$  and  $\mu \otimes \nu$  (Cover, 1999). For any  $\epsilon > 0$ , we define the *Sinkhorn divergence* (Genevay et al., 2018):  $\mathcal{S}_2^\epsilon(\mu, \nu) = OT_2^\epsilon(\mu, \nu) - \frac{1}{2} OT_2^\epsilon(\mu, \mu) - \frac{1}{2} OT_2^\epsilon(\nu, \nu)$ .

The entropy-regularized cost admits a dual formulation given by:

$$\begin{aligned} OT_2^\epsilon(\mu, \nu) &= \sup_{(f, g) \in L^1(\mu) \times L^1(\nu)} \int f(x) d\mu(x) + \int g(y) d\nu(y) \\ &\quad - \epsilon \int \left( e^{-\frac{1}{\epsilon} (\frac{1}{2} \|x - y\|^2 - f(x) - g(y))} - 1 \right) d\mu(x) d\nu(y). \end{aligned} \quad (2)$$

The optimizers are a pair of functions  $(f_{\mu \rightarrow \nu}^\epsilon, g_{\mu \rightarrow \nu}^\epsilon)$ , unique  $\mu$  (resp.  $\nu$ )-a.e. up to translation  $(f_{\mu \rightarrow \nu}^\epsilon, g_{\mu \rightarrow \nu}^\epsilon) \mapsto (f_{\mu \rightarrow \nu}^\epsilon - C, g_{\mu \rightarrow \nu}^\epsilon + C)$ ,  $C \in \mathbb{R}$ , which we refer to as *entropic potentials*. They satisfy:

$$f_{\mu \rightarrow \nu}^\epsilon(x) = -\epsilon \log \left( \int e^{-\frac{1}{\epsilon} (\frac{1}{2} \|x - y\|^2 - g_{\mu \rightarrow \nu}^\epsilon(y))} d\nu(y) \right), \quad (3)$$

$$g_{\mu \rightarrow \nu}^\epsilon(y) = -\epsilon \log \left( \int e^{-\frac{1}{\epsilon} (\frac{1}{2} \|x - y\|^2 - f_{\mu \rightarrow \nu}^\epsilon(x))} d\mu(x) \right), \quad (4)$$

for  $\mu$ -a.e.  $x \in \mathbb{R}^d$  and  $\nu$ -a.e.  $y \in \mathbb{R}^d$  (Nutz, 2021). These equations allow one to define canonical extensions of  $f_{\mu \rightarrow \nu}^\epsilon$  and  $g_{\mu \rightarrow \nu}^\epsilon$  to all  $x, y \in \mathbb{R}^d$  such that (3)

<sup>1</sup>Code to recreate experiments is available at <https://github.com/brendanmallery9/Entropic-Barycenters>

and (4) are satisfied for all  $x, y \in \mathbb{R}^d$ . We refer to these as *extended entropic potentials*. These extensions are unique for a fixed pair of entropic potentials  $f_{\mu \rightarrow \nu}^\epsilon$  and  $g_{\mu \rightarrow \nu}^\epsilon$ .

Henceforth, we will only deal with extended entropic potentials. The extensions of  $f_{\mu \rightarrow \nu}^\epsilon$  and  $g_{\mu \rightarrow \nu}^\epsilon$  are a priori only finite  $\mu$  and  $\nu$ -almost everywhere respectively, but if  $\mu, \nu \in \mathcal{G}(\mathbb{R}^d)$  then  $f_{\mu \rightarrow \nu}^\epsilon$  and  $g_{\mu \rightarrow \nu}^\epsilon$  are finite and smooth everywhere (Mena and Niles-Weed, 2019). However it can be shown that  $f_{\mu \rightarrow \nu}^\epsilon$  is finite and smooth everywhere if  $\nu$  is subgaussian and  $\mu$  is in  $\mathcal{P}_2(\mathbb{R}^d)$ , i.e., only subgaussianity of the “target” measure is required for one of the potentials to be smooth and finite everywhere (see Proposition 2.3)

Let  $\mathcal{V} = \{\nu_j\}_{j=1}^m \subset \mathcal{P}_2(\mathbb{R}^d)$ . Let  $\lambda$  denote the vector of coefficients belonging to the  $(m-1)$ -dimensional simplex  $\Delta^m = \{(\lambda_1, \lambda_2, \dots, \lambda_m) \in \mathbb{R}^m : \sum_{i=1}^m \lambda_i = 1, \lambda_i \geq 0\}$ .

**Definition 1.1.** Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$  and let  $\epsilon > 0$ .

- The  $\epsilon$ -entropic barycenter functional for  $(\lambda, \mathcal{V})$  is defined as  $F_{\lambda, \mathcal{V}}^\epsilon(\mu) := \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu, \nu_j)$ .
- The  $\epsilon$ -Sinkhorn barycenter functional for  $(\lambda, \mathcal{V})$  is defined as:  $S_{\lambda, \mathcal{V}}^\epsilon(\mu) := \sum_{j=1}^m \lambda_j S_2^\epsilon(\mu, \nu_j)$ .

We write  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$  to refer to either  $F_{\lambda, \mathcal{V}}^\epsilon$  or  $S_{\lambda, \mathcal{V}}^\epsilon$ . We note that  $F_{\lambda, \mathcal{V}}^\epsilon$  is convex on  $\mathcal{P}_2(\mathbb{R}^d)$ , and  $S_{\lambda, \mathcal{V}}^\epsilon$  is convex when restricted to the set of subgaussian measures (Proposition 4 in (Janati et al., 2020a)). We now define the synthesis and analysis problems.

Given  $\lambda$  and  $\mathcal{V}$ , the *synthesis* problem refers to solving for:

$$\arg \min_{\mu \in \mathcal{P}_2(\Omega)} \mathcal{F}_{\lambda, \mathcal{V}}^\epsilon(\mu). \quad (5)$$

Given  $\mu$  and  $\mathcal{V}$ , the *analysis* problem refers to solving for:

$$\arg \min_{\lambda \in \Delta^m} OT_2(\mu, \rho_\lambda), \quad (6)$$

where  $\rho_\lambda = \arg \min_{\rho \in \mathcal{P}_2(\Omega)} \mathcal{F}_{\lambda, \mathcal{V}}^\epsilon(\rho)$ . Solving (5) produces a measure that can be viewed as a nonlinear combination of the reference measures  $\mathcal{V}$ . Solving (6) produces coefficients  $\lambda$  that parameterize the “projection” of  $\mu$  onto the set of solutions to (5), which we refer to as the barycentric coefficients of  $\mu$  with respect to the reference measures. The synthesis and analysis problems are in some sense inverse to one another when  $\mu$  in (6) is restricted to the set of minimizers of  $\mathcal{F}_{\lambda_*, \mathcal{V}}^\epsilon$ , for some  $\lambda_* \in \Delta^m$ . Indeed, if  $\mu \in \arg \min_{\rho \in \mathcal{P}_2(\Omega)} \mathcal{F}_{\lambda_*, \mathcal{V}}^\epsilon(\rho)$ , then  $\min_{\lambda \in \Delta^m} OT_2(\mu, \rho_\lambda) = OT_2(\mu, \rho_{\lambda_*}) = 0$ . In other words, the analysis of a  $(\lambda_*, \mathcal{V})$ -barycenter recovers  $\lambda_*$ . In this paper, we show that for this special case it is possible to reduce (6) to a convex, quadratic program (see Section 4).

We end this section with definitions from functional calculus.

**Definition 1.2.** Let  $\mathcal{U} \subseteq \mathcal{P}_2(\mathbb{R}^d)$  be convex. A functional  $\mathcal{F} : \mathcal{U} \rightarrow \mathbb{R}$  admits a derivative at  $\mu \in \mathcal{U}$  (is differentiable at  $\mu$ ) if there exists a continuous function  $\delta_\mu \mathcal{F}(\mu) : \mathbb{R}^d \rightarrow \mathbb{R}$ , such that for any displacement  $\chi := \rho - \mu$ , where  $\rho \in \mathcal{U}$ ,  $\delta_\mu \mathcal{F}(\mu)$  is  $\chi$ -integrable and satisfies:

$$\lim_{t \rightarrow 0^+} \frac{\mathcal{F}(\mu + t\chi) - \mathcal{F}(\mu)}{t} = \int \delta_\mu \mathcal{F}(\mu) d\chi. \quad (7)$$

A functional which is differentiable for all  $\mu \in \mathcal{U}$  is said to be differentiable on  $\mathcal{U}$ .

When the argument of  $\mathcal{F}$  is clear, we will write  $\delta_\mu \mathcal{F}(\mu)$  as  $\delta \mathcal{F}(\mu)$ . Key to our investigation is the notion of critical points of functionals of probability measures.

**Definition 1.3.** Let  $\mathcal{U}$  be either  $\mathcal{P}_2(\mathbb{R}^d)$  or  $\mathcal{G}(\mathbb{R}^d)$ , and let  $\mathcal{F}$  be a functional on  $\mathcal{U}$  which is differentiable at  $\mu \in \mathcal{U}$ . We say that  $\mu$  is a critical point of  $\mathcal{F}$  on  $\mathcal{U}$  if  $\nabla \delta \mathcal{F}(\mu)(x) = 0$  for  $\mu$ -almost every  $x$ .

We will simply say that  $\mu$  is a critical point of  $\mathcal{F}$  if the domain of  $\mathcal{F}$  is clear. Importantly, minimizers of a differentiable functional on  $\mathcal{U}$  are critical points (see Corollary A.11). We remark that under sufficient regularity assumptions on  $\mathcal{F}$  (e.g., as made in (Chizat, 2022)),  $\nabla \delta \mathcal{F}(\mu)$  corresponds to what is referred to as the Wasserstein-2 gradient of the functional  $\mathcal{F}$  at  $\mu$  (Ambrosio et al., 2005).

## 1.2 Main Contributions

### 1.2.1 Critical Points and Optimality Criteria for $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$

In Section 2, we rigorously establish analytic properties of  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$  for subgaussian  $\mathcal{V}$ . First, we establish existence of minima for  $\min_{\mu \in \mathcal{P}_2(\Omega)} F_{\lambda, \mathcal{V}}^\epsilon$  and  $\min_{\mu \in \mathcal{G}_\sigma(\Omega)} S_{\lambda, \mathcal{V}}^\epsilon$  for fixed  $\sigma$ . We then establish Theorem 2.2, in which we prove that the entropic cost  $OT_2^\epsilon(\mu, \nu)$  (as a function of  $\mu$ ) is differentiable when  $\nu$  is subgaussian. This extends Proposition 3 in (Janati et al., 2020a), which establishes this when both  $\mu$  and  $\nu$  are subgaussian, and to our knowledge provides the most general version of this result available in the literature. From this we obtain a first-order characterization of critical points of  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$ . We show in Proposition 2.7 that when  $\mathcal{V}$  is subgaussian, then any critical point is subgaussian. This generalizes results in (Yang et al., 2024) for the case of  $\mathcal{V} = \{\nu\}$ . Finally, we obtain a sufficient criteria to ensure a critical point of  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$  is a minimizer in Corollary 2.5.

### 1.2.2 Sample Complexity for Synthesis and Analysis

In Sections 3 and 4 we investigate the sample complexity of synthesis and analysis for  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$ .

Our main result for synthesis is Theorem 3.1, which establishes that estimating the optimal value of the barycenter functionals using samples from the reference measures can be achieved with *dimension-free* error rates. Our proof is a generalization of the arguments in (Yang et al., 2024) where this result is proved for  $m = 1$ , and contrasts with the Wasserstein-2 case where rates are typically exponential in the dimension (e.g., Theorem 2 in (Chizat et al., 2020)). A similar result for measures supported on a bounded set was achieved in (Luise et al., 2019), though they are also able to guarantee bounds with high probability. We then discuss how a potentially stronger sample complexity result fails to be true in the case of the Sinkhorn divergence barycenter.

Next, we consider the analysis problem. In Proposition 4.1, we establish that when  $\mu$  is an  $\epsilon$ -entropic (resp. Sinkhorn) barycenter for a reference set of measures  $\mathcal{V}$ , then the analysis problem (6) can be solved via a convex, quadratic program. In Theorem 4.2, we use this characterization to show that under weak assumptions, solving the analysis problem for an  $\epsilon$ -entropic barycenter  $\mu$  is no harder than estimating  $\nabla \delta \mathcal{F}_{\lambda, \mathcal{V}}^\epsilon(\mu)$  from samples. We apply this in two important cases. First, if  $\Omega$  is compact,  $\lambda$  may be estimated for  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$  at a rate proportional to  $n^{-1/2}$  via results obtained in (Rigollet and Stromme, 2022). Second, if  $\Omega = \mathbb{R}^d$  and the reference measures are strongly log-concave with mean zero, then  $\lambda$  may be estimated for  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$  at a rate proportional to  $n^{-1/3}$ , via results from (Werenski et al., 2023). Finally, we show that the barycentric coefficients are stable (with respect to the Wasserstein-2 distance) when the analyzed measure is perturbed (Proposition 4.5). We numerically verify the rates in Theorem 4.2 in Section 4.1 with several choices of reference measures, observing even faster convergence rates than established in Theorem 4.2. We provide an additional demonstration of the synthesis and analysis pipeline with probability measures obtained from MNIST digits (LeCun, 1998) in Appendix G.

### 1.2.3 Application to Point Cloud Classification with Barycentric coefficients

We demonstrate the utility of our approach by using the estimated barycentric coefficients as features for point cloud classification. Our method is described in Algorithm 2, that makes use of Proposition 4.1. In our experiments (detailed in Section 5), our method

is shown to achieve high accuracy using a small set of reference measures from each class, even when the test data is corrupted with various forms of noise and occlusions. We compare our method to PointNet (Qi et al., 2017), a well-known neural network architecture designed for point cloud classification. We observe that our method is able to outperform PointNet even when PointNet is trained on a much larger data set than is required by our method, suggesting the effectiveness of our approach in low-data settings. We compare our method to the analogous classification scheme using the unregularized barycenter functional (in which case, our method agrees with the method studied in (Gunsilius et al., 2024)) and the doubly-regularized barycenter functional introduced in (Chizat, 2023). We provide an application to point cloud completion in Appendix I.

### 1.3 Related Work

Synthesis for  $F_{\lambda, \mathcal{V}}^\epsilon$  was first studied in (Cuturi and Doucet, 2014) as a computationally efficient surrogate for Wasserstein-2 barycenters. Synthesis for the functional  $S_{\lambda, \mathcal{V}}^\epsilon$  was first considered in the fixed support setting in (Luise et al., 2018). These barycenters were studied at length in (Janati et al., 2020a) under a sub-gaussian assumption on the reference measures. Several free support synthesis algorithms have been proposed (Luise et al., 2019), (Shen et al., 2020). Recent work of (Chizat, 2023) introduced *doubly-regularized* entropic barycenters, which have good regularity and sample complexity guarantees (Vaškevičius and Chizat, 2023).

The analysis problem has received comparatively less attention. The earliest work in this setting is (Bonneel et al., 2016), where barycentric coefficients are applied to perform regression with histogram data for computer vision applications. In (Schmitz et al., 2018) barycentric coefficients are applied to dictionary learning for histograms, and these ideas are extended in (Mueller et al., 2023) where the focus is on sparse representations. In (Werenski et al., 2022) the analysis problem using unregularized Wasserstein-2 barycenters is considered with finite sample guarantees but under strong regularity assumptions on measures. In (Gunsilius et al., 2024) similar ideas utilizing a model based on the tangential structure of the Wasserstein-2 space are pursued with applications to causal inference, albeit without finite sample guarantees. In (Janati et al., 2020a), barycentric coefficients for the Sinkhorn barycenter functional are used to classify image data, modeled as probability measures on a fixed support. To our knowledge, our work is the first to consider analysis with regularized barycenter functionals in the free support setting under mild assumptions on the measures, with finite sample guarantees as well as provable stability guarantees

under measure perturbations.

## 2 ANALYTIC PROPERTIES OF $\mathcal{F}_{\lambda, \nu}^\epsilon$

We begin by establishing existence results for the functionals  $\mathcal{F}_{\lambda, \nu}^\epsilon$ :

**Proposition 2.1.** *Let  $\mathcal{V} = \{\nu_j\}_{j=1}^m \subset \mathcal{P}_2(\Omega)$ . Then for any  $\epsilon > 0$ ,  $F_{\lambda, \nu}^\epsilon$  admits a minimizer over  $\mathcal{P}_2(\Omega)$ , and for any  $\sigma > 0$ ,  $S_{\lambda, \nu}^\epsilon$  admits a minimizer over  $\mathcal{G}_\sigma(\Omega)$ . If  $\Omega$  is bounded, then the minimizer of  $S_{\lambda, \nu}^\epsilon$  is unique.*

Next, we characterize the derivatives and critical points for the functionals  $F_{\lambda, \nu}^\epsilon$  and  $S_{\lambda, \nu}^\epsilon$  when  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$ . The following is a generalization of Proposition 3 in (Janati et al., 2020a), which characterizes the derivatives of  $OT_2^\epsilon(\mu, \nu)$  when both  $\mu$  and  $\nu$  are subgaussian. We relax this assumption, and show that as long as  $\nu$  is subgaussian,  $\mu \mapsto OT_2^\epsilon(\mu, \nu)$  admits a derivative in the sense of Definition 1.2.

**Theorem 2.2.** *Let  $\nu \in \mathcal{G}(\mathbb{R}^d)$ . Then  $OT_2^\epsilon(\cdot, \nu) : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$  is differentiable, with derivative  $\delta_\mu OT_2^\epsilon(\mu, \nu) = f_{\mu \rightarrow \nu}^\epsilon$ , which is unique up to an additive constant.*

We sketch the proof, which is deferred to Section B.2. We essentially follow the standard argument used in [Proposition B.1, (Janati et al., 2020a)]. The most technical step is establishing pointwise convergence of  $f_{\mu+t\chi \rightarrow \nu}^\epsilon$  to  $f_{\mu \rightarrow \nu}^\epsilon$  as  $t \rightarrow 0^+$  for  $\chi := (\rho - \mu)$  for any  $\rho \in \mathcal{P}_2(\mathbb{R}^d)$  using *only* the subgaussian assumption on  $\nu$ , in contrast to (Janati et al., 2020a) where both measures are assumed to be subgaussian. We achieve this by using recent results from (Nutz and Wiesel, 2022) to establish convergence in probability of the potentials, which can then be upgraded to pointwise convergence using (3) and the subgaussian condition on  $\nu$  that in turn implies that  $|f_{\mu \rightarrow \nu}^\epsilon(x)|$  is at most of quadratic growth in  $x$  when  $\nu$  is subgaussian.

We remark that the proof of Theorem 2.2 also follows by following a strategy pioneered in (Mena and Niles-Weed, 2019) in the context of convergence of potentials associated to empirical measures. Indeed, it can be shown that for each fixed  $x$ ,  $\nabla f_{\mu+t\chi \rightarrow \nu}(x)$  is uniformly bounded over  $t$ . Hence, any family of potentials  $\{f_{\mu+t\chi \rightarrow \nu}\}_{t \in [0,1]}$  considered in the context of Theorem 2.2 is locally equicontinuous. By the Arzela-Ascoli Lemma (Royden and Fitzpatrick, 2010) this implies that there exists a subsequence that converges pointwise to  $f_{\mu \rightarrow \nu}$ . Then the pointwise convergence of the original sequence  $\{f_{\mu_t \rightarrow \nu}\}_{t \in [0,1]}$  can be deduced under the normalization conditions imposed as in our proof of Theorem 2.2. We leave the details to the reader. We note that proof of Theorem 2.2 as supplied does not require the following Proposition 2.3 to establish

the gradients of the potentials, so it is of independent interest and may generalize to other scenarios for other functionals.

We now note the following proposition regarding regularity properties of  $f_{\mu \rightarrow \nu}^\epsilon$ .

**Proposition 2.3.** *Let  $\nu \in \mathcal{G}(\mathbb{R}^d)$  and let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ . Then  $f_{\mu \rightarrow \nu}^\epsilon$  is infinitely differentiable, with gradient  $\nabla f_{\mu \rightarrow \nu}^\epsilon(x) = x - T_{\mu \rightarrow \nu}^\epsilon(x)$ , where*

$$T_{\mu \rightarrow \nu}^\epsilon(x) := \frac{\int y e^{\left(\frac{1}{\epsilon} \left(-\frac{1}{2} \|x-y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y)\right)\right)} d\nu(y)}{\int e^{\left(\frac{1}{\epsilon} \left(-\frac{1}{2} \|x-y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y)\right)\right)} d\nu(y)}. \quad (8)$$

Following (Pooladian and Niles-Weed, 2021), we refer to  $T_{\mu \rightarrow \nu}^\epsilon$  as the *entropic map* from  $\mu$  to  $\nu$ . Proposition 2.3 allows us to state the following result that characterizes the critical points of the barycenter functionals in terms of entropic maps.

**Corollary 2.4.** *Let  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$ .*

- *Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ . Then  $F_{\lambda, \nu}^\epsilon$  admits a derivative at  $\mu$ , given by  $\delta F_{\lambda, \nu}^\epsilon(\mu) = \sum_{j=1}^m \lambda_j f_{\mu \rightarrow \nu_j}^\epsilon$ . Hence  $\mu^\epsilon$  is a critical point for  $F_{\lambda, \nu}^\epsilon$  on  $\mathcal{P}_2(\mathbb{R}^d)$  if and only if  $x - \sum_{j=1}^m \lambda_j T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon(x) = 0$  for  $\mu^\epsilon$ -almost every  $x$ .*
- *Let  $\mu \in \mathcal{G}(\mathbb{R}^d)$ . Then  $S_{\lambda, \nu}^\epsilon$  admits a derivative at  $\mu$ , given by  $\delta S_{\lambda, \nu}^\epsilon(\mu) = \sum_{j=1}^m \lambda_j f_{\mu \rightarrow \nu_j}^\epsilon - f_{\mu \rightarrow \mu}^\epsilon$ . Hence  $\mu^\epsilon \in \mathcal{G}(\mathbb{R}^d)$  is a critical point for  $S_{\lambda, \nu}^\epsilon$  on  $\mathcal{G}(\mathbb{R}^d)$  if and only if  $T_{\mu^\epsilon \rightarrow \mu^\epsilon}^\epsilon(x) - \sum_{j=1}^m \lambda_j T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon(x) = 0$  for  $\mu^\epsilon$ -almost every  $x$ .*

We highlight that Theorem 2.2 allows us to establish differentiability of  $F_{\lambda, \nu}^\epsilon$  on the entirety of  $\mathcal{P}_2(\mathbb{R}^d)$ , but does not allow us to conclude the same for  $S_{\lambda, \nu}^\epsilon$ , which we leave as an open problem.

The critical point conditions in Corollary 2.4 are crucial for our method of solving the analysis problem, as explained in Section 4. We give the following sufficient condition for a critical point to be optimal.

**Corollary 2.5.** *Let  $\epsilon > 0$ , and let  $\nu_j \in \mathcal{G}(\mathbb{R}^d)$   $1 \leq j \leq n$ . Let  $\mu^\epsilon \in \mathcal{P}_2(\mathbb{R}^d)$  (resp.  $\mu^\epsilon \in \mathcal{G}(\mathbb{R}^d)$ ) be a critical point for  $F_{\lambda, \nu}^\epsilon$  (resp.  $S_{\lambda, \nu}^\epsilon$ ). If  $\text{supp}(\mu^\epsilon) = \mathbb{R}^d$ , then  $\mu^\epsilon$  minimizes  $F_{\lambda, \nu}^\epsilon$  (resp. minimizes  $S_{\lambda, \nu}^\epsilon$  over the set  $\mathcal{G}(\mathbb{R}^d)$ ).*

It is unknown what conditions on  $\mathcal{V}$  guarantee the existence of a critical point with support equal to  $\mathbb{R}^d$ . On the other hand, there always exists a singular critical point given by a Dirac mass.

**Lemma 2.6.** *Let  $\nu_j \in \mathcal{G}(\mathbb{R}^d)$  for all  $1 \leq j \leq m$ . Let  $x_\nu := \sum_{j=1}^m \lambda_j \mathbb{E}(\nu_j)$ . Then the Dirac mass  $\delta_{x_\nu}$  is a critical point for  $F_{\lambda, \nu}^\epsilon$ .*

Finally, we establish a control on the growth of critical points of  $F_{\lambda, \mathcal{V}}^\epsilon$ , which will be useful for establishing our sample complexity results in the sequel:

**Proposition 2.7.** *Suppose  $\mathcal{V} \subset \mathcal{G}_\sigma(\mathbb{R}^d)$ , and let  $\mu^\epsilon$  be a critical point for  $F_{\lambda, \mathcal{V}}^\epsilon$ . Then  $\mu^\epsilon$  is  $\sigma$ -subgaussian. In particular,  $\mu^* \in \arg \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \mathcal{V}}(\mu)$  is  $\sigma$ -subgaussian.*

### 3 SAMPLE COMPLEXITY OF ESTIMATING $\min_\mu \mathcal{F}_{\lambda, \mathcal{V}}^\epsilon(\mu)$

In this section, we show that the rate of estimating  $\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}_{\lambda, \mathcal{V}}^\epsilon(\mu)$  as a function of the number of samples from the reference measures does not depend on the dimension.

**Theorem 3.1.** *Let  $\mathcal{V} = \{\nu_1, \dots, \nu_m\} \subset \mathcal{G}_\sigma(\mathbb{R}^d)$  and let  $\hat{\mathcal{V}}^n = \{\hat{\nu}_1^n, \dots, \hat{\nu}_m^n\}$ . Let  $\mu^n$  denote a (random) minimizer of  $\min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} F_{\lambda, \hat{\mathcal{V}}^n}^\epsilon(\mu)$ . Then:*

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \mathcal{V}}^\epsilon(\mu) - F_{\lambda, \mathcal{V}}^\epsilon(\mu^n) \right| \right) \leq m C_{d, \epsilon, \sigma}^* / \sqrt{n}. \quad (9)$$

Suppose that, for any  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$ ,  $U_n := \arg \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu)$  is nonempty and furthermore  $U_n \subset \mathcal{G}_{\tilde{\sigma}}(\mathbb{R}^d)$ , where  $\tilde{\sigma} = \max_{1 \leq j \leq m} \{\|\nu_j\|_{\mathcal{G}}\}$ . Let  $\mu^n \in \arg \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda, \hat{\mathcal{V}}^n}^\epsilon(\mu)$ . Then:

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) - S_{\lambda, \mathcal{V}}^\epsilon(\mu^n) \right| \right) \leq m C_{d, \epsilon, \sigma}^* / \sqrt{n}. \quad (10)$$

Here,  $C_{d, \epsilon, \sigma}^*$  is a constant depending only on  $d, \epsilon$  and  $\sigma$ .

Note that in contrast to  $F_{\lambda, \mathcal{V}}^\epsilon$ , for  $S_{\lambda, \mathcal{V}}^\epsilon$  in (10) we restrict the minimization to the set  $\mathcal{G}_\sigma(\mathbb{R}^d)$ . This choice is justified by Proposition 2.1 where we show existence of minimizers for  $S_{\lambda, \mathcal{V}}^\epsilon$  when restricted to  $\mathcal{G}_\sigma(\mathbb{R}^d)$ . Our statistical results for synthesis in 3.1 are *algorithm independent* and apply when the support of the measures can be unbounded. This improves upon [Theorem 7, (Luise et al., 2019)], which is restricted to the bounded support setting. However, their result comes with an algorithm that provably achieves the bound. It is natural to therefore ask whether there exist algorithms that can achieve the bounds in Theorem 3.1. We leave this to future work.

A natural question is whether a stronger sample complexity result is true. Let  $\mu^*$  denote a minimizer of  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$  on  $\mathcal{P}_2(\Omega)$  with  $\Omega$  bounded. Is it true that some  $\mu_n^* \in \arg \min_{\mu \in \mathcal{P}^n(\Omega)} \mathcal{F}_{\lambda, \hat{\mathcal{V}}^n}^\epsilon$  approximates  $\mu^*$  in  $OT_2$  on  $\mathcal{P}_2(\Omega)$  with parametric sample complexity, i.e.,  $\mathbb{E}(OT_2(\mu^*, \mu_n^*)) \lesssim \frac{1}{\sqrt{n}}$ ? We claim this cannot be true for the functional  $S_{\lambda, \mathcal{V}}^\epsilon$ . Indeed, let  $\mathcal{V} = \{\nu\}$  for any  $\nu \in \mathcal{P}_2(\Omega)$  for some bounded  $\Omega$ . By Proposition 2.1,

$\arg \min_{\mu \in \mathcal{P}_2(\Omega)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) = \nu$ , as  $S_{\lambda, \mathcal{V}}^\epsilon$  is debiased and strictly convex on  $\mathcal{P}_2(\Omega)$ . Similarly, if  $\mathcal{V} = \{\nu^n\}$  for any  $\nu^n \in \mathcal{P}^n(\Omega)$ , then  $\arg \min_{\mu \in \mathcal{P}^n(\Omega)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) = \nu^n$ , again by debiasing and strict convexity. Thus,

$$\begin{aligned} & \mathbb{E}(OT_2(\arg \min_{\mu \in \mathcal{P}_2(\Omega)} S_{\lambda, \{\nu\}}^\epsilon(\mu), \arg \min_{\mu \in \mathcal{P}^n(\Omega)} S_{\lambda, \{\nu^n\}}^\epsilon(\mu))) \\ &= \mathbb{E}(OT_2(\nu, \nu^n)) \gtrsim \frac{1}{n^{1/d}}, \end{aligned}$$

where the inequality is classical (Dudley, 1969).

### 4 SAMPLE COMPLEXITY AND STABILITY FOR THE ANALYSIS PROBLEM

As explained in Section 1.1, if  $\mu^* \in \arg \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}_{\lambda^*, \mathcal{V}}^\epsilon(\mu)$ , solving the analysis problem (6) is equivalent to computing  $\lambda_* \in \Delta^m$ . Using the criteria from Corollary 2.4, we adapt the proof of Proposition 1 in (Werenski et al., 2022) to show that solving the analysis program is equivalent to solving a convex, quadratic program.

**Proposition 4.1.** *Let  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$ .*

- *Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ , and define  $A_\mu^\epsilon \in \mathbb{R}^{m \times m}$  by  $[A_\mu^\epsilon]_{ij} := \int_{\mathbb{R}^d} \langle T_{\mu \rightarrow \nu_i}^\epsilon(x) - x, T_{\mu \rightarrow \nu_j}^\epsilon(x) - x \rangle d\mu(x)$ . Then  $\|\nabla \delta F_{\lambda, \mathcal{V}}^\epsilon\|_{L^2(\mu)}^2 = \lambda^\top A_\mu^\epsilon \lambda$ , and hence  $\mu$  is a critical point iff  $\min_{\lambda \in \Delta^m} \lambda^\top A_\mu^\epsilon \lambda = 0$ .*
- *Let  $\tilde{\mu} \in \mathcal{G}(\mathbb{R}^d)$ , and define  $S_{\tilde{\mu}}^\epsilon \in \mathbb{R}^{m \times m}$  by  $[S_{\tilde{\mu}}^\epsilon]_{ij} := \int_{\mathbb{R}^d} \langle T_{\tilde{\mu} \rightarrow \nu_i}^\epsilon(x) - T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon(x), T_{\tilde{\mu} \rightarrow \nu_j}^\epsilon(x) - T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon(x) \rangle d\tilde{\mu}(x)$ . Then  $\|\nabla \delta S_{\lambda, \mathcal{V}}^\epsilon\|_{L^2(\tilde{\mu})}^2 = \lambda^\top S_{\tilde{\mu}}^\epsilon \lambda$ , and hence  $\tilde{\mu}$  is a critical point of  $S_{\lambda, \mathcal{V}}^\epsilon$  iff  $\min_{\lambda \in \Delta^m} \lambda^\top S_{\tilde{\mu}}^\epsilon \lambda = 0$ .*

Furthermore, if  $\text{supp}(\mu) = \mathbb{R}^d$  (resp.  $\text{supp}(\tilde{\mu}) = \mathbb{R}^d$ ), then  $\mu$  is an entropic barycenter (resp.  $\tilde{\mu}$  is a Sinkhorn barycenter).

Our main result in this section is that the sample complexity of solving the analysis problem for a critical point  $\mu$  is equivalent to the sample complexity of estimating  $\nabla \delta \mathcal{F}_{\lambda, \mathcal{V}}^\epsilon$ .

**Theorem 4.2.** *Let  $\Omega \subseteq \mathbb{R}^d$  and let  $\mathcal{V} \subset \mathcal{G}_\sigma(\Omega)$ . Let  $\mu \in \mathcal{P}_2(\Omega)$  and suppose that  $A_\mu^\epsilon$  has an eigenvalue of 0 with unique eigenvector  $\lambda_* \in \Delta^m$ . Suppose there exists an estimator  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)$  and a nonincreasing function  $\theta(n) : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$  such that*

$$\mathbb{E} \left( \|T_{\mu \rightarrow \nu_j}^\epsilon - \hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)\|_{L^2(\mu)}^2 \right) \leq \theta(n), \quad (11)$$

for all  $1 \leq j \leq m$ . Let  $\hat{\lambda}^n$  be the output of Algorithm 1, applied to  $F_{\lambda, \mathcal{V}}^\epsilon$  at  $\mu$ . Then:

$$\mathbb{E} \left( \|\hat{\lambda}^n - \lambda_*\|_2^2 \right) \leq \frac{Jm^3 d \sigma^2}{\alpha_2} \max \left\{ \frac{1}{\sqrt{n}}, \sqrt{\theta(n) + \theta(n)^2} \right\}, \quad (12)$$

**Algorithm 1: Coefficient Recovery**

**Data:** Input measure  $\mu \in \mathcal{P}(\Omega)$ ;  
 Reference measures  $\nu_1, \dots, \nu_m \in \mathcal{P}(\Omega)$ ;  
 Regularization parameter  $\epsilon > 0$ ;  
 Number of samples  $n$ ;  
 Functional  $\mathcal{F}_{\lambda, \nu}^\epsilon \in \{F_{\lambda, \nu}^\epsilon, S_{\lambda, \nu}^\epsilon\}$ ;  
 Estimators  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)$  for  $T_{\mu \rightarrow \nu_j}^\epsilon, \forall 1 \leq j \leq m$ ;  
 Estimator  $\hat{T}(\hat{\mu}^n)$  for  $T_{\mu \rightarrow \mu}^\epsilon$ .

```

1 Sample  $X_1, X_2, \dots, X_{2n} \sim \mu$ ;
2 for  $1 \leq j \leq m$  do
3   Sample  $Y_1^j, Y_2^j, \dots, Y_n^j \sim \nu_j$ ;
4    $\hat{\nu}_j^n \leftarrow$  uniform measure on  $Y_1^j, Y_2^j, \dots, Y_n^j$ ;
5 end
6  $\hat{\mu}^n \leftarrow$  uniform measure on  $X_1, X_2, \dots, X_n$ ;
7 for  $1 \leq i, j \leq m$  do
8   if  $\mathcal{F}_{\lambda, \nu}^\epsilon == F_{\lambda, \nu}^\epsilon$  then
9      $[\hat{M}_\mu]_{ij} \leftarrow \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}(\hat{\mu}^n, \hat{\nu}_i^n)(X_k) - X_k, \hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)(X_k) - X_k \rangle$ ;
10  end
11  else
12     $[\hat{M}_\mu]_{ij} \leftarrow \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}(\hat{\mu}^n, \hat{\nu}_i^n)(X_k) - \hat{T}(\hat{\mu}^n)(X_k), \hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)(X_k) - \hat{T}(\hat{\mu}^n)(X_k) \rangle$ ;
13  end
14 end
15 Output:  $\hat{\lambda}^n = \arg \min_{\lambda \in \Delta^m} \lambda^\top \hat{M}_\mu^\epsilon \lambda$ ;
    
```

where  $J$  is an absolute constant, and  $\alpha_2$  is the smallest nonzero eigenvalue of  $A_\mu^\epsilon$ .

Let  $\mu \in \mathcal{G}_\sigma(\Omega)$ , and suppose that  $S_\mu^\epsilon$  has an eigenvalue of 0 with multiplicity 1 with eigenvector  $\lambda_* \in \Delta^m$ . Suppose that (11) holds for all  $1 \leq j \leq m$ , and furthermore that there exists an estimator  $\hat{T}(\hat{\mu}^n)$  such that:

$$\mathbb{E} \left( \|T_{\mu \rightarrow \mu}^\epsilon - \hat{T}(\hat{\mu}^n)\|_{L^2(\mu)}^2 \right) \leq \theta(n). \quad (13)$$

Let  $\hat{\lambda}^n$  be the output of Algorithm 1, applied to  $S_{\lambda, \nu}^\epsilon$  at  $\mu$ . Then:

$$\mathbb{E} \left( \|\hat{\lambda}^n - \lambda_*\|_2^2 \right) \leq \frac{\tilde{J} m^3 d \sigma^2}{\alpha_2} \max \left\{ \frac{1}{\sqrt{n}}, \sqrt{\theta(n) + \theta(n)^2} \right\}, \quad (14)$$

where  $\tilde{J}$  is an absolute constant, and  $\alpha_2$  is the smallest nonzero eigenvalue of  $S_\mu^\epsilon$ .

We present two applications of Theorem 4.2 based on (Rigollet and Stromme, 2022) and (Werenski et al., 2023). In both cases we may conclude that the rate of estimation of regularized barycentric coefficients is not cursed by dimensionality.

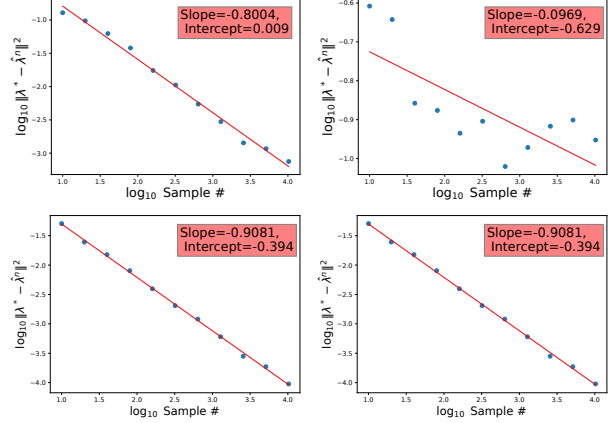


Figure 1: (Top left) Average log  $\ell^2$ -loss for two random 1D Gaussian measures with random weights,  $\epsilon = 2$ . (Top right) Average log  $\ell^2$ -loss for three random 1D Gaussian measures with random weights,  $\epsilon = 2$ . (Bottom left) Average log  $\ell^2$ -loss for three random 5D Gaussian measures,  $\lambda = (0.2201, 0.0269, 0.7530)$ ,  $\epsilon = 1$ . (Bottom right) Average log  $\ell^2$ -loss for three uniform-measures on random 5D cubes,  $\lambda = (0.5112, 0.4477, 0.0411)$ ,  $\epsilon = 0.1$ .

**Corollary 4.3.** Let  $\Omega \subset \mathbb{R}^d$  be bounded, and let  $\mathcal{V} \subset \mathcal{P}_2(\Omega)$ . Let  $X_1, X_2, \dots, X_{2n} \sim \mu$ , and let  $\hat{\mu}^n$  denote the uniform distribution on  $X_1, \dots, X_n$ ,  $\hat{\mu}_1$  denote the uniform distribution on  $X_1, \dots, X_{\lfloor n/2 \rfloor}$  and  $\hat{\mu}_2$  denote the uniform distribution on  $X_{\lfloor n/2 \rfloor + 1}, \dots, X_{2\lfloor n/2 \rfloor}$ . Then Theorem 4.2 holds, with  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n) = T_{\hat{\mu}^n \rightarrow \hat{\nu}_j^n}^\epsilon$ ,  $\hat{T}(\hat{\mu}^n) = T_{\hat{\mu}_1 \rightarrow \hat{\mu}_2}^\epsilon$  and  $\theta(n) = \frac{\tilde{C}_{\Omega, \epsilon}}{n}$ , where  $\tilde{C}_{\Omega, \epsilon}$  is a constant depending only on  $|\Omega|$  and  $\epsilon$ .

**Corollary 4.4.** Let  $\nu_j$  be  $c$ -strongly log-concave with  $\mathbb{E}(\nu_j) = 0$  for all  $1 \leq j \leq m$ . Then there exists an estimator  $\hat{T}$  such that Theorem 4.2 holds for  $F_{\lambda, \nu}^\epsilon$ , with  $\theta(n) = \frac{\tilde{K}_{d, \epsilon, c}}{n^{1/3}}$ , where  $\tilde{K}_{d, \epsilon, c}$  is a constant depending on  $d, \epsilon$  and  $c$ .

Finally, we prove a stability property, which implies that the encoding obtained via solving the analysis problem is stable with respect to perturbations of the measure being analyzed, suggesting an advantage of using regularized barycenter functionals for representation of measures in  $\mathcal{P}_2(\Omega)$ .

**Proposition 4.5.** Let  $\Omega \subset \mathbb{R}^d$  be bounded, and let  $\mu, \rho \in \mathcal{P}(\Omega)$  and  $\mathcal{V} \subset \mathcal{P}(\Omega)$ . Suppose that  $A_\mu^\epsilon$  (resp.  $S_\mu^\epsilon$ ) has an eigenvalue 0 with unique eigenvector  $\lambda_\mu \in \Delta^m$ . Let  $\lambda_\rho \in \arg \min_{\lambda \in \Delta^m} \lambda^\top A_\rho^\epsilon \lambda$  (resp.  $\lambda_\rho \in \arg \min_{\lambda \in \Delta^m} \lambda^\top S_\rho^\epsilon \lambda$ ). Then there exists a constant  $H_{\Omega, \epsilon}$ , depending only on  $\Omega$  and  $\epsilon$ , such that  $\|\lambda_\mu - \lambda_\rho\|_2^2 \leq \frac{m^3 H_{\Omega, \epsilon}}{\alpha_2} OT_2(\mu, \rho)$ , where  $\alpha_2 > 0$  is the smallest nonzero eigenvalue of  $A_\mu^\epsilon$  (resp.  $S_\mu^\epsilon$ ).

#### 4.1 Numerical Verification of Theorem 4.2

We consider several numerical experiments validating Theorem 4.2. In these experiments we focus only on  $F_{\lambda, \mathcal{V}}^\epsilon$  and remark that similar results are achievable for  $S_{\lambda, \mathcal{V}}^\epsilon$ . Our choices of  $\mathcal{V}$  are 1-dimensional Gaussians, 5-dimensional Gaussians, and uniform measures on 5-dimensional cubes. In each experiment, we generate a random set of  $m$  reference measures and weights  $\lambda \in \Delta^m$ . We sample from the  $(\lambda, \mathcal{V})$ -barycenter and the reference measures  $\mathcal{V}$ , and using these samples we compute  $\hat{\lambda}^n$  by applying Algorithm 1. Our choice of estimator for the map  $T_{\mu \rightarrow \nu_j}^\epsilon$  is  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n) = T_{\hat{\mu}^n \rightarrow \hat{\nu}_j^n}^\epsilon$ , for each  $1 \leq j \leq m$ . For 1-dimensional Gaussian reference measures, we may sample from  $\arg \min_{\mu \in \mathcal{P}_2(\mathbb{R})} F_{\lambda, \mathcal{V}}^\epsilon(\mu)$  directly by applying the closed-form expression for entropic barycenters with Gaussian reference measures, given in Theorem 2 in (Janati et al., 2020a). As we do not have access to closed-form expressions for the entropic barycenters in the higher-dimensional examples, we instead apply a free-support synthesis algorithm (see details in Appendix E.1) to compute approximate critical points  $\mu^*$ , which we then sample from to estimate  $\hat{\lambda}^n$ . We then compute  $\|\hat{\lambda}^n - \lambda\|_2^2$ . We repeat this over 100 trials and report the logarithm of the mean-squared error against  $\log(n)$  as  $n$  ranges over  $\{10, 20, 40, \dots, 10240\}$ . The results are plotted in Figure 1. See Appendix E.2 for implementation details.

**1-Dimensional Gaussians:** We see for  $m = 2$ , the errors decays at a rate close to  $n^{-1}$ , which is faster than predicted by Theorem 4.2. In the case of  $m = 3$ , we find that the  $\ell^2$ -distance between the coefficients essentially does not decay in  $n$  (top right). This is due to the fact that in one dimension, it is more likely that the eigenspace of  $A_\mu^\epsilon$  associated to the eigenvalue zero has dimension greater than 1, which violates the conditions in Theorem 4.2, leading to non-uniqueness of coefficients corresponding to the measure being analyzed. Nevertheless, we empirically show (see Appendix F) that the expected  $OT_2^2$ -error between the barycenters associated to  $\lambda$  and  $\hat{\lambda}^n$  rapidly decays as  $n$  increases, implying successful recovery of a *valid* set of barycentric coefficients via Algorithm 1.

**5-Dimensional Gaussians and Uniform Measures:** For the 5-dimensional Gaussian reference measures, we see that the decay in  $\ell^2$  error between coefficients  $\lambda$  and  $\hat{\lambda}^n$  is at a rate of approximately  $n^{-1}$ . We conjecture that the improved rate could be due to the fact that the entropic maps between Gaussians are affine (see Theorem 1 in (Janati et al., 2020b)). We further tested our method with reference measures given by three random translates of  $\text{Unif}([0, 2]^5)$ . We

#### Algorithm 2: Point Cloud Classification Algorithm

---

**Data:**  $m$  classes;  
 $b$  labeled point clouds for each class:  
 $\mathcal{A}^y = \{P_1^y, P_2^y, \dots, P_b^y\}$  for  $1 \leq y \leq m$ ;  
 Unlabelled point cloud:  $Q$ ;  
 Regularization parameter:  $\epsilon$ ;  
 Functional:  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon \in \{F_{\lambda, \mathcal{V}}^\epsilon, S_{\lambda, \mathcal{V}}^\epsilon\}$ .

---

```

1  $\mu \leftarrow$  Empirical measure supported on  $Q$ ;
2 for  $1 \leq y \leq m$  do
3   for  $1 \leq j \leq b$  do
4     Set  $\nu_j^y \leftarrow$  Empirical measure supported
       on  $P_j^y$ ;
5   end
6 end
7  $\mathcal{V} \leftarrow \{\nu_1^1, \dots, \nu_b^1, \nu_1^2, \dots, \nu_b^2, \dots, \nu_1^m, \dots, \nu_b^m\}$ ;
8 for  $1 \leq i_1, i_2 \leq m$  and  $1 \leq j_1, j_2 \leq b$  do
9   if  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon == F_{\lambda, \mathcal{V}}^\epsilon$  then
10     $[M_\mu]_{b(i_1-1)+j_1, b(i_2-1)+j_2} \leftarrow$ 
        $\frac{1}{n} \sum_{x \in \text{supp}(\mu)} \langle T_{\mu \rightarrow \nu_{j_1}^{i_1}}^\epsilon(x) -$ 
        $x, T_{\mu \rightarrow \nu_{j_2}^{i_2}}^\epsilon(x) - x \rangle$ ;
11  end
12  else
13     $[M_\mu]_{b(i_1-1)+j_1, b(i_2-1)+j_2} \leftarrow$ 
        $\frac{1}{n} \sum_{x \in \text{supp}(\mu)} \langle T_{\mu \rightarrow \nu_{j_1}^{i_1}}^\epsilon(x) -$ 
        $T_{\mu \rightarrow \mu}^\epsilon(x), T_{\mu \rightarrow \nu_{j_2}^{i_2}}^\epsilon(x) - T_{\mu \rightarrow \mu}^\epsilon(x) \rangle$ ;
14  end
15 end
16  $\lambda \leftarrow \arg \min_{\lambda \in \Delta^{mb}} \lambda^\top M_\mu \lambda$ ;
17 Define  $B \in \mathbb{R}^{m \times mb}$  such that:

```

$$[B]_{ij} := \begin{cases} 1 & \text{if } j \in \{b(i-1)+1, \dots, bi\}, \\ 0 & \text{otherwise.} \end{cases}$$


---

```

18  $\tilde{\lambda} \leftarrow B\lambda$ 
19 Output: Assign point cloud  $Q$  the class
     $\tilde{y} \leftarrow \arg \max_{1 \leq y \leq m} [\tilde{\lambda}]_y$ 

```

---

see that the error decays at a rate of approximately  $n^{-9/10}$ , again exceeding the rate predicted by our theory, though marginally slower than the Gaussian case.

## 5 POINT CLOUD CLASSIFICATION

We demonstrate the utility of the analysis of measures via regularized barycenters for 3D point cloud-classification. We obtained our data from the PointCloud-C dataset (Ren et al., 2022), a reposi-



	dropout_local_1	dropout_local_2	dropout_global	jitter	add_local	clean
$F_{\lambda,\nu}^\epsilon, \epsilon = 0.75$	$0.397 \pm 0.055$	$0.387 \pm 0.051$	$0.399 \pm 0.057$	$0.392 \pm 0.055$	$0.389 \pm 0.051$	$0.397 \pm 0.057$
$S_{\lambda,\nu}^\epsilon, \epsilon = 0.75$	$0.600 \pm 0.020$	$0.562 \pm 0.0233$	$0.626 \pm 0.017$	$0.758 \pm 0.020$	$0.578 \pm 0.017$	$0.740 \pm 0.010$
$F_{\lambda,\nu}^\epsilon, \epsilon = 0.089$	$0.544 \pm 0.037$	$0.511 \pm 0.036$	$0.559 \pm 0.037$	$0.532 \pm 0.038$	$0.510 \pm 0.037$	$0.558 \pm 0.039$
$S_{\lambda,\nu}^\epsilon, \epsilon = 0.089$	$0.862 \pm 0.019$	$0.81 \pm 0.018$	$0.886 \pm 0.010$	$0.904 \pm 0.010$	$0.846 \pm 0.010$	$0.886 \pm 0.022$
$F_{\lambda,\nu}^\epsilon, \epsilon = 0.009$	$0.908 \pm 0.020$	$0.902 \pm 0.022$	$0.913 \pm 0.018$	$0.914 \pm 0.018$	$0.888 \pm 0.019$	$0.901 \pm 0.017$
$S_{\lambda,\nu}^\epsilon, \epsilon = 0.009$	$0.905 \pm 0.022$	<b><math>0.902 \pm 0.021</math></b>	$0.919 \pm 0.020$	$0.920 \pm 0.018$	<b><math>0.900 \pm 0.018</math></b>	<b><math>0.908 \pm 0.024</math></b>
Unregularized ( $\epsilon = 0$ )	<b><math>0.912 \pm 0.019</math></b>	$0.901 \pm 0.018$	<b><math>0.924 \pm 0.020</math></b>	<b><math>0.921 \pm 0.020</math></b>	$0.897 \pm 0.017$	<b><math>0.908 \pm 0.018</math></b>
Doubly Regularized $\epsilon = 0.009, \tau = 0.01$	$0.874 \pm 0.015$	$0.856 \pm 0.019$	$0.891 \pm 0.019$	$0.896 \pm 0.019$	$0.865 \pm 0.019$	$0.884 \pm 0.022$
PointNet	$0.68 \pm 0.035$	$0.672 \pm 0.036$	$0.661 \pm 0.037$	$0.684 \pm 0.031$	$0.6567 \pm 0.029$	$0.678 \pm 0.029$

Table 1: Comparison of PointNet (trained on 400 point clouds) with classification based on barycentric coefficients (with 15 reference measures) for point cloud classification. Confidence intervals calculated as  $t_{\alpha,4} * \frac{\hat{\sigma}}{\sqrt{n}}$  with  $\hat{\sigma}$  the sample standard deviation and  $t_{\alpha,4}$  is the Student’s t-value with  $\alpha = 0.05$  and 4 degrees of freedom.

tory of point clouds separated into classes and corrupted with various regimes of noise and occlusions. We select 100 clean (uncorrupted) point clouds from five classes (airplanes, beds, guitars, monitors, vases) and perform the following experiment. We run Algorithm 4 (detailed in Subsection E.3 in the Appendix) on these point clouds with  $m = 5$ ,  $q = 100$ ,  $b = 3$ ,  $n_{train} = 80$ ,  $n_{test} = 20$ ,  $K = 5$ , and vary over  $\mathcal{F}_{\lambda,\nu}^\epsilon \in \{F_{\lambda,\nu}^\epsilon, S_{\lambda,\nu}^\epsilon\}$  and  $\epsilon \in \{0.009, 0.089, 0.75\}$ . We also augment our *test* set with 500 corrupted point clouds, which are copies of the 100 clean test point clouds with 5 types of corruptions applied. We obtain the corrupted data from the dropout\_local\_1, dropout\_local\_2, dropout\_global\_4, jitter\_4 and add\_local\_4 datasets; see (Ren et al., 2022) for details. We do this by applying the same train-test splits to the corrupted point clouds, discarding the corrupted training point clouds, and then using the previously selected clean training point clouds as  $\mathcal{A}_{train}^y$  in Algorithm 4.

In Table 1, we display results for classification using  $\mathcal{F}_{\lambda,\nu}^\epsilon$  with different choices of  $\epsilon$ . We compare our results to PointNet (Qi et al., 2017), a convolutional neural network (CNN) used for point cloud classification. Our implementation is adapted from (Karaev and Nikulina, 2023). We also compare against the same classification method with the unregularized barycenter functional (corresponding to  $\mathcal{F}_{\lambda,\nu}^\epsilon$ , with  $\epsilon = 0$ ) and the doubly-regularized barycenter functional (Chizat, 2023), with inner regularization  $\epsilon = 0.009$  and outer-regularization  $\tau = 0.01$ . See Section E.3 for details. We see that classification using barycentric coefficients is significantly more accurate than PointNet. Furthermore, our method required  $< 4\%$  of the data used to train PointNet to achieve this accuracy, suggesting its applicability to problems where training data is rare or expensive to acquire. We also note that the accuracy remained high even in the presence of corrupted data, indicating the robustness of our method. However, we note that PointNet has on the order of  $10^6$  parameters, and hence 400 training examples may be insufficient to truly assess its accuracy on our test set. In general

we see that classification using the debiased functional  $S_{\lambda,\nu}^\epsilon$  outperforms classification using  $F_{\lambda,\nu}^\epsilon$ , with a more noticeable difference in performance at high  $\epsilon$ . In general, the Sinkhorn functional with  $\epsilon = 0.009$  and the unregularized functional outperform all other methods. Both  $F_{\lambda,\nu}^\epsilon$  and  $S_{\lambda,\nu}^\epsilon$  outperformed the doubly regularized functional in the low  $\epsilon$  regime, perhaps due to the additional complications introduced by score estimation in this setting. Indeed, the point clouds under consideration can be viewed as samples from a distribution supported on a two-dimensional surface in  $\mathbb{R}^3$ , which would not admit a Lebesgue density.

In Appendix I, we further highlight the applicability of the analysis coefficients with a point cloud completion method, which we test on the PointCloud-C dataset.

## 6 CONCLUSIONS AND OPEN PROBLEMS

We presented new functional analytic, statistical, and stability results for the synthesis and analysis of probability measures with entropy-regularized optimal transport barycenter functionals with applications to sample-efficient classification of point cloud data. Several open problems for future work are to: (a) characterize the *projection* error in analyzing a measure (6) when  $\mu$  is not a barycenter; (b) derive a stronger sample complexity result for synthesis of the barycenter for  $F_{\lambda,\nu}^\epsilon$ , and (c) improve the statistical estimation rates in Theorem 4.2 to match the observed rates in the experiments.

**Acknowledgments:** BM, JM, and SA acknowledge partial support from NSF DMS-2309519. BM and JM acknowledge partial support from NSF DMS-2318894. BM was also supported in part by CCF-1553075. JM acknowledges partial support from a Tufts Springboard grant. SA acknowledges partial support from the NSF under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, <http://iaifi.org/>).

## References

- Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.
- Luigi Ambrosio and Giuseppe Savaré. Gradient flows of probability measures. *Handbook of differential equations: evolutionary equations*, 3:1–136, 2006.
- Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.
- Jordan Bell. Semicontinuous functions and convexity, 2014. URL <http://www.individual.utoronto.ca/jordanbell/notes/semincontinuous.pdf>.
- Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2): 5120–5150, 2019.
- Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2nd edition, 1999.
- Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71–1, 2016.
- Elsa Cazelles, Arnaud Robert, and Felipe Tobar. The Wasserstein-Fourier distance for stationary time series. *IEEE Transactions on Signal Processing*, 69: 709–721, 2020.
- Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.
- Kevin Cheng, Shuchin Aeron, Michael C. Hughes, and Eric L. Miller. Dynamical Wasserstein barycenters for time-series modeling. *Advances in Neural Information Processing Systems*, 34:27991–28003, 2021.
- Kevin C. Cheng, Eric L. Miller, Michael C. Hughes, and Shuchin Aeron. Non-parametric and regularized dynamical Wasserstein barycenters for sequential observations. *IEEE Transactions on Signal Processing*, 71:3164–3178, 2023.
- Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.
- Lénaïc Chizat. Doubly regularized entropic Wasserstein barycenters. *arXiv preprint arXiv:2303.11844*, 2023.
- Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.
- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331, 2020.
- Samuel Cohen, Michael Arbel, and Marc Peter Deisenroth. Estimating barycenters of measures in high dimensions. *arXiv preprint arXiv:2007.07105*, 2020.
- Donald L Cohn. *Measure theory*, volume 2. Springer, 2013.
- Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.
- Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.
- Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.
- Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.
- Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Tight stability bounds for entropic Brenier maps. *arXiv preprint arXiv:2404.02855*, 2024.
- Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, Cicero Dos Santos, and Tom Sercu. Wasserstein barycenter model ensembling. *arXiv preprint arXiv:1902.04999*, 2019.
- Richard Mansfield Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.
- Stephan Eckstein and Marcel Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn’s algorithm. *SIAM Journal on Mathematical Analysis*, 54(6):5922–5948, 2022.
- Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trounev, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.
- Remi Flamary and Nicolas Courty. Python optimal transport. <https://pythonot.github.io/>, 2023.
- Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

- Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.
- Florian Gunsilius, Meng Hsuan Hsieh, and Myung Jin Lee. Tangential Wasserstein projections. *Journal of Machine Learning Research*, 25(69):1–41, 2024.
- Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.
- Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In *International Conference on Machine Learning*, pages 4692–4701, 2020a.
- Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced Gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 33:10468–10479, 2020b.
- Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.
- Nikita Karaev and Irina Nikulina. nikitakaraevv/pointnet. <https://github.com/nikitakaraevv/pointnet>, 2023.
- Parnian Kassraie, Aram-Alexandre Pooladian, Michal Klein, James Thornton, Jonathan Niles-Weed, and Marco Cuturi. Progressive entropic optimal transport solvers. *arXiv preprint arXiv:2406.05061*, 2024.
- Yann LeCun. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13, 2000.
- Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with Wasserstein distance. *Advances in Neural Information Processing Systems*, 31, 2018.
- Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. *Advances in Neural Information Processing Systems*, 32, 2019.
- Shoaib Bin Masud, Matthew Werenski, James M Murphy, and Shuchin Aeron. Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling. *Journal of Machine Learning Research*, 24(160):1–65, 2023.
- Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Marshall Mueller, Shuchin Aeron, James M Murphy, and Abiy Tasissa. Geometrically regularized Wasserstein dictionary learning. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 384–403, 2023.
- Saad Nadeem, Travis Hollmann, and Allen Tannenbaum. Multimarginal Wasserstein barycenter for stain normalization and augmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 362–371. Springer, 2020.
- Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.
- Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1):401–424, 2022.
- Marcel Nutz and Johannes Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn’s algorithm. *The Annals of Probability*, 51(2):699–722, 2023.
- Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
- Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.
- Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.
- Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference*, pages 435–446. Springer, 2012.
- Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International Conference on Machine Learning*, pages 18559–18575, 2022.
- Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

- Philippe Rigollet and Austin J. Stromme. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.
- Halsey Royden and Patrick Michael Fitzpatrick. *Real analysis*. China Machine Press, 2010.
- Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of variations, pdes, and modeling*. Birkhäuser, 2015.
- Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.
- Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, and Hamed Hassani. Sinkhorn barycenter via functional gradient descent. *Advances in Neural Information Processing Systems*, 33:986–996, 2020.
- Dror Simon and Aviad Aberdam. Barycenters of natural images constrained Wasserstein barycenters for image morphing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7910–7919, 2020.
- Richard Sinkhorn and Paul Knopp. Concerning non-negative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.
- Austin Stromme. Sampling from a Schrödinger bridge. In *International Conference on Artificial Intelligence and Statistics*, pages 4058–4067, 2023.
- Tomas Vaškevičius and Lénaïc Chizat. Computational guarantees for doubly entropic Wasserstein barycenters via damped Sinkhorn iterations. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 12363–12388, 2023.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.
- Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.
- Matthew Werenski, James M Murphy, and Shuchin Aeron. Estimation of entropy-regularized optimal transport maps between non-compactly supported measures. *arXiv preprint arXiv:2311.11934*, 2023.
- Matthew Werenski, Brendan Mallery, Shuchin Aeron, and James M Murphy. Linearized wasserstein barycenters: Synthesis, analysis, representational capacity, and applications. *arXiv preprint arXiv:2410.23602*, 2024a.
- Matthew Werenski, Shoaib Bin Masud, James M Murphy, and Shuchin Aeron. On rank energy statistics via optimal transport: Continuity, convergence, and change point detection. *IEEE Transactions on Information Theory*, 70:5168 – 5190, 2024b.
- Matthew E Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, and James M Murphy. Measure estimation in the barycentric coding model. In *International Conference on Machine Learning*, pages 23781–23803, 2022.
- Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled Wasserstein learning for word embedding and topic modeling. *Advances in Neural Information Processing Systems*, 31, 2018.
- Yibo Yang, Stephan Eckstein, Marcel Nutz, and Stephan Mandt. Estimating the rate-distortion function by Wasserstein gradient descent. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *International Conference on Machine Learning*, pages 11513–11522, 2020.

## Checklist

1. For all models and algorithms presented, check if you include:
  - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
  - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
  - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]
2. For any theoretical claim, check if you include:
  - (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
  - (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
  - (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]
3. For all figures and tables that present empirical results, check if you include:
  - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
  - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
  - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
  - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/**No**/Not Applicable]  
 Explanation: Some experiments were conducted on a personal laptop, some were conducted on a small server, and rest were conducted on University HPC server. The run-times and memory therefore varied from set-up to set-up and we haven't conducted everything on a single machine to answer this Q concretely.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
  - (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
  - (b) The license information of the assets, if applicable. [**Yes**/No/**Not Applicable**]
  - (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**/No/Not Applicable]
  - (d) Information about consent from data providers/curators. [**Yes**/**No**/Not Applicable]  
 Explanation: All data and implementations used are publically available on Github, which we systematically cite.
  - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Yes**/No/**Not Applicable**]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
  - (a) The full text of instructions given to participants and screenshots. [**Yes**/No/**Not Applicable**]
  - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Yes**/No/**Not Applicable**]
  - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Yes**/No/**Not Applicable**]

## A BACKGROUND

### A.1 Background on Subgaussian Measures

**Definition A.1.** A  $\mathbb{R}^d$ -valued random variable  $X \sim \mu$  is subgaussian if any of the following equivalent properties holds (Vershynin, 2018):

1. There exists  $\sigma \geq 0$  such that  $\sup_{v \in S^{d-1}} \mathbb{E}_{X \sim \mu} \left( e^{\frac{|\langle X, v \rangle|^2}{\sigma^2}} \right) \leq 2$ , where  $S^{d-1}$  is the unit sphere in  $\mathbb{R}^d$ .
2. There exists  $\sigma \geq 0$  and an absolute constant  $c_G$  such that for all  $t \geq 0$  we have:

$$\sup_{v \in S^{d-1}} \mathbb{P}_{X \sim \mu}(|\langle X, v \rangle| \geq t) \leq 2 \exp \left( -\frac{c_G t^2}{\sigma^2} \right).$$

3. There exists  $\sigma \geq 0$  and an absolute constant  $C_G$  such that for all  $p \geq 1$ , we have:

$$\sup_{v \in S^{d-1}} \mathbb{E}_{X \sim \mu} (|\langle X, v \rangle|^p)^{1/p} \leq C_G \sigma \sqrt{p}.$$

The smallest  $\sigma$  such that condition 1. holds is referred to as the subgaussian norm of  $X$ , denoted  $\|X\|_{\mathcal{G}}$ . Up to the absolute constants  $c_G$  and  $C_G$ ,  $\|X\|_{\mathcal{G}}$  is the smallest constant that makes any of the above inequalities valid.

We say that  $\mu$  is subgaussian if  $X \sim \mu$  is subgaussian, and define the subgaussian norm of  $\mu$  to be  $\|\mu\|_{\mathcal{G}} := \|X\|_{\mathcal{G}}$ . The set of probability measures on  $\mathbb{R}^d$  with subgaussian norm bounded above  $\sigma$  is referred to as the set of  $\sigma$ -subgaussian measures, and denoted  $\mathcal{G}_{\sigma}(\mathbb{R}^d)$ .

**Lemma A.2.** (Lemma 1 from (Jin et al., 2019)) Let  $\mu \in \mathcal{G}_{\sigma}(\mathbb{R}^d)$  and let  $X \sim \mu$ . Then there exists an absolute constant  $q_G$  such that  $\|X\|$  is  $q_G \sqrt{d} \sigma$ -subgaussian, i.e.,  $\mathbb{E} \left( \exp \left( \frac{\|X\|^2}{2 q_G^2 d \sigma^2} \right) \right) \leq 2$ .

We remark that  $\mathcal{G}_{\sigma}(\mathbb{R}^d)$  is closed with respect to the weak topology on probability measures:

**Definition A.3.** We say that  $\{\mu_n\}_{n=1}^{\infty} \subset \mathcal{P}(\mathbb{R}^d)$  converges weakly (or converges in the weak topology) to  $\mu \in \mathcal{P}(\mathbb{R}^d)$  if  $\int \phi d\mu_n \rightarrow \int \phi d\mu$  as  $n \rightarrow \infty$  for all bounded, continuous functions  $\phi$ .

**Lemma A.4.** Let  $\{\mu_n\}_{n=1}^{\infty} \subset \mathcal{G}_{\sigma}(\mathbb{R}^d)$  be a sequence weakly converging to  $\mu \in \mathcal{P}(\mathbb{R}^d)$  as  $n \rightarrow \infty$ . Then:

1.  $\mu \in \mathcal{G}_{\sigma}(\mathbb{R}^d)$ ;
2.  $\int \|x\|^p d\mu_n \rightarrow \int \|x\|^p d\mu$  as  $n \rightarrow \infty$  for all  $p \geq 1$ .

*Proof.* Fix  $v \in S^{d-1}$  and let  $Z_n^v := |\langle v, X_n \rangle|$  with  $X_n \sim \mu_n$  and  $Z^v := |\langle v, X \rangle|$  with  $X \sim \mu$ . Since  $\mu_n$  converges to  $\mu$  weakly,  $X_n$  converges to  $X$  in distribution by the Portmanteau Lemma [Theorem 2.1 in (Billingsley, 1999)], and by the continuous mapping theorem [Theorem 2.7 in (Billingsley, 1999)],  $Z_n^v$  converges to  $Z^v$  in distribution. By the layer-cake decomposition, we may write for any  $p \geq 1$ :

$$\begin{aligned} \mathbb{E}(|Z_n^v|^p) &= \int_0^{\infty} \mathbb{P}(|Z_n^v|^p \geq t) dt \\ &= \int_0^{\infty} p t^{p-1} \mathbb{P}(|Z_n^v| \geq t) dt. \end{aligned}$$

Hence we have:

$$\lim_{n \rightarrow \infty} \mathbb{E}(|Z_n^v|^p) = \lim_{n \rightarrow \infty} \int_0^{\infty} p t^{p-1} \mathbb{P}(|Z_n^v| \geq t) dt. \quad (15)$$

To interchange the limit and integral in (15), we apply the concentration bound on  $\mathbb{P}(|Z_n^v| \geq t)$  (i.e., characterization 2. in Definition A.1) to get that for all  $t \in [0, \infty)$  and all  $n$ ,

$$p t^{p-1} \mathbb{P}(|Z_n^v| \geq t) \leq 2 p t^{p-1} \exp \left( -\frac{c_G t^2}{\sigma^2} \right),$$

which is integrable for  $p \geq 1$ . Therefore

$$\begin{aligned} (\mathbb{E}(|Z^v|^p))^{\frac{1}{p}} &= \left( \int_0^\infty pt^{p-1} \mathbb{P}(|Z^v| \geq t) dt \right)^{\frac{1}{p}} \\ &= \left( \int_0^\infty \lim_{n \rightarrow \infty} pt^{p-1} \mathbb{P}(|Z_n^v| \geq t) dt \right)^{\frac{1}{p}} \end{aligned} \quad (16)$$

$$\begin{aligned} &= \left( \lim_{n \rightarrow \infty} \int_0^\infty pt^{p-1} \mathbb{P}(|Z_n^v| \geq t) dt \right)^{\frac{1}{p}} \\ &= \left( \lim_{n \rightarrow \infty} \mathbb{E}(|Z_n^v|^p) \right)^{\frac{1}{p}} \\ &\leq C_G \sigma \sqrt{p}, \end{aligned} \quad (17)$$

where (16) follows by convergence in distribution of  $|Z_n^v|$  to  $|Z^v|$  and (17) follows by the Dominated Convergence Theorem. Since this holds for all  $v \in S^{d-1}$ , we conclude  $\mu$  is  $\sigma$ -subgaussian.

The proof of the second statement is similar. Note that since  $\mu$  and  $\{\mu_n\}_{n=1}^\infty$  are  $\sigma$ -subgaussian, Lemma A.2 establishes that both  $\|X\|$  and  $\{\|X_n\|\}_{n=1}^\infty$  are  $q_G \sqrt{d} \sigma$ -subgaussian. Hence we may apply characterization 2 in Definition A.1 to conclude for all  $t \in [0, \infty)$ , and  $n$ :

$$pt^p \mathbb{P}(\|X_n\| \geq t) \leq 2pt^{p-1} \exp\left(-\frac{c_G t^2}{q_G^2 d \sigma^2}\right)$$

which is integrable for all  $p \geq 1$ . Hence we may once again apply the layer-cake decomposition, the dominated convergence theorem, the convergence in distribution of  $X_n$  to  $X$  (and therefore by the continuous mapping theorem of  $\|X_n\|$  to  $\|X\|$ ) to conclude that:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}(\|X_n\|^p) &= \lim_{n \rightarrow \infty} \int_0^\infty pt^{p-1} \mathbb{P}(\|X_n\| \geq t) dt \\ &= \int_0^\infty pt^{p-1} \lim_{n \rightarrow \infty} \mathbb{P}(\|X_n\| \geq t) dt \\ &= \int_0^\infty pt^{p-1} \mathbb{P}(\|X\| \geq t) dt = \mathbb{E}(\|X\|^p). \end{aligned}$$

□

The following variance bound holds uniformly for  $\sigma$ -subgaussian measures:

**Lemma A.5.** *Let  $\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)$ . Then  $\text{Var}(\mu) \leq \tilde{q}_G d \sigma^2$ , where  $\tilde{q}_G$  is an absolute constant.*

*Proof.* For  $X \sim \mu$ , we have  $\text{Var}(\mu) \leq M_2(\mu) = \mathbb{E}_{X \sim \mu} \|X\|^2$ . By Lemma A.2,  $\|X\|$  is  $q_G \sqrt{d} \sigma$ -subgaussian. Applying characterization 3. in Definition A.1 gives  $\mathbb{E}_{X \sim \mu} \|X\|^2 \leq 2q_G^2 d \sigma^2$  as desired. □

Finally, we record a bound showing that centering a random variable can only increase its subgaussian norm by an absolute constant factor:

**Lemma A.6.** *(Lemma 2.6.8 in (Vershynin, 2018)) Let  $\mu$  be  $\sigma$ -subgaussian, and let  $X \sim \mu$ . Let  $\tilde{\mu}$  be the distribution of  $X - \mathbb{E}_{X \sim \mu}(X)$ . Then  $\tilde{\mu}$  is  $\tilde{C}_G \sigma$ -subgaussian, where  $\tilde{C}_G$  is an absolute constant.*

## A.2 Background on Optimization of Functionals on $\mathcal{P}_2(\mathbb{R}^d)$

**Definition A.7.** *Let  $\mathcal{U} \subseteq \mathcal{P}_2(\mathbb{R}^d)$  be a convex subset (i.e., for all  $\mu, \mu' \in \mathcal{U}$  and  $t \in [0, 1]$ ,  $t\mu + (1-t)\mu' \in \mathcal{U}$ ). Then a functional  $\mathcal{F} : \mathcal{U} \rightarrow \mathbb{R}$  is convex if for all  $\mu, \mu' \in \mathcal{U}$  and  $t \in [0, 1]$  we have:*

$$\mathcal{F}((1-t)\mu + t\mu') \leq (1-t)\mathcal{F}(\mu) + t\mathcal{F}(\mu').$$

*It is strictly convex if the inequality is strict for  $t \in (0, 1)$ .*

If  $\mathcal{F}$  is differentiable on a convex subset  $\mathcal{U} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ , one has an alternative characterization of convexity:

**Proposition A.8.** (Proposition 6 in (Janati et al., 2020a)) *A differentiable functional  $\mathcal{F}$  is convex over a convex subset  $\mathcal{U} \subseteq \mathcal{P}_2(\mathbb{R}^d)$  iff for all  $\mu, \mu' \in \mathcal{U}$  we have:*

$$\mathcal{F}(\mu) \geq \mathcal{F}(\mu') + \langle \delta \mathcal{F}(\mu'), \mu - \mu' \rangle. \quad (18)$$

*It is strictly convex if the inequality is strict for all  $\mu \neq \mu'$ .*

We note that Proposition 6 in (Janati et al., 2020a) establishes the above for  $\mathcal{U} = \mathcal{P}_2(\mathbb{R}^d)$  but the proof is the same for a general convex  $\mathcal{U} \subset \mathcal{P}_2(\mathbb{R}^d)$ .

We note the following equivalent characterization of optimizers of a convex functional:

**Proposition A.9.** (Proposition 7 in (Janati et al., 2020a)) *Let  $\mathcal{F}$  be a convex, differentiable functional on a convex subset  $\mathcal{U} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ . Then  $\mu^*$  is minimal for  $\mathcal{F}$  on  $\mathcal{U}$  iff  $\langle \delta \mathcal{F}(\mu^*), \mu - \mu^* \rangle \geq 0$  for all  $\mu \in \mathcal{U}$ .*

Next, we show if  $\mathcal{F}$  is differentiable and admits a minimizer  $\mu^*$ , then the derivative at  $\mu^*$  is constant  $\mu^*$ -almost everywhere. The following is essentially Proposition 7.20 in (Santambrogio, 2015):

**Lemma A.10.** *Let  $\mathcal{U}$  be either  $\mathcal{P}_2(\Omega)$  or  $\mathcal{G}(\Omega)$ , and let  $\mathcal{F} : \mathcal{U} \rightarrow \mathbb{R}$  be differentiable on  $\mathcal{U}$ . Suppose that  $\mathcal{F}$  admits a minimizer  $\mu^* \in \mathcal{U}$ . Let  $c^* = \inf_{x \in \Omega} \delta \mathcal{F}(\mu^*)(x)$ . Then:*

- $c^* > -\infty$ ;
- $\delta \mathcal{F}(\mu^*)(x) = c^*$  for all  $x \in \text{supp}(\mu^*)$ .

*Proof.* Let  $\mathcal{U} = \mathcal{P}_2(\Omega)$  (the proof for  $\mathcal{G}(\Omega)$  is identical). Since  $\mu^* \in \mathcal{P}_2(\Omega)$  is minimal, we have for any  $\rho \in \mathcal{P}_2(\Omega)$  and  $t \in (0, 1]$ :

$$\frac{\mathcal{F}(\mu^* + t(\rho - \mu^*)) - \mathcal{F}(\mu^*)}{t} \geq 0.$$

Taking the limit as  $t \rightarrow 0^+$ , we have that

$$0 \leq \lim_{t \rightarrow 0^+} \frac{\mathcal{F}(\mu^* + t(\rho - \mu^*)) - \mathcal{F}(\mu^*)}{t} = \int \delta \mathcal{F}(\mu^*) d(\rho - \mu^*) = \langle \delta \mathcal{F}(\mu^*), \rho - \mu^* \rangle \quad (19)$$

by definition. Assume that  $c^* = \inf \delta \mathcal{F}(\mu^*)(x) = -\infty$ . For any  $\epsilon \in \mathbb{R}$ , define  $V^\epsilon := \{x \in \Omega \mid \delta \mathcal{F}(\mu^*)(x) > \epsilon\}$ . These sets are measurable by continuity of  $\delta \mathcal{F}(\mu^*)$ , and by definition of the infimum,  $(V^\epsilon)^c$  is nonempty for any  $\epsilon$ . By definition  $\{x \in \Omega \mid \delta \mathcal{F}(\mu^*)(x) = -\infty\} = \emptyset$ . Towards a contradiction, assume that  $\mu^*(V^{\bar{\epsilon}}) > 0$  for some fixed  $\bar{\epsilon}$ , and define  $A := \text{supp}(\mu^*) \cap V^{\bar{\epsilon}}$ , and denote  $\Omega \setminus A$  by  $A^c \neq \emptyset$ . We define  $\mu_A^*(B) := \mu^*(A \cap B)$  and  $\mu_{A^c}^*(B) = \mu^*(A^c \cap B)$ . Then for any  $\nu \in \mathcal{P}_2(A^c)$  and  $t \in (0, 1)$ , we may define  $\mu_t := (1-t)\mu_A^* + \mu_{A^c}^* + t\mu^*(A)\nu$ , which lies in  $\mathcal{P}_2(\Omega)$ . Using the fact that  $\mu^* = \mu_A^* + \mu_{A^c}^*$ ,

$$\begin{aligned} \langle \delta \mathcal{F}(\mu^*), \mu_t - \mu^* \rangle &= \langle \delta \mathcal{F}(\mu^*), (1-t)\mu_A^* + \mu_{A^c}^* + t\mu^*(A)\nu - \mu_A^* - \mu_{A^c}^* \rangle \\ &= \langle \delta \mathcal{F}(\mu^*), t\mu^*(A)\nu - t\mu_A^* \rangle \\ &= t(\langle \delta \mathcal{F}(\mu^*), \mu^*(A)\nu \rangle - \langle \delta \mathcal{F}(\mu^*), \mu_A^* \rangle) \end{aligned} \quad (20)$$

By construction,

$$\langle \delta \mathcal{F}(\mu^*), \mu_A^* \rangle > \bar{\epsilon} \mu^*(A),$$

whereas

$$\langle \delta \mathcal{F}(\mu^*), \mu^*(A)\nu \rangle = \mu^*(A) \langle \delta \mathcal{F}(\mu^*), \nu \rangle \leq \bar{\epsilon} \mu^*(A),$$

and hence (20) is less than 0, contradicting (19). Hence we must conclude that  $\mu^*(V^{\bar{\epsilon}}) = 0$  for any fixed  $\bar{\epsilon}$ . But this implies that  $\mu^*$  must be supported on  $\Omega \setminus \bigcup_{\epsilon \in \mathbb{R}} V^\epsilon$ , which is empty, giving us a contradiction. We conclude that  $c^* = \inf \delta \mathcal{F}(\mu^*)(x) > -\infty$ .

We prove the second claim similarly. We define for any  $\epsilon > 0$ ,

$$U^\epsilon := \{x \in \Omega \mid \delta \mathcal{F}(\mu^*)(x) > c^* + \epsilon\}.$$



which is measurable by continuity of  $\delta\mathcal{F}(\mu^*)$ . By definition of  $c^*$ , the set  $\Omega \setminus U^\epsilon$  is nonempty for all  $\epsilon > 0$ . By way of contradiction, suppose there exists an  $\bar{\epsilon} > 0$  such that  $\mu^*(U^{\bar{\epsilon}}) > 0$ . Let  $A := \text{supp}(\mu^*) \cap U^{\bar{\epsilon}}$  and let  $\mu_A^*$  and  $\mu_{A^c}^*$  be as above. Let  $\nu \in \mathcal{P}_2(A^c)$ , and for any  $t \in (0, 1)$ , define  $\mu_t = (1 - t)\mu_A^* + \mu_{A^c}^* + t\mu^*(A)\nu$ , which lies in  $\mathcal{P}_2(\Omega)$ . Then:

$$\begin{aligned} \langle \delta\mathcal{F}(\mu^*), \mu_t - \mu^* \rangle &= \langle \delta\mathcal{F}(\mu^*), (1 - t)\mu_A^* + \mu_{A^c}^* + t\mu^*(A)\nu - \mu_A^* - \mu_{A^c}^* \rangle \\ &= \langle \delta\mathcal{F}(\mu^*), t\mu^*(A)\nu - t\mu_A^* \rangle \\ &= t(\langle \delta\mathcal{F}(\mu^*), \mu^*(A)\nu \rangle - \langle \delta\mathcal{F}(\mu^*), \mu_A^* \rangle). \end{aligned} \quad (21)$$

By construction,

$$\langle \delta\mathcal{F}(\mu^*), \mu_A^* \rangle > (c^* + \bar{\epsilon})\mu^*(A),$$

whereas

$$\langle \delta\mathcal{F}(\mu^*), \mu^*(A)\nu \rangle = \mu^*(A)\langle \delta\mathcal{F}(\mu^*), \nu \rangle \leq (c^* + \bar{\epsilon})\mu^*(A),$$

and hence (21) is less than 0, contradicting (19), hence it must be the case that  $\mu^*(\text{supp}(\mu^*) \cap U^\epsilon) = 0$  for all  $\epsilon > 0$ , i.e., that  $\mu^* \in \mathcal{P}(\Omega \setminus U^\epsilon)$  for all  $\epsilon > 0$ . Hence for  $\mu^*$ -almost every  $x$  and all  $\epsilon > 0$ ,  $\delta\mathcal{F}(\mu^*)(x) < c^* + \epsilon$ . Hence for  $\mu^*$ -almost every  $x$ ,

$$c^* \leq \delta\mathcal{F}(\mu^*)(x) < \liminf_{\epsilon \rightarrow 0^+} c^* + \epsilon$$

which implies that  $\delta\mathcal{F}(\mu^*)(x) = c^*$ . We may immediately promote this to  $\delta\mathcal{F}(\mu^*)(x) = c^*$  everywhere on  $\text{supp}(\mu^*)$  by continuity of  $\delta\mathcal{F}(\mu^*)$ .  $\square$

**Corollary A.11.** *Let  $\mathcal{U}$  be either  $\mathcal{P}_2(\mathbb{R}^d)$  or  $\mathcal{G}(\mathbb{R}^d)$ . Let  $\mathcal{F} : \mathcal{U} \rightarrow \mathbb{R}$  be a differentiable, convex functional which admits a minimizer  $\mu^* \in \mathcal{U}$ . Furthermore, suppose that  $\delta\mathcal{F}(\mu^*)$  is differentiable. Then  $\mu^*$  is a critical point of  $\mathcal{F}$  on  $\mathcal{U}$ .*

*Proof.* From Lemma A.10, we know that  $\delta\mathcal{F}(\mu^*)(x) = c^*$  for all  $x \in \text{supp}(\mu^*)$ , where  $c^* := \min_{x \in \mathbb{R}^d} \delta\mathcal{F}(\mu^*)(x)$ . Hence any  $x \in \text{supp}(\mu^*)$  minimizes a differentiable function  $\delta\mathcal{F}(\mu^*)$ , and hence  $\nabla \delta\mathcal{F}(\mu^*)(x) = 0$ . Thus  $\mu^*$  is a critical point of  $\mathcal{F}$ .  $\square$

## B PROOFS FOR SECTION 2

### B.1 Proof of Proposition 2.1

We begin by proving existence. We will require Prokhorov's Theorem, which provides a sufficient condition for convergence of an infimizing subsequence for our functionals:

**Definition B.1.** *Let  $(\mathcal{X}, d)$  be a separable metric space. A sequence of probability measures  $\{\mu_n\}_{n=1}^\infty \subseteq \mathcal{P}(\mathcal{X})$  is tight if for all  $\epsilon > 0$  there exists a compact set  $K_\epsilon$  such that  $\mu_n(\mathcal{X} \setminus K_\epsilon) < \epsilon$  for all  $n$ .*

**Theorem B.2.** *(Prokhorov's Theorem, Theorems 5.1 and 5.2 in (Billingsley, 1999)) Let  $(\mathcal{X}, d)$  be a complete separable metric space. Then a sequence of probability measures  $\{\mu_n\}_{n=1}^\infty \subseteq \mathcal{P}(\mathcal{X})$  is tight if and only if every subsequence of  $\{\mu_n\}_{n=1}^\infty$  contains a further subsequence which weakly converges to a limit in  $\mathcal{P}(\mathcal{X})$ .*

**Lemma B.3.** *Let  $\mathcal{V} \subset \mathcal{P}_2(\Omega)$ . Any infimizing sequence for  $F_{\lambda, \mathcal{V}}^\epsilon$  is tight.*

*Proof.* Let  $\{\mu_n\}_{n=1}^\infty$  be an infimizing sequence for  $F_{\lambda, \mathcal{V}}^\epsilon$ . We claim that  $\|\mathbb{E}(\mu_n)\|$  is uniformly bounded over  $n$ . Let  $\rho \in \mathcal{P}_2(\Omega)$  and define  $L := F_{\lambda, \mathcal{V}}^\epsilon(\rho) < \infty$ . As  $\{\mu_n\}_{n=1}^\infty$  is infimizing, there exists some  $N$  such that for all  $n > N$ ,

$F_{\lambda, \nu}^\epsilon(\mu_n) \leq L$ . Since  $OT_2(\mu, \nu)^2 \leq OT_2^\epsilon(\mu, \nu)$ , we have:

$$\begin{aligned}
 L &\geq F_{\lambda, \nu}^\epsilon(\mu_n) \\
 &\geq \sum_{j=1}^m \lambda_j OT_2(\mu_n, \nu_j)^2 \\
 &= \sum_{j=1}^m \lambda_j \inf_{\zeta_j \in \Pi(\mu_n, \nu_j)} \int \frac{1}{2} \|x - y\|^2 d\zeta_j(x, y) \\
 &\geq \frac{1}{2} \sum_{j=1}^m \lambda_j \inf_{\zeta_j \in \Pi(\mu_n, \nu_j)} \left\| \int (x - y) d\zeta_j(x, y) \right\|^2 \\
 &= \frac{1}{2} \sum_{j=1}^m \lambda_j \left\| \int x d\mu_n(x) - \int y d\nu_j(y) \right\|^2 \\
 &= \frac{1}{2} \sum_{j=1}^m \lambda_j \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|^2 \\
 &\geq \frac{1}{2} \min_{1 \leq j \leq m} \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|^2
 \end{aligned} \tag{22}$$

where we applied Jensen's inequality at (22), and hence  $\min_{1 \leq j \leq m} \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\| \leq \sqrt{2L}$  for all large enough  $n$ . We have thus shown that  $\min_{1 \leq j \leq m} \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|$  is uniformly bounded over all  $n$ , which implies that  $\|\mathbb{E}(\mu_n)\|$  is uniformly bounded over all  $n$  as well by some constant  $M$ .

Next, we claim that  $\mathbf{Var}(\mu_n)$  is uniformly bounded for all  $n$ . To see this, note that for any Dirac mass  $\delta_x$  and  $\mu \in \mathcal{P}_2(\Omega)$ ,  $OT_2^\epsilon(\delta_x, \mu) = OT_2^2(\delta_x, \mu)$ , as the only coupling between  $\delta_x$  and  $\mu$  is  $\delta_x \otimes \mu$ , and hence the KL term vanishes. We lower bound:  $OT_2^2(\delta_x, \mu) \geq \min_z OT_2^2(\delta_z, \mu)$ , then observe that:

$$\begin{aligned}
 \min_z OT_2^2(\delta_z, \mu) &= \frac{1}{2} \min_z \int \|z - y\|^2 d\mu(y) \\
 &= \frac{1}{2} \int \|\mathbb{E}(\mu) - y\|^2 d\mu(y) \\
 &= \frac{1}{2} \mathbf{Var}(\mu),
 \end{aligned}$$

which follows from the characterization of  $\mathbb{E}(\mu)$  as the minimizer for the mean-squared error of  $\mu$ . Now, fix some  $x_0 \in \mathbb{R}^d$ . Then by the triangle inequality for  $OT_2$  and the bound  $L \geq \sum_{j=1}^m \lambda_j OT_2(\mu_n, \nu_j)^2$ , for all large enough  $n$  there exists an index  $1 \leq j_n \leq m$  such that:

$$\begin{aligned}
 \sqrt{\mathbf{Var}(\mu_n)} &\leq \sqrt{2} OT_2(\delta_{x_0}, \mu_n) \\
 &\leq \sqrt{2} OT_2(\delta_{x_0}, \delta_{\mathbb{E}(\nu_{j_n})}) + \sqrt{2} OT_2(\delta_{\mathbb{E}(\nu_{j_n})}, \nu_{j_n}) + \sqrt{2} OT_2(\nu_{j_n}, \mu_n) \\
 &\leq \sqrt{2} \|x_0 - \mathbb{E}(\nu_{j_n})\| + \sqrt{2 \mathbf{Var}(\nu_{j_n})} + \sqrt{2L}
 \end{aligned}$$

Hence  $\mathbf{Var}(\mu_n) \leq \max_{1 \leq j \leq m} (\sqrt{2} \|x_0 - \mathbb{E}(\nu_j)\| + \sqrt{2 \mathbf{Var}(\nu_j)} + \sqrt{2L})^2 := Z$  for all large enough  $n$ , which implies a uniform bound for all  $n$ . As we have uniformly bounded  $\mathbf{Var}(\mu_n)$  for all  $n$ , we may apply Chebyshev's inequality to conclude that:

$$\mathbb{P}_{X \sim \mu_n}(\|X - \mathbb{E}(\mu_n)\| > t) < \frac{Z}{t^2} \tag{24}$$

for some  $Z \geq 0$  and for all  $n$  and  $t > 0$ . As  $\|\mathbb{E}(\mu_n)\| \leq M$  for all  $n$ ,  $\mathbb{E}(\mu_n) \in \overline{B_M(0)}$ , and hence for any  $t > 0$  and  $n \in \mathbb{N}$  we have

$$A_n^t := \{x \in \Omega \mid \|x - \mathbb{E}(\mu_n)\| \leq t\} \subseteq \overline{B_{t+M}(0)} \cap \Omega.$$

By this and (24) we have:

$$\mathbb{P}_{X \sim \mu_n}(X \in \overline{B_{t+M}(0)} \cap \Omega) \geq \mathbb{P}_{X \sim \mu_n}(X \in A_n^t) > 1 - \frac{Z}{t^2}$$

for all  $n$  and  $t > 0$ . Since  $\Omega$  is closed,  $\overline{B_{t+M}(0)} \cap \Omega$  is compact, and hence  $K(t) := \overline{B_{t+M}(0)} \cap \Omega$  forms a family of compact sets such that, for any  $\epsilon > 0$  and  $n \in \mathbb{N}$ ,  $\mu_n(K(t)) > 1 - \epsilon$  for all  $t > \sqrt{Z/\epsilon}$ , establishing that  $\{\mu_n\}_{n=1}^\infty$  is tight.  $\square$

We now show that for fixed  $\sigma > 0$ , a  $\sigma$ -subgaussian infimizing sequence for  $S_{\lambda, \mathcal{V}}^\epsilon$  is tight.

**Lemma B.4.** *Let  $\sigma > 0$ . Suppose that  $\{\mu_n\}_{n=1}^\infty \subset \mathcal{G}_\sigma(\Omega)$  is an infimizing sequence for  $S_{\lambda, \mathcal{V}}^\epsilon$ , where  $\mathcal{V} \subset \mathcal{P}_2(\Omega)$ . Then  $\|\mathbb{E}(\mu_n)\|$  is uniformly bounded and  $\{\mu_n\}_{n=1}^\infty$  is tight.*

*Proof.* Let  $\lim_{n \rightarrow \infty} S_{\lambda, \mathcal{V}}^\epsilon(\mu_n) = \inf_{\mu \in \mathcal{G}_\sigma(\Omega)} S_{\lambda, \mathcal{V}}^\epsilon(\mu)$ . We construct a lower bound on  $S_{\lambda, \mathcal{V}}^\epsilon$  along our infimizing sequence as follows. For all  $n$ , we have:

$$\begin{aligned} S_{\lambda, \mathcal{V}}^\epsilon(\mu_n) &= \sum_{j=1}^m \lambda_j (OT_2^\epsilon(\mu_n, \nu_j) - \frac{1}{2} OT_2^\epsilon(\nu_j, \nu_j)) - \frac{1}{2} OT_2^\epsilon(\mu_n, \mu_n) \\ &\geq \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu_n, \nu_j) - J - \frac{1}{2} OT_2^\epsilon(\mu_n, \mu_n) \\ &\geq \sum_{j=1}^m \lambda_j OT_2(\mu_n, \nu_j)^2 - J - \frac{1}{2} OT_2^\epsilon(\mu_n, \mu_n) \end{aligned} \quad (25)$$

$$\geq \sum_{j=1}^m \lambda_j OT_2(\mu_n, \nu_j)^2 - J - \frac{1}{2} \text{Var}(\mu_n) \quad (26)$$

$$\geq \sum_{j=1}^m \lambda_j OT_2(\mu_n, \nu_j)^2 - J - \frac{1}{2} \tilde{q}_G d\sigma^2 \quad (27)$$

where  $J = \max_{1 \leq j \leq m} OT_2^\epsilon(\nu_j, \nu_j)$ , and  $\tilde{q}_G$  is an absolute constant. In (25) we used the fact that  $OT_2^\epsilon(\rho_1, \rho_2) \geq OT_2(\rho_1, \rho_2)^2$  for any  $\rho_1, \rho_2 \in \mathcal{P}_2(\Omega)$  and in (26), we used:

$$\begin{aligned} OT_2^\epsilon(\mu_n, \mu_n) &= \inf_{\zeta \in \Pi(\mu_n, \mu_n)} \int \frac{1}{2} \|x - y\|^2 d\zeta(x, y) + \epsilon KL(\zeta | \mu_n \otimes \mu_n) \\ &\leq \int \frac{1}{2} \|x - y\|^2 d\mu_n \otimes \mu_n(x, y) + \epsilon KL(\mu_n \otimes \mu_n | \mu_n \otimes \mu_n) \\ &= \int \frac{1}{2} \|x - y\|^2 d\mu_n \otimes \mu_n(x, y) \\ &= \int \|x - \mathbb{E}(\mu_n)\|^2 d\mu_n = \text{Var}(\mu_n). \end{aligned}$$

In (27) we applied Lemma A.5.

Using the same arguments to establish (23), we have

$$\begin{aligned} \sum_{j=1}^m \lambda_j OT_2(\mu_n, \nu_j)^2 &= \sum_{j=1}^m \lambda_j \inf_{\zeta_j \in \Pi(\mu_n, \nu_j)} \int \frac{1}{2} \|x - y\|^2 d\zeta_j(x, y) \\ &\geq \frac{1}{2} \sum_{j=1}^m \lambda_j \inf_{\zeta_j \in \Pi(\mu_n, \nu_j)} \left\| \int (x - y) d\zeta_j(x, y) \right\|^2 \\ &= \frac{1}{2} \sum_{j=1}^m \lambda_j \left\| \int x d\mu_n(x) - \int y d\nu_j(y) \right\|^2 \\ &= \frac{1}{2} \sum_{j=1}^m \lambda_j \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|^2 \\ &\geq \frac{1}{2} \min_{1 \leq j \leq m} \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|^2 \end{aligned}$$

and thus we may lower bound  $S_{\lambda, \nu}^\epsilon(\mu_n)$  by:

$$S_{\lambda, \nu}^\epsilon(\mu_n) \geq \frac{1}{2} \min_{1 \leq j \leq m} \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|^2 - J - \frac{1}{2} \tilde{q}_G d \sigma^2. \quad (28)$$

Since  $\mu_n$  is infimizing and  $\inf_{\mu \in \mathcal{G}_\sigma(\Omega)} S_{\lambda, \nu}^\epsilon(\mu_n) < \infty$ , there exists  $B > 0$  such that  $\sup_n S_{\lambda, \nu}^\epsilon(\mu_n) < B$ . Hence  $\frac{1}{2} \min_{1 \leq j \leq m} \|\mathbb{E}(\mu_n) - \mathbb{E}(\nu_j)\|^2 < B + J + \frac{1}{2} \tilde{q}_G d \sigma^2$  for all  $n$ , and we can conclude that  $\|\mathbb{E}(\mu_n)\|$  must be uniformly bounded in  $n$ . To conclude, we have established a uniform bound on  $\|\mathbb{E}(\mu_n)\|$ , and assumed a uniform bound on  $\text{Var}(\mu_n)$  via Lemma A.5. We may thus repeat the argument in Lemma B.3 to establish that  $\{\mu_n\}_{n=1}^\infty$  is tight.  $\square$

We will also need the following continuity properties of  $OT_2^\epsilon$ :

**Theorem B.5.** (Theorem 3.7 in (Eckstein and Nutz, 2022)) Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ . Then:

$$\begin{aligned} & |OT_2^\epsilon(\mu, \nu) - OT_2^\epsilon(\mu', \nu')| \\ & \leq \sqrt{2} \left( \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} + \sqrt{M_2(\nu)} + \sqrt{M_2(\nu')} \right) \sqrt{OT_2(\mu, \mu')^2 + OT_2(\nu, \nu')^2}. \end{aligned}$$

This immediately implies a continuity result for  $\mathcal{F}_{\lambda, \nu}^\epsilon$ :

**Corollary B.6.** Let  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$ . For any  $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$  and  $\epsilon > 0$ ,  $F_{\lambda, \nu}^\epsilon(\mu)$  satisfies:

$$|F_{\lambda, \nu}^\epsilon(\mu') - F_{\lambda, \nu}^\epsilon(\mu)| \leq \sqrt{2} \left( 2 \max_{1 \leq j \leq m} \sqrt{M_2(\nu_j)} + \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} \right) OT_2(\mu, \mu').$$

Similarly,  $S_{\lambda, \nu}^\epsilon(\mu)$  satisfies:

$$|S_{\lambda, \nu}^\epsilon(\mu') - S_{\lambda, \nu}^\epsilon(\mu)| \leq 3\sqrt{2} \left( \max_{1 \leq j \leq m} \sqrt{M_2(\nu_j)} + \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} \right) OT_2(\mu, \mu').$$

*Proof.* Theorem B.5 establishes:

$$|OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu', \nu_j)| \leq \sqrt{2} \left( 2\sqrt{M_2(\nu_j)} + \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} \right) OT_2(\mu, \mu')$$

for each  $j$ . We then bound:

$$\begin{aligned} & |F_{\lambda, \nu}^\epsilon(\mu) - F_{\lambda, \nu}^\epsilon(\mu')| \\ & = \left| \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu, \nu_j) - \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu', \nu_j) \right| \\ & \leq \sum_{j=1}^m \lambda_j |OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu', \nu_j)| \\ & \leq \sum_{j=1}^m \lambda_j \sqrt{2} \left( \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} + 2\sqrt{M_2(\nu_j)} \right) OT_2(\mu, \mu') \\ & = \sqrt{2} (\sqrt{M_2(\mu)} + \sqrt{M_2(\mu')}) OT_2(\mu, \mu') + 2\sqrt{2} \sum_{j=1}^m \lambda_j \sqrt{M_2(\nu_j)} OT_2(\mu, \mu') \\ & \leq \sqrt{2} (\sqrt{M_2(\mu)} + \sqrt{M_2(\mu')}) OT_2(\mu, \mu') + 2\sqrt{2} \max_{1 \leq j \leq m} \sqrt{M_2(\nu_j)} OT_2(\mu, \mu') \\ & = \sqrt{2} \left( 2 \max_{1 \leq j \leq m} \sqrt{M_2(\nu_j)} + \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} \right) OT_2(\mu, \mu'). \end{aligned}$$

The proof for  $S_{\lambda, \nu}^\epsilon$  is similar:

$$\begin{aligned} & |S_{\lambda, \nu}^\epsilon(\mu') - S_{\lambda, \nu}^\epsilon(\mu)| \\ &= \left| \sum_{j=1}^m \lambda_j (OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu', \nu_j)) + \frac{1}{2} (OT_2^\epsilon(\mu', \mu') - OT_2^\epsilon(\mu, \mu)) \right| \\ &\leq \left| \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu', \nu_j) \right| + \frac{1}{2} |OT_2^\epsilon(\mu', \mu') - OT_2^\epsilon(\mu, \mu)|. \end{aligned}$$

By Theorem B.5, we may bound the first term by

$$\sqrt{2} \left( 2 \max_{1 \leq j \leq m} \sqrt{M_2(\nu_j)} + \sqrt{M_2(\mu)} + \sqrt{M_2(\mu')} \right) OT_2(\mu, \mu'),$$

and the second term by  $2\sqrt{2}(\sqrt{M_2(\mu)} + \sqrt{M_2(\mu')})OT_2(\mu, \mu')$ .  $\square$

**Proof of Proposition 2.1:** We first establish existence of the minimizers for  $F_{\lambda, \nu}^\epsilon$ . For all  $j$ , the functional  $OT_2^\epsilon(\mu, \nu_j)$  is non-negative and lower semicontinuous w.r.t. the weak convergence of probability measures. This can be seen by noting (2) implies that  $OT_2^\epsilon(\mu, \nu_j)$  is the supremum of a family of continuous linear functionals, and hence is lower semicontinuous (Bell, 2014). This implies that  $F_{\lambda, \nu}^\epsilon$  is lower semicontinuous and bounded below. As  $F_{\lambda, \nu}^\epsilon$  is bounded below,  $\inf_{\mu \in \mathcal{P}_2(\Omega)} F_{\lambda, \nu}^\epsilon(\mu)$  is finite. Let  $\{\mu_n\}_{n=1}^\infty$  be an infimizing sequence for  $F_{\lambda, \nu}^\epsilon$ , which by Lemma B.3 is tight. As  $\Omega$  is closed, it is a complete metric space w.r.t.  $d(x, y) = \|x - y\|$ . Hence by Theorem B.2, the sequence  $\{\mu_n\}_{n=1}^\infty$  contains a subsequence  $\{\mu'_n\}_{n=1}^\infty \subset \{\mu_n\}_{n=1}^\infty$  which converges weakly to a limit  $\mu' \in \mathcal{P}_2(\Omega)$ . By lower semicontinuity,  $F_{\lambda, \nu}^\epsilon(\mu') \leq \liminf_{n \rightarrow \infty} F_{\lambda, \nu}^\epsilon(\mu'_n) = \inf_{\mu \in \mathcal{P}_2(\Omega)} F_{\lambda, \nu}^\epsilon(\mu)$  as  $\{\mu'_n\}_{n=1}^\infty$  is a subsequence of an infimizing sequence, and since  $\inf_{\mu \in \mathcal{P}_2(\Omega)} F_{\lambda, \nu}^\epsilon(\mu) \leq F_{\lambda, \nu}^\epsilon(\mu')$  it implies that  $\mu'$  achieves the infimum.

We now establish existence for  $S_{\lambda, \nu}^\epsilon$ . By (27),  $S_{\lambda, \nu}^\epsilon$  is uniformly lower bounded on  $\mathcal{G}_\sigma(\Omega)$ , and hence  $\inf_{\mu \in \mathcal{G}_\sigma(\Omega)} S_{\lambda, \nu}^\epsilon(\mu) > -\infty$ . By definition of infimum, we may find a sequence  $\{\mu_n\}_{n=1}^\infty \subseteq \mathcal{G}_\sigma(\Omega)$  so that  $\lim_{n \rightarrow \infty} S_{\lambda, \nu}^\epsilon(\mu_n) = \inf_{\mu \in \mathcal{G}_\sigma(\Omega)} S_{\lambda, \nu}^\epsilon(\mu)$ . By Lemma B.4,  $\{\mu_n\}_{n=1}^\infty$  is tight, and hence by Prokhorov's theorem it has a subsequence  $\{\mu'_n\}_{n=1}^\infty$  weakly converging to a limit  $\mu'$ . We now claim that:

$$S_{\lambda, \nu}^\epsilon(\mu') = \lim_{n \rightarrow \infty} S_{\lambda, \nu}^\epsilon(\mu'_n). \quad (29)$$

By Lemma A.4,  $\mu' \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , and  $\lim_{n \rightarrow \infty} M_2(\mu'_n) = M_2(\mu')$ . Weak convergence and convergence of second moments implies that  $OT_2(\mu'_n, \mu') \rightarrow 0$  as  $n \rightarrow \infty$  (Theorem 5.11 in (Santambrogio, 2015)). We now appeal to Corollary B.6, which establishes that:

$$\begin{aligned} |S_{\lambda, \nu}^\epsilon(\mu'_n) - S_{\lambda, \nu}^\epsilon(\mu')| &\leq 3\sqrt{2} \left( \max_{1 \leq j \leq m} \sqrt{M_2(\nu_j)} + \sqrt{M_2(\mu'_n)} + \sqrt{M_2(\mu')} \right) OT_2(\mu'_n, \mu') \\ &\leq 18\sqrt{d}q_G\sigma OT_2(\mu'_n, \mu') \rightarrow 0, \quad n \rightarrow \infty, \end{aligned}$$

where we applied the bound  $\sqrt{M_2(\rho)} \leq \sqrt{2d}q_G\sigma$  for all  $\rho \in \mathcal{G}_\sigma(\mathbb{R}^d)$  (item 3 in Definition A.1 and Lemma A.2). As  $\{S_{\lambda, \nu}^\epsilon(\mu'_n)\}_{n=1}^\infty$  is a subsequence of a convergent sequence, we may conclude that

$$S_{\lambda, \nu}^\epsilon(\mu') = \lim_{n \rightarrow \infty} S_{\lambda, \nu}^\epsilon(\mu'_n) = \lim_{n \rightarrow \infty} S_{\lambda, \nu}^\epsilon(\mu_n) = \inf_{\mu \in \mathcal{G}_\sigma(\Omega)} S_{\lambda, \nu}^\epsilon(\mu),$$

and thus  $\mu'$  minimizes  $S_{\lambda, \nu}^\epsilon$  on  $\mathcal{G}_\sigma(\Omega)$ .

We now establish uniqueness of the minimizer of  $S_{\lambda, \nu}^\epsilon$  when  $\Omega$  is bounded. By Proposition 4 in (Feydy et al., 2019), the functional  $-\frac{1}{2}OT_2^\epsilon(\mu, \mu) : \mathcal{P}_2(\Omega) \rightarrow \mathbb{R}$  is strictly convex. As  $\sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu, \nu_j)$  is convex,  $S_{\lambda, \nu}^\epsilon$  is strictly convex on  $\mathcal{P}_2(\Omega)$ , from which we conclude uniqueness of the minimizer.  $\square$

## B.2 Proof of Theorem 2.2

First, we state general bounds for subgaussian random variables:

**Lemma B.7.** (Lemma 1.5 in (Rigollet and Hütter, 2023)) Suppose  $\mu$  is  $\sigma$ -subgaussian in  $\mathbb{R}^d$  and  $\mathbb{E}(\mu) = 0$ . Then, for any  $s > 0$  we have:

$$\sup_{v \in S^{d-1}} \int \exp(s|\langle x, v \rangle|) d\mu(x) \leq \exp(4\sigma^2 s^2).$$

We obtain an immediate corollary:

**Corollary B.8.** Suppose  $\mu$  is  $\sigma$ -subgaussian in  $\mathbb{R}^d$  and  $z \in \mathbb{R}^d$ . Then:

$$\int \exp(|\langle x, z \rangle|) d\mu(x) \leq \exp(4\tilde{C}_G^2 \sigma^2 \|z\|^2 + \|\mathbb{E}(\mu)\| \|z\|).$$

where  $\tilde{C}_G$  is an absolute constant.

*Proof.* For any  $z$ ,

$$\begin{aligned} \int \exp(|\langle x, z \rangle|) d\mu(x) &= \int \exp(|\langle x - \mathbb{E}(\mu) + \mathbb{E}(\mu), z \rangle|) d\mu(x) \\ &\leq \int \exp(|\langle x - \mathbb{E}(\mu), z \rangle| + |\langle \mathbb{E}(\mu), z \rangle|) d\mu(x) \end{aligned} \quad (30)$$

$$\begin{aligned} &= \exp(|\langle \mathbb{E}(\mu), z \rangle|) \int \exp(|\langle x - \mathbb{E}(\mu), z \rangle|) d\mu(x) \\ &= \exp(|\langle \mathbb{E}(\mu), z \rangle|) \int \exp\left(\left|\left\langle x - \mathbb{E}(\mu), \frac{z}{\|z\|} \right\rangle\right| \|z\|\right) d\mu(x) \\ &\leq \exp(\|\mathbb{E}(\mu)\| \|z\|) \sup_{v \in S^{d-1}} \int \exp(|\langle x - \mathbb{E}(\mu), v \rangle| \|z\|) d\mu(x) \end{aligned} \quad (31)$$

$$\leq \exp(4\|\tilde{\mu}\|_{\mathcal{G}}^2 \|z\|^2 + \|\mathbb{E}(\mu)\| \|z\|). \quad (32)$$

where  $\tilde{\mu}$  is the distribution of the mean zero random variable  $X - \mathbb{E}(\mu)$ . In (30) we applied the triangle inequality, in (31) we applied the Cauchy-Schwarz inequality, and (32) follows from Lemma B.7. We conclude by applying Lemma A.6 to  $\tilde{\mu}$ , giving us a final bound of  $\exp(4\tilde{C}_G^2 \sigma^2 \|z\|^2 + \|\mathbb{E}(\mu)\| \|z\|)$ .  $\square$

**Lemma B.9.** (Lemma 4.9 from (Nutz, 2021)) Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ , and choose entropic potentials that solve (2) such that  $\int f_{\mu \rightarrow \nu}^\epsilon d\mu \geq 0$  and  $\int g_{\mu \rightarrow \nu}^\epsilon d\nu \geq 0$ . Then:

$$\inf_{y \in \mathbb{R}^d} \left\{ \frac{1}{2} \|x - y\|^2 - g_{\mu \rightarrow \nu}^\epsilon(y) \right\} \leq f_{\mu \rightarrow \nu}^\epsilon(x) \leq \frac{1}{2} \int \|x - y\|^2 d\nu(y),$$

$$\inf_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|x - y\|^2 - f_{\mu \rightarrow \nu}^\epsilon(x) \right\} \leq g_{\mu \rightarrow \nu}^\epsilon(y) \leq \frac{1}{2} \int \|x - y\|^2 d\mu(x),$$

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

The lower bounds in Lemma B.9, while general, can be difficult to apply. We derive more straightforward bounds in the special case when one of the measures is subgaussian. The lemma below can be extracted from Proposition A.1 in (Mena and Niles-Weed, 2019).

**Lemma B.10.** Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $\nu \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , and choose entropic potentials that solve (2) such that  $\int f_{\mu \rightarrow \nu}^\epsilon d\mu \geq 0$  and  $\int g_{\mu \rightarrow \nu}^\epsilon d\nu \geq 0$ . Then there exists a constant  $C_{\epsilon, \sigma, \|\mathbb{E}(\mu)\|, M_2(\mu), \|\mathbb{E}(\nu)\|} \geq 0$ , depending quadratically on  $\sigma$ ,  $\|\mathbb{E}(\mu)\|$ , and linearly on  $M_2(\mu)$ ,  $\frac{1}{\epsilon}$  and  $\|\mathbb{E}(\nu)\|$  such that for all  $x \in \mathbb{R}^d$ :

$$-C_{\epsilon, \sigma, \|\mathbb{E}(\mu)\|, M_2(\mu), \|\mathbb{E}(\nu)\|} (\|x\|^2 + 1) \leq f_{\mu \rightarrow \nu}^\epsilon(x).$$

*Proof.* By (3), we may write  $-f_{\mu \rightarrow \nu}^\epsilon(x) = \epsilon \log \int \exp \left( \frac{1}{\epsilon} (g_{\mu \rightarrow \nu}^\epsilon(y) - \frac{1}{2} \|x - y\|^2) \right) d\nu(y)$  for all  $x \in \mathbb{R}^d$ . We then bound:

$$\begin{aligned} & -f_{\mu \rightarrow \nu}^\epsilon(x) \\ &= \epsilon \log \int \exp \left( \frac{1}{\epsilon} \left( g_{\mu \rightarrow \nu}^\epsilon(y) - \frac{1}{2} \|x - y\|^2 \right) \right) d\nu(y) \\ &\leq \epsilon \log \int \exp \left( \frac{1}{\epsilon} \left( \frac{1}{2} \int \|z - y\|^2 d\mu(z) - \frac{1}{2} \|x - y\|^2 \right) \right) d\nu(y) \end{aligned} \quad (33)$$

$$= \epsilon \log \int \exp \left( \frac{1}{\epsilon} \left( \frac{1}{2} M_2(\mu) - \langle y, \mathbb{E}(\mu) \rangle + \frac{1}{2} \|y\|^2 - \frac{1}{2} \|x\|^2 + \langle x, y \rangle - \frac{1}{2} \|y\|^2 \right) \right) d\nu(y) \quad (34)$$

$$\begin{aligned} &= \epsilon \log \int \exp \left( \frac{1}{\epsilon} \left( \frac{1}{2} M_2(\mu) - \langle y, \mathbb{E}(\mu) - x \rangle - \frac{1}{2} \|x\|^2 \right) \right) d\nu(y) \\ &= \frac{1}{2} M_2(\mu) - \frac{1}{2} \|x\|^2 + \epsilon \log \int \exp \left( \frac{1}{\epsilon} (-\langle y, \mathbb{E}(\mu) - x \rangle) \right) d\nu(y) \\ &\leq \frac{1}{2} M_2(\mu) - \frac{1}{2} \|x\|^2 + \epsilon \log \int \exp \left( \frac{1}{\epsilon} |\langle y, \mathbb{E}(\mu) - x \rangle| \right) d\nu(y) \\ &\leq \frac{1}{2} M_2(\mu) - \frac{1}{2} \|x\|^2 + \epsilon \log \exp \left( \frac{4\tilde{C}_G^2 \sigma^2}{\epsilon^2} \|x - \mathbb{E}(\mu)\|^2 + \frac{1}{\epsilon} \|x - \mathbb{E}(\mu)\| \|\mathbb{E}(\nu)\| \right) \end{aligned} \quad (35)$$

$$\begin{aligned} &= \frac{1}{2} M_2(\mu) - \frac{1}{2} \|x\|^2 + \frac{4\tilde{C}_G^2 \sigma^2}{\epsilon} \|x - \mathbb{E}(\mu)\|^2 + \|x - \mathbb{E}(\mu)\| \|\mathbb{E}(\nu)\| \\ &\leq \frac{1}{2} M_2(\mu) - \frac{1}{2} \|x\|^2 + \frac{8\tilde{C}_G^2 \sigma^2}{\epsilon} \|x\|^2 + \frac{8\tilde{C}_G^2 \sigma^2}{\epsilon} \|\mathbb{E}(\mu)\|^2 + \|\mathbb{E}(\nu)\| (\|x\| + \|\mathbb{E}(\mu)\|) \\ &\leq \frac{1}{2} M_2(\mu) - \frac{1}{2} \|x\|^2 + \frac{8\tilde{C}_G^2 \sigma^2}{\epsilon} \|x\|^2 + \frac{8\tilde{C}_G^2 \sigma^2}{\epsilon} \|\mathbb{E}(\mu)\|^2 + \|\mathbb{E}(\nu)\| (1 + \|\mathbb{E}(\mu)\|) + \|\mathbb{E}(\nu)\| \|x\|^2, \end{aligned} \quad (36)$$

where in (33) we applied Lemma B.9 to  $g_{\mu \rightarrow \nu}^\epsilon$ , in (34) we expanded  $\|\cdot\|^2$ , in (35) we applied Corollary B.8, and in (36) we applied Young's inequality and the triangle inequality. Combining various constants and multiplying both sides by  $-1$ , we obtain our result.  $\square$

We combine Lemma B.9 and Lemma B.10 to give a simple bound on the absolute value of  $f_{\mu \rightarrow \nu}^\epsilon$  with  $\nu \in \mathcal{G}(\mathbb{R}^d)$ :

**Corollary B.11.** *Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ , let  $\nu \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , and choose entropic potentials that solve (2) such that  $\int f_{\mu \rightarrow \nu}^\epsilon d\mu \geq 0$  and  $\int g_{\mu \rightarrow \nu}^\epsilon d\nu \geq 0$ . Then there exists a constant  $\tilde{C}_{\epsilon, \sigma, \|\mathbb{E}(\mu)\|, M_2(\mu), M_2(\nu)} > 0$  only depending quadratically on  $\sigma$ ,  $\|\mathbb{E}(\mu)\|$  and linearly on  $\frac{1}{\epsilon}$ ,  $\|\mathbb{E}(\nu)\|$ ,  $M_2(\mu)$ ,  $M_2(\nu)$ , such that:*

$$|f_{\mu \rightarrow \nu}^\epsilon(x)| \leq \tilde{C}_{\epsilon, \sigma, \|\mathbb{E}(\mu)\|, \|\mathbb{E}(\nu)\|, M_2(\mu), M_2(\nu)} (\|x\|^2 + 1)$$

*Proof.* From Lemma B.9, we have

$$\begin{aligned} f_{\mu \rightarrow \nu}^\epsilon(x) &\leq \frac{1}{2} \int \|x - y\|^2 d\nu(y) \\ &\leq \|x\|^2 + \int \|y\|^2 d\nu(y) \\ &= \|x\|^2 + M_2(\nu), \end{aligned}$$

where we applied Young's inequality. From this and Lemma B.10, we have

$$|f_{\mu \rightarrow \nu}^\epsilon(x)| \leq \max\{C_{\epsilon, \sigma, \|\mathbb{E}(\mu)\|, M_2(\mu), \|\mathbb{E}(\nu)\|} (\|x\|^2 + 1), \|x\|^2 + M_2(\nu)\},$$

where  $C_{\epsilon, \sigma, \|\mathbb{E}(\mu)\|, M_2(\mu)}$  depends quadratically on  $\sigma$ ,  $\|\mathbb{E}(\mu)\|$  and linearly on  $\frac{1}{\epsilon}$ ,  $M_2(\mu)$  and  $\|\mathbb{E}(\nu)\|$ . We conclude by combining constants.  $\square$

We now prove the differentiability of  $OT_2^\epsilon(\mu, \nu)$  as a function of  $\mu$ . A crucial step is establishing the continuity of the entropic potentials along perturbations of  $\mu$ :

**Proposition B.12.** *Let  $\mu, \rho \in \mathcal{P}_2(\mathbb{R}^d)$  and  $\nu \in \mathcal{G}(\mathbb{R}^d)$ , let  $\chi := \rho - \mu$ , and let  $\mu_t := \mu + t\chi$ . For all  $t \in (0, 1]$ , let  $f_{\mu_t \rightarrow \nu}^\epsilon$  be the unique entropic potential such that  $\int f_{\mu_t \rightarrow \nu}^\epsilon d\mu_t = \frac{1}{2}OT_2^\epsilon(\mu_t, \nu)$ . Then  $f_{\mu_t \rightarrow \nu}^\epsilon$  converges pointwise to  $f_{\mu \rightarrow \nu}^\epsilon$  with  $\int f_{\mu \rightarrow \nu}^\epsilon d\mu = \frac{1}{2}OT_2^\epsilon(\mu, \nu)$  as  $t \rightarrow 0^+$ .*

We defer its proof to Subsection B.3.

**Proof of Theorem 2.2:** For convenience, set  $\epsilon = 1$ . Let  $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $\chi$  and  $\mu_t$  be as in the statement of Proposition B.12. We write  $(f_t, g_t) := (f_{\mu_t \rightarrow \nu}^\epsilon, g_{\mu_t \rightarrow \nu}^\epsilon)$  and  $(f, g) := (f_{\mu \rightarrow \nu}^\epsilon, g_{\mu \rightarrow \nu}^\epsilon)$ . Let  $\Delta_t := \frac{1}{t}(OT_2^\epsilon(\mu_t, \nu) - OT_2^\epsilon(\mu, \nu))$ . By suboptimality of  $(f, g)$  for the dual formulation of  $OT_2^\epsilon(\mu_t, \nu)$  and of  $(f_t, g_t)$  for  $OT_2^\epsilon(\mu, \nu)$ , we have the inequalities:

$$OT_2^\epsilon(\mu_t, \nu) \geq \int f d\mu_t + \int g d\nu - \int \int \left( \exp \left( f(x) + g(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\mu_t(x) d\nu(y), \quad (37)$$

$$OT_2^\epsilon(\mu, \nu) \geq \int f_t d\mu + \int g_t d\nu - \int \int \left( \exp \left( f_t(x) + g_t(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\mu(x) d\nu(y). \quad (38)$$

Replacing  $OT_2^\epsilon(\mu_t, \nu)$  with the lower bound in (37), we obtain a lower bound on  $\Delta_t$ :

$$\begin{aligned} \Delta_t &= \frac{1}{t}(OT_2^\epsilon(\mu_t, \nu) - OT_2^\epsilon(\mu, \nu)) \\ &\geq \frac{1}{t} \left( \int f d\mu_t + \int g d\nu - \int \int \left( \exp \left( f(x) + g(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\mu_t(x) d\nu(y) \right. \\ &\quad \left. - \int f_t d\mu - \int g_t d\nu + \int \int \left( \exp \left( f(x) + g(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\mu(x) d\nu(y) \right) \\ &= \int f d\chi - \int \int \left( \exp \left( f(x) + g(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\chi(x) d\nu(y). \end{aligned} \quad (39)$$

Similarly, replacing  $OT_2^\epsilon(\mu, \nu)$  with the lower bound in (38), we obtain:

$$\begin{aligned} \Delta_t &= \frac{1}{t}(OT_2^\epsilon(\mu_t, \nu) - OT_2^\epsilon(\mu, \nu)) \\ &\leq \frac{1}{t} \left( \int f_t d\mu_t + \int g_t d\nu - \int \int \left( \exp \left( f_t(x) + g_t(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\mu_t(x) d\nu(y) \right. \\ &\quad \left. - \int f_t d\mu - \int g_t d\nu + \int \int \left( \exp \left( f_t(x) + g_t(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\mu(x) d\nu(y) \right) \\ &= \int f_t d\chi - \int \int \left( \exp \left( f_t(x) + g_t(y) - \frac{1}{2}\|x - y\|^2 \right) - 1 \right) d\chi(x) d\nu(y). \end{aligned} \quad (40)$$

Subtracting (39) from (40), we obtain:

$$\begin{aligned} &\int (f_t - f) d\chi - \int \int \exp \left( f_t(x) + g_t(y) - \frac{1}{2}\|x - y\|^2 \right) d\chi(x) d\nu(y) \\ &\quad + \int \int \exp \left( f(x) + g(y) - \frac{1}{2}\|x - y\|^2 \right) d\chi(x) d\nu(y), \end{aligned} \quad (41)$$

which by construction is an upper bound on  $\limsup_{t \rightarrow 0^+} \Delta_t - \liminf_{t \rightarrow 0^+} \Delta_t$ , and hence if we show (41) converges to 0 then we will have established existence of  $\lim_{t \rightarrow 0^+} \Delta_t$ . We now show that the first term converges to zero. For any  $x \in \mathbb{R}^d$ , we bound:

$$|f_t(x) - f(x)| \leq |f_t(x)| + |f(x)| \leq C_t^*(\|x\|^2 + 1),$$

where we applied the triangle inequality followed by Corollary B.11 and where  $C_t^*$  is a constant quadratically depending on  $\sigma, \|\mathbb{E}(\mu_t)\|, \|\mathbb{E}(\mu)\|$  and linearly depending on  $\frac{1}{\epsilon}, \|\mathbb{E}(\nu)\|, M_2(\mu_t), M_2(\mu), M_2(\nu)$ . As we have



$\|\mathbb{E}(\mu_t)\| \leq \|\mathbb{E}(\mu)\| + \|\mathbb{E}(\rho)\|$  and  $M_2(\mu_t) \leq M_2(\mu) + M_2(\rho)$ , we can further upper bound  $\sup_{t \in [0,1]} C_t^* \leq C^*$ , where  $C^*$  depends only on  $\frac{1}{\epsilon}, \sigma, \|\mathbb{E}(\mu)\|, \|\mathbb{E}(\nu)\|, \|\mathbb{E}(\rho)\|, M_2(\mu), M_2(\rho)$ , and  $M_2(\nu)$ , and is independent of  $t$  and  $x$ . Since  $\mu, \rho \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $C^*(\|x\|^2 + 1)$  is  $\chi$ -integrable as a function of  $x$ . Hence by the Dominated Convergence Theorem and Proposition B.12, we have that  $\lim_{t \rightarrow 0^+} \int |f_t - f| d\chi = \int \lim_{t \rightarrow 0^+} |f_t - f| d\chi = 0$ .

Next, we claim that

$$\int \int \exp \left( f_t(x) + g_t(y) - \frac{1}{2} \|x - y\|^2 \right) d\chi(x) d\nu(y) = 0$$

for all  $t \in (0, 1]$ . Applying the Fubini-Tonelli Theorem (Proposition 5.2.1 in (Cohn, 2013)), which applies since the integrand is nonnegative and measurable, we have, for all  $t \in (0, 1]$ :

$$\begin{aligned} & \int \int \exp \left( f_t(x) + g_t(y) - \frac{1}{2} \|x - y\|^2 \right) d\chi(x) d\nu(y) \\ &= \int \exp(f_t(x)) \int \exp \left( g_t(y) - \frac{1}{2} \|x - y\|^2 \right) d\nu(y) d\chi(x) \\ &= \int \exp(f_t(x)) \exp(-f_t(x)) d\chi(x) \\ &= \int 1 d\chi(x) = 0, \end{aligned} \tag{42}$$

where we applied (3) and (4) and the fact that  $\chi = \rho - \mu$  has total mass zero. By the same reasoning,

$$\begin{aligned} & \int \int \exp \left( f(x) + g(y) - \frac{1}{2} \|x - y\|^2 \right) d\chi(x) d\nu(y) \\ &= \int \exp(f(x)) \int \exp \left( g(y) - \frac{1}{2} \|x - y\|^2 \right) d\nu(y) \\ &= \int \exp(f(x)) \exp(-f(x)) d\chi(x) \\ &= \int 1 d\chi(x) = 0. \end{aligned} \tag{43}$$

By (42) and (43), we have established that (41) equals  $\int (f_t - f) d\chi$  for all  $t \in (0, 1]$ , and hence converges to zero as  $t \rightarrow 0^+$  as desired. This establishes that  $\lim_{t \rightarrow 0^+} \Delta_t$  exists. and we compute this limit using (39) and (4) which reduces to  $\int f d\chi$ , by (42), and we conclude that  $\lim_{t \rightarrow 0^+} \Delta_t = \int f d\chi$ . Thus we have established that  $f$  is a derivative for  $OT_2^\epsilon(\mu, \nu)$  as a function of  $\mu$ .  $\square$

**Remark B.13.** *The above proof hinges on the fact that the Sinkhorn relation (3) holds for  $f_{\mu \rightarrow \nu}^\epsilon$  for all  $x \in \mathbb{R}^d$  (and the same for  $f_{\mu_t \rightarrow \nu}^\epsilon$ ). In particular, this is crucial in establishing (43). Without (3), one would need to argue that*

$$\begin{aligned} & \int \int \exp \left( f_t(x) + g_t(y) - \frac{1}{2} \|x - y\|^2 \right) d\chi(x) d\nu(y) \\ & \rightarrow \int \int \exp \left( f(x) + g(y) - \frac{1}{2} \|x - y\|^2 \right) d\chi(x) d\nu(y) \end{aligned}$$

as  $t \rightarrow 0^+$  directly, which is challenging as  $\exp(f_t(x) - \frac{1}{2} \|x - y\|^2)$  is not a priori uniformly bounded over  $t > 0$  by a  $\chi$ -integrable function. Hence it is not obvious whether a naive application of the Dominated Convergence Theorem will allow for the verification of this limit.

### B.3 Proof of Proposition B.12

To establish Proposition B.12, we will need to show that the sequence of potentials converges in probability:

**Proposition B.14.** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ , let  $\chi := \rho - \mu$  and let  $\mu_t = \mu + t\chi$ . For all  $t \in [0, 1]$ , let  $(f_{\mu_t \rightarrow \nu}^\epsilon, g_{\mu_t \rightarrow \nu}^\epsilon)$  be the unique entropic potentials such that  $\int f_{\mu_t \rightarrow \nu}^\epsilon d\mu_t = \int g_{\mu_t \rightarrow \nu}^\epsilon d\nu = \frac{1}{2} OT_2^\epsilon(\mu_t, \nu)$ . Then:*

- $f_{\mu_t \rightarrow \nu}^\epsilon$  converges in  $\mu$ -probability as  $t \rightarrow 0^+$  to the entropic potential  $f_{\mu \rightarrow \nu}^\epsilon$  with  $\int f_{\mu \rightarrow \nu}^\epsilon d\mu = \frac{1}{2} OT_2^\epsilon(\mu, \nu)$ ;

- $g_{\mu_t \rightarrow \nu}^\epsilon$  converges in  $\nu$ -probability as  $t \rightarrow 0^+$  to the entropic potential  $g_{\mu \rightarrow \nu}^\epsilon$  with  $\int g_{\mu \rightarrow \nu}^\epsilon d\nu = \frac{1}{2} OT_2^\epsilon(\mu, \nu)$ .

**Remark B.15.** We note that the normalization  $\int f_{\mu_t \rightarrow \nu}^\epsilon d\mu_t = \int g_{\mu_t \rightarrow \nu}^\epsilon d\nu = \frac{1}{2} OT_2^\epsilon(\mu_t, \nu) \geq 0$  is convenient as it allows us to apply Lemma B.9, and also varies smoothly as a function of  $t$  (see e.g., Lemma B.18).

We postpone the proof of Proposition B.14 to Subsection B.4, and continue with the proof of Proposition B.12.

**Proof of Proposition B.12:** For convenience, we will assume that  $\epsilon = 1$ . Let  $f_t := f_{\mu_t \rightarrow \nu}^\epsilon$  and  $g_t := g_{\mu_t \rightarrow \nu}^\epsilon$  for any  $t \in (0, 1]$  and let  $f := f_{\mu \rightarrow \nu}^\epsilon$  and  $g := g_{\mu \rightarrow \nu}^\epsilon$  be the unique entropic potentials such that  $\frac{1}{2} OT_2^\epsilon(\mu, \nu) = \int g d\nu$ . Let  $T := \{t_i\}_{i \in \mathbb{N}} \subset [0, 1)$ , with  $t_i \rightarrow 0$  as  $i \rightarrow \infty$ . By Proposition B.14, the sequence  $\{g_{t_i}\}_{t_i \in T}$  converges in  $\nu$ -probability to  $g$ . Since  $\{g_{t_i}\}_{t_i \in T}$  converges in probability, there is a subsequence  $T' \subset T$  such that  $\{g_{t'_i}\}_{t'_i \in T'}$  converges  $\nu$ -almost-surely to  $g$ . We now show that this, together with subgaussian assumption on  $\nu$ , is enough to prove pointwise convergence everywhere of  $\{f_{t'_i}\}_{t'_i \in T'}$  to  $f$ . Indeed, by (3), for any  $x \in \mathbb{R}^d$  we may write:

$$\lim_{i \rightarrow \infty} \exp(-f_{t'_i}(x)) = \lim_{i \rightarrow \infty} \int \exp\left(-\frac{1}{2}\|x - y\|^2 + g_{t'_i}(y)\right) d\nu(y).$$

By Lemma B.9, we may bound uniformly:

$$\begin{aligned} & \exp\left(-\frac{1}{2}\|x - y\|^2 + g_{t'_i}(y)\right) \\ & \leq \exp\left(-\frac{1}{2}\|x\|^2 + \langle x, y \rangle - \frac{1}{2}\|y\|^2 + \int \frac{1}{2}\|z - y\|^2 d\mu_{t'_i}(z)\right) \\ & = \exp\left(-\frac{1}{2}\|x\|^2 + \langle x, y \rangle - \frac{1}{2}\|y\|^2 + \int \left(\frac{1}{2}\|z\|^2 - \langle y, z \rangle + \frac{1}{2}\|y\|^2\right) d\mu_{t'_i}(z)\right) \\ & = \exp\left(-\frac{1}{2}\|x\|^2 + \frac{1}{2}M_2(\mu_{t'_i}) + \langle x - \mathbb{E}(\mu_{t'_i}), y \rangle\right) \\ & \leq \exp\left(-\frac{1}{2}\|x\|^2 + \frac{1}{2}M_2((1 - t'_i)\mu + t'_i\rho) + |\langle x, y \rangle| + |\langle \mathbb{E}((1 - t'_i)\mu + t'_i\rho), y \rangle|\right) \\ & \leq \exp\left(-\frac{1}{2}\|x\|^2 + \frac{1}{2}M_2(\mu) + \frac{1}{2}M_2(\rho) + |\langle x, y \rangle| + |\langle \mathbb{E}(\mu), y \rangle| + |\langle \mathbb{E}(\rho), y \rangle|\right) \end{aligned} \quad (44)$$

where we applied the triangle inequality twice to obtain  $|\langle x - \mathbb{E}(\mu_t), y \rangle| \leq |\langle x, y \rangle| + |\langle \mathbb{E}(\mu_t), y \rangle|$  and  $|\langle \mathbb{E}(\mu_t), y \rangle| \leq |\langle \mathbb{E}(\mu), y \rangle| + |\langle \mathbb{E}(\rho), y \rangle|$ , and also used that  $0 < t'_i \leq 1$ . We now show that the function in (44) is  $\nu$ -integrable in  $y$ . By two applications of the Cauchy-Schwarz inequality, we bound:

$$\begin{aligned} & \int \exp(|\langle x, y \rangle| + |\langle \mathbb{E}(\mu), y \rangle| + |\langle \mathbb{E}(\rho), y \rangle|) d\nu(y) \\ & = \int \exp(|\langle x, y \rangle|) \exp(|\langle \mathbb{E}(\mu), y \rangle|) \exp(|\langle \mathbb{E}(\rho), y \rangle|) d\nu(y) \\ & \leq \left(\int \exp(2|\langle x, y \rangle|) d\nu(y)\right)^{1/2} \left(\int \exp(4|\langle \mathbb{E}(\mu), y \rangle|) d\nu(y) \int \exp(4|\langle \mathbb{E}(\rho), y \rangle|) d\nu(y)\right)^{1/4} \end{aligned}$$

and by Corollary B.8 this last expression is finite. Hence (44) is  $\nu$ -integrable, and we conclude that for each  $x \in \mathbb{R}^d$ ,  $\exp(-\frac{1}{2}\|x - y\|^2 + g_{t'_i}(y))$  may be uniformly (in  $t'_i$ ) upper bounded by a  $\nu$ -integrable function of  $y$ . By the Dominated Convergence Theorem, we have for all  $x \in \mathbb{R}^d$ :

$$\begin{aligned} \lim_{i \rightarrow \infty} \exp(-f_{t'_i}(x)) & = \int \lim_{i \rightarrow \infty} \exp\left(-\frac{1}{2}\|x - y\|^2 + g_{t'_i}(y)\right) d\nu(y) \\ & = \int \exp\left(-\frac{1}{2}\|x - y\|^2 + g(y)\right) d\nu(y) \\ & = \exp(-f(x)) \end{aligned}$$

by  $\nu$ -almost sure convergence of  $g_{t'_i}$  to  $g$ , which, along with continuity of  $\log$ , establishes pointwise everywhere convergence of  $\{f_{t'_i}\}_{t'_i \in T'}$ .

We claim this is enough to conclude pointwise everywhere convergence of  $\{f_{t_i}\}_{t_i \in T}$  to  $f$ . Indeed, suppose  $\{f_{t_i}\}_{t_i \in T}$  does not converge pointwise everywhere to  $f$ . Then there exists an  $x_0 \in \mathbb{R}^d$ , a  $\delta > 0$ , and a subsequence  $\tilde{T} \subset T$  with  $\tilde{t}_i \rightarrow 0$  as  $i \rightarrow \infty$ , such that  $\|f(x_0) - f_{\tilde{t}_i}(x_0)\| > \delta$  for all  $\tilde{t}_i \in \tilde{T}$ . But  $\{g_{\tilde{t}_i}\}_{\tilde{t}_i \in \tilde{T}}$  converges in  $\nu$ -probability to  $g$ , and thus the sequence  $\tilde{T}$  contains a subsequence  $\tilde{T}' \subset \tilde{T}$  along which  $\{g_{\tilde{t}_i'}\}_{\tilde{t}_i' \in \tilde{T}'}$  converges  $\nu$ -almost surely to  $g$ . Hence we may apply the above argument to conclude that  $f_{\tilde{t}_i'}$  converges pointwise to  $f$ , contradicting our assumption. We conclude that  $\{f_{t_i}\}_{t_i \in T}$  converges pointwise to  $f$  as  $i \rightarrow \infty$ .  $\square$

#### B.4 Proof of Proposition B.14

To prove Proposition B.14, we will first need to verify that our sequence  $\mu_t$  contains a subsequence satisfying the conditions in the lemma below:

**Lemma B.16.** (*Corollary 2.4 in (Nutz and Wiesel, 2023)*) *Let  $T := \{t_i\}_{i \in \mathbb{N}} \subset [0, 1]$  be any sequence such that  $t_i \rightarrow 0^+$  as  $i \rightarrow \infty$ . Let  $\{\mu_{t_i}\}_{t_i \in T} \subset \mathcal{P}_2(\mathbb{R}^d)$  be a sequence of probability measures such that:*

1.  $\mu_{t_i}$  converges weakly to  $\mu$  as  $i \rightarrow \infty$ ;
2.  $\mu \ll \mu_{t_i}$  for all  $t_i \in [0, 1]$ ;
3.  $OT_2^\epsilon(\mu_{t_i}, \nu) < \infty$  for all  $t_i \in [0, 1]$ ;
4.  $\sup_i \int \max\{f_{\mu_{t_i} \rightarrow \nu}^\epsilon, 0\} d\mu_{t_i} < \infty$ ,  $\sup_i \int \max\{g_{\mu_{t_i} \rightarrow \nu}^\epsilon, 0\} d\nu < \infty$ ;
5.  $\lim_{i \rightarrow \infty} \int \arctan(f_{\mu_{t_i} \rightarrow \nu}^\epsilon) d\mu_{t_i} = L$  for some  $L \in [-\pi/2, \pi/2]$ ;
6.  $\limsup_{C \rightarrow \infty} \limsup_{i \rightarrow \infty} \int \mathbf{1}_{\frac{d\mu}{d\mu_{t_i}}(x) \geq C} d\mu(x) = 0$ .

Then:

- $f_{\mu_{t_i} \rightarrow \nu}^\epsilon$  converges in  $\mu$ -probability to  $f_{\mu \rightarrow \nu}^{\epsilon, L}$ , the unique entropic potential with  $\int \arctan(f_{\mu \rightarrow \nu}^{\epsilon, L}) d\mu = L$ ;
- $g_{\mu_{t_i} \rightarrow \nu}^\epsilon$  converges in  $\nu$ -probability to  $g_{\mu \rightarrow \nu}^{\epsilon, L}$ , such that  $g_{\mu \rightarrow \nu}^{\epsilon, L}$  is the unique entropic potential paired with  $f_{\mu \rightarrow \nu}^{\epsilon, L}$ .

We remark that, for any  $f \in L^1(\mu)$ ,  $\int \arctan(f + a) d\mu$  is a strictly increasing function of  $a \in \mathbb{R}$ . Hence if there are two entropic potentials  $f_{\mu \rightarrow \nu}^\epsilon, \tilde{f}_{\mu \rightarrow \nu}^\epsilon$  such that:

- $f_{\mu \rightarrow \nu}^\epsilon + c = \tilde{f}_{\mu \rightarrow \nu}^\epsilon$  for some  $c \in \mathbb{R}$ ;
- $\int \arctan(f_{\mu \rightarrow \nu}^\epsilon) d\mu = \int \arctan(\tilde{f}_{\mu \rightarrow \nu}^\epsilon) d\mu$ ;

then  $c = 0$ , and consequently  $f_{\mu \rightarrow \nu}^\epsilon = \tilde{f}_{\mu \rightarrow \nu}^\epsilon$ . This justifies the uniqueness claimed for  $f_{\mu \rightarrow \nu}^{\epsilon, L}$  (and hence for  $g_{\mu \rightarrow \nu}^{\epsilon, L}$ ) in Lemma B.16.

To verify item 6 we will apply the following Lemma:

**Lemma B.17.** (*Lemma 2.5 in (Nutz and Wiesel, 2023)*) *Let  $T := \{t_i\}_{i \in \mathbb{N}} \subset [0, 1]$  be any sequence such that  $t_i \rightarrow 0^+$  as  $i \rightarrow \infty$ , and let  $\{\mu_{t_i}\}_{t_i \in T} \subset \mathcal{P}_2(\mathbb{R}^d)$  be a sequence of probability measures such that  $\mu \ll \mu_{t_i}$  for all  $t_i \in T$ . Suppose that  $\mu_{t_i} \rightarrow \mu$  in total variation as  $i \rightarrow \infty$ . Then:*

$$\limsup_{C \rightarrow \infty} \limsup_{i \rightarrow \infty} \int \mathbf{1}_{\frac{d\mu}{d\mu_{t_i}}(x) \geq C} d\mu(x) = 0.$$

After verifying the conditions in Lemma B.16, we will need to establish that the limiting potentials satisfy  $\int f_{\mu \rightarrow \nu}^{\epsilon, L} d\mu = \frac{1}{2} OT_2^\epsilon(\mu, \nu)$ ,  $\int g_{\mu \rightarrow \nu}^{\epsilon, L} d\mu = \frac{1}{2} OT_2^\epsilon(\mu, \nu)$ . The result immediately follows from the Lemma below under the stated assumption in Proposition B.14 that  $\int f_t d\mu_t = \int g_t d\nu = \frac{1}{2} OT_2^\epsilon(\mu_t, \nu)$ .

**Lemma B.18.** (Theorem 2.1 (iii) in (Nutz and Wiesel, 2023)) Let  $T := \{t_i\}_{i \in \mathbb{N}} \subset [0, 1]$  be any sequence such that  $t_i \rightarrow 0^+$  as  $i \rightarrow \infty$ , and let  $\{\mu_{t_i}\}_{t_i \in T} \subset \mathcal{P}_2(\mathbb{R}^d)$  be a sequence of probability measures, converging weakly to  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ , and let  $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ . Let  $(f_{t_i}, g_{t_i}) := (f_{\mu_{t_i} \rightarrow \nu}^\epsilon, g_{\mu_{t_i} \rightarrow \nu}^\epsilon)$  be solutions to (2), and suppose that:

1.  $\sup_i \{\int f_{t_i}^\epsilon d\mu_{t_i}, \int g_{t_i}^\epsilon d\mu_{t_i}\} < \infty$ .
2.  $\{(f_{t_i}, g_{t_i})\}_{t_i \in T}$  are uniformly integrable with respect to  $(\mu_{t_i}, \nu)$ :

$$\lim_{C \rightarrow \infty} \sup_{i \in \mathbb{N}} \int f_{t_i}(x) \mathbf{1}_{f_{t_i} > C}(x) d\mu_{t_i}(x) = 0,$$

$$\lim_{C \rightarrow \infty} \sup_{i \in \mathbb{N}} \int g_{t_i}(y) \mathbf{1}_{g_{t_i} > C}(y) d\nu(y) = 0.$$

3.  $\lim_{i \rightarrow \infty} \int \arctan(f_{t_i}) d\mu_{t_i} = \int \arctan(f) d\mu$ .

Then:

- $OT_2^\epsilon(\mu_{t_i}, \nu) \rightarrow OT_2^\epsilon(\mu, \nu)$  as  $i \rightarrow \infty$ .
- $\int f_{t_i} d\mu_{t_i} \rightarrow \int f d\mu$  as  $i \rightarrow \infty$ .
- $\int g_{t_i} d\nu \rightarrow \int g d\nu$  as  $i \rightarrow \infty$ .

To verify item 2 in Lemma B.18, we will require the following result:

**Lemma B.19.** Let  $\mu, \rho, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ . We let  $\chi = \rho - \mu$ , and define a sequence of probability measures  $\mu_{t_i} := \mu + t_i \chi$  where  $\{t_i\}_{i=1}^\infty =: T \subset [0, 1]$  is a sequence of nonnegative real numbers such that  $t_i \rightarrow 0^+$  as  $i \rightarrow \infty$ . Then the sequence of potentials  $\{(f_{\mu_{t_i} \rightarrow \nu}^\epsilon, g_{\mu_{t_i} \rightarrow \nu}^\epsilon)\}_{i=1}^\infty$  satisfies:

$$\lim_{C \rightarrow \infty} \sup_{i \in \mathbb{N}} \int f_{\mu_{t_i} \rightarrow \nu}^\epsilon(x) \mathbf{1}_{f_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(x) d\mu_{t_i}(x) = 0, \quad (45)$$

$$\lim_{C \rightarrow \infty} \sup_{i \in \mathbb{N}} \int g_{\mu_{t_i} \rightarrow \nu}^\epsilon(y) \mathbf{1}_{g_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(y) d\nu(y) = 0. \quad (46)$$

*Proof.* Without loss of generality, let  $\epsilon = 1$ . We bound:

$$\begin{aligned} & \int f_{\mu_{t_i} \rightarrow \nu}^\epsilon(x) \mathbf{1}_{f_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(x) d\mu_{t_i}(x) \\ & \leq \int \left( \int \frac{1}{2} \|x - y\|^2 d\nu(y) \right) \mathbf{1}_{f_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(x) d\mu_{t_i}(x) \\ & \leq \int \left( \int \|x\|^2 + \|y\|^2 d\nu(y) \right) \mathbf{1}_{f_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(x) d\mu_{t_i}(x) \\ & = \int (\|x\|^2 + M_2(\nu)) \mathbf{1}_{f_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(x) d\mu_{t_i}(x) \end{aligned}$$

where we applied Lemma B.9, the triangle inequality, and Young's inequality. We also note that:

$$\begin{aligned} \mathbf{1}_{f_{\mu_{t_i} \rightarrow \nu}^\epsilon > C}(x) & \leq \mathbf{1}_{\frac{1}{2} \int \|x - y\|^2 d\nu(y) > C}(x) \\ & \leq \mathbf{1}_{\|\cdot\|^2 + M_2(\nu) > C}(x) \\ & = \mathbf{1}_{\|\cdot\|^2 > C - M_2(\nu)}(x) \end{aligned}$$

where we applied Lemma B.9, the triangle inequality, and Young's inequality. As  $2\|x\|^2 \geq \|x\|^2 + M_2(\nu)$  for all  $\|x\|^2 \geq M_2(\nu)$  we have:

$$\begin{aligned}
 & \lim_{C \rightarrow \infty} \sup_{i \in \mathbb{N}} \int (\|x\|^2 + M_2(\nu)) \mathbf{1}_{\|x\|^2 > C - M_2(\nu)}(x) d\mu_{t_i}(x) \\
 & \leq \lim_{D \rightarrow \infty} \sup_{i \in \mathbb{N}} \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\mu_{t_i}(x) \\
 & = 2 \lim_{D \rightarrow \infty} \sup_{i \in \mathbb{N}} \left( (1 - t_i) \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\mu(x) + t_i \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\rho(x) \right) \\
 & \leq 2 \lim_{D \rightarrow \infty} \left( \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\mu(x) + \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\rho(x) \right) \tag{47}
 \end{aligned}$$

Since  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ ,

$$M_2(\mu) = \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\mu + \int \|x\|^2 \mathbf{1}_{\|x\|^2 \leq D}(x) d\mu < \infty,$$

and by the monotone convergence theorem  $\lim_{D \rightarrow \infty} \int \|x\|^2 \mathbf{1}_{\|x\|^2 \leq D}(x) d\mu = M_2(\mu)$ , and hence  $\limsup_{D \rightarrow \infty} \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\mu = 0$ . Similarly, since  $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $\lim_{D \rightarrow \infty} \int \|x\|^2 \mathbf{1}_{\|x\|^2 > D}(x) d\rho = 0$ , and (47) is zero.

The proof for  $g_{\mu_{t_i} \rightarrow \nu}^\epsilon$  follows similarly.  $\square$

**Proof of Proposition B.14:** For convenience, we will assume that  $\epsilon = 1$ . Let  $T := \{t_i\}_{i \in \mathbb{N}} \subset [0, 1)$  be any sequence such that  $t_i \rightarrow 0^+$  as  $i \rightarrow \infty$ . We begin by verifying that the conditions in Lemma B.16 hold along a subsequence  $T' \subset T$ . Items 1 and 2 clearly hold for any subsequence of  $T$  by definition of  $\mu_{t_i} = (1 - t_i)\mu + t_i\rho$ .

We may establish item 4 for any subsequence of  $T$  via Lemma B.9 and Young's inequality:

$$\begin{aligned}
 \int \max\{f_{\mu_{t_i} \rightarrow \nu}^\epsilon, 0\} d\mu_{t_i}(x) & \leq \int \int \frac{1}{2} \|x - y\|^2 d\nu(y) d\mu_{t_i}(x) \\
 & \leq \int (\|x\|^2 + M_2(\nu)) d\mu_{t_i}(x) \\
 & = M_2(\mu_{t_i}) + M_2(\nu) \\
 & \leq M_2(\mu) + M_2(\rho) + M_2(\nu)
 \end{aligned}$$

and hence  $\sup_i \int \max\{f_{\mu_{t_i} \rightarrow \nu}^\epsilon(x), 0\} d\mu_{t_i}(x) < \infty$ , and the same holds for  $\sup_i \int \max\{g_{\mu_{t_i} \rightarrow \nu}(y), 0\} d\nu(y)$ . This also verifies item 3, since  $\int f_{\mu_{t_i} \rightarrow \nu}^\epsilon d\mu_{t_i} + \int g_{\mu_{t_i} \rightarrow \nu}^\epsilon d\nu = OT_2^\epsilon(\mu_{t_i}, \nu)$ .

To see item 5, since  $\arctan$  is bounded,  $\int \arctan(f_{\mu_{t_i} \rightarrow \nu}^\epsilon) d\mu_{t_i}$  is uniformly bounded, and hence there exists a convergent subsequence  $T'$  with  $\lim_{i \rightarrow \infty} \int \arctan(f_{\mu_{t'_i} \rightarrow \nu}^\epsilon) d\mu_{t'_i} = L$  for some  $L \in [-\pi/2, \pi/2]$ . To verify item 6, the sequence  $\mu_{t_i}$  clearly converges to  $\mu$  in total variation<sup>2</sup>, and hence we may apply Lemma B.17 to conclude that:

$$\limsup_{C \rightarrow \infty} \limsup_{i \rightarrow \infty} \int \mathbf{1}_{\frac{d\mu}{d\mu_{t_i}}(x) \geq C} d\mu(x) = 0.$$

We have thus established that Lemma B.16 holds for the subsequence indexed by  $T'$ , we conclude that  $\{f_{\mu_{t'_i} \rightarrow \nu}^\epsilon\}_{t'_i \in T'}$  converges in  $\mu$ -probability to  $f_{\mu \rightarrow \nu}^{\epsilon, L}$ , and similarly  $\{g_{\mu_{t'_i} \rightarrow \nu}^\epsilon\}_{t'_i \in T'}$  converges in  $\nu$ -probability to  $g_{\mu \rightarrow \nu}^{\epsilon, L}$ .

Next, we verify the conditions in Lemma B.18 for the subsequence  $(f_{\mu_{t'_i} \rightarrow \nu}^\epsilon, g_{\mu_{t'_i} \rightarrow \nu}^\epsilon)$ . Indeed, item 1 has already been established in the verification of the conditions of B.16 above, and by definition of the subsequence, item 3 holds with limit potentials  $(f_{\mu \rightarrow \nu}^{\epsilon, L}, g_{\mu \rightarrow \nu}^{\epsilon, L})$ . By Lemma B.19, this subsequence also satisfies (45) and (46), establishing item 2. Hence we have verified Lemma B.18 for our subsequence, and can conclude that  $\int f_{\mu_{t'_i} \rightarrow \nu}^\epsilon d\mu_{t'_i} \rightarrow \int f_{\mu \rightarrow \nu}^{\epsilon, L} d\mu$

<sup>2</sup>The *total variation* distance between  $\mu$  and  $\nu$  is  $d_{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$  where the supremum is over all measurable sets  $A$ .

and  $\int g_{\mu_{t'_i} \rightarrow \nu}^\epsilon d\nu \rightarrow \int g_{\mu \rightarrow \nu}^{\epsilon, L} d\nu$ . By assumption, we have that  $\int f_{\mu_{t'_i} \rightarrow \nu}^\epsilon d\mu_{t'_i} = \int g_{\mu_{t'_i} \rightarrow \nu}^\epsilon d\nu = \frac{1}{2} OT_2^\epsilon(\mu_{t'_i}, \nu)$ , and again by Lemma B.18 we have  $\lim_{i \rightarrow \infty} OT_2^\epsilon(\mu_{t'_i}, \nu) = OT_2^\epsilon(\mu, \nu)$ . We conclude that

$$\int f_{\mu \rightarrow \nu}^{\epsilon, L} d\mu = \int g_{\mu \rightarrow \nu}^{\epsilon, L} d\nu = \frac{1}{2} OT_2^\epsilon(\mu, \nu). \quad (48)$$

Hence we have identified  $(f_{\mu \rightarrow \nu}^{\epsilon, L}, g_{\mu \rightarrow \nu}^{\epsilon, L}) = (f_{\mu \rightarrow \nu}^\epsilon, g_{\mu \rightarrow \nu}^\epsilon)$ , the unique pair of potentials satisfying (48).

Since our sequence  $T$  was arbitrary, we can conclude as follows. Let  $T = \{t_i\}_{i=1}^\infty$  be a sequence with  $t_i \rightarrow 0^+$  as  $i \rightarrow \infty$ , and suppose that  $f_{\mu_{t_i} \rightarrow \nu}$  does not converge in  $\mu$ -probability to  $f_{\mu \rightarrow \nu}^\epsilon$ . Since convergence in probability is metrizable, there exists a metric  $d$  and a  $\kappa > 0$  such that there exists a subsequence  $S$  of  $T$  with the property that  $d(f_{\mu_{s_i} \rightarrow \nu}^\epsilon, f_{\mu \rightarrow \nu}^\epsilon) > \kappa$  for all  $s_i \in S$ . The above proof also applies to  $S$ , and hence we may find a subsequence  $S' \subset S$  so that  $f_{\mu_{s'_i} \rightarrow \nu}$  converges to  $f_{\mu \rightarrow \nu}^\epsilon$  in  $\mu$ -probability, and hence  $d(f_{\mu_{s'_i} \rightarrow \nu}^\epsilon, f_{\mu \rightarrow \nu}^\epsilon) \rightarrow 0$  as  $i \rightarrow \infty$  which is a contradiction. Thus we may conclude that  $f_{\mu_{t_i} \rightarrow \nu}^\epsilon$  converges to  $f_{\mu \rightarrow \nu}^\epsilon$  in  $\mu$ -probability. The same argument allow us to conclude for  $g_{\mu_{t_i} \rightarrow \nu}^\epsilon$ .  $\square$

### B.5 Proof of Proposition 2.3

For convenience, we set  $\epsilon = 1$ . Recall that we may write

$$f_{\mu \rightarrow \nu}^\epsilon(x) = -\log \left( \int \exp \left( -\frac{1}{2} \|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y) \right) d\nu(y) \right).$$

We define  $h(x, y) := \exp \left( -\frac{1}{2} \|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y) \right)$ , and will show that  $\nabla_x \int h(x, y) d\nu(y)$  exists. Define

$$h_{t,i}(x, y) := \frac{1}{t} \left( \exp \left( -\frac{1}{2} \|x + te_i - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y) \right) - \exp \left( -\frac{1}{2} \|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y) \right) \right)$$

where  $e_i$  is a standard basis vector in  $\mathbb{R}^d$ . We then define  $h_t = (h_{t,1}, \dots, h_{t,d})$ , and observe that:

$$\begin{aligned} \lim_{t \rightarrow 0^+} h_t(x, y) &= \nabla_x h(x, y) \\ &= -(x - y) \exp \left( -\frac{1}{2} \|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y) \right). \end{aligned}$$

At each  $x$ , we apply the mean value theorem to write  $h_t(x, y)$  as  $\|\nabla_x h(\theta_x^t, y)\|$  for some  $\theta_x^t \in B_t(x) := \{y \in \mathbb{R}^d \mid \|x - y\| < t\}$ . We upper bound:

$$\begin{aligned} |h_t(x, y)| &= \|\nabla_x h(\theta_x^t, y)\| \\ &= \left\| (\theta_x^t - y) \exp \left( -\frac{1}{2} \|\theta_x^t - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y) \right) \right\| \\ &\leq \|\theta_x^t - y\| \left| \exp \left( -\frac{1}{2} \|\theta_x^t - y\|^2 + \int \frac{1}{2} \|z - y\|^2 d\mu(z) \right) \right| \\ &= \|\theta_x^t - y\| \left| \exp \left( -\frac{1}{2} \|\theta_x^t\|^2 + \frac{1}{2} M_2(\mu) + \langle \theta_x^t - \mathbb{E}(\mu), y \rangle \right) \right| \end{aligned} \quad (49)$$

where in (49) we applied the bound on  $g_{\mu \rightarrow \nu}^\epsilon(y)$  from Lemma B.9. Since  $\theta_x^t \in B_t(x)$ , we may assume WLOG that  $\|\theta_x^t - x\| \leq 1$  for all small enough  $t \leq 1$ . Hence we may uniformly (in  $y$  and  $t$ ) bound:

$$|h_t(x, y)| \leq \sup_{z \in \bar{B}_1(x)} \|z - y\| \exp \left( -\frac{1}{2} \|z\|^2 + \frac{1}{2} M_2(\mu) + \langle z, y \rangle - \langle \mathbb{E}(\mu), y \rangle \right).$$

We may then bound:

$$\begin{aligned} & \int |h_t(x, y)| d\nu(y) \\ & \leq \int \sup_{z \in \overline{B_1}(x)} \|z - y\| \exp\left(-\frac{1}{2}\|z\|^2 + \frac{1}{2}M_2(\mu) + \langle z - \mathbb{E}(\mu), y \rangle\right) d\nu(y) \end{aligned} \quad (50)$$

$$\begin{aligned} & \leq \left( \int \sup_{z \in \overline{B_1}(x)} \|z - y\|^2 \exp(-\|z\|^2 + M_2(\mu)) d\nu(y) \right)^{1/2} \left( \int \sup_{z \in \overline{B_1}(x)} \exp(2\langle z - \mathbb{E}(\mu), y \rangle) d\nu(y) \right)^{1/2} \\ & < \infty, \end{aligned} \quad (51)$$

where we applied the Cauchy-Schwarz inequality in (50) and Corollary B.8 to conclude that (51) is finite. Thus for each  $x \in \mathbb{R}^d$  and all  $t \in [0, 1)$ ,  $h_t(x, y)$  is uniformly bounded by an integrable function of  $y$  in a neighborhood of  $x$ , and hence we may apply the Dominated Convergence Theorem to conclude that  $\lim_{t \rightarrow 0^+} \int h_t(x, y) d\nu(y) = \int \lim_{t \rightarrow 0^+} h_t(x, y) d\nu(y)$ . Equivalently, we have shown:

$$\begin{aligned} \nabla_x \int \exp\left(-\frac{1}{2}\|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y)\right) d\nu(y) &= \int \nabla_x \exp\left(-\frac{1}{2}\|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y)\right) d\nu(y) \\ &= \int (y - x) \exp\left(-\frac{1}{2}\|x - y\|^2 + g_{\mu \rightarrow \nu}^\epsilon(y)\right) d\nu(y) \end{aligned}$$

for all  $x \in \mathbb{R}^d$ . Since  $f_{\mu \rightarrow \nu}^\epsilon(x) = -\log\left(\int h(x, y) d\nu(y)\right)$ ,  $f_{\mu \rightarrow \nu}^\epsilon(x)$  is a composition of two differentiable functions and is hence differentiable by the chain rule. One then easily verifies (8). The proof that the higher derivatives of  $f_{\mu \rightarrow \nu}^\epsilon$  exist follows similarly.  $\square$

## B.6 Proof of Corollary 2.4

*Proof.* The result immediately follows from Theorem 2.2 followed by Proposition 2.3 and linearity of derivatives.  $\square$

## B.7 Proof of Corollary 2.5

Let  $\mu^\epsilon$  be a critical point for  $F_{\lambda, \nu}^\epsilon$  with  $\text{supp}(\mu^\epsilon) = \mathbb{R}^d$ . By Theorem 2.2,  $\delta F_{\lambda, \nu}^\epsilon$  is smooth, and hence  $\nabla \delta F_{\lambda, \nu}^\epsilon(x) = 0$  for  $\mu^\epsilon$ -almost every  $x$  implies that  $\nabla \delta F_{\lambda, \nu}^\epsilon(x) = 0$  for all  $x \in \mathbb{R}^d$ . Since  $\mathbb{R}^d$  is connected, this implies that there exists a constant  $c$  such that  $\delta F_{\lambda, \nu}^\epsilon(x) = c$  for all  $x \in \mathbb{R}^d$ . Thus  $\langle \delta F_{\lambda, \nu}^\epsilon(\mu^\epsilon), \mu - \mu^\epsilon \rangle = 0$  for any  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$  satisfying the optimality criteria in Proposition A.9. The same proof applies to  $S_{\lambda, \nu}^\epsilon$ .  $\square$

## B.8 Proof of Lemma 2.6

For each  $1 \leq j \leq m$  there is a unique coupling between  $\delta_{x_\nu}$  and  $\nu_j$ , given by  $\delta_{x_\nu} \otimes \nu_j$ . We then compute the entropic map  $T_{\delta_{x_\nu} \rightarrow \nu_j}^\epsilon(x_\nu) = \mathbb{E}_{(X, Y) \sim \delta_{x_\nu} \otimes \nu_j}(Y | X = x_\nu) = \mathbb{E}(\nu_j)$ . This holds for each  $j$ , and hence we may compute  $\sum_{j=1}^m \lambda_j T_{\delta_{x_\nu} \rightarrow \nu_j}^\epsilon(x_\nu) = x_\nu$ .

Hence  $\delta_{x_\nu}$  satisfies the fixed point condition in Corollary 2.4, and thus is a critical point.  $\square$

## B.9 Proof of Proposition 2.7

The proof is essentially an immediate consequence of Proposition B.23 and Lemma B.21 stated below. In order to state the results, we first recall the following definition of convex ordering and a useful lemma.

**Definition B.20.** Two probability measures  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  are in convex order, denoted  $\mu \leq_c \nu$ , if  $\int \phi d\mu \leq \int \phi d\nu$  for all convex  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ .

**Lemma B.21.** *Let  $\nu$  be  $\sigma$ -subgaussian and  $\mu \leq_c \nu$ . Then  $\mu$  is  $\sigma$ -subgaussian.*

*Proof.* Let  $X \sim \mu$  and  $Y \sim \nu$ , and let  $v \in S^{d-1}$ . Then we define the function  $f_v(x) := \exp\left(\frac{|\langle x, v \rangle|^2}{\sigma^2}\right)$ , which is convex in  $x$ . Since  $\mu \leq_c \nu$ , we have  $\mathbb{E}_{X \sim \mu}(f_v(X)) \leq \mathbb{E}_{Y \sim \nu}(f_v(Y))$  for any  $v$ , and since  $\nu$  is subgaussian, we have that  $\sup_{v \in S^{d-1}} \mathbb{E}_{Y \sim \nu}(f_v(Y)) \leq 2$ . Hence we may conclude that  $\sup_{v \in S^{d-1}} \mathbb{E}_{X \sim \mu}(f_v(X)) \leq 2$ , so that  $\mu$  is  $\sigma$ -subgaussian.  $\square$

We note the following result from (Yang et al., 2024).

**Lemma B.22.** *(Lemma 9.2 in (Yang et al., 2024)) Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ . Then  $(T_{\mu \rightarrow \nu}^\epsilon)_\#(\mu) \leq_c \nu$ .*

We apply the above lemma to obtain the following result for critical points of  $F_{\lambda, \mathcal{V}}^\epsilon$ :

**Proposition B.23.** *Suppose  $\{\nu_1, \dots, \nu_m\} = \mathcal{V} \subset \mathcal{G}_\sigma(\mathbb{R}^d)$ , and let  $\mu^\epsilon$  be a critical point for  $F_{\lambda, \mathcal{V}}^\epsilon$ . Then  $\mu^\epsilon \leq_c \sum_{j=1}^m \lambda_j \nu_j$ .*

*Proof.* Let  $\mu^\epsilon$  be a critical point of  $F_{\lambda, \mathcal{V}}^\epsilon$ . By Lemma B.22,  $\tilde{\mu}_j := (T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon)_\#(\mu^\epsilon) \leq_c \nu_j$  for all  $1 \leq j \leq m$ . As  $\mu^\epsilon$  is a critical point, we have:  $(\sum_{j=1}^m \lambda_j T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon)_\#(\mu^\epsilon) = \mu^\epsilon$  by Corollary 2.4. Therefore, for any convex function  $\phi(x)$ :

$$\begin{aligned} \int \phi(x) d\mu^\epsilon(x) &= \int \phi(x) d \left( \sum_{j=1}^m \lambda_j T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon \right)_\#(\mu^\epsilon)(x) \\ &= \int \phi \left( \sum_{j=1}^m \lambda_j T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon(x) \right) d\mu^\epsilon(x) \\ &\leq \int \sum_{j=1}^m \lambda_j \phi(T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon(x)) d\mu^\epsilon(x) \end{aligned} \tag{52}$$

$$\begin{aligned} &= \sum_{j=1}^m \lambda_j \int \phi(x) d(T_{\mu^\epsilon \rightarrow \nu_j}^\epsilon)_\#(\mu^\epsilon)(x) \\ &= \sum_{j=1}^m \lambda_j \int \phi(x) d\tilde{\mu}_j(x) \\ &\leq \sum_{j=1}^m \lambda_j \int \phi(x) d\nu_j(x) \\ &= \int \phi(x) d \left( \sum_{j=1}^m \lambda_j \nu_j \right)(x), \end{aligned} \tag{53}$$

where we applied Jensen's inequality in (52) and Lemma B.22 to assert that  $\tilde{\mu}_j \leq_c \nu_j$  in (53). Thus  $\mu^\epsilon \leq_c \sum_{j=1}^m \lambda_j \nu_j$ .  $\square$

We conclude the proof of Proposition 2.7 by observing that

$$\left\| \sum_{j=1}^m \lambda_j \nu_j \right\|_{\mathcal{G}} \leq \max_{1 \leq j \leq m} \|\nu_j\|_{\mathcal{G}} \leq \sigma.$$

$\square$



## C PROOF OF THEOREM 3.1

The proof of Theorem 3.1 rests on the following propositions.

**Proposition C.1.** *Let  $\mathcal{V} = \{\nu_1, \dots, \nu_m\} \subset \mathcal{G}_\sigma(\mathbb{R}^d)$  and let  $\hat{\mathcal{V}}^n = \{\hat{\nu}_1^n, \dots, \hat{\nu}_m^n\}$ . Then:*

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \mathcal{V}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \hat{\mathcal{V}}^n}^\epsilon(\mu) \right| \right) \leq \frac{m^{1/2} R_{d, \epsilon, \sigma}^*}{\sqrt{n}}, \quad (54)$$

$$\left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \mathcal{V}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} F_{\lambda, \mathcal{V}}^\epsilon(\mu) \right| \leq \frac{R_{d, \epsilon, \sigma}^*}{\sqrt{n}}, \quad (55)$$

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \mathcal{V}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} F_{\lambda, \hat{\mathcal{V}}^{n_2}}^\epsilon(\mu) \right| \right) \leq m R_{d, \epsilon, \sigma}^* \left( \frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right), \quad (56)$$

where  $R_{d, \epsilon, \sigma}^*$  is a constant depending only on  $d, \epsilon$  and  $\sigma$ .

We note the different dependencies on  $m$  appearing in (54), (55) and (56), which is possibly due to our method of proof. We leave the question of whether the dependence on  $m$  can be improved to future work. We give an analogous proposition for the Sinkhorn barycenter functional.

**Proposition C.2.** *Suppose that, for any  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$ ,  $U_n := \arg \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu)$  is nonempty and furthermore  $U_n \subset \mathcal{G}_{\tilde{\sigma}}(\mathbb{R}^d)$ , where  $\tilde{\sigma} = \max_{1 \leq j \leq m} \{\|\nu_j\|_{\mathcal{G}}\}$ . Let  $\mathcal{V} = \{\nu_1, \dots, \nu_m\} \subset \mathcal{G}_\sigma(\mathbb{R}^d)$  and let  $\hat{\mathcal{V}}^n = \{\hat{\nu}_1^n, \dots, \hat{\nu}_m^n\}$ . Then:*

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) - \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \hat{\mathcal{V}}^n}^\epsilon(\mu) \right| \right) \leq \frac{m^{1/2} R_{d, \epsilon, \sigma}^*}{\sqrt{n}}, \quad (57)$$

$$\left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) \right| \leq \frac{R_{d, \epsilon, \sigma}^*}{\sqrt{n}}, \quad (58)$$

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \mathcal{V}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \hat{\mathcal{V}}^{n_2}}^\epsilon(\mu) \right| \right) \leq m R_{d, \epsilon, \sigma}^* \left( \frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right), \quad (59)$$

We remark that the additional assumptions present in Proposition C.2 are to guarantee that the minimizer of the Sinkhorn barycenter functional has subgaussian constant bounded above by the maximal subgaussian constant over the reference measures. This is not needed in Proposition C.1, as this is implied by Proposition B.23. The proofs of Propositions C.1 and Proposition C.2 rely on three technical results established in (Mena and Niles-Weed, 2019).

**Proposition C.3.** *(Proposition 2 in (Mena and Niles-Weed, 2019)) Let  $\mu, \nu$  and  $\hat{\nu}^n$  be  $\sigma$ -subgaussian measures, where  $\sigma \in [0, \infty)$  is (possibly) random. Then:*

$$|OT_2^\epsilon(\mu, \hat{\nu}^n) - OT_2^\epsilon(\mu, \nu)| \leq 2 \sup_{u \in \mathcal{F}_{\epsilon, \sigma}} \left| \int u d\nu - \int u d\hat{\nu}^n \right|,$$

almost-surely, where  $\mathcal{F}_{\epsilon, \sigma}$  is defined as the set of functions  $u$  such that, for any multi-index  $\alpha$  with  $|\alpha| = k$ :

$$|D^\alpha(u - \frac{1}{2}\|\cdot\|^2)(x)| \leq C_{\epsilon, k, d} \begin{cases} 1 + \sigma^4 & \text{if } k = 0 \\ \sigma^k(\sigma + \sigma^2)^k & \text{otherwise} \end{cases}$$

if  $\|x\| \leq \sqrt{d}\sigma$ , and

$$|D^\alpha(u - \frac{1}{2}\|\cdot\|^2)(x)| \leq C_{\epsilon, k, d} \begin{cases} 1 + (1 + \sigma^2)\|x\|^2 & \text{if } k = 0 \\ \sigma^k(\sqrt{\sigma\|x\|} + \sigma\|x\|)^k & \text{otherwise} \end{cases}$$

if  $\|x\| > \sqrt{d}\sigma$ , where  $C_{\epsilon, k, d}$  is a constant depending only on  $\epsilon, k, d$ .

**Proposition C.4.** *(Proof of Theorem 2, (Mena and Niles-Weed, 2019)) Let  $\nu$  be  $\sigma$ -subgaussian and let  $\tilde{\sigma}$  be a random variable such that both  $\nu$  and  $\hat{\nu}^n$  are  $\tilde{\sigma}$ -subgaussian. Then:*

$$\mathbb{E}_{\hat{\nu}^n, \tilde{\sigma}} \left( \sup_{u \in \mathcal{F}_{\epsilon, \tilde{\sigma}}} \left| \int u d\nu - \int u d\hat{\nu}^n \right| \right) \leq \frac{C_{d, \epsilon}}{\sqrt{n}} (\mathbb{E}_{\tilde{\sigma}}((1 + \tilde{\sigma}^{3s})^2))^{1/2} (1 + \sigma^{d+2}), \quad (60)$$

where  $s = \lceil \frac{d}{2} \rceil + 1$  and  $C_{d, \epsilon}$  is a constant depending on  $d, \epsilon$ .

As Proposition C.4 is not stated exactly as a result in (Mena and Niles-Weed, 2019), we give a sketch of the proof:

*Proof.* For any  $q \geq 2$ , let  $\mathcal{F}^{\epsilon, q}$  to be the set of all functions with  $|f(x)| \leq C_{d, \epsilon}(1 + \|x\|^2)$  and  $|D^\alpha f(x)| \leq C_{d, \epsilon}(1 + \|x\|^q)$  for  $|\alpha| \leq q$ , where  $C_{d, \epsilon}$  is an absolute constant that can depend on  $d$  and  $\epsilon$ . Then it can be shown that if  $u \in \mathcal{F}_{\epsilon, \tilde{\sigma}}$  (the function class from Proposition C.3) then  $\frac{u}{1 + \tilde{\sigma}^{3s}} \in \mathcal{F}^{\epsilon, s}$ , almost-surely. This implies the left hand side of (60) can be upper bound by

$$\begin{aligned} & \mathbb{E}_{\tilde{\sigma}, \hat{\nu}^n} \left( (1 + \tilde{\sigma}^{3s}) \sup_{u \in \mathcal{F}^{\epsilon, s}} \left| \int u d\nu - \int u d\hat{\nu}^n \right| \right) \\ & \leq (\mathbb{E}_{\tilde{\sigma}}((1 + \tilde{\sigma}^{3s})^2))^{1/2} \left( \mathbb{E}_{\hat{\nu}^n} \left( \sup_{u \in \mathcal{F}^{\epsilon, s}} \left| \int u d\nu - \int u d\hat{\nu}^n \right| \right)^2 \right)^{1/2}, \end{aligned}$$

where we applied the Cauchy-Schwarz inequality. Following the empirical process theory argument as in the proof of Theorem 2 in (Mena and Niles-Weed, 2019), one is then able to bound

$$\mathbb{E}_{\hat{\nu}^n} \left( \left( \sup_{u \in \mathcal{F}^{\epsilon, s}} \left| \int u d\nu - \int u d\hat{\nu}^n \right| \right)^2 \right)$$

by  $\frac{C_{d, \epsilon}}{n}(1 + \sigma^{2d+4})$ . We conclude by applying the bound  $(1 + \sigma^{2d+4})^{1/2} \leq 1 + \sigma^{d+2}$ , which holds since  $\sigma \geq 0$ .  $\square$

**Lemma C.5.** (Lemma A.4 in (Mena and Niles-Weed, 2019)) *Let  $\nu$  be  $\sigma$ -subgaussian. Then for all integers  $k \geq 1$ ,*

$$\begin{aligned} \mathbb{E}(\|\hat{\nu}^n\|_{\mathcal{G}}^{2k}) & \leq L_k \sigma^{2k}, \\ \mathbb{E}(\|\hat{\nu}^n\|_{\mathcal{G}}^{2k-1}) & \leq \max\{L_k \sigma^{2k}, 1\} \end{aligned}$$

where  $L_k$  is a constant depending on  $k$ .

We remark that Lemma A.4 in (Mena and Niles-Weed, 2019) establishes  $\mathbb{E}(\|\hat{\nu}^n\|_{\mathcal{G}}^{2k}) \leq L_k \sigma^{2k}$ . To extend this to odd powers, note that  $\mathbb{E}(\|X\|^k) \leq \mathbb{E}(\|X\|^{k+1})^{\frac{k}{k+1}}$  for any random variable  $X$  by Jensen's inequality. If  $\mathbb{E}(\|X\|^{k+1}) \geq 1$  then  $\mathbb{E}(\|X\|^{k+1})^{\frac{k}{k+1}} \leq \mathbb{E}(\|X\|^{k+1})$ , and hence  $\mathbb{E}(\|X\|^k) \leq \mathbb{E}(\|X\|^{k+1})$ , and if  $\mathbb{E}(\|X\|^{k+1}) \leq 1$  then  $\mathbb{E}(\|X\|^k) \leq 1$ , from which we establish Lemma C.5. Also note that from [Lemma A.2, (Mena and Niles-Weed, 2019)], the random variable  $\|\hat{\nu}^n\|_{\mathcal{G}}$  is finite almost surely.

**Lemma C.6.** *For all  $1 \leq j \leq m$ , let  $\|\nu_j\|_{\mathcal{G}} \leq \sigma$ , and let  $\tilde{\sigma} := \max\{\sigma, \|\hat{\nu}_1^n\|_{\mathcal{G}}, \|\hat{\nu}_2^n\|_{\mathcal{G}}, \dots, \|\hat{\nu}_m^n\|_{\mathcal{G}}\}$ . Then if  $\|\mu\|_{\mathcal{G}} \leq \tilde{\sigma}$ ,*

$$\mathbb{E}(|OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n)|) \leq \frac{m^{1/2} P_{d, \epsilon}(\sigma)}{\sqrt{n}}, \quad \forall 1 \leq j \leq m, \quad (61)$$

for some polynomial  $P_{d, \epsilon}(\sigma)$  with coefficients depending on  $d, \epsilon$  and degree depending on  $d$ .

*Proof.* As  $\mu, \nu_j$  and  $\hat{\nu}_j^n$  are all  $\tilde{\sigma}$  subgaussian, by Proposition C.3 we have:

$$\mathbb{E}(|OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n)|) \leq \mathbb{E} \left( 2 \sup_{u \in \mathcal{F}_{\epsilon, \tilde{\sigma}}} \left| \int u d\nu_j - \int u d\hat{\nu}_j^n \right| \right),$$

where the expectation is with respect to the empirical measure  $\hat{\nu}_j^n$  and  $\tilde{\sigma}$ . By Proposition C.4 and Young's inequality we may upper bound:

$$\begin{aligned} \mathbb{E} \left( 2 \sup_{u \in \mathcal{F}_{\epsilon, \tilde{\sigma}}} \left| \int u d\nu_j - \int u d\hat{\nu}_j^n \right| \right) & \leq \frac{2C_{d, \epsilon}}{\sqrt{n}} (\mathbb{E}_{\tilde{\sigma}}((1 + \tilde{\sigma}^{3s})^2))^{1/2} (1 + \sigma^{d+2}) \\ & \leq \frac{2C_{d, \epsilon}}{\sqrt{n}} (2 + 2\mathbb{E}_{\tilde{\sigma}}(\tilde{\sigma}^{6s}))^{1/2} (1 + \sigma^{d+2}) \end{aligned} \quad (62)$$

where  $s = \lceil \frac{d}{2} \rceil + 1$ . It remains to bound  $\mathbb{E}_{\tilde{\sigma}}(\tilde{\sigma}^{6s})$ .

$$\begin{aligned} \mathbb{E}_{\tilde{\sigma}}(\tilde{\sigma}^{6s}) &= \mathbb{E}(\max\{\sigma, \|\hat{\nu}_1^n\|_{\mathcal{G}}, \|\hat{\nu}_2^n\|_{\mathcal{G}}, \dots, \|\hat{\nu}_m^n\|_{\mathcal{G}}\}^{6s}) \\ &\leq \mathbb{E}\left(\sigma^{6s} + \sum_{j=1}^m \|\hat{\nu}_j^n\|_{\mathcal{G}}^{6s}\right) \\ &= \sigma^{6s} + \sum_{j=1}^m \mathbb{E}(\|\hat{\nu}_j^n\|_{\mathcal{G}}^{6s}) \\ &\leq \sigma^{6s} + mL_{6s}\sigma^{6s} \end{aligned}$$

by Lemma C.5. Therefore:

$$\begin{aligned} \frac{2C_{d,\epsilon}}{\sqrt{n}}(2 + 2\mathbb{E}_{\tilde{\sigma}}(\tilde{\sigma}^{6s}))^{1/2}(1 + \sigma^{d+2}) &\leq \frac{2C_{d,\epsilon}}{\sqrt{n}}(2 + 2(mL_{6s} + 1)\sigma^{6s})^{1/2}(1 + \sigma^{d+2}) \\ &\leq \frac{2m^{1/2}C_{d,\epsilon}}{\sqrt{n}}(2 + 2(L_{6s} + 1)\sigma^{6s})^{1/2}(1 + \sigma^{d+2}) \\ &\leq \frac{2m^{1/2}C_{d,\epsilon}}{\sqrt{n}}(\sqrt{2} + \sqrt{2(L_{6s} + 1)}\sigma^{3s})(1 + \sigma^{d+2}). \end{aligned}$$

We conclude that  $\mathbb{E}(|OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n)|) \leq \frac{m^{1/2}P_{d,\epsilon}(\sigma)}{\sqrt{n}}$  for some polynomial  $P_{d,\epsilon}(\sigma)$  with coefficients depending on  $d, \epsilon$  and degree depending on  $d$ .  $\square$

**Lemma C.7.** *Let  $\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)$ . Then:*

$$\mathbb{E}(|OT_2^\epsilon(\mu, \mu) - OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n)|) \leq \frac{2P_{d,\epsilon}(\sigma)}{\sqrt{n}},$$

for some polynomial  $P_{d,\epsilon}(\sigma)$  with coefficients depending on  $d, \epsilon$  and degree depending on  $d$ , and the expectation is over the empirical measure  $\hat{\mu}^n$ .

*Proof.* We apply the triangle inequality to bound:

$$\begin{aligned} &\mathbb{E}(|OT_2^\epsilon(\mu, \mu) - OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n)|) \\ &= \mathbb{E}(|OT_2^\epsilon(\mu, \mu) - OT_2^\epsilon(\hat{\mu}^n, \mu) + OT_2^\epsilon(\hat{\mu}^n, \mu) - OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n)|) \\ &\leq \mathbb{E}(|OT_2^\epsilon(\mu, \mu) - OT_2^\epsilon(\hat{\mu}^n, \mu)|) + \mathbb{E}(|OT_2^\epsilon(\hat{\mu}^n, \mu) - OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n)|). \end{aligned} \tag{63}$$

The first term is bounded by  $\frac{P_{d,\epsilon}(\sigma)}{\sqrt{n}}$  following a similar argument as used in the proof of Lemma C.6 with  $m = 1$ . We bound the second term by conditioning:

$$\begin{aligned} &\mathbb{E}(|OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n) - OT_2^\epsilon(\hat{\mu}^n, \mu)|) \\ &= \mathbb{E}\left(\mathbb{E}\left(|OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n) - OT_2^\epsilon(\hat{\mu}^n, \mu)| \middle| \hat{\mu}^n\right)\right) \\ &\leq \mathbb{E}\left(\mathbb{E}\left(2 \sup_{u \in \mathcal{F}_{\epsilon, \tilde{\sigma}}} \left| \int u d\mu - \int u d\hat{\mu}^n \right| \middle| \hat{\mu}^n\right)\right) \\ &= \mathbb{E}\left(2 \sup_{u \in \mathcal{F}_{\epsilon, \tilde{\sigma}}} \left| \int u d\mu - \int u d\hat{\mu}^n \right| \right) \end{aligned} \tag{64}$$

where we applied Proposition C.3 to the inner expectation in (64) with  $\tilde{\sigma} = \max\{\sigma, \|\hat{\mu}^n\|_{\mathcal{G}}\}$ . We then apply Proposition C.4 followed by Lemma C.5 to obtain the bound:

$$\mathbb{E}(|OT_2^\epsilon(\hat{\mu}^n, \hat{\mu}^n) - OT_2^\epsilon(\hat{\mu}^n, \mu)|) \leq \frac{P_{d,\epsilon}(\sigma)}{\sqrt{n}}.$$

$\square$

**Proof of Proposition C.1:** We begin by proving (54). Let  $\mu^*$  be a minimizer for  $F_{\lambda, \nu}^\epsilon$ , and  $\mu^{*,n}$  be a random minimizer for  $F_{\lambda, \hat{\nu}^n}^\epsilon$ , so that

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^n}^\epsilon(\mu) \right| \right) = \mathbb{E} \left( \left| F_{\lambda, \nu}^\epsilon(\mu^*) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) \right| \right).$$

Suppose that  $F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) \geq F_{\lambda, \nu}^\epsilon(\mu^*)$ . Then:

$$\begin{aligned} \left| F_{\lambda, \nu}^\epsilon(\mu^*) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) \right| &= F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) - F_{\lambda, \nu}^\epsilon(\mu^*) \\ &\leq F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^*) - F_{\lambda, \nu}^\epsilon(\mu^*). \end{aligned}$$

Similarly, if  $F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) \leq F_{\lambda, \nu}^\epsilon(\mu^*)$ , we have

$$\left| F_{\lambda, \nu}^\epsilon(\mu^*) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) \right| \leq F_{\lambda, \nu}^\epsilon(\mu^{*,n}) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n})$$

and hence we may conclude that

$$\mathbb{E} \left( \left| F_{\lambda, \nu}^\epsilon(\mu^*) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu^{*,n}) \right| \right) \leq \mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left| F_{\lambda, \nu}^\epsilon(\mu) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu) \right| \right).$$

We then bound:

$$\begin{aligned} &\mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left| F_{\lambda, \nu}^\epsilon(\mu) - F_{\lambda, \hat{\nu}^n}^\epsilon(\mu) \right| \right) \\ &= \mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left| \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu, \nu_j) - \sum_{j=1}^m \lambda_j OT_2^\epsilon(\mu, \hat{\nu}_j^n) \right| \right) \\ &\leq \mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \sum_{j=1}^m \lambda_j \left| OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n) \right| \right) \\ &\leq \mathbb{E} \left( \sum_{j=1}^m \lambda_j \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left| OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n) \right| \right) \\ &= \sum_{j=1}^m \lambda_j \mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left| OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n) \right| \right) \end{aligned} \tag{65}$$

$$\leq \sum_{j=1}^m \lambda_j \mathbb{E} \left( \left| OT_2^\epsilon(\mu^*, \nu_j) - OT_2^\epsilon(\mu^*, \hat{\nu}_j^n) \right| \right) + \sum_{j=1}^m \lambda_j \mathbb{E} \left( \left| OT_2^\epsilon(\mu^{*,n}, \nu_j) - OT_2^\epsilon(\mu^{*,n}, \hat{\nu}_j^n) \right| \right) \tag{66}$$

To bound  $\mathbb{E}(|OT_2^\epsilon(\mu^*, \nu_j) - OT_2^\epsilon(\mu^*, \hat{\nu}_j^n)|)$ , note by Proposition 2.7,  $\mu^*$  is  $\sigma$ -subgaussian, and  $\mu^{*,n}$  is  $\max_{1 \leq j \leq m} \{\|\hat{\nu}_j^n\|_{\mathcal{G}}\}$ -subgaussian, both  $\mu^*$  and  $\mu^{*,n}$  are  $\max\{\sigma, \|\hat{\nu}_1^n\|_{\mathcal{G}}, \dots, \|\hat{\nu}_m^n\|_{\mathcal{G}}\}$ -subgaussian, and we may apply Lemma C.6 to conclude:

$$\mathbb{E}(|OT_2^\epsilon(\mu^*, \nu_j) - OT_2^\epsilon(\mu^*, \hat{\nu}_j^n)|) \leq \frac{m^{1/2} P_{d,\epsilon}(\sigma)}{\sqrt{n}}, \quad \mathbb{E}(|OT_2^\epsilon(\mu^{*,n}, \nu_j) - OT_2^\epsilon(\mu^{*,n}, \hat{\nu}_j^n)|) \leq \frac{m^{1/2} P_{d,\epsilon}(\sigma)}{\sqrt{n}}, \tag{67}$$

where  $P_{d,\epsilon}(\sigma)$  is a polynomial in  $\sigma$ , depending on  $d$  and  $\epsilon$ . Summing over  $1 \leq j \leq m$ , we establish (54) with a bound of  $\frac{2m^{1/2} P_{d,\epsilon}(\sigma)}{\sqrt{n}}$ .

Next we prove (55). Let  $\mu^*$  minimize  $F_{\lambda, \nu}^\epsilon$ . Note that the empirical measure  $\hat{\mu}^{*,n}$  associated to  $\mu^*$  is an element

of  $\mathcal{P}^n(\mathbb{R}^d)$ , and hence

$$\begin{aligned}
 \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) \right| &= \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) \\
 &\leq \mathbb{E} \left( F_{\lambda, \nu}^\epsilon(\hat{\mu}^{*n}) - F_{\lambda, \nu}^\epsilon(\mu^*) \right) \\
 &= \mathbb{E} \left( |F_{\lambda, \nu}^\epsilon(\hat{\mu}^{*n}) - F_{\lambda, \nu}^\epsilon(\mu^*)| \right) \\
 &\leq \sum_{j=1}^m \lambda_j \mathbb{E} \left( |OT_2^\epsilon(\hat{\mu}^{*n}, \nu_j) - OT_2^\epsilon(\mu^*, \nu_j)| \right) \tag{68}
 \end{aligned}$$

by the triangle inequality. We bound each term

$$\mathbb{E} \left( |OT_2^\epsilon(\hat{\mu}^{*n}, \nu_j) - OT_2^\epsilon(\mu^*, \nu_j)| \right) \leq \frac{P_{\epsilon, d}(\sigma)}{\sqrt{n}}$$

by applying Lemma C.6 with  $m = 1$ . This applies because  $\|\mu^*\|_{\mathcal{G}} \leq \sigma$ , by Proposition 2.7. This gives (55).

We now prove (56):

$$\begin{aligned}
 &\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) \\
 &\leq \mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) + \mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right)
 \end{aligned}$$

We apply (54) to the first term to produce a bound of  $\frac{2m^{1/2}P_{d, \epsilon}(\sigma)}{\sqrt{n_2}}$ . Let

$$\tilde{\sigma} = \max_{1 \leq j \leq m} \|\hat{\nu}_j^{n_2}\|_{\mathcal{G}}.$$

Then, by conditioning on the samples  $\hat{\nu}^{n_2}$ , we have:

$$\begin{aligned}
 &\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) \\
 &= \mathbb{E} \left( \mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} F_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \middle| \hat{\nu}^{n_2} \right) \right) \\
 &\leq \mathbb{E} \left( \mathbb{E} \left( \frac{2P_{d, \epsilon}(\tilde{\sigma})}{\sqrt{n_1}} \middle| \hat{\nu}^{n_2} \right) \right) \tag{69}
 \end{aligned}$$

$$= \frac{2}{\sqrt{n_1}} \mathbb{E}(P_{\sigma, \epsilon}(\tilde{\sigma})). \tag{70}$$

where we applied (55) conditionally on the samples  $\hat{\nu}^{n_2}$ . By linearity of expectation:

$$\mathbb{E}(P_{d, \epsilon}(\tilde{\sigma})) = \mathbb{E} \left( \sum_{i=1}^{r_d} c_{d, \epsilon, i} \tilde{\sigma}^i \right) \tag{71}$$

$$= \sum_{i=1}^{r_d} c_{d, \epsilon, i} \mathbb{E}(\tilde{\sigma}^i) \tag{72}$$

for some maximum degree  $r_d$  depending on  $d$ , and (deterministic) coefficients  $c_{d, \epsilon, i}$ , depending on  $d$  and  $\epsilon$ . Since  $\tilde{\sigma}^i \leq \sum_{j=1}^m \|\hat{\nu}_j\|_{\mathcal{G}}^i$ , we have:

$$\begin{aligned}
 \mathbb{E}(P_{d,\epsilon}(\tilde{\sigma})) &\leq \sum_{i=1}^{r_d} c_{d,\epsilon,i} \sum_{j=1}^m \mathbb{E}(\|\hat{\nu}_j^n\|_{\mathcal{G}}^i) \\
 &\leq \sum_{i=1}^{r_d} c_{d,\epsilon,i} \sum_{j=1}^m \max\{L_{\lceil \frac{i}{2} \rceil} \sigma^{2\lceil \frac{i}{2} \rceil}, 1\} \\
 &\leq m \sum_{i=1}^{r_d} c_{d,\epsilon,i} (L_{\lceil \frac{i}{2} \rceil} \sigma^{2\lceil \frac{i}{2} \rceil} + 1) \\
 &=: m\tilde{P}_{d,\epsilon}(\sigma),
 \end{aligned}$$

where we applied Lemma C.5. Thus we may conclude that:

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda,\nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} F_{\lambda,\hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) \leq (2m^{1/2}P_{d,\epsilon}(\sigma) + m\tilde{P}_{d,\epsilon}(\sigma)) \left( \frac{1}{\sqrt{n_2}} + \frac{1}{\sqrt{n_1}} \right),$$

which implies (56).

Labeling the maximum over all constants in the three statements in Proposition C.1 as  $R_{d,\epsilon,\sigma}^*$ , we obtain our result.  $\square$

**Proof of Proposition C.2:** We first show (57). Let  $\mu^* \in \arg \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon$  and  $\mu^{*,n} \in \arg \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\hat{\nu}^n}^\epsilon$ . Then we bound:

$$\begin{aligned}
 &\mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\hat{\nu}^n}^\epsilon(\mu) \right| \right) \\
 &\leq \mathbb{E} \left( \sum_{j=1}^m \lambda_j \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left| OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n) - \frac{1}{2} OT_2^\epsilon(\nu_j, \nu_j) + \frac{1}{2} OT_2^\epsilon(\hat{\nu}_j^n, \hat{\nu}_j^n) \right| \right) \\
 &\leq \mathbb{E} \left( \sum_{j=1}^m \lambda_j \max_{\mu \in \{\mu^*, \mu^{*,n}\}} \left( \left| OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n) \right| + \left| \frac{1}{2} OT_2^\epsilon(\nu_j, \nu_j) - \frac{1}{2} OT_2^\epsilon(\hat{\nu}_j^n, \hat{\nu}_j^n) \right| \right) \right) \\
 &= \sum_{j=1}^m \lambda_j \left( \mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} |OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n)| \right) + \mathbb{E} \left( \left| \frac{1}{2} OT_2^\epsilon(\nu_j, \nu_j) - \frac{1}{2} OT_2^\epsilon(\hat{\nu}_j^n, \hat{\nu}_j^n) \right| \right) \right)
 \end{aligned}$$

where we have applied the triangle inequality. Using the assumptions  $\mu^*, \mu^{*,n} \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , we bound:

$$\sum_{j=1}^m \lambda_j \mathbb{E} \left( \max_{\mu \in \{\mu^*, \mu^{*,n}\}} |OT_2^\epsilon(\mu, \nu_j) - OT_2^\epsilon(\mu, \hat{\nu}_j^n)| \right) \leq \frac{2m^{1/2}P_{d,\epsilon}(\sigma)}{\sqrt{n}} \quad (73)$$

by applying Lemma C.6. To bound  $\frac{1}{2} \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\nu_j, \nu_j) - OT_2^\epsilon(\hat{\nu}_j^n, \hat{\nu}_j^n)|)$ , we apply Lemma C.7, giving us a bound of  $\frac{P_{d,\epsilon}(\sigma)}{\sqrt{n}}$ . We conclude:

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\hat{\nu}^n}^\epsilon(\mu) \right| \right) \leq \frac{(2m^{1/2} + 1)P_{d,\epsilon}(\sigma)}{\sqrt{n}},$$

which establishes (57).

Similarly, to establish (58), we let  $\mu^*$  again be the minimizer of  $S_{\lambda,\nu}^\epsilon$  on  $\mathcal{G}_\sigma(\mathbb{R}^d)$  and let  $\hat{\mu}^{*,n}$  be the empirical measure for  $\mu^*$ . By assumption, we have for any  $\rho \in \arg \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon(\mu)$ ,  $\rho \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , and hence:

$$\begin{aligned}
 &\left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon(\mu) \right| \\
 &= \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda,\nu}^\epsilon(\mu) - S_{\lambda,\nu}^\epsilon(\mu^*).
 \end{aligned}$$

Since  $\hat{\mu}^{*n} \in \mathcal{P}^n(\mathbb{R}^d)$ , we may upper bound this by:

$$\begin{aligned} & \mathbb{E} \left( S_{\lambda, \nu}^\epsilon(\hat{\mu}^{*n}) - S_{\lambda, \nu}^\epsilon(\mu^*) \right) \\ &= \mathbb{E} \left( \left| S_{\lambda, \nu}^\epsilon(\hat{\mu}^{*n}) - S_{\lambda, \nu}^\epsilon(\mu^*) \right| \right) \\ &\leq \sum_{j=1}^m \lambda_j \mathbb{E} \left( |OT_2^\epsilon(\hat{\mu}^{*n}, \nu_j) - OT_2^\epsilon(\mu^*, \nu_j)| \right) + \frac{1}{2} \mathbb{E} \left( |OT_2^\epsilon(\hat{\mu}^{*n}, \hat{\mu}^{*n}) - OT_2^\epsilon(\mu^*, \mu^*)| \right) \end{aligned}$$

by the triangle inequality. By the assumption that  $\mu^* \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , we may apply Lemma C.6 to bound each summand in the first term by  $\frac{P_{d, \epsilon}(\sigma)}{\sqrt{n}}$  and the second term can be bounded by  $\frac{P_{d, \epsilon}(\sigma)}{\sqrt{n}}$  via Lemma C.7. Combining these two bounds, we have:

$$\left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) \right| \leq \frac{2P_{d, \epsilon}(\sigma)}{\sqrt{n}}. \quad (74)$$

We conclude with the proof of (59):

$$\begin{aligned} & \mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) \\ &\leq \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) \right| + \mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right). \end{aligned}$$

by the triangle inequality. We may apply (74) to bound the first term by  $\frac{2P_{d, \epsilon}(\sigma)}{\sqrt{n_1}}$ . To bound the second term, we apply a similar argument as used to bound (57). Let  $\mu^{n_1} \in \arg \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu)$  and  $\mu_{n_2}^{n_1} \in \arg \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu)$ . Then:

$$\begin{aligned} & \mathbb{E} \left( \left| \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) \\ &\leq \mathbb{E} \left( \max_{\mu^* \in \{\mu^{n_1}, \mu_{n_2}^{n_1}\}} \left| S_{\lambda, \nu}^\epsilon(\mu^*) - S_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu^*) \right| \right) \\ &\leq \mathbb{E} \left( \max_{\mu^* \in \{\mu^{n_1}, \mu_{n_2}^{n_1}\}} \sum_{j=1}^m \lambda_j |OT_2^\epsilon(\mu^*, \nu_j) - OT_2^\epsilon(\mu^*, \hat{\nu}_j^{n_2})| \right) + \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\nu_j, \nu_j) - OT_2^\epsilon(\hat{\nu}_j^{n_2}, \hat{\nu}_j^{n_2})|) \end{aligned}$$

by the triangle inequality. The second term can be bounded by  $\frac{P_{d, \epsilon}(\sigma)}{\sqrt{n}}$  by Lemma C.7. To deal with the first term, note that by assumption,  $\mu^{n_1} \in \mathcal{G}_\sigma(\mathbb{R}^d)$ , and  $\mu_{n_2}^{n_1} \in \mathcal{G}_{\tilde{\sigma}}(\mathbb{R}^d)$  with  $\tilde{\sigma} = \max_{1 \leq j \leq m} \{\|\hat{\nu}_j^{n_2}\|\}$ , and hence  $\{\mu^{n_1}, \mu_{n_2}^{n_1}\} \subset \mathcal{G}_{\sigma^*}(\mathbb{R}^d)$  where  $\sigma^* = \max\{\sigma, \tilde{\sigma}\}$ . Hence we can bound:

$$\begin{aligned} & \mathbb{E} \left( \max_{\mu^* \in \{\mu^{n_1}, \mu_{n_2}^{n_1}\}} \sum_{j=1}^m \lambda_j |OT_2^\epsilon(\mu^*, \nu_j) - OT_2^\epsilon(\mu^*, \hat{\nu}_j^{n_2})| \right) \\ &\leq \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\mu^{n_1}, \nu_j) - OT_2^\epsilon(\mu^{n_1}, \hat{\nu}_j^{n_2})|) + \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\mu_{n_2}^{n_1}, \nu_j) - OT_2^\epsilon(\mu_{n_2}^{n_1}, \hat{\nu}_j^{n_2})|) \\ &\leq \sum_{j=1}^m \lambda_j \frac{m^{1/2} P_{d, \epsilon}(\sigma)}{\sqrt{n_2}} + \sum_{j=1}^m \lambda_j \frac{m^{1/2} P_{d, \epsilon}(\sigma)}{\sqrt{n_2}} \\ &= \frac{2m^{1/2} P_{d, \epsilon}(\sigma)}{\sqrt{n_2}}, \end{aligned}$$

where we applied Lemma C.6 in the final inequality. Thus we have bounded

$$\mathbb{E} \left( \left| \min_{\mu \in \mathcal{G}_\sigma(\mathbb{R}^d)} S_{\lambda, \nu}^\epsilon(\mu) - \min_{\mu \in \mathcal{P}^{n_1}(\mathbb{R}^d)} S_{\lambda, \hat{\nu}^{n_2}}^\epsilon(\mu) \right| \right) \leq 2(m^{1/2} + 1)P_{d, \epsilon}(\sigma) \left( \frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right).$$

Taking the maximum over all constants as  $R_{d,\epsilon,\sigma}^*$ , we conclude.

**Proof of Theorem 3.1:** Let  $\mu^* = \arg \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_{\lambda,\nu}^\epsilon(\mu)$ , and let  $\mu^n$  denote a random minimizer of  $F_{\lambda,\hat{\nu}^n}^\epsilon$  over  $\mathcal{P}^n(\mathbb{R}^d)$ . We bound:

$$\mathbb{E} \left( \left| F_{\lambda,\nu}^\epsilon(\mu^*) - F_{\lambda,\nu}^\epsilon(\mu^n) \right| \right) \leq \mathbb{E} \left( \left| F_{\lambda,\nu}^\epsilon(\mu^*) - F_{\lambda,\hat{\nu}^n}^\epsilon(\mu^n) \right| \right) + \mathbb{E} \left( \left| F_{\lambda,\hat{\nu}^n}^\epsilon(\mu^n) - F_{\lambda,\nu}^\epsilon(\mu^n) \right| \right).$$

By Proposition C.1, specifically (56), the first term is upper bounded by  $\frac{mR_{d,\epsilon,\sigma}^*}{\sqrt{n}}$ . We bound the second term:

$$\begin{aligned} \mathbb{E} \left( \left| F_{\lambda,\hat{\nu}^n}^\epsilon(\mu^n) - F_{\lambda,\nu}^\epsilon(\mu^n) \right| \right) &\leq \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\mu^n, \hat{\nu}_j^n) - OT_2^\epsilon(\mu^n, \nu_j)|) \\ &\leq \sum_{j=1}^m \lambda_j \frac{m^{1/2} P_{d,\epsilon}(\sigma)}{\sqrt{n}} \\ &= \frac{m^{1/2} P_{d,\epsilon}(\sigma)}{\sqrt{n}}. \end{aligned}$$

Above, we applied the triangle inequality, followed by (61) from Lemma C.6 to each summand, which applies since  $\|\mu^n\|_{\mathcal{G}} \leq \max_{1 \leq j \leq m} \|\hat{\nu}_j^n\|_{\mathcal{G}}$  almost-surely. From this we conclude (9), with a final bound of  $\frac{1}{\sqrt{n}}(m^{1/2} P_{d,\epsilon}(\sigma) + mR_{d,\epsilon,\sigma}^*)$ .

Similar arguments apply to  $S_{\lambda,\nu}^\epsilon$ . By assumption, for  $\mu^n \in \arg \min_{\mu \in \mathcal{P}^n(\mathbb{R}^d)} S_{\lambda,\hat{\nu}^n}^\epsilon(\mu)$  we have  $\|\mu^n\|_{\mathcal{G}} \leq \max_{1 \leq j \leq m} \{\|\hat{\nu}_j^n\|_{\mathcal{G}}, \sigma\} =: \tilde{\sigma}$ . We bound:

$$\mathbb{E} \left( \left| S_{\lambda,\nu}^\epsilon(\mu^*) - S_{\lambda,\nu}^\epsilon(\mu^n) \right| \right) \leq \mathbb{E} \left( \left| S_{\lambda,\nu}^\epsilon(\mu^*) - S_{\lambda,\hat{\nu}^n}^\epsilon(\mu^n) \right| \right) + \mathbb{E} \left( \left| S_{\lambda,\hat{\nu}^n}^\epsilon(\mu^n) - S_{\lambda,\nu}^\epsilon(\mu^n) \right| \right).$$

We apply (59) to bound the first term by  $\frac{mR_{d,\epsilon,\sigma}^*}{\sqrt{n}}$ . We bound the second term:

$$\begin{aligned} &\mathbb{E} \left( \left| S_{\lambda,\hat{\nu}^n}^\epsilon(\mu^n) - S_{\lambda,\nu}^\epsilon(\mu^n) \right| \right) \\ &\leq \sum_{j=1}^m \lambda_j \mathbb{E} \left( |OT_2^\epsilon(\mu^n, \hat{\nu}_j^n) - OT_2^\epsilon(\mu^n, \nu_j)| + \frac{1}{2} |OT_2^\epsilon(\nu_j, \nu_j) - OT_2^\epsilon(\hat{\nu}_j^n, \hat{\nu}_j^n)| \right) \\ &= \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\mu^n, \hat{\nu}_j^n) - OT_2^\epsilon(\mu^n, \nu_j)|) + \frac{1}{2} \sum_{j=1}^m \lambda_j \mathbb{E} (|OT_2^\epsilon(\nu_j, \nu_j) - OT_2^\epsilon(\hat{\nu}_j^n, \hat{\nu}_j^n)|) \end{aligned}$$

by the triangle inequality. To bound the first term, we apply (61) from Lemma C.6 to each summand, which applies since  $\mu^n \in \mathcal{G}_{\tilde{\sigma}}(\mathbb{R}^d)$ , giving us a bound of  $\frac{m^{1/2} P_{d,\epsilon}(\sigma)}{\sqrt{n}}$ . To bound the second term, we apply Lemma C.7, giving a bound of  $\frac{P_{d,\epsilon}(\sigma)}{\sqrt{n}}$ . Adding these bounds together, we have:

$$\mathbb{E} \left( \left| S_{\lambda,\nu}^\epsilon(\mu^*) - S_{\lambda,\nu}^\epsilon(\mu^n) \right| \right) \leq \frac{1}{\sqrt{n}}(mR_{d,\epsilon,\sigma}^* + (m^{1/2} + 1)P_{d,\epsilon}(\sigma)).$$

Taking the maximum over all constants depending on  $d, \epsilon$  and  $\sigma$  as  $C_{d,\epsilon,\sigma}^*$ , we conclude.  $\square$



## D PROOFS FOR SECTION 4

**Proof of Proposition 4.1:** The assumption that  $\mathcal{V} \subset \mathcal{G}(\mathbb{R}^d)$  ensures that  $\nabla \delta F_{\lambda, \mathcal{V}}^\epsilon(\mu)$  exists and equals  $Id - \sum_{j=1}^m \lambda_j T_{\mu \rightarrow \nu_j}^\epsilon$ . We compute:

$$\begin{aligned} \|\nabla \delta F_{\lambda, \mathcal{V}}^\epsilon(\mu)\|_{L^2(\mu)}^2 &= \|Id - \sum_{j=1}^m \lambda_j T_{\mu \rightarrow \nu_j}^\epsilon\|_{L^2(\mu)}^2 \\ &= \int \langle Id - \sum_{j=1}^m \lambda_j T_{\mu \rightarrow \nu_j}^\epsilon, Id - \sum_{k=1}^m \lambda_k T_{\mu \rightarrow \nu_k}^\epsilon \rangle d\mu \\ &= \int \sum_{j,k=1}^m \lambda_j \lambda_k \langle Id - T_{\mu \rightarrow \nu_j}^\epsilon, Id - T_{\mu \rightarrow \nu_k}^\epsilon \rangle d\mu \\ &= \sum_{j,k=1}^m \lambda_j \lambda_k [A_\mu^\epsilon]_{jk} = \lambda^\top A_\mu^\epsilon \lambda. \end{aligned}$$

Similarly:

$$\begin{aligned} \|\nabla \delta S_{\lambda, \mathcal{V}}^\epsilon(\tilde{\mu})\|_{L^2(\tilde{\mu})}^2 &= \|T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon - \sum_{j=1}^m \lambda_j T_{\tilde{\mu} \rightarrow \nu_j}^\epsilon\|_{L^2(\tilde{\mu})}^2 \\ &= \int \langle T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon - \sum_{j=1}^m \lambda_j T_{\tilde{\mu} \rightarrow \nu_j}^\epsilon, T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon - \sum_{k=1}^m \lambda_k T_{\tilde{\mu} \rightarrow \nu_k}^\epsilon \rangle d\tilde{\mu} \\ &= \int \sum_{j,k=1}^m \lambda_j \lambda_k \langle T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon - T_{\tilde{\mu} \rightarrow \nu_j}^\epsilon, T_{\tilde{\mu} \rightarrow \tilde{\mu}}^\epsilon - T_{\tilde{\mu} \rightarrow \nu_k}^\epsilon \rangle d\tilde{\mu} \\ &= \sum_{j,k=1}^m \lambda_j \lambda_k [S_\mu^\epsilon]_{jk} = \lambda^\top S_\mu^\epsilon \lambda. \end{aligned}$$

Hence we see that  $\lambda^\top A_\mu^\epsilon \lambda \geq 0$ , and  $\lambda^\top A_\mu^\epsilon \lambda = 0$  is equivalent to  $\|\nabla \delta F_{\lambda, \mathcal{V}}^\epsilon(\mu)\|_{L^2(\mu)}^2 = 0$ , which is itself equivalent to  $\nabla \delta F_{\lambda, \mathcal{V}}^\epsilon(\mu) = 0$   $\mu$ -a.e. The analogous result holds for  $S_\mu^\epsilon$ . Finally, if  $\text{supp}(\mu) = \mathbb{R}^d$  (resp.  $\text{supp}(\tilde{\mu}) = \mathbb{R}^d$ ), then we may apply the optimality criteria from Corollary 2.5. □

### D.1 Proof of Theorem 4.2

**Lemma D.1.** *Let  $\mu \in \mathcal{P}_2(\Omega)$  and  $\mathcal{V} \subset \mathcal{G}_\sigma(\Omega)$ , and suppose that  $A_\mu^\epsilon$  has an eigenvalue of 0 with unique eigenvector  $\lambda_* \in \Delta^m$ . Suppose there exists an estimator  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)$  satisfying (11) for  $1 \leq j \leq m$ . Then:*

$$\mathbb{E}(|[A_\mu^\epsilon]_{ij} - [\hat{A}_\mu^\epsilon]_{ij}|) \leq J d \sigma^2 \max \left\{ \frac{1}{\sqrt{n}}, \sqrt{\theta(n)} + \sqrt{\theta(n) + \theta(n)^2} \right\}$$

where  $\hat{A}_\mu$  is the matrix  $\hat{M}_\mu$  in Algorithm 1 with  $\mathcal{F}_{\lambda, \mathcal{V}}^\epsilon = F_{\lambda, \mathcal{V}}^\epsilon$ , and  $J$  is an absolute constant.

*Proof.* For ease of notation, we will write  $T_j^\epsilon$  for  $T_{\mu \rightarrow \nu_j}^\epsilon$  and  $\hat{T}_j^\epsilon$  for  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)$ . We begin by bounding  $\mathbb{E}(|[A_\mu^\epsilon]_{ij} -$

$[\hat{A}_\mu^\epsilon]_{ij})$ :

$$\begin{aligned} & \mathbb{E}(|[A_\mu^\epsilon]_{ij} - [\hat{A}_\mu^\epsilon]_{ij}|) \\ &= \mathbb{E} \left( \left| \int \langle T_j^\epsilon - Id, T_i^\epsilon - Id \rangle d\mu - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_j^\epsilon(X_k) - X_k, \hat{T}_i^\epsilon(X_k) - X_k \rangle \right| \right) \\ &\leq \mathbb{E} \left( \left| \int \langle T_j^\epsilon - Id, T_i^\epsilon - Id \rangle d\mu - \frac{1}{n} \sum_{k=n+1}^{2n} \langle T_j^\epsilon(X_k) - X_k, T_i^\epsilon(X_k) - X_k \rangle \right| \right) \end{aligned} \quad (75)$$

$$+ \mathbb{E} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} \langle T_j^\epsilon(X_k) - X_k, T_i^\epsilon(X_k) - X_k \rangle - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_j^\epsilon(X_k) - X_k, \hat{T}_i^\epsilon(X_k) - X_k \rangle \right| \right) \quad (76)$$

The term (75) may be bounded by observing that  $\langle T_j^\epsilon(X_k) - X_k, T_i^\epsilon(X_k) - X_k \rangle := Z_k$  are i.i.d. samples from the random variable  $\langle T_j^\epsilon(X) - X, T_i^\epsilon(X) - X \rangle := Z$  with  $X \sim \mu$ . Using Jensen's inequality and the independence of  $Z_{n+1}, Z_{n+2}, \dots, Z_{2n}$ , we bound:

$$\begin{aligned} \mathbb{E}_{Z_{n+1}, \dots, Z_{2n}} \left( \left| \mathbb{E}_Z(Z) - \frac{1}{n} \sum_{k=n+1}^{2n} Z_k \right| \right) &= \mathbb{E}_{Z_{n+1}, \dots, Z_{2n}} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} (\mathbb{E}_Z(Z) - Z_k) \right| \right) \\ &\leq \sqrt{\mathbb{E}_{Z_{n+1}, \dots, Z_{2n}} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} (\mathbb{E}_Z(Z) - Z_k) \right|^2 \right)} \\ &= \sqrt{\frac{1}{n^2} \mathbb{E}_{Z_{n+1}, \dots, Z_{2n}} \left( \sum_{k=n+1}^{2n} \sum_{k'=n+1}^{2n} (\mathbb{E}_Z(Z) - Z_k)(\mathbb{E}_Z(Z) - Z_{k'}) \right)} \\ &= \sqrt{\frac{1}{n^2} \mathbb{E}_{Z_{n+1}, \dots, Z_{2n}} \left( \sum_{k=n+1}^{2n} (\mathbb{E}_Z(Z) - Z_k)^2 \right)} \\ &= \sqrt{\frac{\mathbb{E}_{Z_{n+1}}((\mathbb{E}_Z(Z) - Z_{n+1})^2)}{n}} \end{aligned}$$

where the outer expectation in the last line is over the sample  $Z_{n+1}$ . Observe that  $\mathbb{E}_{Z_{n+1}}(\mathbb{E}_Z(Z) - Z_{n+1})^2 = \mathbb{E}_Z(\mathbb{E}_Z(Z) - Z)^2$  is the variance of the random variable  $Z$ . Hence we may bound:

$$\mathbb{E}_Z((\mathbb{E}_Z(Z) - Z)^2) \quad (77)$$

$$\begin{aligned} &\leq \mathbb{E}_Z(Z^2) \\ &= \mathbb{E}(\langle T_i^\epsilon(X) - X, T_j^\epsilon(X) - X \rangle^2) \\ &\leq \mathbb{E}(\|T_i^\epsilon(X) - X\|^2 \|T_j^\epsilon(X) - X\|^2) \\ &\leq (\mathbb{E}(\|T_i^\epsilon(X) - X\|^4) \mathbb{E}(\|T_j^\epsilon(X) - X\|^4))^{1/2} \\ &\leq (\mathbb{E}((\|T_i^\epsilon(X)\| + \|X\|)^4) \mathbb{E}((\|T_j^\epsilon(X)\| + \|X\|)^4))^{1/2} \\ &= \left( 2^4 \mathbb{E} \left( \left( \frac{\|T_i^\epsilon(X)\|}{2} + \frac{\|X\|}{2} \right)^4 \right) \cdot 2^4 \mathbb{E} \left( \left( \frac{\|T_j^\epsilon(X)\|}{2} + \frac{\|X\|}{2} \right)^4 \right) \right)^{1/2} \\ &\leq \left( 2^8 \mathbb{E} \left( \frac{1}{2} (\|T_i^\epsilon(X)\|^4 + \|X\|^4) \right) \mathbb{E} \left( \frac{1}{2} (\|T_j^\epsilon(X)\|^4 + \|X\|^4) \right) \right)^{1/2} \\ &= (2^6 (\mathbb{E}(\|T_i^\epsilon(X)\|^4) + \mathbb{E}(\|X\|^4)) (\mathbb{E}(\|T_j^\epsilon(X)\|^4) + \mathbb{E}(\|X\|^4)))^{1/2} \\ &= 8 ((\mathbb{E}(\|T_i^\epsilon(X)\|^4) + \mathbb{E}(\|X\|^4)) (\mathbb{E}(\|T_j^\epsilon(X)\|^4) + \mathbb{E}(\|X\|^4)))^{1/2} \end{aligned} \quad (78)$$

where we applied Jensen's inequality, two applications of the Cauchy-Schwarz inequality, the triangle inequality, and Jensen's inequality. By Lemma B.22, we have that

$$\mathbb{E}_{X \sim \mu}(\|T_i^\epsilon(X)\|^4) \leq \mathbb{E}_{Y \sim \nu_i}(\|Y\|^4), \quad \forall 1 \leq i \leq m.$$

Furthermore, by Proposition 4.1,  $\mu$  is a critical point, we may apply Proposition B.23 to bound

$$\mathbb{E}_{X \sim \mu}(\|X\|^4) \leq \mathbb{E}_{Y \sim \sum_{j=1}^m \lambda_j \nu_j}(\|Y\|^4)$$

and combining these we have upper bounded (78) by  $16 \max_k \mathbb{E}_{Y \sim \nu_k}(\|Y\|^4)$ . As  $\mathcal{V} \subset \mathcal{G}_\sigma(\Omega)$ , by Lemma A.2,  $\|Y\|$  is  $q_G \sqrt{d} \sigma$ -subgaussian. Hence by applying item 3 in Definition A.1, we have the upper bound

$$\begin{aligned} 16 \max_k \mathbb{E}_{Y \sim \nu_k}(\|Y\|^4) &\leq 16(2C_G q_G \sqrt{d} \sigma)^4 \\ &= 2^8 C_G^4 q_G^4 d^2 \sigma^4 \end{aligned}$$

Hence we have upper bounded 75:

$$\mathbb{E} \left( \left| \mathbb{E}(Z) - \frac{1}{n} \sum_{k=n+1}^{2n} Z_k \right| \right) \leq \frac{16C_G^2 q_G^2 d \sigma^2}{\sqrt{n}}. \quad (79)$$

We now bound (76):

$$\begin{aligned} &\mathbb{E} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} \langle T_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i^\epsilon(X_k) - X_k, \hat{T}_j^\epsilon(X_k) - X_k \rangle \right| \right) \\ &\leq \frac{1}{n} \sum_{k=n+1}^{2n} \mathbb{E} \left( \left| \langle T_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle - \langle \hat{T}_i^\epsilon(X_k) - X_k, \hat{T}_j^\epsilon(X_k) - X_k \rangle \right| \right) \\ &= \mathbb{E} \left( \left| \langle T_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle - \langle \hat{T}_i^\epsilon(X_k) - X_k, \hat{T}_j^\epsilon(X_k) - X_k \rangle \right| \right) \end{aligned}$$

for an arbitrary  $k \in \{n+1, n+2, \dots, 2n\}$ , where we have used the triangle inequality and the fact that  $X_k$  are identically distributed. Now,

$$\begin{aligned} &\mathbb{E} \left( \left| \langle T_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle - \langle \hat{T}_i^\epsilon(X_k) - X_k, \hat{T}_j^\epsilon(X_k) - X_k \rangle \right| \right) \\ &\leq \mathbb{E} \left( \left| \langle T_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle - \langle \hat{T}_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle \right| \right) \\ &+ \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - X_k \rangle - \langle \hat{T}_i^\epsilon(X_k) - X_k, \hat{T}_j^\epsilon(X_k) - X_k \rangle \right| \right) \\ &= \mathbb{E} \left( \left| \langle T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k), T_j^\epsilon(X_k) - X_k \rangle \right| \right) + \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X_k) - X_k, T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k) \rangle \right| \right) \\ &\leq \mathbb{E} \left( \|T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k)\| \|T_j^\epsilon(X_k) - X_k\| \right) + \mathbb{E} \left( \|\hat{T}_i^\epsilon(X_k) - X_k\| \|T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k)\| \right) \\ &\leq \sqrt{\mathbb{E} \left( \|T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k)\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X_k) - X_k\|^2 \right)} + \sqrt{\mathbb{E} \left( \|\hat{T}_i^\epsilon(X_k) - X_k\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k)\|^2 \right)} \\ &\leq \sqrt{2\mathbb{E} \left( \|T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k)\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X_k)\|^2 + \|X_k\|^2 \right)} + \sqrt{\mathbb{E} \left( \|\hat{T}_i^\epsilon(X_k) - X_k\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k)\|^2 \right)} \\ &\leq \sqrt{2\mathbb{E} \left( \|T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k)\|^2 \right) (M_2(\nu_j) + M_2(\mu))} + \sqrt{\mathbb{E} \left( \|\hat{T}_i^\epsilon(X_k) - X_k\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k)\|^2 \right)} \quad (80) \end{aligned}$$

In the above, we applied the triangle inequality, two applications of the Cauchy-Schwarz inequality, the triangle

inequality, Young's inequality and Lemma B.22 to bound  $\mathbb{E}(\|T_j^\epsilon(X_k)\|^2) \leq M_2(\nu_j)$ . We now control:

$$\begin{aligned}
 & \mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - X_k\|^2) \\
 & \leq \mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - T_i^\epsilon(X_k)\| + \|T_i^\epsilon(X_k) - X_k\|)^2 \\
 & \leq 2\mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - T_i^\epsilon(X_k)\|^2 + \|T_i^\epsilon(X_k) - X_k\|^2) \\
 & \leq 2\mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - T_i^\epsilon(X_k)\|^2) + 4\mathbb{E}(\|T_i^\epsilon(X_k)\|^2 + \|X_k\|^2) \\
 & = 2\mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - T_i^\epsilon(X_k)\|^2) + 4M_2(\nu_i) + 4M_2(\mu)
 \end{aligned}$$

where we applied the triangle inequality and Young's inequality twice, and the fact that  $\mathbb{E}(\|T_i^\epsilon(X_k)\|^2) \leq M_2(\nu_i)$ . Since  $\mu$  is a critical point of  $F_{\lambda, \nu}^\epsilon$ , we may apply Proposition B.23 to further bound  $M_2(\mu) \leq \max_{1 \leq k \leq m} M_2(\nu_k)$ . Summarizing, we have bounded (80) by:

$$\begin{aligned}
 & \sqrt{2\mathbb{E}(\|T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k)\|^2) \left( M_2(\nu_j) + \max_{1 \leq k \leq m} M_2(\nu_k) \right)} \\
 & + \sqrt{\left( 2\mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - T_i^\epsilon(X_k)\|^2) + 4M_2(\nu_i) + 4 \max_{1 \leq k \leq m} M_2(\nu_k) \right) \mathbb{E}(\|T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k)\|^2)} \\
 & \leq \sqrt{4C_G^2 q_G^2 d \sigma^2 \mathbb{E}(\|T_i^\epsilon(X_k) - \hat{T}_i^\epsilon(X_k)\|^2)} \\
 & + \sqrt{2 \left( \mathbb{E}(\|\hat{T}_i^\epsilon(X_k) - T_i^\epsilon(X_k)\|^2) + 16C_G^2 q_G^2 d \sigma^2 \right) \mathbb{E}(\|T_j^\epsilon(X_k) - \hat{T}_j^\epsilon(X_k)\|^2)}
 \end{aligned}$$

where we have applied Lemma A.2 followed by item 3 in Definition A.1 to upper bound the second moments.

Since  $X_k \sim \mu$  for  $k \geq n+1$  is independent of the samples  $X_1, \dots, X_n$  used to compute the maps  $\hat{T}_i^\epsilon$  and  $\hat{T}_j^\epsilon$ , we may apply (11) to obtain the upper bound (80) by:

$$2C_G q_G \sqrt{d\sigma} \sqrt{\theta(n)} + \sqrt{2\theta(n)^2 + 16C_G^2 q_G^2 d \sigma^2 \theta(n)}.$$

Combining this with the bound on (75), we have

$$\mathbb{E} \left( \left| [A_\mu^\epsilon]_{ij} - [\hat{A}_\mu^\epsilon]_{ij} \right| \right) \leq J d \sigma^2 \max \left\{ \frac{1}{\sqrt{n}}, \sqrt{\theta(n)} + \sqrt{\theta(n) + \theta(n)^2} \right\}, \quad (81)$$

where  $J$  is an absolute constant. □

We prove a similar lemma for  $S_\mu^\epsilon$ :

**Lemma D.2.** *Let  $\mu \in \mathcal{G}_\sigma(\Omega)$  and  $\mathcal{V} \subset \mathcal{G}_\sigma(\Omega)$ , and suppose that  $S_\mu^\epsilon$  has an eigenvalue of 0 with unique eigenvector  $\lambda_* \in \Delta^m$ . Suppose there exists an estimator  $\hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)$  satisfying (11) for all  $1 \leq j \leq m$ , and an estimator  $\hat{T}(\hat{\mu}^n)$  satisfying (13). Then:*

$$\mathbb{E}(|[S_\mu^\epsilon]_{ij} - [\hat{S}_\mu^\epsilon]_{ij}|) \leq \tilde{J} d \sigma^2 \max \left\{ \frac{1}{\sqrt{n}}, \sqrt{\theta(n)} + \sqrt{\theta(n) + \theta(n)^2} \right\}$$

where  $\hat{S}_\mu$  is the matrix  $\hat{M}_\mu$  in Algorithm 1 with  $\mathcal{F}_{\lambda, \nu}^\epsilon = S_{\lambda, \nu}^\epsilon$ , and  $\tilde{J}$  is an absolute constant.

*Proof.* For ease of notation, we write  $T_j^\epsilon := T_{\mu \rightarrow \nu_j}^\epsilon$ ,  $\hat{T}_j^\epsilon := \hat{T}(\hat{\mu}^n, \hat{\nu}_j^n)$ ,  $T^\epsilon := T_{\mu \rightarrow \mu}^\epsilon$  and  $\hat{T}^\epsilon := \hat{T}(\hat{\mu}^n)$ . We bound:

$$\begin{aligned}
 & \mathbb{E} \left( \left| \int \langle T_i^\epsilon - T^\epsilon, T_j^\epsilon - T^\epsilon \rangle d\mu - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i^\epsilon(X_k) - \hat{T}^\epsilon(X_k), \hat{T}_j^\epsilon(X_k) - \hat{T}^\epsilon(X_k) \rangle \right| \right) \\
 & \leq \mathbb{E} \left( \left| \int \langle T_i^\epsilon - T^\epsilon, T_j^\epsilon - T^\epsilon \rangle d\mu - \frac{1}{n} \sum_{k=n+1}^{2n} \langle T_i^\epsilon(X_k) - T^\epsilon(X_k), T_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle \right| \right) \\
 & + \mathbb{E} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} \langle T_i^\epsilon(X_k) - T^\epsilon(X_k), T_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i^\epsilon(X_k) - T^\epsilon(X_k), \hat{T}_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle \right| \right) \\
 & + \mathbb{E} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i^\epsilon(X_k) - T^\epsilon(X_k), \hat{T}_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i^\epsilon(X_k) - \hat{T}^\epsilon(X_k), \hat{T}_j^\epsilon(X_k) - \hat{T}^\epsilon(X_k) \rangle \right| \right). \quad (82)
 \end{aligned}$$

By the same sequence of steps as used to bound (75) by (78), we may upper bound the first term in (82) by

$$\frac{1}{\sqrt{n}} \sqrt{8 \left( (\mathbb{E}(\|T_i^\epsilon(X)\|^4) + \mathbb{E}(\|T^\epsilon(X)\|^4)) (\mathbb{E}(\|T_j^\epsilon(X)\|^4) + \mathbb{E}(\|T^\epsilon(X)\|^4)) \right)^{1/2}}. \quad (83)$$

By Lemma B.22,  $(T_i^\epsilon)_\#(\mu) \leq_c \nu_i \in \mathcal{G}_\sigma(\Omega)$  for all  $1 \leq i \leq m$ , and  $(T^\epsilon)_\#(\mu) \leq_c \mu \in \mathcal{G}_\sigma(\Omega)$ . Hence we may apply Lemma B.21, Lemma A.2 and item 3 in Definition A.1 as in the proof of Lemma D.1 to conclude that (83) is upper bounded by  $\frac{16C_G^2 d_G^2 d\sigma^2}{\sqrt{n}}$ .

The second term in (82) may be bounded in the same manner as (76). By the triangle inequality and noting that  $X_1, X_2, \dots, X_{2n}$  are all i.i.d. we have:

$$\begin{aligned}
 & \mathbb{E} \left( \left| \frac{1}{n} \sum_{k=n+1}^{2n} \langle T_i^\epsilon(X_k) - T^\epsilon(X_k), T_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i^\epsilon(X_k) - T^\epsilon(X_k), \hat{T}_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle \right| \right) \\
 & \leq \frac{1}{n} \sum_{k=n+1}^{2n} \mathbb{E} \left( \left| \langle T_i^\epsilon(X_k) - T^\epsilon(X_k), T_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle - \langle \hat{T}_i^\epsilon(X_k) - T^\epsilon(X_k), \hat{T}_j^\epsilon(X_k) - T^\epsilon(X_k) \rangle \right| \right) \\
 & \leq \mathbb{E} \left( \left| \langle T_i^\epsilon(X) - T^\epsilon(X), T_j^\epsilon(X) - T^\epsilon(X) \rangle - \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), \hat{T}_j^\epsilon(X) - T^\epsilon(X) \rangle \right| \right)
 \end{aligned}$$

where  $X \sim \mu$  is independent of  $X_1, X_2, \dots, X_n$ . By the triangle inequality, we may upper bound this by:

$$\begin{aligned}
 & \mathbb{E} \left( \left| \langle T_i^\epsilon(X) - T^\epsilon(X), T_j^\epsilon(X) - T^\epsilon(X) \rangle - \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), T_j^\epsilon(X) - T^\epsilon(X) \rangle \right| \right) \\
 & + \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), T_j^\epsilon(X) - T^\epsilon(X) \rangle - \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), \hat{T}_j^\epsilon(X) - T^\epsilon(X) \rangle \right| \right) \\
 & \leq \mathbb{E} \left( \left| \langle T_i^\epsilon(X) - \hat{T}_i^\epsilon(X), T_j^\epsilon(X) - T^\epsilon(X) \rangle \right| \right) + \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), T_j^\epsilon(X) - \hat{T}_j^\epsilon(X) \rangle \right| \right) \\
 & \leq \mathbb{E} \left( \|T_i^\epsilon(X) - \hat{T}_i^\epsilon(X)\| \|T_j^\epsilon(X) - T^\epsilon(X)\| \right) + \mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T^\epsilon(X)\| \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\| \right) \\
 & \leq \mathbb{E} \left( \|T_i^\epsilon(X) - \hat{T}_i^\epsilon(X)\| (\|T_j^\epsilon(X)\| + \|T^\epsilon(X)\|) \right) + \mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T^\epsilon(X)\| \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\| \right) \\
 & \leq \sqrt{\mathbb{E} \left( \|T_i^\epsilon(X) - \hat{T}_i^\epsilon(X)\|^2 \right) (2\mathbb{E}(\|T_j^\epsilon(X)\|^2) + 2\mathbb{E}(\|T^\epsilon(X)\|^2))} \\
 & \quad + \sqrt{\mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T^\epsilon(X)\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\|^2 \right)} \\
 & \leq \sqrt{2\theta(n) (M_2(\nu_j) + \mathbb{E}(\|T^\epsilon(X)\|^2))} \\
 & \quad + \sqrt{\mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T^\epsilon(X)\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\|^2 \right)} \\
 & \leq \sqrt{2\theta(n) (M_2(\nu_j) + \mathbb{E}(\|T^\epsilon(X)\|^2))} \\
 & \quad + \sqrt{2\mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T_i^\epsilon(X)\|^2 + \|T_i^\epsilon(X) - T^\epsilon(X)\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\|^2 \right)} \\
 & \leq \sqrt{2\theta(n) (M_2(\nu_j) + \mathbb{E}(\|T^\epsilon(X)\|^2))} \\
 & \quad + \sqrt{2\mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T_i^\epsilon(X)\|^2 + 2\|T_i^\epsilon(X)\|^2 + 2\|T^\epsilon(X)\|^2 \right) \mathbb{E} \left( \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\|^2 \right)} \\
 & \leq \sqrt{2\theta(n) (M_2(\nu_j) + M_2(\mu))} \\
 & \quad + \sqrt{\left( 2\mathbb{E} \left( \|\hat{T}_i^\epsilon(X) - T_i^\epsilon(X_n)\|^2 \right) + 4M_2(\nu_\ell) + 4M_2(\mu) \right) \mathbb{E} \left( \|T_j^\epsilon(X) - \hat{T}_j^\epsilon(X)\|^2 \right)} \\
 & \leq \max_{1 \leq \ell \leq m} \sqrt{2(M_2(\nu_j) + M_2(\mu))\theta(n)} + \sqrt{(2\theta(n) + 4M_2(\nu_\ell) + 4M_2(\mu))\theta(n)}.
 \end{aligned}$$

By the triangle inequality and using the fact that  $X_k, k \in \{1, \dots, 2n\}$  are all i.i.d, the third term in (82) is upper bounded by:

$$\begin{aligned}
 & \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), \hat{T}_j^\epsilon(X) - T^\epsilon(X) \rangle - \langle \hat{T}_i^\epsilon(X) - \hat{T}^\epsilon(X), \hat{T}_j^\epsilon(X) - \hat{T}^\epsilon(X) \rangle \right| \right) \\
 & = \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X) - T^\epsilon(X), \hat{T}_j^\epsilon(X) - T^\epsilon(X) \rangle - \langle \hat{T}_i^\epsilon(X) - \hat{T}^\epsilon(X), \hat{T}_j^\epsilon(X) - T^\epsilon(X) \rangle \right. \right. \\
 & \quad \left. \left. + \langle \hat{T}_i^\epsilon(X) - \hat{T}^\epsilon(X), \hat{T}_j^\epsilon(X) - T^\epsilon(X) \rangle - \langle \hat{T}_i^\epsilon(X) - \hat{T}^\epsilon(X), \hat{T}_j^\epsilon(X) - \hat{T}^\epsilon(X) \rangle \right| \right)
 \end{aligned}$$

where  $X \sim \mu$  is independent of  $X_1, X_2, \dots, X_n$ . By the triangle inequality, we may bound this as:

$$\begin{aligned}
 & \mathbb{E} \left( \left| \langle \hat{T}^\epsilon(X) - T^\epsilon(X), \hat{T}_j^\epsilon(X) - T_j^\epsilon(X) \rangle \right| \right) + \mathbb{E} \left( \left| \langle \hat{T}_i^\epsilon(X) - \hat{T}^\epsilon(X), \hat{T}^\epsilon(X) - T^\epsilon(X) \rangle \right| \right) \\
 & \leq \mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\| \left\| \hat{T}_j^\epsilon(X) - T_j^\epsilon(X) \right\| \right) + \mathbb{E} \left( \left\| \hat{T}_i^\epsilon(X) - \hat{T}^\epsilon(X) \right\| \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\| \right) \\
 & = \mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\| \left\| \hat{T}_j^\epsilon(X) - T_j^\epsilon(X) + T_j^\epsilon(X) - T^\epsilon(X) \right\| \right) \\
 & \quad + \mathbb{E} \left( \left\| \hat{T}_i^\epsilon(X) - T_i^\epsilon(X) + T_i^\epsilon(X) - \hat{T}^\epsilon(X) \right\| \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\| \right) \\
 & \leq \mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\| \left( \left\| \hat{T}_j^\epsilon(X) - T_j^\epsilon(X) \right\| + \left\| T_j^\epsilon(X) - T^\epsilon(X) \right\| \right) \right) \\
 & \quad + \mathbb{E} \left( \left( \left\| \hat{T}_i^\epsilon(X) - T_i^\epsilon(X) \right\| + \left\| T_i^\epsilon(X) - \hat{T}^\epsilon(X) \right\| \right) \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\| \right) \\
 & \leq \sqrt{\mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\|^2 \right) \mathbb{E} \left( 2 \left\| \hat{T}_j^\epsilon(X) - T_j^\epsilon(X) \right\|^2 + 2 \left\| T_j^\epsilon(X) - T^\epsilon(X) \right\|^2 \right)} \\
 & \quad + \sqrt{\mathbb{E} \left( 2 \left\| \hat{T}_i^\epsilon(X) - T_i^\epsilon(X) \right\|^2 + 2 \left\| T_i^\epsilon(X) - \hat{T}^\epsilon(X) \right\|^2 \right) \mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\|^2 \right)} \\
 & \leq \sqrt{\mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\|^2 \right) \mathbb{E} \left( 2 \left\| \hat{T}_j^\epsilon(X) - T_j^\epsilon(X) \right\|^2 + 4 \left\| T_j^\epsilon(X) \right\|^2 + 4 \left\| T^\epsilon(X) \right\|^2 \right)} \\
 & \quad + \sqrt{\mathbb{E} \left( 2 \left\| \hat{T}_i^\epsilon(X) - T_i^\epsilon(X) \right\|^2 + 2 \left\| T_i^\epsilon(X) - \hat{T}^\epsilon(X) \right\|^2 \right) \mathbb{E} \left( \left\| \hat{T}^\epsilon(X) - T^\epsilon(X) \right\|^2 \right)} \\
 & \leq \sqrt{2\theta(n) (2\theta(n) + 4M_2(\nu_j) + 4M_2(\mu))} \\
 & \quad + \sqrt{\mathbb{E} \left( \left( 2 \left\| \hat{T}_i^\epsilon(X) - T_i^\epsilon(X) \right\|^2 + 4 \left\| T_i^\epsilon(X) - T^\epsilon(X) \right\|^2 + 4 \left\| T^\epsilon(X) - \hat{T}^\epsilon(X) \right\|^2 \right) \theta(n) \right)} \\
 & \leq \sqrt{2\theta(n) (2\theta(n) + 4M_2(\nu_j) + 4M_2(\mu))} \\
 & \quad + \sqrt{\mathbb{E} \left( \left( 2 \left\| \hat{T}_i^\epsilon(X) - T_i^\epsilon(X) \right\|^2 + 8 \left\| T_i^\epsilon(X) \right\|^2 + 8 \left\| T^\epsilon(X) \right\|^2 + 4 \left\| T^\epsilon(X) - \hat{T}^\epsilon(X) \right\|^2 \right) \theta(n) \right)} \\
 & \leq \sqrt{2\theta(n) (2\theta(n) + 4M_2(\nu_j) + 4M_2(\mu))} + \sqrt{(2\theta(n) + 8M_2(\nu_i) + 8M_2(\mu) + 4\theta(n)) \theta(n)}.
 \end{aligned}$$

Applying the bound  $M_2(\mu) \leq 2C_G^2 q_G^2 \sigma^2 d$ , which applies since  $\mu$  is  $\sigma$  subgaussian, and combining constants not depending on  $\sigma$  and  $d$ , we obtain:

$$\begin{aligned}
 & \mathbb{E} \left( \left| \int \langle T_j^\epsilon - T^\epsilon, T_i^\epsilon - T^\epsilon \rangle d\mu - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_j^\epsilon(X_k) - \hat{T}^\epsilon(X_k), \hat{T}_i^\epsilon(X_k) - \hat{T}^\epsilon(X_k) \rangle \right| \right) \\
 & \leq \tilde{J} d \sigma^2 \max \left\{ \frac{1}{\sqrt{n}}, \sqrt{\theta(n)} + \sqrt{\theta(n) + \theta(n)^2} \right\}.
 \end{aligned}$$

□

Theorem 4.2 immediately follows from Lemmas D.1 and D.2 and the following lemma, which allows us to apply (expected) bounds between the entries of two matrices to the (expected) distance between their eigenvectors, which we have extracted from Corollary 2 in (Werenski et al., 2022):

**Lemma D.3.** *Let  $M \in \mathbb{R}^{m \times m}$  be a positive semidefinite matrix with eigenvalue 0 with multiplicity 1 and associated unique eigenvector  $\lambda_* \in \Delta^m$ . Suppose that  $\hat{M} \in \mathbb{R}^{m \times m}$  is a random positive semidefinite matrix such that*

$$\mathbb{E}(|[M]_{ij} - [\hat{M}]_{ij}|) \leq \mathcal{Q}, \text{ for all } i, j = 1, \dots, m. \quad (84)$$

Let  $\hat{\lambda} := \arg \min_{\lambda \in \Delta^m} \lambda^\top \hat{M} \lambda$ . Then:

$$\mathbb{E}(\|\hat{\lambda} - \lambda_*\|^2) \leq \frac{8m^3 \mathcal{Q}}{\alpha_2},$$

where  $\alpha_2$  is the second smallest eigenvalue of  $M$ .

*Proof.* We decompose  $\hat{\lambda}$  as  $\hat{\beta}\lambda_* + \hat{\lambda}_\perp$ , where  $\hat{\beta} \in \mathbb{R}$  and  $\hat{\lambda}_\perp$  is orthogonal to  $\lambda_*$ .

$$\begin{aligned} \mathbb{E} \left( \|\hat{\lambda} - \lambda_*\|_2^2 \right) &= \mathbb{E} \left( \|\hat{\lambda} - \hat{\beta}\lambda_* + \hat{\beta}\lambda_* - \lambda_*\|_2^2 \right) \\ &\leq 2\mathbb{E} \left( \|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2 \right) + 2\mathbb{E} \left( \|(\hat{\beta} - 1)\lambda_*\|_2^2 \right) \end{aligned} \quad (85)$$

Observe that  $\hat{\lambda}_\perp$  is in the span of the orthonormal eigenvectors  $v_2, \dots, v_m$  of  $M$ , which we take to be ordered by their eigenvalues  $0 \leq \alpha_2 \leq \dots \leq \alpha_m$ . Furthermore, as  $\lambda_* \in \Delta^m$  is the unique eigenvector with eigenvalue 0 by assumption,  $0 < \alpha_2$ . Then:

$$\begin{aligned} \mathbb{E} \left( \|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2 \right) &= \mathbb{E} \left( \|\hat{\lambda}_\perp\|_2^2 \right) \\ &= \mathbb{E} \left( \sum_{j=2}^m |v_j^\top \hat{\lambda}_\perp|^2 \right) \\ &\leq \frac{1}{\alpha_2} \mathbb{E} \left( \sum_{j=2}^m \alpha_j |v_j^\top \hat{\lambda}_\perp|^2 \right) \\ &= \frac{1}{\alpha_2} \mathbb{E} \left( |(\hat{\lambda}_\perp)^\top M \hat{\lambda}_\perp| \right) \\ &= \frac{1}{\alpha_2} \mathbb{E} \left( |(\hat{\lambda})^\top M \hat{\lambda}| \right) \end{aligned}$$

We bound:

$$\begin{aligned} \mathbb{E} \left( \hat{\lambda}^\top M \hat{\lambda} \right) &= \mathbb{E} \left( \hat{\lambda}^\top (M - \hat{M}) \hat{\lambda} \right) + \mathbb{E} \left( \hat{\lambda}^\top \hat{M} \hat{\lambda} \right) \\ &\leq \mathbb{E} \left( |\hat{\lambda}^\top (M - \hat{M}) \hat{\lambda}| \right) + \mathbb{E} \left( \lambda_*^\top \hat{M} \lambda_* \right) \\ &= \mathbb{E} \left( |\hat{\lambda}^\top (M - \hat{M}) \hat{\lambda}| \right) + \mathbb{E} \left( \lambda_*^\top (\hat{M} - M) \lambda_* \right) \end{aligned}$$

where we used the minimality of  $\hat{\lambda}$  and the fact that  $\lambda_*^\top M \lambda_* = 0$ . We may rewrite these as the following sums:

$$\begin{aligned} &\mathbb{E} \left( \left| \sum_{i,j=1}^m [\hat{\lambda}]_i [\hat{\lambda}]_j [M]_{ij} - [\hat{M}]_{ij} \right| \right) + \mathbb{E} \left( \sum_{i,j=1}^m [\lambda_*]_i [\lambda_*]_j ([\hat{M}]_{ij} - [M]_{ij}) \right) \\ &\leq \sum_{i,j=1}^m \mathbb{E} \left( |[\hat{\lambda}]_i [\hat{\lambda}]_j| | [M]_{ij} - [\hat{M}]_{ij} | \right) + \sum_{i,j=1}^m [\lambda_*]_i [\lambda_*]_j \mathbb{E} \left( |[\hat{M}]_{ij} - [M]_{ij}| \right) \\ &\leq 2m^2 \mathcal{Q}, \end{aligned} \quad (86)$$

where we used the triangle inequality, the fact that  $\hat{\lambda}, \lambda_* \in \Delta^m$  and (84). This lets us bound (85) by:

$$\begin{aligned} &\frac{4m^2}{\alpha_2} \mathcal{Q} + 2\mathbb{E} \left( (\hat{\beta} - 1)^2 \|\lambda_*\|_2^2 \right) \\ &\leq \frac{4m^2}{\alpha_2} \mathcal{Q} + 2\mathbb{E} \left( (\hat{\beta} - 1)^2 \right), \end{aligned}$$

where we used the fact that  $\|\lambda_*\|_2^2 \leq 1$ . Since  $\hat{\lambda} = \hat{\beta}\lambda_* + \hat{\lambda}_\perp$  with  $\hat{\lambda}, \lambda_* \in \Delta^m$ , we have that:

$$\begin{aligned} \mathbb{E}((\hat{\beta} - 1)^2) &= \mathbb{E} \left( \left( \sum_{j=1}^m [\hat{\lambda}_\perp]_j \right)^2 \right) \\ &\leq m\mathbb{E}(\|\hat{\lambda}_\perp\|_2^2) \end{aligned}$$



by the Cauchy-Schwarz inequality, and hence  $\mathbb{E}((1 - \hat{\beta})^2) \leq m\mathbb{E}(\|\hat{\lambda}_\perp\|_2^2) \leq \frac{2m^3}{\alpha_2} \mathcal{Q}$  by another application of (86). We conclude that

$$\mathbb{E}(\|\hat{\lambda} - \lambda_*\|_2^2) \leq \frac{4m^2}{\alpha_2} \mathcal{Q} + \frac{4m^3}{\alpha_2} \mathcal{Q} \leq \frac{8m^3}{\alpha_2} \mathcal{Q}.$$

□

**Proof of Theorem 4.2:** Immediately follows from substituting the bounds from Lemmas D.1 and D.2 as  $\mathcal{Q}$  in Lemma D.3. □

**Proof of Corollary 4.3:** The proof follows from the following result:

**Theorem D.4.** (Theorem 4 in (Rigollet and Stromme, 2022)) Let  $\Omega$  be a bounded set in  $\mathbb{R}^d$ ,  $\epsilon > 0$ , and  $\mu, \nu \in \mathcal{P}_2(\Omega)$ . Then

$$\mathbb{E}(\|T_{\mu \rightarrow \nu}^\epsilon - T_{\hat{\mu}^n \rightarrow \hat{\nu}^n}^\epsilon\|_{L^2(\mu)}^2) < \frac{C_{\Omega, \epsilon}}{n}$$

where  $C_{\Omega, \epsilon}$  is a constant only depending on  $|\Omega|$  and  $\epsilon$ .

By Theorem D.4, we bound

$$\mathbb{E}(\|T_{\mu \rightarrow \nu}^\epsilon - T_{\hat{\mu}^n \rightarrow \hat{\nu}^n}^\epsilon\|_{L^2(\mu)}^2) \leq \frac{C_{\Omega, \epsilon}}{n}.$$

Similarly, we may bound

$$\mathbb{E}(\|T_{\mu \rightarrow \mu}^\epsilon - T_{\hat{\mu}_1 \rightarrow \hat{\mu}_2}^\epsilon\|_{L^2(\mu)}^2) \leq \frac{3C_{\Omega, \epsilon}}{n}.$$

Hence we may apply Theorem 4.2 with  $\theta(n) = \frac{1}{n}$ . As  $\sqrt{\frac{1}{n} + \frac{1}{n^2}} \leq \sqrt{\frac{2}{n}}$  for  $n \geq 1$ , we may bound  $\max\left\{\frac{1}{\sqrt{n}}, \sqrt{\frac{1}{n} + \frac{1}{n^2}}\right\} \leq \frac{2}{\sqrt{n}}$ . The second moments of  $\nu_1, \dots, \nu_m, \mu$  are uniformly bounded by  $|\Omega|^2$ , and hence we may collapse the various constants into a single constant  $\tilde{C}_{\Omega, \epsilon}$ . □

**Proof of Corollary 4.4:** It is well known that  $c$ -strongly log-concave measures are  $\frac{L}{\sqrt{c}}$ -subgaussian for some universal constant  $L$  (see e.g., Theorem 5 in (Werenski et al., 2023)). Similarly to Corollary 4.3, the proof follows from the following result:

**Theorem D.5.** (Theorem 7 in (Werenski et al., 2023)) Let  $\mu$  be  $\sigma$ -subgaussian and let  $\nu$  be  $c$ -strongly log concave with  $\mathbb{E}(\nu) = 0$ . Then there exists an estimator  $\mathcal{L}$  such that:

$$\mathbb{E}(\|T_{\mu \rightarrow \nu}^\epsilon - \mathcal{L}(\hat{\mu}^n, \hat{\nu}^n)\|_{L^2(\mu)}^2) \leq \frac{K_{d, \sigma, \epsilon, c}}{n^{1/3}},$$

where  $K_{d, \sigma, \epsilon, c}$  is a constant depending on  $c, \sigma, d$  and  $\epsilon$ .

Since  $\mu$  is a critical point of  $F_{\lambda, \nu}^\epsilon$ , it is  $\frac{L}{\sqrt{c}}$ -subgaussian by Proposition 2.7, and hence the constant factors only depend on  $c, d$  and  $\epsilon$ . □

## D.2 Proof of Proposition 4.5

To establish Proposition 4.5, we will require the following technical results:

**Lemma D.6.** (Lemma A.1 in (Werenski et al., 2024b)) Let  $\Omega \subset \mathbb{R}^d$  be bounded, and let  $\mu, \nu \in \mathcal{P}_2(\Omega)$ . Then:

$$\|T_{\mu \rightarrow \nu}^\epsilon(x) - T_{\mu \rightarrow \nu}^\epsilon(y)\| \leq M_{\Omega, \epsilon} \|x - y\|,$$

for all  $x, y \in \Omega$ , where  $M_{\Omega, \epsilon} = \frac{4|\Omega|^2}{\epsilon}$ .

**Proposition D.7.** (Proposition 4 in (Kassraie et al., 2024) and Corollary 3.3 in (Divol et al., 2024)) Let  $\Omega \subset \mathbb{R}^d$  be bounded, and let  $\mu_1, \mu_2, \nu \in \mathcal{P}_2(\Omega)$ . Then:

$$\|T_{\nu \rightarrow \mu_1}^\epsilon - T_{\nu \rightarrow \mu_2}^\epsilon\|_{L^2(\nu)} \leq K_{\Omega, \epsilon} OT_2(\mu_1, \mu_2) \quad (87)$$

$$\|T_{\mu_1 \rightarrow \nu}^\epsilon - T_{\mu_2 \rightarrow \nu}^\epsilon\|_{L^2(\mu_1)} \leq K_{\Omega, \epsilon} OT_2(\mu_1, \mu_2) \quad (88)$$

where  $K_{\Omega, \epsilon} = \frac{3^{1/2}|\Omega|}{\epsilon^{1/2}} + 1$ .

From Proposition D.7 we obtain the following lemma:

**Lemma D.8.** Suppose  $\Omega \subset \mathbb{R}^d$  is bounded. Let  $A_\mu^\epsilon$  and  $S_\mu^\epsilon$  be defined as in Proposition 4.1. Then for any  $1 \leq i, j \leq m$ :

$$\begin{aligned} |[A_\mu^\epsilon]_{ij} - [A_\rho^\epsilon]_{ij}| &\leq H_{\Omega, \epsilon} OT_2(\rho, \mu) \\ |[S_\mu^\epsilon]_{ij} - [S_\rho^\epsilon]_{ij}| &\leq H_{\Omega, \epsilon} OT_2(\mu, \rho) \end{aligned}$$

where  $H_{\Omega, \epsilon}$  depends on  $|\Omega|$  and  $\epsilon$ .

*Proof.* We begin by bounding  $|[A_\mu^\epsilon]_{ij} - [A_\rho^\epsilon]_{ij}|$  for any  $1 \leq i, j \leq m$ :

$$\begin{aligned} &|[A_\mu^\epsilon]_{ij} - [A_\rho^\epsilon]_{ij}| \\ &= |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle Id - T_{\rho \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\ &\leq |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \\ &\quad + |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\ &\quad + |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)} - \langle Id - T_{\rho \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\ &= |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \nu_j}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \\ &\quad + |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\ &\quad + |\langle T_{\mu \rightarrow \nu_i}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \end{aligned} \quad (89)$$

where we applied the triangle inequality. We bound the first term in (89):

$$\begin{aligned} |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \nu_j}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| &\leq \|Id - T_{\mu \rightarrow \nu_i}^\epsilon\|_{L^2(\mu)} \|T_{\mu \rightarrow \nu_j}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon\|_{L^2(\mu)} \\ &\leq |\Omega| \|T_{\mu \rightarrow \nu_j}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon\|_{L^2(\mu)} \\ &\leq |\Omega| K_{\Omega, \epsilon} OT_2(\mu, \rho), \end{aligned}$$

where we applied the Cauchy-Schwarz inequality, the diameter bound on  $\Omega$  and (88) to control  $\|T_{\mu \rightarrow \nu_j}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon\|_{L^2(\mu)}$ . The third term can be bounded identically.

To bound the second term, we first observe that, by Lemma D.6,  $T_{\mu \rightarrow \nu_i}^\epsilon$  and  $T_{\rho \rightarrow \nu_j}^\epsilon$  are  $M_{\Omega, \epsilon}$ -Lipschitz vector-valued maps from  $\Omega$  to the convex hull of  $\Omega$ . As  $\Omega$  is bounded,  $x \mapsto \langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x), x - T_{\rho \rightarrow \nu_j}^\epsilon(x) \rangle$  is  $2|\Omega|(1 + M_{\Omega, \epsilon})$ -Lipschitz as a function of  $x \in \Omega$ . Indeed, let  $x, y \in \Omega$ . Then:

$$\begin{aligned} &|\langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x), x - T_{\rho \rightarrow \nu_j}^\epsilon(x) \rangle - \langle y - T_{\mu \rightarrow \nu_i}^\epsilon(y), y - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\ &\leq |\langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x), x - T_{\rho \rightarrow \nu_j}^\epsilon(x) \rangle - \langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x), y - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\ &\quad + |\langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x), y - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle - \langle y - T_{\mu \rightarrow \nu_i}^\epsilon(y), y - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \end{aligned} \quad (90)$$

$$\begin{aligned} &= |\langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x), x - T_{\rho \rightarrow \nu_j}^\epsilon(x) - y + T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\ &\quad + |\langle x - T_{\mu \rightarrow \nu_i}^\epsilon(x) - y + T_{\mu \rightarrow \nu_i}^\epsilon(y), y - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\ &\leq \|x - T_{\mu \rightarrow \nu_i}^\epsilon(x)\| \|x - T_{\rho \rightarrow \nu_j}^\epsilon(x) - y + T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \\ &\quad + \|x - T_{\mu \rightarrow \nu_i}^\epsilon(x) - y + T_{\mu \rightarrow \nu_i}^\epsilon(y)\| \|y - T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \end{aligned} \quad (91)$$

$$\leq |\Omega| \left( \|x - T_{\mu \rightarrow \nu_i}^\epsilon(x) - y + T_{\mu \rightarrow \nu_i}^\epsilon(y)\| + \|x - T_{\rho \rightarrow \nu_j}^\epsilon(x) - y + T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \right) \quad (92)$$

$$\leq |\Omega| \left( 2\|x - y\| + \|T_{\mu \rightarrow \nu_i}^\epsilon(x) - T_{\mu \rightarrow \nu_i}^\epsilon(y)\| + \|T_{\rho \rightarrow \nu_j}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \right) \quad (93)$$

$$\leq 2|\Omega|(1 + M_{\Omega, \epsilon})\|x - y\|. \quad (94)$$

In (90) we applied the triangle inequality, in (91) the Cauchy-Schwarz inequality, in (92) we used the bounds  $\|x - T_{\mu \rightarrow \nu_i}^\epsilon(x)\|, \|y - T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \leq |\Omega|$ , in (93) the triangle inequality and finally in (94) we applied Lemma D.6. Since  $\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle$  is  $2|\Omega|(1 + M_{\Omega, \epsilon})$ -Lipschitz, we may apply Kantorovich-Rubinstein duality for the Wasserstein-1 distance (see e.g., 5.16 in (Villani, 2009)) to produce the bound:

$$\begin{aligned} & |\langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle Id - T_{\mu \rightarrow \nu_i}^\epsilon, Id - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\ & \leq 2|\Omega|(1 + M_{\Omega, \epsilon})OT_1(\mu, \rho) \\ & \leq 2|\Omega|(1 + M_{\Omega, \epsilon})OT_2(\mu, \rho) \end{aligned}$$

where in the last line we applied Jensen's inequality. Combining the bounds on these three terms, we may conclude that  $|[A_\mu^\epsilon]_{ij} - [A_\rho^\epsilon]_{ij}| \leq 2|\Omega|(M_{\Omega, \epsilon} + K_{\Omega, \epsilon} + 1)OT_2(\mu, \rho)$ .

We now treat the Sinkhorn case. For any  $\mu \in \mathcal{P}_2(\Omega)$  and  $1 \leq i, j \leq m$ , we have:

$$\begin{aligned} & |[S_\mu^\epsilon]_{ij} - [S_\rho^\epsilon]_{ij}| \\ & = |\langle T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\ & \leq |\langle T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \end{aligned} \quad (95)$$

$$+ |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \quad (96)$$

$$+ |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \quad (97)$$

$$+ |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \quad (98)$$

$$+ |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \quad (99)$$

We first bound (95):

$$\begin{aligned} & |\langle T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \\ & = |\langle T_{\mu \rightarrow \mu}^\epsilon - T_{\rho \rightarrow \rho}^\epsilon, T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \\ & \leq \|T_{\mu \rightarrow \mu}^\epsilon - T_{\rho \rightarrow \rho}^\epsilon\|_{L^2(\mu)} \|T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon\|_{L^2(\mu)} \\ & \leq \|T_{\mu \rightarrow \mu}^\epsilon - T_{\rho \rightarrow \rho}^\epsilon\|_{L^2(\mu)} |\Omega| \\ & \leq (\|T_{\mu \rightarrow \mu}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon\|_{L^2(\mu)} + \|T_{\mu \rightarrow \nu_i}^\epsilon - T_{\rho \rightarrow \rho}^\epsilon\|_{L^2(\mu)}) |\Omega| \\ & \leq 2K_{\Omega, \epsilon} |\Omega| OT_2(\mu, \rho), \end{aligned}$$

where we applied the Cauchy-Schwarz inequality, the fact that the image of  $T_{\mu \rightarrow \mu}^\epsilon$  and  $T_{\mu \rightarrow \nu_j}^\epsilon$  are contained in  $\Omega$ , followed by (87) and (88). The same bound applies to (96).

A similar argument applies to (97):

$$\begin{aligned} & |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \\ & = |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)}| \\ & \leq \|T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_i}^\epsilon\|_{L^2(\mu)} \|T_{\rho \rightarrow \rho}^\epsilon - T_{\mu \rightarrow \nu_j}^\epsilon\|_{L^2(\mu)} \\ & \leq |\Omega| K_{\Omega, \epsilon} OT_2(\mu, \rho) \end{aligned}$$

where we applied (88) and the fact that  $\|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\mu \rightarrow \nu_i}^\epsilon(x)\| \leq |\Omega|$  for all  $x \in \Omega$ . The same bound applies to (98).

Finally, to bound (99), we observe that  $x \rightarrow \langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x), T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(x) \rangle$  is  $4|\Omega|M_{\Omega,\epsilon}$ -Lipschitz:

$$\begin{aligned}
 & |\langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x), T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(x) \rangle - \langle T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_i}^\epsilon(y), T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\
 & \leq |\langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x), T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(x) \rangle - \langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x), T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\
 & \quad + |\langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x), T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle - \langle T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_i}^\epsilon(y), T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\
 & = |\langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x), T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y) + T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\
 & \quad + |\langle T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y) + T_{\rho \rightarrow \nu_i}^\epsilon(y), T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_j}^\epsilon(y) \rangle| \\
 & \leq \|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x)\| \|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y) + T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \\
 & \quad + \|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y) + T_{\rho \rightarrow \nu_i}^\epsilon(y)\| \|T_{\rho \rightarrow \rho}^\epsilon(y) - T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \\
 & \leq |\Omega| \left( \|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y) + T_{\rho \rightarrow \nu_i}^\epsilon(y)\| + \|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y) + T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \right) \\
 & \leq |\Omega| \left( 2\|T_{\rho \rightarrow \rho}^\epsilon(x) - T_{\rho \rightarrow \rho}^\epsilon(y)\| + \|T_{\rho \rightarrow \nu_i}^\epsilon(x) - T_{\rho \rightarrow \nu_i}^\epsilon(y)\| + \|T_{\rho \rightarrow \nu_j}^\epsilon(x) - T_{\rho \rightarrow \nu_j}^\epsilon(y)\| \right) \\
 & \leq 4|\Omega|M_{\Omega,\epsilon}\|x - y\|,
 \end{aligned}$$

where we applied Lemma D.6 in the final line. Thus we may apply Kantorovich-Rubinstein duality once more to conclude that

$$\begin{aligned}
 & |\langle T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\mu)} - \langle T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_i}^\epsilon, T_{\rho \rightarrow \rho}^\epsilon - T_{\rho \rightarrow \nu_j}^\epsilon \rangle_{L^2(\rho)}| \\
 & \leq 4|\Omega|M_{\Omega,\epsilon}OT_1(\mu, \rho) \\
 & \leq 4|\Omega|M_{\Omega,\epsilon}OT_2(\mu, \rho).
 \end{aligned}$$

Combining the bounds on (95)-(99), we have that:

$$|[S_\mu^\epsilon]_{ij} - [S_\rho^\epsilon]_{ij}| \leq |\Omega|(6K_{\Omega,\epsilon} + 4M_{\Omega,\epsilon})OT_2(\mu, \rho).$$

Taking the maximum of these to be  $H_{\Omega,\epsilon}$ , we conclude.  $\square$

**Proof of Proposition 4.5:** We apply Lemma D.3 with  $M = A_\mu^\epsilon$  and  $\hat{M} = A_\rho^\epsilon$  (note this is deterministic) and can take  $\mathcal{Q} = H_{\Omega,\epsilon}OT_2(\mu, \rho)$  by Lemma D.8. The same holds for  $S_\mu^\epsilon$ .  $\square$

## E IMPLEMENTATION DETAILS

### E.1 Synthesis Implementation

#### Algorithm 3: Wasserstein gradient descent for $\phi$ (WGD)

**Data:** Initial measure  $\mu_0 \in \mathcal{P}_2(\Omega)$ ;

Functional  $\phi : \mathcal{P}_2(\Omega) \rightarrow \mathbb{R}$ ;

Step-size  $\eta > 0$ ;

Iterations  $k$ .

```

1  $\mathcal{T}_{[0:0]} := \mathcal{T}_0 \leftarrow Id - \eta \nabla \delta \phi(\mu_0)$  ; // Compute Wasserstein-2 gradient of  $\phi$  at  $\mu_0$ ;
2  $\mu_1 \leftarrow [\mathcal{T}_{[0:0]}]^\#(\mu_0)$  ; // Transport  $\mu_0$  according to  $\mathcal{T}_{[0:0]}$ 
3 for  $0 \leq i \leq k-1$  do
4    $\mathcal{T}_i \leftarrow Id - \eta \nabla \delta \phi(\mu_i)$  ; // Compute Wasserstein-2 gradient of  $\phi$  at  $\mu_i$ 
5    $\mathcal{T}_{[0:i]} \leftarrow \mathcal{T}_i \circ \mathcal{T}_{[0:i-1]}$  ; // Update by composing maps
6    $\mu_{i+1} \leftarrow [\mathcal{T}_{[0:i]}]^\#(\mu_0)$  ; // Transport  $\mu_0$  according to composed map
7 end

```

**8 Output:**  $\mu_k$ .

We detail the implementation of the free support synthesis algorithm used in Section 4.1 (see Algorithm 3 for pseudocode). For the case of  $F_{\lambda,\nu}^\epsilon$ , this corresponds to a modified version of Algorithm 3 in (Cuturi and Doucet,

2014), without the density update step. More generally, we may view Algorithm 3 as a forward Euler discretization of the Wasserstein-2 gradient flow (Ambrosio and Savaré, 2006) for a functional  $\phi : \mathcal{P}_2(\Omega) \rightarrow \mathbb{R}$  with some fixed step-size  $\eta$ . It is also implemented for  $F_{\lambda, \nu}^\epsilon$  for the popular optimal transport package Python Optimal Transport (POT) (Flamary and Courty, 2023). See Theorem 4.1 in (Shen et al., 2020) for a proof of convergence in the bounded setting.

We use Algorithm 3 to compute (approximate) critical points of  $\mathcal{F}_{\lambda, \nu}^\epsilon$ . In each trial, we construct  $\hat{\nu}_j^n$  by sampling  $n = 20000$  samples according to  $\nu_j$ . We construct  $\mu_0$  by sampling an empirical measure (on  $n = 20000$  samples) from  $Unif([0, 1]^5)$ . We apply Algorithm 3 to  $F_{\lambda, \hat{\nu}^n}^\epsilon$  with initial measure  $\mu_0$  and  $\hat{\nu}^n := \{\hat{\nu}_j^n\}_{j=1}^m$ , with stopping criterion  $\|Id - \eta \nabla \delta F_{\lambda, \hat{\nu}^n}^\epsilon(\mu_i)\|_{L^2(\mu_i)} < 10^{-7}$ . We denote the output by  $\mu^*$ .

## E.2 Sample Complexity Experiment Implementation Details

**1-Dimensional Gaussians:** We build a random set of reference measures

$$\mathcal{V} = \{\mathcal{N}(M_1, \sigma_1^2), \mathcal{N}(M_2, \sigma_2^2), \dots, \mathcal{N}(M_m, \sigma_m^2)\}$$

for  $m \geq 1$  where  $M_j$  are i.i.d.  $\sim Unif([a_j, b_j])$  and  $\sigma_i$  i.i.d.  $\sim Unif([c_j, d_j])$ . We then sample  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_m$  i.i.d.  $\sim Unif([0, 1])$ , and define  $\lambda_j := \tilde{\lambda}_j / \sum_{i=1}^m \tilde{\lambda}_i$  for all  $j$ . We then compute the exact mean  $\bar{M}$  and variance  $S$  of the true minimizer  $\arg \min_{\mu \in \mathcal{P}_2(\mathbb{R})} F_{\lambda, \nu}^\epsilon(\mu)$ , which is known in closed form via Theorem 2 in (Janati et al., 2020a) (we solve for the variance numerically using `scipy.optimize.least_squares`). We then run Algorithm 1 over  $n$  samples, with  $\mu = \mathcal{N}(\bar{M}, S^2)$  and  $\mathcal{V}$ ,  $\lambda$  and  $\epsilon$  chosen as above to recover a set of estimated coefficients  $\hat{\lambda}^n$ . We compute the  $\ell^2$  error between  $\hat{\lambda}^n$  and  $\lambda$ .

In the case of  $m = 2$ , we sample  $M_1 \sim Unif([0, 2])$ ,  $M_2 \sim Unif([3, 5])$  and  $\sigma_1, \sigma_2 \sim Unif([2, 4])$ . In the case of  $m = 3$ , we sample  $M_1, M_2, \sigma_1, \sigma_2$  as before, and sample  $M_3 \sim Unif([-2, 0])$  and  $\sigma_3 \sim Unif([2, 4])$ . We found that the  $\ell^2$ -distance between the coefficients essentially does not decay in  $n$  (top right). This is due to the fact that in one dimension, the matrix  $A_\mu^\epsilon$  is often singular, which violates the conditions in Theorem 4.2, leading to problems with non-unique recovery of coefficients. Despite this, we empirically show that the expected  $OT_2^2$ -error between the barycenters associated to  $\lambda$  and  $\hat{\lambda}^n$  rapidly decays as  $n$  increases, implying successful recovery of a *valid* set of barycentric coefficients even when they are non-unique (see Appendix F).

**5-Dimensional Gaussians:** We work with a fixed, random choice of  $\lambda$ , which we obtain by sampling  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_m$  i.i.d.  $\sim Unif([0, 1])$ , and define  $\lambda_j := \tilde{\lambda}_j / \sum_{i=1}^m \tilde{\lambda}_i$ . For each trial, we set  $\epsilon = 1$ , and generate each Gaussian  $\nu_j := \mathcal{N}(M_j, A_j)$  as follows. We sample  $M_1 \sim Unif([0.2, 0.4]^5)$ ,  $M_2 \sim Unif([0.4, 0.6]^5)$  and  $M_3 \sim Unif([0.6, 0.8]^5)$ .

For each  $j$ , we generate covariance matrix  $\Sigma_j = R v_j v_j^\top + 0.7 * Id$ , where  $v_j \in \mathbb{R}^5$  is generated by sampling  $[v_j]_i \sim Unif([0.2, 0.5])$  for  $1 \leq i \leq 5$  and  $R$  is a uniformly random  $d$ -dimensional rotation matrix. We work with a fixed choice of coefficients  $\lambda$  across all trials, which we generate by sampling  $\tilde{\lambda}_i \sim Unif([0, 1])$  and setting  $\lambda_i = \tilde{\lambda}_i / \sum_{j=1}^m \tilde{\lambda}_j$ . Unlike the 1D case, we no longer have access to the closed-form solution for the entropic barycenter. We implement a free support synthesis algorithm (see details in Appendix E.1) to compute an approximate fixed point  $\mu^*$ . We then sample  $2n$  samples from  $\mu^*$  and resample  $n$  samples from  $\{\nu_j\}_{j=1}^m$  and use these as inputs for Algorithm 1, as we let  $n$  range over  $[10, 20, 40, \dots, 10240]$ .

**5-Dimensional Cubes:** We work with a fixed, random choice of  $\lambda$ , which we obtain by sampling  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_m$  i.i.d.  $\sim Unif([0, 1])$ , and define  $\lambda_j := \tilde{\lambda}_j / \sum_{i=1}^m \tilde{\lambda}_i$ . For each trial, we set  $\epsilon = 1$ , and generate our reference measures using random translations of  $Unif([0, 2]^5)$ . Specifically, we let  $\nu_j = Unif([0, 2]^5 + M_j)$  where  $M_1 \sim Unif([0.2, 0.4]^5)$ ,  $M_2 \sim Unif([0.4, 0.6]^5)$  and  $M_3 \sim Unif([0.6, 0.8]^5)$ .

We solve the minimization problem  $\min_{\lambda \in \Delta^3} \lambda^\top \hat{A}_{\mu^*}^\epsilon \lambda$  using the `cp.Minimize` function from the convex optimization module `cvx.py`, and denote the output  $\hat{\lambda}^n$ .

## E.3 Point Cloud Classification Implementation Details

**Analysis for the barycenter functional:** We compute the barycentric coefficients for the unregularized barycenter functional by computing a set of maps  $\{T_{\mu \rightarrow \nu_j}^0\}_{j=1}^m$ , where  $T_{\mu \rightarrow \nu_j}^0 := \mathbb{E}_{\zeta_j}(Y|X=x)$  is the conditional

**Algorithm 4:** Classification Experiment

---

**Data:**  $m$  classes;  
 Labeled point clouds for each class:  $\mathcal{A}^y = \{P_1^y, P_2^y, \dots, P_q^y\}$  for  $1 \leq y \leq m$ ;  
 Number of reference measures per class:  $b$ ;  
 Train-test split numbers:  $n_{train} + n_{test} = q$ ;  
 Regularization parameter:  $\epsilon$ ;  
 Functional:  $\mathcal{F}_{\lambda, \nu}^\epsilon \in \{F_{\lambda, \nu}^\epsilon, S_{\lambda, \nu}^\epsilon\}$ ;  
 Number of iterations:  $K$ ;

---

```

1 for  $1 \leq i \leq K$  do
2   for  $1 \leq y \leq m$  do
3     Partition  $\mathcal{A}^y$  into training and testing sets,  $\mathcal{A}_{train}^y$  and  $\mathcal{A}_{test}^y$ , with  $|\mathcal{A}_{train}^y| = n_{train}$  and
4      $|\mathcal{A}_{test}^y| = n_{test}$ ;
5     Randomly sample  $b$  point clouds  $\mathcal{V}^y := \{P_{j_1}^y, P_{j_2}^y, \dots, P_{j_b}^y\} \subset \mathcal{A}_{train}^y$ ;
6   end
7   for  $1 \leq y \leq m$  do
8     for  $Q \in \mathcal{A}_{test}^y$  do
9        $\tilde{y}_Q \leftarrow$  Output of Algorithm 2 with inputs: unlabelled point cloud  $Q$ ,  $m$  classes, labeled point
10      clouds  $\{\mathcal{V}^y\}_{y=1}^m$ , regularization parameter  $\epsilon$ , and functional  $\mathcal{F}_{\lambda, \nu}^\epsilon$ ;
11       $s_{i, Q}^y \leftarrow \begin{cases} 1 & \text{if } \tilde{y}_Q = y \\ 0 & \text{otherwise} \end{cases}$ ;
12    end
13  end
14   $s_i \leftarrow \frac{1}{mn_{test}} \sum_{y=1}^m \sum_{Q \in \mathcal{A}_{test}^y} s_{i, Q}^y$ 
15 end
16  $s \leftarrow \frac{1}{K} \sum_{i=1}^K s_i$ ;
17 Output: Overall estimated classification accuracy  $s$ .
    
```

---

expectation of an optimal coupling  $\zeta_j \in \Pi(\mu, \nu_j)$  for all  $1 \leq j \leq m$ , which we use as inputs to define a matrix  $A_\mu^0$ , defined by  $[A_\mu^0]_{ij} = \langle Id - T_{\mu \rightarrow \nu_i}^0, Id - T_{\mu \rightarrow \nu_j}^0 \rangle_{L^2(\mu)}$ . We solve the corresponding optimization problem  $\min_{\lambda \in \Delta^m} \lambda^\top A_\mu^0 \lambda$ .

**Analysis for the doubly-regularized functional:** For the doubly regularized problem, we must assume that  $\mu$  admits a density  $P_\mu$ . Then the corresponding matrix is  $D_\mu^{\epsilon, \tau}$ , defined by  $[D_\mu^{\epsilon, \tau}]_{ij} = \langle Id - T_{\mu \rightarrow \nu_i}^\epsilon + \tau \nabla \log P_\mu, Id - T_{\mu \rightarrow \nu_j}^\epsilon + \tau \nabla \log P_\mu \rangle_{L^2(\mu)}$ ; see Appendix H for details. In practice, we estimate  $\nabla \log P_\mu$  using the  $\nu$ -method implemented in (Zhou et al., 2020). Our choice of kernel is the curl-free inverse multiquadric kernel with bandwidth 5.0, and regularization strength  $10^{-5}$  (see (Zhou et al., 2020) and the associated implementation for details). We additionally convolved  $\mu$  with isotropic Gaussian noise with standard deviation 0.05, which empirically improved the performance of our method.

## F RECOVERY WITH NON-UNIQUE BARYCENTRIC COEFFICIENTS

We further study the analysis problem for three one-dimensional Gaussians. In Section 4.1, we showed that coefficient recovery fails in this setting. We claim this is due to the fact that there are multiple sets of coefficients giving rise to the same barycenter, i.e.,  $A_\mu^\epsilon$  is often singular.

Let  $\mu^*$  be the barycenter with respect to the original coefficients  $\lambda$  and  $\hat{\mu}^n$  be the barycenter with the recovered coefficients  $\hat{\lambda}^n$  as  $n$  ranges over the same set of values. To estimate  $\mathbb{E}(OT_2^2(\mu^*, \hat{\mu}^n))$ , we estimate the mean and variance of  $\hat{\mu}^n$  (denoted  $\hat{m}, \hat{\sigma}$ ) and applying the closed-form formula for the optimal transport distance between two Gaussians (Chen et al., 2018), which specializes in this case to  $OT_2^2(\mu^*, \hat{\mu}^n) = (\bar{\mu} - \hat{m})^2 + S^2 + \hat{\sigma}^2 - 2\hat{\sigma}S$ . We average this quantity over 100 trials to estimate  $\mathbb{E}(OT_2^2(\mu^*, \hat{\mu}^n))$ . We plot our results in Figure 2, and see

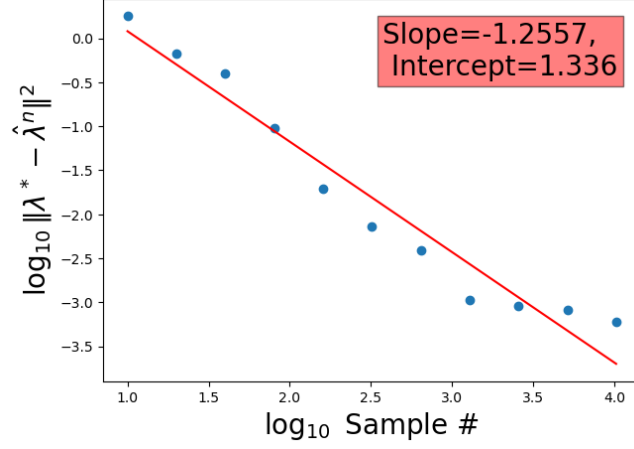


Figure 2: Average log squared- $OT_2$ -loss for three random 1D Gaussian measures with random weights,  $\epsilon = 2$ .

that  $\mathbb{E}(OT_2^2(\mu^*, \hat{\mu}^n))$  decays faster than  $n^{-1}$ . Hence from the perspective of barycentric coefficients, our method is successful even in the case of non-unique recovery.

## G SYNTHESIS AND ANALYSIS WITH MNIST

We give an in-depth example of both synthesis and analysis using the entropic barycenter functional and probability measures constructed from the MNIST data set (LeCun, 1998). Given images of digits  $D_1, D_2, \dots, D_m$ , we construct reference measures  $\nu_1, \nu_2, \dots, \nu_m$  using Algorithm 5. We fix once and for all  $\epsilon = 0.0002$ . For a given  $\lambda \in \Delta^m$  and an initial measure  $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ , we run Algorithm 3 ( $k = 1000$ ,  $\eta = 0.025$ ) to optimize  $F_{\lambda, \nu}^\epsilon$ , outputting an approximate critical point  $\mu_k^\lambda$ . We then compute the matrix  $A_{\mu_k}^\epsilon$  (see Proposition 4.1) and perform the minimization  $\min_{\lambda \in \Delta^3} \lambda^\top A_{\mu_k}^\epsilon \lambda$  using the `cp.Minimize` function from the convex optimization module `cvx.py`, and denote the output  $\hat{\lambda}^i$ . We perform this for  $\{\lambda^i\}_{i=1}^{21}$  ranging over 21 evenly spaced points in the simplex, which correspond to all tuples of the form  $\{(\frac{i_1}{5}, \frac{i_2}{5}, \frac{i_3}{5}) \mid i_1, i_2, i_3 \in \mathbb{Z}_{\geq 0}, i + j + k = 5\}$ . In Figure 3 we visualize our results. We convert  $\mu_k^{\lambda^i}$  to an image using Algorithm 6 which we position at  $\lambda^i$  on the left hand triangle. At the corners of the triangle, we display the original images which were converted to reference measures. In the right triangle, we plot  $\lambda^i$  (in red) and  $\hat{\lambda}^i$  (in blue).

From Figure 3 we see that the fixed points produced by our synthesis method intuitively interpolate between the reference images. We note that there is some degree of shrinkage compared to the original digits, which could be due to the fact that fixed points are dominated in convex order by  $\sum_{j=1}^3 \lambda_j \nu_j$  by Proposition B.23, where  $\{\nu_j\}_{j=1}^3$  are the reference measures. From the triangle on the right we see that the recovered coefficients  $\hat{\lambda}^i$  closely match the true coefficients  $\lambda^i$ . To verify this, we compute the average  $\ell^2$ -error  $\frac{1}{21} \sum_{i=1}^{21} \|\lambda^i - \hat{\lambda}^i\|_2$  and find that it is approximately  $1.269 \times 10^{-5}$ .

### G.1 Image-Measure Processing

To convert images to measures and to visualize measures in  $\mathbb{R}^2$  as images, we use Algorithms 5 and 6 respectively, which are adapted from the methods in (Werenski et al., 2024a). Algorithm 5 produces a measure by simply normalizing the intensities of an image (represented as a matrix). Algorithm 6 first applies a Gaussian kernel with bandwidth  $b = 0.1$  to each of the data points, and builds a matrix  $K$  of shape  $rd \times rd$ , where  $r = 5$  is a resolution parameter and  $d = 28$  is the desired output size such that the  $K_{ij}$  is the density at  $(\frac{i}{rd}, \frac{j}{rd})$ . The matrix is then reduced by removing entries below  $\max_{i,j} K_{ij} * \ell$ , where  $\ell = 0.0002$  is a tolerance parameter. We then reduce  $K$  to a  $d \times d$  matrix  $I$  and normalize intensities to return an image.

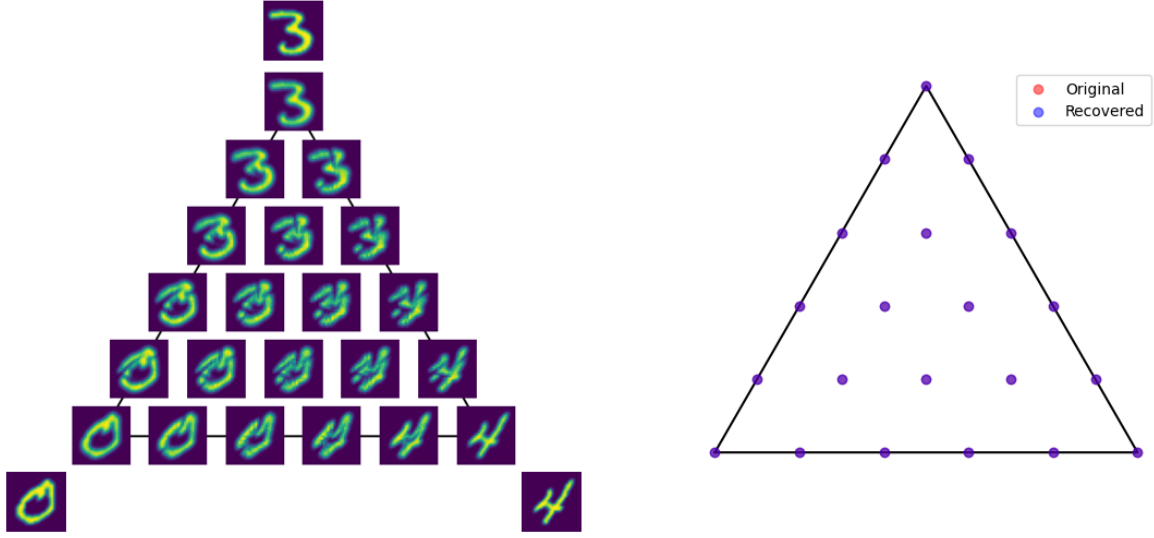


Figure 3: We show synthesized entropy-regularized barycenters on the left, with the reference measures on the corners. On the right, original coefficients are shown as well as what our analysis algorithm recovers, showing good recovery.

**Algorithm 5:** Image to Measure Conversion

**Data:** Matrix representation of an image  $\mathbf{I} \in \mathbb{R}_+^{d \times d}$ .

1 Set  $s \leftarrow \sum_{i,j=1}^d \mathbf{I}_{ij}$ ;

2 **Output:**  $\frac{1}{s} \sum_{i,j=1}^d \mathbf{I}_{ij} \delta_{(i/d, j/d)}$ ;

## H DERIVATION OF QUADRATIC PROGRAM FOR DOUBLY-REGULARIZED FUNCTIONAL

In this section, we derive the form of the quadratic program used for solving the analysis problem for the  $(\epsilon, \tau)$ -doubly-regularized barycenter functional,  $\mathcal{D}_{\lambda, \nu}^{\epsilon, \tau} := F_{\lambda, \nu}^{\epsilon} + \tau H(\mu)$ , under the assumption that the measure  $\mu$  being analyzed is a critical point of the doubly-regularized functional. Here,  $H$  denotes the negative differential entropy:

$$H(\mu) := \begin{cases} \int_{\mathbb{R}^d} \log(P_{\mu}(x)) P_{\mu}(x) dx, & \text{if } \mu \text{ is absolutely continuous with respect to the Lebesgue measure on } \Omega \\ +\infty, & \text{otherwise} \end{cases}$$

where  $P_{\mu}$  denotes the density of  $\mu$ .

Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$  be absolutely continuous with connected support. For convenience, we will also assume that  $\mu$  has  $C^\infty(\mathbb{R}^d)$  density  $P_{\mu}$ . Then we compute:

$$\begin{aligned} \delta \mathcal{D}_{\lambda, \nu}^{\epsilon, \tau}(\mu) &= \delta F_{\lambda, \nu}^{\epsilon}(\mu) + \delta \tau H(\mu) \\ &= \sum_{j=1}^m \lambda_j f_{\mu \rightarrow \nu_j}^{\epsilon} + \tau \log P_{\mu}. \end{aligned}$$



**Algorithm 6:** Measure to Image Conversion

**Data:** Discrete measure  $\nu \in \mathcal{P}_2([0, 1]^2)$   
 Output size  $d$ ;  
 Resolution  $r$ ;  
 Bandwidth  $b$ ;  
 Tolerance  $\ell$ .

1 Define  $S = \{S_1, S_2, \dots, S_n\} \leftarrow \text{supp}(\nu)$ ,  $S_i \in [0, 1]^2$ ;  
 2 Define  $b_k \leftarrow \nu(S_k)$ ;  
 3 Create  $K \in \mathbb{R}_+^{rd \times rd}$  with

$$K_{ij} \leftarrow \sum_{k=1}^n b_k \exp\left(-\frac{\|(i/rd, j/rd) - S_k\|_2^2}{b^2}\right)$$

4 Set  $K \leftarrow \left(\sum_{i,j=1}^{rd} K_{ij}\right)^{-1} \cdot K$ ;  
 5 **for**  $1 \leq i, j \leq rd$  **do**  
 6 |  $K_{ij} \leftarrow K_{ij} \cdot \mathbf{1}[K_{ij} > \max_{i,j} K_{ij} * \ell]$ ;  
 7 **end**  
 8 Create  $I \in \mathbb{R}^{d \times d}$  with

$$I_{ij} \leftarrow \sum_{k,l=1}^r K_{(i-1)*r+k, (l-1)*r+l}$$

9 **Output:**  $\left(\sum_{i,j=1}^d I_{ij}\right)^{-1} \cdot I$ ;

Since  $\mu$  is a critical point, we have that  $\|\nabla \delta \mathcal{D}_{\lambda, \nu}^{\epsilon, \tau}(\mu)\|_{L^2(\mu)}^2 = 0$ , which implies that:

$$Id - \sum_{j=1}^m \lambda_j T_{\mu \rightarrow \nu_j}^\epsilon + \tau \nabla \log P_\mu = 0$$

$\mu$ -almost everywhere. We may then compute:

$$\begin{aligned} \|\nabla \delta \mathcal{D}_{\lambda, \nu}^{\epsilon, \tau}(\mu)\|_{L^2(\mu)}^2 &= \|Id - \sum_{j=1}^m \lambda_j T_{\mu \rightarrow \nu_j}^\epsilon + \tau \nabla \log P_\mu\|_{L^2(\mu)}^2 \\ &= \int \langle Id - \sum_{j=1}^m \lambda_j T_{\mu \rightarrow \nu_j}^\epsilon + \tau \nabla \log P_\mu, Id - \sum_{k=1}^m \lambda_k T_{\mu \rightarrow \nu_k}^\epsilon + \tau \nabla \log P_\mu \rangle d\mu \\ &= \int \sum_{j,k=1}^m \lambda_j \lambda_k \langle Id - T_{\mu \rightarrow \nu_j}^\epsilon + \tau \nabla \log P_\mu, Id - T_{\mu \rightarrow \nu_k}^\epsilon + \tau \nabla \log P_\mu \rangle d\mu \\ &= \sum_{j,k=1}^m \lambda_j \lambda_k [D_\mu^{\epsilon, \tau}]_{jk} = \lambda^\top D_\mu^{\epsilon, \tau} \lambda, \end{aligned}$$

where  $D_\mu^{\epsilon, \tau}$  is defined by  $[D_\mu^{\epsilon, \tau}]_{jk} := \langle Id - T_{\mu \rightarrow \nu_j}^\epsilon + \tau \nabla \log P_\mu, Id - T_{\mu \rightarrow \nu_k}^\epsilon + \tau \nabla \log P_\mu \rangle_{L^2(\mu)}$ . Hence if  $\mu$  is a critical point for  $\mathcal{D}_{\lambda, \nu}^{\epsilon, \tau}$ , then solving the analysis problem is equivalent to solving  $0 = \lambda^\top D_\mu^{\epsilon, \tau} \lambda$  for  $\lambda \in \Delta^m$ . Note that the assumed smoothness of  $P_\mu$  and the guaranteed smoothness of the potentials, together with the connectedness of  $\text{supp}(\mu)$  imply the reverse implication as well:  $\|\nabla \delta \mathcal{D}_{\lambda, \nu}^{\epsilon, \tau}(\mu)\|_{L^2(\mu)} = 0$  implies  $\mu$  is a critical point.

## I APPLICATION: POINT CLOUD COMPLETION

We include a preliminary demonstration of an application of the synthesis and analysis methods to point cloud completion. We work with the PointCloud-C data set (Ren et al., 2022) which consists of point cloud

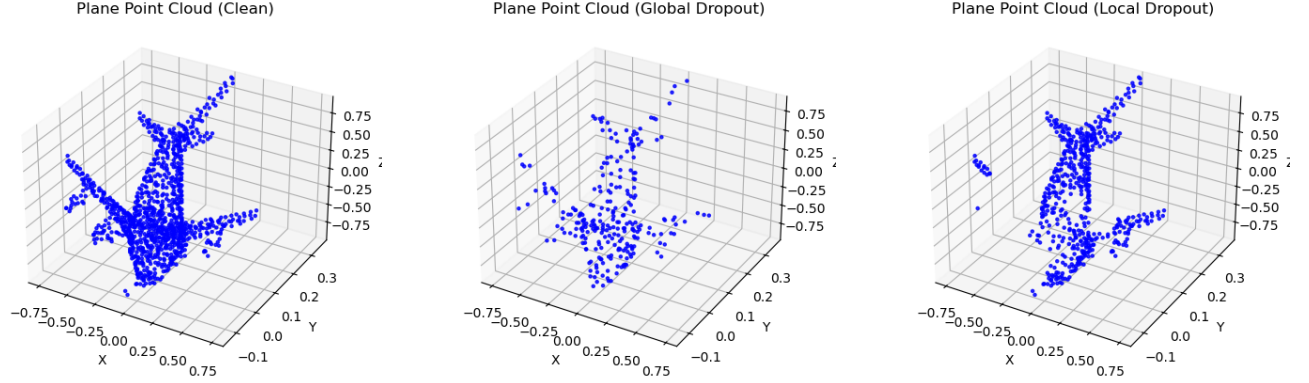


Figure 4: (Left to right) Point clouds from `clean`, `dropout_global_4` and `dropout_local_4`.

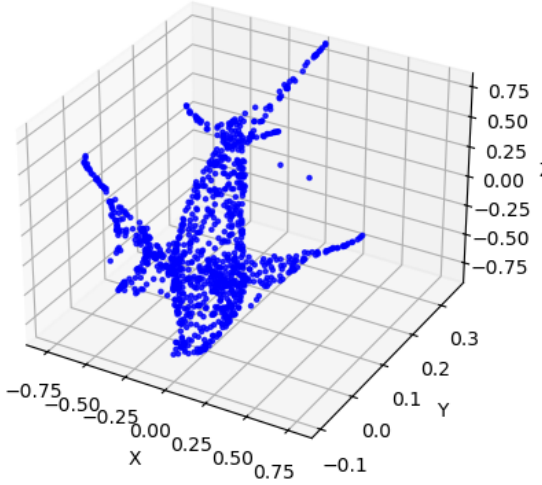
representations of objects corrupted by various regimes of noise and occlusions. In this demonstration, we use data from the `clean`, `dropout_global_4` and `dropout_local_4` datasets. The `clean` dataset consists of point clouds with 1024 points. The `dropout_global_4` dataset is obtained by selecting point clouds from the `clean` dataset and (uniformly) randomly removing 75% of the data points. The `dropout_local_4` dataset by selecting a random set of points  $x_1, x_2, \dots, x_\ell$  ( $\ell \leq 8$ ) and removing the  $N_i$  nearest points to each  $x_i$ , where  $\sum_{i=1}^\ell N_i = 500$  (see (Ren et al., 2022) for more details). We give some examples in Figure 4. The object of point cloud completion is to take a partial point cloud and to generate a new point cloud which as closely as possible matches the original point cloud. Unlike most methods in the literature, our method does not require explicit training.

Our method proceeds as follows. We let  $\mu_0$  be the uniform measure on a corrupted point cloud. We then select a set of random clean point clouds  $\nu_1, \dots, \nu_5$  from the same class of objects. We solve the analysis problem for  $\mu_0$  with Algorithm 1, producing a set of coefficients  $\lambda \in \Delta^5$ . We then use these coefficients and reference measures to synthesize a new point cloud  $\bar{\mu}_0$ , using Algorithm 3, initialized at an empirical measure (sampled from a uniform measure on the 3D cube) supported on 1024 points. We remark that this method is similar to one developed in (Werenski et al., 2022) for image completion, with a key difference being that their method operates in a fixed-support setting, as opposed to free-support like ours.

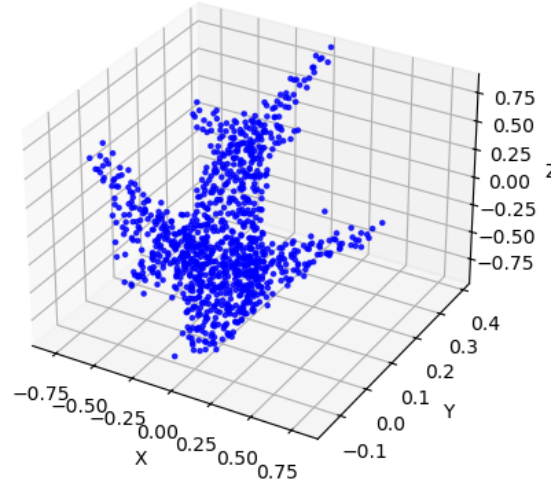
In Figure 5 we show results for our completion method using both  $F_{\lambda, \nu}^\epsilon$  and  $S_{\lambda, \nu}^\epsilon$ . We see that in all cases our methods are able to produce a completed point cloud that roughly match the original. The reconstructions using  $F_{\lambda, \nu}^\epsilon$  appear to be more singular than the original, whereas  $S_{\lambda, \nu}^\epsilon$  appear to be more diffuse. However some fine details of the original point cloud are not recovered, such as the second turbine on the left wing.

To give a quantitative comparison, we compute the  $OT_2$ -distance between the (uniform measures on the) clean point cloud and the reconstruction. We see that the Sinkhorn reconstruction method performs slightly better in both cases with respect to this metric.

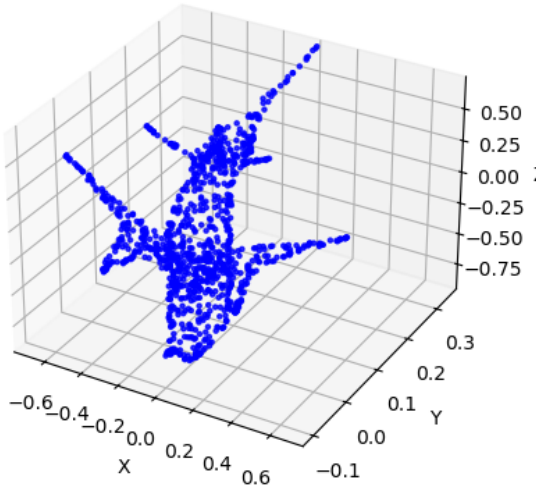
Reconstructed Pointcloud with Global Dropout  
(Entropic OT,  $\varepsilon = 0.003$ ,  $OT_2$ -cost=0.0534)



Reconstructed Pointcloud with Global Dropout  
(Sinkhorn,  $\varepsilon = 0.003$ ,  $OT_2$ -cost=0.0520)



Reconstructed Pointcloud with Local Dropout  
(Entropic OT,  $\varepsilon = 0.003$ ,  $OT_2$ -cost=0.0929)



Reconstructed Pointcloud with Local Dropout  
(Sinkhorn,  $\varepsilon = 0.003$ ,  $OT_2$ -cost=0.0890)

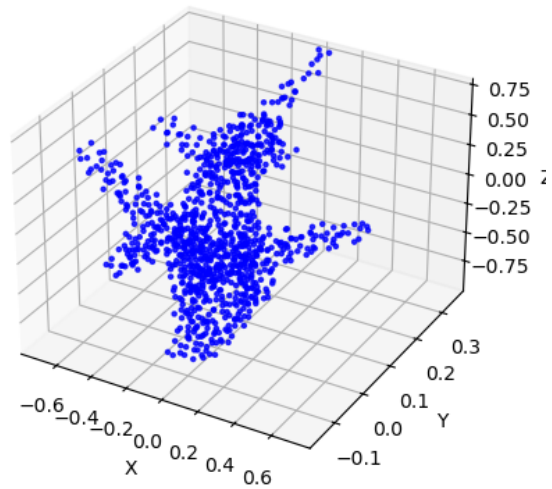


Figure 5: (Top left) Global dropout,  $F_{\lambda, \nu}^\varepsilon$  reconstruction, (top right) global dropout  $S_{\lambda, \nu}^\varepsilon$  reconstruction, (bottom left) local dropout,  $F_{\lambda, \nu}^\varepsilon$  reconstruction, (bottom right) local dropout  $S_{\lambda, \nu}^\varepsilon$  reconstruction.  $OT_2$ -cost denotes the  $OT_2$  distance between the (uniform measures on the) clean point cloud and the reconstruction.