# On the Finite-Sample Bias of Minimizing Expected Wasserstein Loss Between Empirical Distributions

**Cheongjae Jang**
Hanyang University
cjjang@hanyang.ac.kr

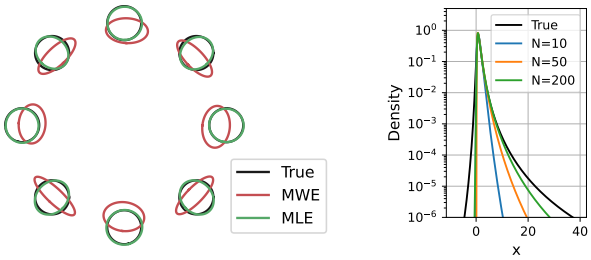
**Yung-Kyun Noh**
Hanyang University
Korea Institute for Advanced Study
nohyung@hanyang.ac.kr

## Abstract

We show that minimizing the expected Wasserstein loss between empirical distributions can lead to biased parameter estimates in the finite-sample regime. Remarkably, such bias arises even in well-specified settings where both empirical distributions are drawn from the same parametric family: unlike maximum likelihood estimation—understood here as maximizing the expected log-likelihood—optimizing one parameter while fixing another fails to recover the true fixed value. We derive closed-form expressions for the expected Wasserstein loss in one dimension and, focusing on location–scale models, provide an analytic characterization of the bias. This analysis reveals that finite-sample bias occurs whenever the expected loss varies along the diagonal subspace where parameter values coincide, and we propose a simple correction scheme that removes this effect. We extend our analysis to misspecified models and the Sinkhorn divergence, demonstrating that finite-sample bias persists in more practical settings. Experiments on synthetic and real data confirm that stochastic optimization of Wasserstein-based objectives converges to biased solutions, and validate the effectiveness of the proposed correction scheme.

## 1 INTRODUCTION

Recent advances in computational optimal transport have made Wasserstein distances a powerful tool for quantifying differences between probability distributions, with growing impact across a wide range of machine learning and statistical applications (Peyré et al., 2019). In particular, parameter estimation methods based on minimizing the Wasserstein distance have drawn increasing attention as an alternative to classical likelihood-based approaches (Bassetti et al., 2006; Bernton et al., 2019b). Unlike maximum likelihood estimation—which can be considered as minimizing the KL divergence between the empirical distribution and a parametric likelihood model—Wasserstein-based methods offer robustness when the support of the distributions differs, and can be applied even when the likelihood function is intractable but sampling is possible. These properties have led to the widespread adoption of Wasserstein-based loss functions in modern inference (Marin et al., 2012; Bonneel et al., 2015; Bernton et al., 2019a; Nadjahi et al., 2020) and generative modeling frameworks (Genevay et al., 2018; Arjovsky et al., 2017; Deshpande et al., 2018; Kolouri et al., 2018).




(a) Gaussian mixture model (b) Tukey $g$-and-$h$

Figure 1: Illustrations of finite-sample bias in minimizing Wasserstein loss (well-specified case). In (a), for a Gaussian mixture model, stochastic optimization with batch size 64 yields a biased estimate when using the Wasserstein loss (MWE) compared to the log-likelihood (MLE). In (b), for a Tukey $g$-and-$h$ model, minimizing expected Wasserstein loss between empirical distributions underestimates the density's tail behavior.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. 

Theoretical properties of the minimum Wasserstein estimator, which minimizes the Wasserstein distance between a parametric model and an empirical distribution, have attracted growing interest. This estimator is consistent, converging to the parameter value that minimizes the distance between the model and the underlying distribution as the sample size increases (Bassetti et al., 2006; Bernton et al., 2019b). In finite-sample regimes, however, sample-based Wasserstein objectives are known to exhibit biased gradients and may lead to biased minima relative to their population counterparts (Bellemare et al., 2017). This contrasts with the log-likelihood, for which the expected empirical objective coincides exactly with the population objective, so no such bias arises.

While most theoretical analyses have focused on asymptotic or one-sided empirical settings, less attention has been given to the regime most relevant in practice—minimizing the expected Wasserstein loss between two empirical distributions, both constructed from finite samples. This loss corresponds to a plug-in estimator of the Wasserstein distance, whose statistical properties have been studied in Chizat et al. (2020); in particular, bias in the distance estimate and its reduction have been examined by Papp and Sherlock (2025). Prior work has also analyzed learning with minibatch Wasserstein losses and emphasized that they differ from using the true distance between the underlying distributions (Fatras et al., 2020). However, the impact of these losses on optimization outcomes, particularly the presence and characterization of bias in parameter estimation, remains largely unexplored.

In this paper, we show that minimizing the *expected Wasserstein loss between two empirical distributions* can lead to biased parameter estimates in finite-sample regimes. Remarkably, such bias arises even in well-specified settings where both empirical distributions are drawn from the same parametric family: unlike maximum likelihood estimation—understood here as maximizing the expected log-likelihood—optimizing one parameter while fixing another generally fails to recover the true fixed value (see Figure 1).

To make this phenomenon analytically tractable, following Jang et al. (2026), we focus on *one-dimensional settings* where closed-form expressions for optimal transport are available (Villani et al., 2009; Santambrogio, 2015; Peyré et al., 2019; Bobkov and Ledoux, 2019). We derive the expected empirical Wasserstein loss in closed form for representative models like location-scale families, and explicitly demonstrate the resulting bias.

Furthermore, we show that if the expected loss is non-constant along the diagonal (where the two parameters coincide), the gradient of the expected loss with respect to the variable parameter, evaluated at the fixed parameter value, is nonzero, shifting the minimizer away from it. To address this, we propose a *simple yet effective correction scheme* that eliminates the bias in well-specified cases.

We then extend our analysis to misspecified settings and to Sinkhorn divergences (Genevay et al., 2018), which arise from entropic regularization of optimal transport (OT), correct the resulting entropic bias, and are positive-definite, while also being widely used for their strong convexity and computational efficiency. Experiments on synthetic and real data, including more practical settings involving neural network generators, show that stochastic optimization of these Wasserstein-based objectives still converges to biased solutions in finite-sample regimes and that the proposed bias correction scheme effectively mitigates this effect.

These findings challenge the intuitive anticipation that the expected loss should be minimized at the true parameter when both distributions are drawn from the same model class and have equal sample sizes, suggesting an inherent symmetry in the problem. Instead, we reveal a *structural limitation* of Wasserstein-based objectives for parametric inference in the finite-sample regime. In particular, our results suggest that common design choices, such as batch size or sample count, can systematically distort the information captured by empirical Wasserstein distances.

The paper is organized as follows. Section 2 provides background on Wasserstein distance and minimum Wasserstein estimation. Section 3 characterizes the finite-sample bias that arises when minimizing expected Wasserstein loss between empirical distributions in the well-specified case. Section 4 extends the analysis to the misspecified setting, and Section 5 addresses Sinkhorn divergences.

## 2 BACKGROUND

The $p$-Wasserstein distance (denoted $W_p$) between two probability density functions (PDFs) $\mu$ and $\nu$ over $\mathbb{R}^d$ is defined as

$$W_p^p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-y\|^p d\gamma(x,y), \quad (1)$$

where $\Gamma(\mu,\nu)$ denotes the set of all joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ that have respective marginals $\mu$ and $\nu$.

In the one-dimensional case ($d = 1$), this admits a closed-form expression using the cumulative distribution functions (CDFs) $P$ and $Q$ of $\mu$ and $\nu$, respectively (Dall'Aglio, 1956; Peyré et al., 2019):

$$W_p^p(\mu,\nu) = \int_0^1 |P^{-1}(u) - Q^{-1}(u)|^p du. \quad (2)$$

We primarily focus on the one-dimensional case, following Jang et al. (2026), which allows for a more precise analytic characterization of the distance and reveals core phenomena that persist in higher dimensions.

Let $\hat{\mu}_N$ and $\hat{\nu}_N$ denote empirical distributions constructed from i.i.d. samples $x_1, \ldots, x_N \sim \mu$ and $y_1, \ldots, y_N \sim \nu$, respectively: $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ and $\hat{\nu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i}$, where $\delta_x$ is a Dirac measure at $x \in \mathbb{R}$.

In one dimension, let $x_{(1)} \leq \cdots \leq x_{(N)}$ and $y_{(1)} \leq \cdots \leq y_{(N)}$ be the ordered samples. Then the empirical $W_p$ distance between $\hat{\mu}_N$ and $\hat{\nu}_N$ becomes

$$W_p^p(\hat{\mu}_N, \hat{\nu}_N) = \frac{1}{N}\sum_{i=1}^{N} |x_{(i)} - y_{(i)}|^p. \tag{3}$$

It is well known that under mild conditions, $W_p(\hat{\mu}_N, \mu) \to 0$ in expectation as $N \to \infty$ (Boissard and Le Gouic, 2014; Fournier and Guillin, 2015; Weed and Bach, 2019). By the triangle inequality, this implies $\mathbb{E}\left[W_p(\hat{\mu}_N, \hat{\nu}_N)\right] \to W_p(\mu, \nu)$.

Let $\{f_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^m\}$ be a parametric family of distributions. The minimum Wasserstein estimator minimizes the Wasserstein distance between an empirical distribution and a model distribution:

$$\hat{\theta}_N = \arg\min_\theta W_p(\hat{\mu}_N, f_\theta). \tag{4}$$

Under suitable conditions, this estimator is known to be consistent (Bassetti et al., 2006; Bernton et al., 2019b): $\hat{\theta}_N \to \theta^* = \arg\min_\theta W_p(\mu, f_\theta)$ as $N \to \infty$.

However, the convergence to $\theta^*$ holds only asymptotically, and Wasserstein-based estimators can behave quite differently in finite-sample settings. For instance, in a Bernoulli model, minimizing the expected empirical Wasserstein loss yields a biased estimate (Bellemare et al., 2017), i.e., in general, $\hat{\theta}_N = \arg\min_\theta \mathbb{E}[W_p^p(\hat{\mu}_N, f_\theta)] \neq \arg\min_\theta W_p^p(\mu, f_\theta)$.

In practice, Wasserstein-based objectives are often minimized between two empirical distributions, each constructed from finite samples (Deshpande et al., 2018; Genevay et al., 2018; Kolouri et al., 2018). Although consistency results have been established for such estimators under suitable asymptotic conditions (Bernton et al., 2019b), their behavior in the finite-sample regime remains poorly understood.

**Our Focus** We study this finite-sample regime directly. Let $\hat{f}_{\theta^*,N}$ and $\hat{f}_{\theta,N}$ denote empirical distributions independently drawn from the parametric model at parameter values $\theta^*$ (fixed) and $\theta$ (variable), respectively. We consider the expected loss $J_N(\theta^*, \theta) = \mathbb{E}[W_p^p(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})]$, and ask whether its minimizer recovers the fixed parameter, i.e., $\hat{\theta}_N = \arg\min_\theta J_N(\theta^*, \theta) \stackrel{?}{=} \theta^*$.

As we show, this equality fails to hold in general—even under well-specified models with equal sample sizes, where the problem exhibits inherent symmetry. We then extend our analysis to misspecified cases and to the Sinkhorn divergence. Our results characterize when finite-sample bias arises and how it can be corrected, offering new insights into the limitations of Wasserstein-based objectives for parameter estimation.

# 3 FINITE-SAMPLE BIAS IN WELL-SPECIFIED MODELS

In this section, we present analytic and numerical results demonstrating that the minimizer of the expected Wasserstein loss between two empirical distributions can be biased even in well-specified settings, by focusing on location-scale models. We then propose a simple method to correct this bias.

## 3.1 Expected $W_2$ Loss in One Dimension

Let $\hat{f}_{\theta^*,N} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ and $\hat{f}_{\theta,N} = \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i}$, where $x_1, \ldots, x_N \sim f_{\theta^*}$ and $y_1, \ldots, y_N \sim f_\theta$. Here $\theta^*$ denotes the fixed target parameter, whereas $\theta$ is the optimization variable. Setting $p = 2$ in (3), our loss is

$$\begin{aligned} J_N(\theta^*, \theta) &= \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(x_{(i)} - y_{(i)})^2]. \end{aligned} \tag{5}$$

Each term is $\mathbb{E}[(x_{(i)} - y_{(i)})^2] = \int_{\mathbb{R}}\int_{\mathbb{R}} (x_{(i)} - y_{(i)})^2 p_{X_{(i)}}(x_{(i)}) p_{Y_{(i)}}(y_{(i)}) dx_{(i)} dy_{(i)}$, where $p_{X_{(i)}}(x_{(i)})$ and $p_{Y_{(i)}}(y_{(i)})$ are the densities of the $i$-th order statistics $x_{(i)}$ and $y_{(i)}$ from $f_{\theta^*}$ and $f_\theta$, respectively (Bobkov and Ledoux, 2019).

The loss in (5) can be expressed as follows:

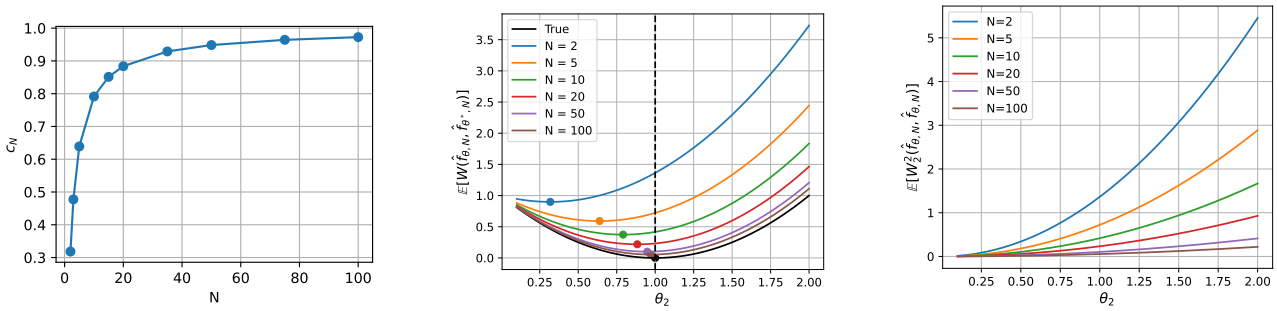
**Lemma 3.1.** *The expected squared $W_2$ loss between $\hat{f}_{\theta^*,N}$ and $\hat{f}_{\theta,N}$ is*

$$J_N(\theta^*, \theta) = m_2(\theta^*) + m_2(\theta) - \frac{2}{N}\sum_{i=1}^{N} m_{1,i}(\theta^*) m_{1,i}(\theta), \tag{6}$$

*where*

$$m_{1,i}(\theta) \equiv \int_0^1 F_\theta^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)}, \tag{7}$$

$$m_2(\theta) \equiv \int_{\mathbb{R}} y^2 f_\theta(y) dy. \tag{8}$$




(a) The constant $c_N$ (b) $J_N(\theta^*, \theta)$ for $\theta_2^* = 1$ (c) $J_N(\theta, \theta)$

Figure 2: The constant $c_N$ and expected Wasserstein loss $J_N(\theta^*, \theta)$ for Gaussian distributions, shown across different values of $N$. In (a), the value of $c_N$ is computed for $N \in [2, 100]$. In (b), colored solid curves represent $J_N(\theta^*, \theta)$, with minimizers $\hat{\theta}_{N,2}$ marked by dots. In (c), $J_N(\theta^*, \theta)$ are evaluated along the diagonal $\theta^* = \theta$.

*Here $F_\theta$ denotes the CDF associated with $f_\theta$ and $p_{U_{(i)}}(u_{(i)}) = \frac{N!}{(i-1)!(N-i)!} u_{(i)}^{i-1}(1-u_{(i)})^{N-i}$ is the probability density function of the $i$-th order statistic of a* Uniform$(0,1)$ *distribution.*

***Proof sketch.*** *We compute the expectation by integrating the loss with respect to the joint density of the order statistics. We then simplify the resulting sum of second moments using the permutation invariance of i.i.d. samples, which reduces the expectation to terms involving only the first moments of the order statistics and the second moment of the model distribution. The full proof is provided in Appendix A.1.*

In (6), the last term averages the products of the first moments of corresponding order statistics and depends on $N$. It can deviate from the infinite-sample limit, indicating the presence of finite-sample bias in the loss.

### 3.2 Bias in Location-Scale Models

We analyze the bias in the location-scale family $f_\theta(y) = \frac{1}{\theta_2} f_0\left(\frac{y-\theta_1}{\theta_2}\right)$, where $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$, $\theta_1$ is the location parameter, and $\theta_2 > 0$ is the scale parameter. The reference density $f_0$ is normalized to have zero mean and unit variance without loss of generality.

The expected Wasserstein loss in (6) and its minimizer can be derived analytically:

**Proposition 3.2.** *Assume that the first and second moments of the reference distribution $f_0$ are finite. Then the expected squared $W_2$ loss is given by*

$$\begin{aligned} J_N(\theta^*, \theta) &= \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] \\ &= (\theta_1 - \theta_1^*)^2 + \theta_2^2 - 2c_N\theta_2\theta_2^* + \theta_2^{*2}, \end{aligned} \tag{9}$$

*where $c_N = \frac{1}{N}\sum_{i=1}^N \left(\int_0^1 F_0^{-1}(u_{(i)})\, p_{U_{(i)}}(u_{(i)})\, du_{(i)}\right)^2$ and $F_0$ is the CDF of the reference density $f_0$. The optimal location and scale parameters, $\hat{\theta}_N = (\hat{\theta}_{N,1}, \hat{\theta}_{N,2})$ that minimize the expected loss in* (9) *are given by $\hat{\theta}_{N,1} = \theta_1^*$ and $\hat{\theta}_{N,2} = c_N\theta_2^*$.*

***Proof sketch.*** *This follows directly from Lemma 3.1 by substituting the closed-form moments of the location-scale family into the general expression. The full proof is provided in Appendix A.2.*

Therefore, when $N$ is finite, we generally have $\hat{\theta}_{N,2} \neq \theta_2^*$, indicating that the fixed scale parameter cannot be recovered by minimizing the expected Wasserstein loss. This contrasts with maximum likelihood estimation, where the finite-sample expected log-likelihood coincides with the population objective and thus recovers the fixed parameter under a well-specified model.

**Convergence Rates** We also summarize how the finite-sample bias decays with the sample size $N$ (see Appendix A.3 for details). In one dimension, $\mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] - W_2^2(f_{\theta^*}, f_\theta)$ is of order $O(1/N)$ for distributions with lighter-than-Gaussian tails, and $J_N(\theta^*, \theta)$ and $c_N$ in (9) converge at the same rate. For heavier-tailed distributions, the convergence is slower, with rate $O(\log N/N)$ for exponential distributions and $O(\log\log N/N)$ for Gaussian distributions.

**Numerical Examples** We illustrate the finite-sample bias using the Gaussian distribution. Since Proposition 3.2 shows no bias in the location parameter, we focus on the scale parameter (i.e., the standard deviation).

Figure 2(a) shows that $c_N$ in (9) deviates significantly from 1 for small $N$ and approaches 1 as $N$ increases, indicating that the optimal scale parameter tends to be underestimated when $N$ is small, but converges to the true value as $N$ grows. This behavior aligns with the consistency results of minimum Wasserstein estimation (Bassetti et al., 2006; Bernton et al., 2019b).

In Figure 2(b), we set $\theta_1 = \theta_1^* = 0$ and $\theta_2^* = 1$, then plot $J_N(\theta^*, \theta)$ for various $N$. Compared to the true Wasserstein loss (the $N \to \infty$ limit), the empirical loss remains strictly higher, does not vanish at $\theta = \theta^*$, and its minimum is biased when $N$ is finite.

### 3.3 Bias Characterization and Correction Scheme

Because the objective $J_N(\cdot, \cdot)$ is symmetric in its two arguments, we have $\frac{\partial J_N(\theta,\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = 2 \cdot \frac{\partial J_N(\theta^*,\theta)}{\partial \theta}\Big|_{\theta=\theta^*}$ as shown in Appendix A.4. This identity leads directly to the following remark on the condition under which minimizing the expected Wasserstein loss exhibits finite-sample bias:

*Remark* 3.3. If the expected empirical Wasserstein loss $J_N(\theta^*, \theta)$ with finite $N$ has a non-zero gradient along the diagonal subspace $\theta^* = \theta$—that is, if $\frac{\partial J_N(\theta,\theta)}{\partial \theta} \neq 0$—then $\frac{\partial J_N(\theta^*,\theta)}{\partial \theta}\Big|_{\theta=\theta^*} \neq 0$ and there exists a parameter $\theta \neq \theta^*$ that achieves a lower loss than $\theta = \theta^*$.

This condition applies to the Gaussian density example (see Figure 2(c)). The fact that $J_N(\theta, \theta)$ increases with $\theta_2$ aligns with the downward bias of $\hat{\theta}_{N,2}$ observed in Figure 2(b), as further explained in Appendix A.4.

**A Simple Bias Correction Scheme** Accordingly, we define a bias-corrected loss by subtracting a self-distance term:

$$\tilde{J}_N(\theta^*, \theta) = J_N(\theta^*, \theta) - \frac{1}{2} J_N(\theta, \theta). \tag{10}$$

Differentiating (10) with respect to $\theta$ gives

$$\frac{\partial \tilde{J}_N(\theta^*,\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = 0, \text{ since } \frac{\partial J_N(\theta,\theta)}{\partial \theta} = 2\frac{\partial J_N(\theta^*,\theta)}{\partial \theta}.$$

Therefore, $\theta = \theta^*$ is always a stationary point of $\tilde{J}_N(\theta^*, \theta)$, for any $\theta^*$ and $N$.

As an example, for the location–scale model in Proposition 3.2, the bias-corrected loss becomes

$$\tilde{J}_N(\theta^*, \theta) = (\theta_1 - \theta_1^*)^2 + c_N \theta_2^2 - 2c_N \theta_2 \theta_2^* + {\theta_2^*}^2, \tag{11}$$

which is minimized at $(\theta_1, \theta_2) = (\theta_1^*, \theta_2^*)$.

Note that Remark 3.3 and the bias-corrected loss in (10) extend naturally to the case of higher-dimensional data, since the loss $J_N(\theta^*, \theta) = \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})]$ remains symmetric in its two arguments (see Appendix A.4).

**Numerical Examples** We first evaluate the modified loss $\tilde{J}_N(\theta^*, \theta)$ for Gaussian density examples in Figure 3(a). Unlike the biased minimizers of $J_N(\theta^*, \theta)$ in Figure 2(b), the minimizers of $\tilde{J}_N(\theta^*, \theta)$ coincide with those of the true Wasserstein loss between the underlying densities.

Figures 3(b) and (c) show the results of stochastic gradient descent (SGD) applied to $J_N(\theta^*, \theta)$ and $\tilde{J}_N(\theta^*, \theta)$ for Gaussian models and multivariate Gaussian mixture models (GMMs), respectively. Experimental details are provided in Appendices B.1 and B.2.

In Figure 3(b), for finite $N$, the scale parameter minimizing $J_N(\theta^*, \theta)$ converges to a biased solution, which matches the minimizers characterized in Proposition 3.2. As $N$ increases, this bias diminishes, and the solution approaches the fixed parameter. In contrast, minimizing the modified loss $\tilde{J}_N(\theta^*, \theta)$, yields convergence to the fixed parameter even for small $N$, demonstrating the effectiveness of the bias correction scheme.

Figure 3(c) reports fitting errors (defined in Appendix B.2) for multivariate GMMs with four mixtures as a function of batch size $N$. Finite-sample bias is evident in all cases, becomes more pronounced in higher dimensions, and decreases as $N$ grows. The bias correction scheme consistently reduces this error, with the largest improvements for small $N$ and large $d$. Appendix B.5.2 for the Tukey $g$-and-$h$ model (Tukey, 1977) and Appendix B.6.1 for an affine PDF model defined in Appendix A.9.

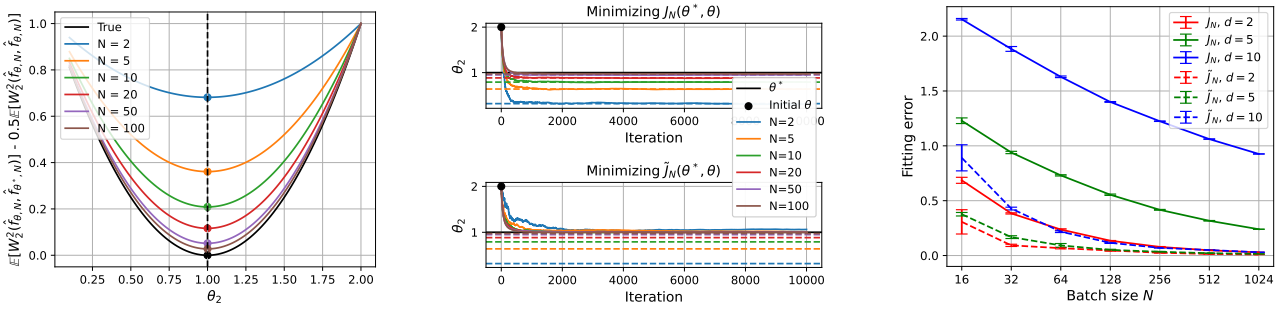
### 3.4 Discussion on Higher-Dimensional Settings

Although our experiments demonstrate finite-sample bias and the effectiveness of the proposed correction scheme in higher-dimensional settings, an exact characterization of this bias, analogous to our one-dimensional analysis, is likely intractable, since a closed-form expression for the Wasserstein loss is generally unavailable when $d > 1$. Instead, we discuss the existence of the bias and its asymptotic convergence behavior for $d > 1$ through a perturbation analysis, suggesting that the parameter estimate converges at a rate governed by that of the empirical Wasserstein distance (see Appendix A.5 for details).

We also note that, for higher-dimensional distributions, the sliced Wasserstein distance (SWD) is defined as the expectation of one-dimensional Wasserstein distances over projections along arbitrary directions (Bonneel et al., 2015). Since our one-dimensional analysis applies to each projection, this suggests that finite-sample bias should also persist in the expected SWD loss.

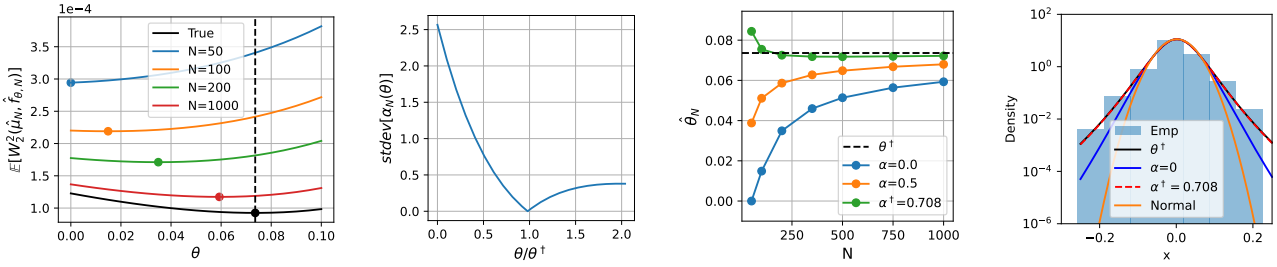
## 4 FINITE-SAMPLE BIAS IN MISSPECIFIED MODELS

We now extend the analysis to the misspecified case, a setting often encountered in practice when the data distribution $\mu$ does not lie in the parametric family.




(a) $\tilde{J}_N(\theta^*, \theta)$ for $\theta_2^* = 1$ (b) $\hat{\theta}_{N,2}$ trajectory (Gaussian) (c) Fitting error (GMM)

Figure 3: Modified loss and stochastic optimization results of empirical Wasserstein losses for different sample sizes $N$. In (a), colored solid curves represent $\tilde{J}_N(\theta^*, \theta)$ for Gaussian distributions, with minimizers $\hat{\theta}_{N,2}$ marked by dots. In (b), we apply SGD to minimize $J_N(\theta^*, \theta)$ (top) and $\tilde{J}_N(\theta^*, \theta)$ (bottom) for Gaussian distributions. Colored solid lines represent the parameter trajectories, and dashed lines indicate the corresponding minimizers of $J_N(\theta^*, \theta)$. In (c), we depict the fitting error for Gaussian mixture models in dimensions $d = 2, 5, 10$.




(a) $J_N^{\mathrm{MS}}(\theta)$ (b) Stdev. of $\alpha_N(\theta)$ (c) Bias reduction (d) PDFs

Figure 4: The expected Wasserstein loss $J_N^{\mathrm{MS}}(\theta)$ and the application of the bias correction scheme in the misspecified case for Tukey $g$-and-$h$ models with $\theta = h$, using the BTC-USD dataset, across different sample sizes $N$. In (a), $J_N^{\mathrm{MS}}(\theta)$ is shown. Colored solid curves depict $J_N^{\mathrm{MS}}(\theta)$, with minimizers $\hat{\theta}_N$ marked by dots. In (b), for each $\theta$, the standard deviation (Stdev.) of $\{\alpha_N(\theta) \mid N \in \{200, 500, 1000\}\}$ is computed. In (c), solid curves represent the minimizers of the bias-corrected losses for $\alpha = 0, 0.5$, and $\alpha^\dagger$. The black dashed line indicates the infinite-sample solution $\theta^\dagger$. In (d), we show the empirical histogram of the data and the PDFs at $\theta^\dagger$, the biased ($\alpha = 0$) and bias-corrected ($\alpha = \alpha^\dagger$) estimates with $N = 200$, and the normal distribution, all on a log scale.

### 4.1 Expected $W_2$ Loss in the Misspecified Setting

Let the empirical distribution of the data $x_1, \ldots, x_N \sim \mu$ be $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, and $x_{(1)} \le \ldots \le x_{(N)}$ be the ordered samples. The following proposition is the analogue of Lemma 3.1 and Proposition 3.2.

**Proposition 4.1.** *The expected squared $W_2$ loss between $\hat{\mu}_N$ and $\hat{f}_{\theta,N}$ is*

$$J_N^{\mathrm{MS}}(\theta) = \mathbb{E}[W_2^2(\hat{\mu}_N, \hat{f}_{\theta,N})] = \mathbb{E}[x^2] + m_2(\theta) - \frac{2}{N}\sum_{i=1}^N \mathbb{E}[x_{(i)}] m_{1,i}(\theta). \tag{12}$$

*When $f_\theta$ belongs to the location-scale model discussed in Section 3.2,* (12) *becomes as follows:*

$$J_N^{MS}(\theta) = \mathbb{E}[x^2] + \theta_1^2 + \theta_2^2 - 2\mathbb{E}[x]\theta_1 - \frac{2}{N}\sum_{i=1}^N \mathbb{E}[x_{(i)}] b_i \theta_2, \tag{13}$$

*where $b_i = \int_0^1 F_0^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)}$. The optimal location and scale parameters minimizing* (13) *are $\hat{\theta}_{N,1} = \mathbb{E}[x]$ and $\hat{\theta}_{N,2} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[x_{(i)}] b_i$, respectively.*

***Proof.*** *The argument parallels that of Lemma 3.1 and Proposition 3.2, with $m_2(\theta^*)$ and $m_{1,i}(\theta^*)$ replaced by $\mathbb{E}[x^2]$ and $\mathbb{E}[x_{(i)}]$, respectively.* □

In (12) and (13), the last term depends on $N$ and can deviate from the infinite-sample limit, indicating the presence of bias in the loss. Moreover, for finite

$N$, $\hat{\theta}_N$ can deviate from its limit, reconciling with the observation of Bellemare et al. (2017).

This formulation naturally extends to real-data settings where $\mu$ is represented by a finite dataset $\mathcal{D} = \{x_1, \ldots, x_M\}$. In this case, with $\mu = \frac{1}{M}\sum_{i=1}^{M} \delta_{x_i}$, $\hat{\mu}_N$ is constructed by sampling $N$ points with replacement from $\mathcal{D}$. The expectations $\mathbb{E}[x_{(i)}]$ can be expressed analogously to (7), using an inverse CDF constructed from $\mu$. This results in weighted sums of the data points, with weights given by incomplete Beta functions.

**Numerical Examples** We illustrate finite-sample bias in the misspecified case using the Tukey $g$–$h$ family, where the parameters $\xi$, $\omega$, $g$, and $h$ control location, scale, skewness, and tail heaviness, respectively (Tukey, 1977) (see Appendix B.5.1). For the data distribution $\mu$, we use the empirical distribution of BTC-USD log-returns data obtained from Yahoo Finance via the `yfinance` library (Aroussi, 2019). Since the Tukey $g$–$h$ density lacks a closed form, maximum likelihood estimation is infeasible, making Wasserstein-based methods a natural alternative. We focus on estimating the tail-heaviness parameter $h$, fixing $g = 0$ and setting $(\xi, \omega)$ to the sample mean and standard deviation. The infinite-sample solution is denoted by $\theta^\dagger = \hat{\theta}_\infty$.

Figure 4(a) plots $J_N^{\mathrm{MS}}(\theta)$ for several $N$. As in the Gaussian case (Figure 2(b)), the finite-sample loss is inflated relative to the $N \to \infty$ limit and yields biased minimizers: for small $N$, $\hat{\theta}_N$ underestimates tail heaviness, but converges to $\theta^\dagger$ as $N$ grows.

### 4.2 Bias Correction in the Misspecified Setting

In the misspecified case, the bias-correction scheme in (10) cannot be used directly since the symmetry argument of Section 3.3 does not apply. However, for bias correction in a one-dimensional parameter, we can modify the objective by subtracting $\alpha J_N(\theta, \theta)$ with a suitable choice of $\alpha$, thereby shifting the stationary point to the infinite-sample solution, denoted by $\theta^\dagger$.

Specifically, define

$$\tilde{J}_N^{\mathrm{MS}}(\theta; \alpha) = J_N^{\mathrm{MS}}(\theta) - \alpha J_N(\theta, \theta). \tag{14}$$

When $\frac{\partial J_N(\theta,\theta)}{\partial \theta}\Big|_{\theta=\theta^\dagger}$ is not zero, with an appropriate choice of $\alpha = \alpha^\dagger$, the stationary point of this modified objective coincides with $\theta^\dagger$, i.e., $\frac{\partial \tilde{J}_N^{\mathrm{MS}}(\theta;\alpha^\dagger)}{\partial \theta}\Big|_{\theta=\theta^\dagger} = 0$. By solving this equation for $\alpha^\dagger$, we obtain

$$\alpha^\dagger = \left.\frac{\partial J_N^{\mathrm{MS}}(\theta)/\partial\theta}{\partial J_N(\theta,\theta)/\partial\theta}\right|_{\theta=\theta^\dagger}. \tag{15}$$

That is, minimizing $\tilde{J}_N^{\mathrm{MS}}(\theta; \alpha^\dagger)$ corrects the bias and recovers $\theta^\dagger$ even in the finite-sample regime.

When the underlying data density is unknown, the infinite-sample solution $\theta^\dagger$ is not available, and $\alpha^\dagger$ cannot be obtained directly. We therefore propose a heuristic to approximate both, assuming that $J_N^{\mathrm{MS}}(\theta)$ and $J_N(\theta, \theta)$ converge at comparable rates as $N$ grows.

Note that as $N \to \infty$, the loss in (12) converges to $J_\infty^{\mathrm{MS}}(\theta)$, while the self-distance term $J_N(\theta, \theta)$ vanishes. Assume both terms and their gradients converge at rate $O(f(N))$.[1] Then, for a given $N$, we can define

$$\alpha_N(\theta) = \frac{\partial J_N^{\mathrm{MS}}(\theta)/\partial\theta}{\partial J_N(\theta,\theta)/\partial\theta} = \frac{\partial J_\infty^{\mathrm{MS}}(\theta)/\partial\theta + c(\theta)O(f(N))}{d(\theta)O(f(N))}, \tag{16}$$

where $c(\theta)$ and $d(\theta)$ are convergence coefficients depending on $\theta$. Thus, in general, $\alpha_N(\theta)$ depends on $N$.

At $\theta = \theta^\dagger$, however, $\partial J_\infty^{\mathrm{MS}}(\theta)/\partial\theta = 0$, and hence $\alpha_N(\theta) = c(\theta)/d(\theta)$, independent of $N$. This suggests the following procedure to obtain $\theta^\dagger$ and $\alpha^\dagger$: compute $\alpha_N(\theta)$ for several values of $N$, and identify the value of $\theta$ for which $\alpha_N(\theta)$ remains nearly constant across $N$. We then treat this $\theta$ as a proxy for $\theta^\dagger$,[2] and the corresponding averaged value of $\alpha_N(\theta^\dagger)$ across different $N$ as a proxy for $\alpha^\dagger$, which we use to construct the bias-corrected loss in (14).

**Numerical Examples** To validate the proposed bias-correction heuristic, we conduct experiments with the Tukey $g$-and-$h$ model and the BTC-USD dataset introduced in Section 4.1.

Figure 4(b) shows the standard deviation of $\alpha_N(\theta)$ (in (16)) across different values of $N$ as a function of $\theta$. This deviation is minimized when $\theta/\theta^\dagger$ is close to one.

Figure 4(c) presents optimization results obtained by minimizing the bias-corrected loss in (14) with different choices of $\alpha$. For $N$ ranging from 50 to 1000, setting $\alpha = \alpha^\dagger$ as defined in (15) effectively reduces finite-sample bias and yields solutions closer to the infinite-sample optimum, even when $N$ is small. This approach consistently outperforms both the case without bias correction ($\alpha = 0$) and the direct application of the scheme from Section 3.3 ($\alpha = 0.5$). Figure 4(d) shows that the correction improves estimation of the tail-heaviness of the data distribution. Together, these results confirm that the bias-correction scheme is effective even in practically relevant misspecified settings.

[1] For a discussion about the convergence speed of the empirical Wasserstein distance, we refer the reader to Boissard and Le Gouic (2014); Fournier and Guillin (2015); Weed and Bach (2019).

[2] This proxy for $\theta^\dagger$ may itself serve as a useful estimate of the infinite-sample minimizer. Moreover, analyzing the ratio of gradient norms of $J_N^{\mathrm{MS}}(\theta)$ and $J_N(\theta, \theta)$ can potentially extend this idea of estimation to higher-dimensional parameter spaces. This extension is left for future work.

Additional results for the SPY dataset (obtained via `yfinance`) and the Diamonds dataset (Wickham, 2016) are provided in Appendix B.5.3, and results for Gaussian models appear in Appendix B.4.1.

We also note that the semi-discrete case, where the expected Wasserstein loss between the empirical distribution $\hat{\mu}_N$ and the density $f_\theta$ is minimized, admits an analogous bias correction scheme (see Appendix A.6); detailed results for Gaussian and affine models are given in Appendices B.4.2 and B.6.2.

### 4.3 A Neural Network Generator Example

To demonstrate that finite-sample bias persists in a more practical setting and to examine whether our correction scheme can mitigate it, we train a neural network (NN) generator, $z \in \mathbb{R}^{10} \mapsto x = \mathrm{NN}_\theta(z) \in \mathbb{R}^d$, to fit multivariate Gaussian mixture data by minimizing the expected Wasserstein loss between empirical distributions constructed from minibatches. Although this is a misspecified setting, the high-dimensional parameter space makes the heuristics in (15)–(16) not directly applicable. As a useful alternative, we consider the bias-corrected objective with $\alpha = 0.5$, as in (10). Experimental details are provided in Appendix B.3.

To quantify performance, we use two metrics computed from $10^5$ samples: (i) the squared 2-sliced Wasserstein distance ($SW_2^2$), estimated with 1,000 projections, between generated samples and test data, which we use as a more stable evaluation metric in high dimensions; and (ii) the covariance trace ratio between generated and test data, which serves as a proxy for the overall scale of the distribution.

Figure 5 shows that finite-sample bias persists for NN generators, becomes more pronounced in higher dimensions, and decreases as the batch size grows. Samples generated by the NN trained with bias correction are qualitatively more similar to the data than those generated without correction, and they also achieve substantially lower test $SW_2^2$ across all batch sizes. Moreover, the covariance trace ratio shows that training with bias correction better recovers the scale of the distribution, whereas training without correction exhibits severe shrinkage, consistent with the downward scale bias seen in Section 3.2. These results demonstrate the relevance of our finite-sample bias analysis and correction scheme in practical settings involving neural network generators.

## 5 FINITE-SAMPLE BIAS IN SINKHORN DIVERGENCE

We next examine finite-sample bias in minimizing the expected Sinkhorn divergence. The Sinkhorn divergence is derived from entropic regularization of optimal transport (OT), which augments the transport cost with an entropy term on the transport plan. For $p = 2$, entropically regularized OT problem is

$$\begin{aligned}\mathrm{OT}_\varepsilon(\mu,\nu) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-y\|^2 d\gamma(x,y) \\ + \varepsilon KL(\gamma(x,y)||\mu(x)\nu(y)),\end{aligned} \tag{17}$$

where $\varepsilon > 0$ is the regularization weight, and $KL(\gamma(x,y)||\mu(x)\nu(y)) = \int_{\mathbb{R}^d} \log\left(\frac{\gamma(x,y)}{\mu(x)\nu(y)}\right) d\gamma(x,y)$ is the Kullback–Leibler divergence between the transport plan and the product measure.

Although this regularization makes the problem strongly convex and computationally efficient in the discrete case, it also introduces entropic bias, and the resulting value in (17) is no longer positive-definite. In particular, even for identical distributions, $\mathrm{OT}_\varepsilon(\mu,\mu) \neq 0$. The Sinkhorn divergence remedies this by subtracting the corresponding self-terms (Genevay et al., 2018):

$$S_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu). \tag{18}$$

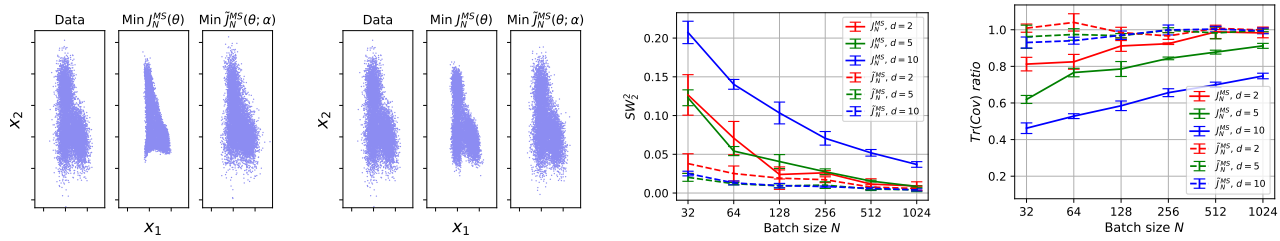
On compact sample spaces, this divergence is positive-definite (Feydy et al., 2019).

Following the well-specified settings of Section 3.1, we define the expected Sinkhorn divergence between empirical measures $\hat{f}_{\theta^*,N}$ and $\hat{f}_{\theta,N}$ as

$$S_{N,\varepsilon}(\theta^*,\theta) = \mathbb{E}[S_\varepsilon(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})]. \tag{19}$$

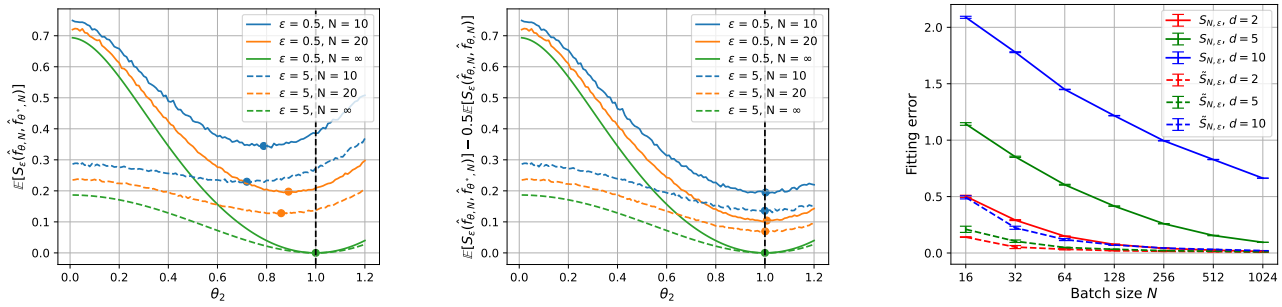
Minimizing this objective introduces bias in the estimated parameter, and a simple correction scheme can reduce it, as observed in Sections 3.3 and 4.2. Since an analytic form of (19) is intractable, we rely on numerical evaluations obtained by averaging 10,000 empirical estimates computed with the `POT` library (Flamary et al., 2021).

Figure 6(a) plots the expected Sinkhorn divergence with respect to the scale parameter in the Gaussian model. For finite $N$, the empirical value is strictly larger than the true value between Gaussians (available in closed form from Janati et al. (2020)), and, as in the earlier sections, we observe a downward bias in the minimizer of the scale parameter $\hat{\theta}_{N,\varepsilon,2}$. Moreover, when the regularization parameter $\varepsilon$ gets larger, the finite-sample bias becomes more severe in this case. In fact, for the location–scale model, as $\varepsilon \to \infty$, $\hat{\theta}_{N,\varepsilon,2}$ converges to zero for finite $N$ (see Appendix A.8). This finite-sample bias also persists under model misspecification, as illustrated in Appendix B.4.3.




(a) Batch size = 64 (b) Batch size = 1024 (c) $SW_2^2$ (d) Tr(Cov) ratio

Figure 5: Results for neural network (NN) generators fitted to Gaussian mixture data by stochastic optimization of $J_N^{\mathrm{MS}}$ and $\tilde{J}_N^{\mathrm{MS}}$ with $\alpha = 0.5$ for different batch sizes. In (a) and (b), we show two-dimensional projections of the data and of samples generated by NNs trained without and with bias correction in ten-dimensional space, for batch sizes 64 and 1024, respectively. In (c) and (d), performance is evaluated using the sliced Wasserstein distance (lower is better) and the covariance trace ratio (closer to 1.0 is better), respectively, for $d = 2, 5$, and 10.




(a) $S_{N,\varepsilon}(\theta^*, \theta)$ for $\theta_2^* = 1$ (b) $\tilde{S}_{N,\varepsilon}(\theta^*, \theta)$ for $\theta_2^* = 1$ (c) Fitting error (GMM)

Figure 6: Expected Sinkhorn divergences and stochastic optimization results for different sample sizes $N$. In (a), the curves show $S_{N,\varepsilon}(\theta^*, \theta)$ for Gaussian distributions with $\varepsilon = 0.5, 5$. In (b), the curves show the bias-corrected objective $\tilde{S}_{N,\varepsilon}(\theta^*, \theta)$ under the same setting. Minimizers $\hat{\theta}_{N,\varepsilon,2}$ are marked by dots. In (c), we depict the fitting error for Gaussian mixture models in dimensions $d = 2, 5, 10$ with regularization weight $\varepsilon = 1$.

**A Simple Bias Correction Scheme** Thanks to the symmetry of the objective in (19), the same correction scheme as in Section 3.3 can be applied:

$$\tilde{S}_{N,\varepsilon}(\theta^*, \theta) = S_{N,\varepsilon}(\theta^*, \theta) - \frac{1}{2} S_{N,\varepsilon}(\theta, \theta). \qquad (20)$$

Figure 6(b) shows that this correction restores $\theta = \theta^*$ as a stationary point regardless of the regularization weight. In the misspecified case, the discussion in Section 4.2 applies, and a similar one-dimensional correction can be carried out, as confirmed in Appendix B.4.3.

Finally, to examine the impact of finite-sample bias and the effectiveness of the correction scheme in stochastic optimization, we fit a Gaussian mixture model by minimizing the expected Sinkhorn divergence. In this experiment, we consider the well-specified setting where the data distribution is exactly a Gaussian mixture and apply the correction scheme in (20). Figure 6(c) shows that, consistent with the results of Section 3.3, the bias is very large for small batch sizes and higher dimensions, and that the proposed correction scheme effectively mitigates it. Additional Sinkhorn-based results for neural network generators fitted to Gaussian mixture data are provided in Appendix B.7.

## 6 CONCLUSION

In this paper, we have analyzed finite-sample bias in minimizing expected Wasserstein loss for parameter estimation. Our results show that such bias arises even in well-specified settings and can be mitigated by a simple modification of the objective. The analysis extends naturally to misspecified models and to Sinkhorn divergences, where we confirm the persistence of finite-sample bias and demonstrate the effectiveness of the correction scheme on both synthetic and real datasets. These findings highlight the importance of accounting for finite-sample bias, with correction whenever possible, in Wasserstein-based learning and inference methods widely used in statistics and machine learning.

**Acknowledgements**

C. Jang was partly supported by NRF/ME (RS-2023-00249714, RS-2025-25427337). Y.-K. Noh was partly supported by IITP/MSIT (IITP-2021-0-02068, RS-2020-II201373, RS-2023-00220628).

**References**

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Aroussi, R. (2019). yfinance: Yahoo! finance market data downloader. `https://github.com/ranaroussi/yfinance`.

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019a). Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019b). On parameter estimation with the wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Bobkov, S. and Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society.

Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in neural information processing systems*, 33:2257–2269.

Dall'Aglio, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 10(1-2):35–74.

Deshpande, I., Zhang, Z., and Schwing, A. G. (2018). Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. (2020). Learning with minibatch wasserstein: asymptotic and gradient properties. In *International Conference on Artificial Intelligence and Statistics*, pages 2131–2141. PMLR.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 2681–2690. PMLR.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Gradshteyn, I. S. and Ryzhik, I. M. (2014). *Table of integrals, series, and products*. Academic press.

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020). Entropic optimal transport between unbalanced gaussian measures has a closed form. *Advances in neural information processing systems*, 33:10468–10479.

Jang, C., Won, J., Jun, S., Chung, C. K., Joo, K., and Noh, Y.-K. (2026). On the information processing of one-dimensional wasserstein distances with finite samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22137–22145.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. (2018). Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180.

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601.

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., and Şimşekli, U. (2020). Approximate bayesian computation with the sliced-wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5470–5474. IEEE.

Papp, T. P. and Sherlock, C. (2025). Centered plug-in estimation of wasserstein distances.

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

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

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

Tukey, J. W. (1977). Modern techniques in data analysis. In *Proceedings of the NSF-Sponsored Regional Research Conference*, volume 7. Southeastern Massachusetts University North Dartmouth, MA, USA.

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

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance.

Wickham, H. (2016). Data analysis. In *ggplot2: elegant graphics for data analysis*, pages 189–201. Springer.

## 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] See Sections 2, 3.1, 3.2, 4.1, and 5.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] See the supplemental codes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Sections 3.1, 3.2, and 4.1.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix A.
   (c) Clear explanations of any assumptions. [Yes] See Sections 3.1, 3.2, and 4.1.

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] See the supplemental codes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Appendices B.1 and B.2.
   (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] See Appendix B.2.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

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] See Sections 4 and 5.
   (b) The license information of the assets, if applicable. [No]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [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. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# On the Finite-Sample Bias of Minimizing Expected Wasserstein Loss Between Empirical Distributions: Supplementary Materials

## A MATHEMATICAL DERIVATIONS AND PROOFS

### A.1 Proof of Lemma 3.1

Let $\hat{f}_{\theta^*,N} = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ and $\hat{f}_{\theta,N} = \frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}$, where $x_1,\dots,x_N \sim f_{\theta^*}$ and $y_1,\dots,y_N \sim f_\theta$. Let $x_{(1)} \le \cdots \le x_{(N)}$ and $y_{(1)} \le \cdots \le y_{(N)}$ be the ordered samples. Then the expected squared $W_2$ loss is expressed as

$$J_N(\theta^*,\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(x_{(i)} - y_{(i)})^2] \tag{21}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}}\int_{\mathbb{R}}(x_{(i)} - y_{(i)})^2 \; p_{X_{(i)}}(x_{(i)}) \; p_{Y_{(i)}}(y_{(i)}) \; dx_{(i)} \; dy_{(i)}, \tag{22}$$

where $p_{X_{(i)}}(x_{(i)}) = \frac{N!}{(i-1)!(N-i)!} f_{\theta^*}(x_{(i)}) F_{\theta^*}(x_{(i)})^{i-1}(1 - F_{\theta^*}(x_{(i)}))^{N-i}$ is the density of the $i$-th order statistic from $f_{\theta^*}$ (Bobkov and Ledoux, 2019). An analogous expression holds for $p_{Y_{(i)}}(y_{(i)})$, the $i$-th order statistic from $f_\theta$.

Applying the change of variables $x_{(i)} \mapsto u_{(i)} = F_{\theta^*}(x_{(i)})$ and $y_{(i)} \mapsto v_{(i)} = F_\theta(y_{(i)})$, the loss in (22) can be written as

$$J_N(\theta^*,\theta) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1\int_0^1 (F^{-1}_{\theta^*}(u_{(i)}) - F^{-1}_{\theta}(v_{(i)}))^2 \; p_{U_{(i)}}(u_{(i)}) p_{U_{(i)}}(v_{(i)}) \; dv_{(i)} \; du_{(i)}, \tag{23}$$

where $p_{U_{(i)}}(u_{(i)}) = \frac{N!}{(i-1)!(N-i)!} u_{(i)}^{i-1}(1-u_{(i)})^{N-i}$ is the probability density function of the $i$-th order statistic of a Uniform$(0,1)$ distribution.

Using $m_{1,i}(\theta)$ in (7) and $m_{2,i}(\theta) = \int_0^1 (F^{-1}_\theta(u_{(i)}))^2 p_{U_{(i)}}(u_{(i)}) du_{(i)}$, the expected loss in (23) simplifies to

$$J_N(\theta^*,\theta) = \frac{1}{N}\sum_{i=1}^{N} m_{2,i}(\theta^*) - 2m_{1,i}(\theta^*)m_{1,i}(\theta) + m_{2,i}(\theta) \tag{24}$$

$$= m_2(\theta^*) + m_2(\theta) - \frac{2}{N}\sum_{i=1}^{N} m_{1,i}(\theta^*)m_{1,i}(\theta), \tag{25}$$

where we have used $\frac{1}{N}\sum_{i=1}^{N} m_{2,i}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 (F^{-1}_\theta(u_{(i)}))^2 p_{U_{(i)}}(u_{(i)}) du_{(i)} = \int_0^1 (F^{-1}_\theta(u))^2 du = \int_{\mathbb{R}} y^2 f_\theta(y) dy$ to derive (25), which follows from the permutation-invariance of the sum and the change of variable $u \mapsto y = F^{-1}_\theta(u)$. □

### A.2 Proof of Proposition 3.2

By Lemma 3.1, it remains to compute the terms $m_{1,i}(\theta)$ and $m_2(\theta)$ in (7)-(8) for the location-scale model in order to obtain (9).

Since the inverse CDF of the location-scale model is $F^{-1}_\theta(v) = \theta_1 + \theta_2 \cdot F^{-1}_0(v)$, where $F_0$ is the CDF of the reference density $f_0$, the assumptions that $f_0$ has zero mean and unit variance yield

$$m_{1,i}(\theta) = \theta_1 + \theta_2 \cdot b_i, \tag{26}$$

$$m_2(\theta) = \theta_1^2 + \theta_2^2, \tag{27}$$

where $b_i = \int_0^1 F_0^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)}$.

Substituting (26) and (27) into (6) and simplifying, we obtain

$$J_N(\theta^*, \theta) = \theta_1^2 + \theta_2^2 + \theta_1^{*2} + \theta_2^{*2} - \frac{2}{N}\sum_{i=1}^{N}(\theta_1^* + \theta_2^* \cdot b_i)(\theta_1 + \theta_2 \cdot b_i) \tag{28}$$

$$= (\theta_1 - \theta_1^*)^2 + \theta_2^2 - 2c_N \cdot \theta_2\theta_2^* + \theta_2^{*2}, \tag{29}$$

where $c_N = \frac{1}{N}\sum_{i=1}^{N} b_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\int_0^1 F_0^{-1}(u_{(i)})\ p_{U_{(i)}}(u_{(i)})\ du_{(i)}\right)^2$. Here we also used the fact that $\frac{1}{N}\sum_{i=1}^{N} b_i = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 F_0^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)} = \int_{-\infty}^{\infty} z f_0(z) dz = 0$, since $f_0$ has zero mean.

The resulting objective is quadratic in $(\theta_1, \theta_2)$. Solving the first-order necessary condition $\frac{\partial J_N(\theta^*,\theta)}{\partial \theta} = 0$ yields $\hat{\theta}_{1,N} = \theta_1^*$ and $\hat{\theta}_{2,N} = c_N\theta_2^*$. □

### A.3 Convergence Rates of $J_N(\theta^*, \theta)$ and $c_N$

We derive the convergence rates of $J_N(\theta^*, \theta)$ and $c_N$ in (9) with respect to $N$ by leveraging established results on expected Wasserstein distances.

(i) General case (lighter-than-Gaussian tails): For parametric models with lighter-than-Gaussian tails where the second derivative of the inverse CDF is uniformly bounded, it is established that $\mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] - W_2^2(f_{\theta^*}, f_\theta)$ is of order $O(1/N)$ (see Theorem 5.1 of Bobkov and Ledoux (2019) and Proposition 5 of Papp and Sherlock (2025)).
This can be justified by the following identity (for detailed derivation, please refer to Bobkov and Ledoux (2019); Papp and Sherlock (2025)):

$$\mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] - W_2^2(f_{\theta^*}, f_\theta) = \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}[F_\theta^{-1}(u_{(i)}), F_{\theta^*}^{-1}(u_{(i)})]. \tag{30}$$

Using the Taylor expansion of the quantile functions (e.g., $F_\theta^{-1}$) around the mean of $u_{(i)}$ and noting that the variance of $u_{(i)}$ is of order $O(1/N)$, we can prove that the convergence rate is $O(1/N)$. Consequently, $c_N$ in (9) converges at the same rate.

(ii) Gaussian and exponential cases: For distributions with heavier tails, the proof strategy above is challenging due to tail behaviors. In such cases, we utilize the known convergence rates of empirical measures to the true measure. According to Corollaries 6.12 and 6.14 of Bobkov and Ledoux (2019), the convergence rate of $\mathbb{E}[W_2^2(\hat{f}_{\theta,N}, f_\theta)]$ to zero is $O(\log N/N)$ for exponential distributions and $O(\log\log N/N)$ for Gaussian distributions. We can use this to derive the rate for $c_N$. For a standard location-scale model (location 0, scale 1), the relationship $J_N(\theta, \theta) = \mathbb{E}[W_2^2(\hat{f}_{\theta,N}, \hat{f}'_{\theta,N})] = 2(1 - c_N)$ holds, where $\hat{f}_{\theta,N}$ and $\hat{f}'_{\theta,N}$ are empirical distributions drawn from $f_\theta$ independently of each other. Using the triangle inequality $W_2(\hat{f}_{\theta,N}, \hat{f}'_{\theta,N}) \le W_2(\hat{f}_{\theta,N}, f_\theta) + W_2(\hat{f}'_{\theta,N}, f_\theta)$, squaring both sides via $(a+b)^2 \le 2(a^2+b^2)$, and taking expectations, we find that $J_N(\theta, \theta)$, and hence $c_N$ and $J_N(\theta^*, \theta)$, converges at the same rate as $\mathbb{E}[W_2^2(\hat{f}_{\theta,N}, f_\theta)]$.

(iii) Unequal sample sizes: Similarly, we can analyze the convergence rate for location-scale models with unequal sample sizes. Using a similar proof technique to Lemma 3.1 and Proposition 3.2, the expected Wasserstein loss between $\hat{f}_{\theta^*,N^*}$ and $\hat{f}_{\theta,N}$ becomes

$$\begin{aligned} J_{N^*,N}(\theta^*, \theta) &= \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N^*}, \hat{f}_{\theta,N})] \\ &= (\theta_1 - \theta_1^*)^2 + \theta_2^2 + \theta_2^{*2} - 2c_{N^*,N}\theta_2\theta_2^*, \end{aligned} \tag{31}$$

where $c_{N^*,N}$ is a constant depending on $N$ and $N^*$. (For example, when $N^* = rN$ with $r \in \mathbb{N}$, we have $c_{N^*,N} = \frac{1}{rN}\sum_{i=1}^{rN}\mathbb{E}[F_0^{-1}(u_{(i)})]\mathbb{E}[F_0^{-1}(v_{(\lceil i/r \rceil)})]$, where $u_{(i)}$ and $v_{(i)}$ denote the $i$-th order statistics of $rN$ and $N$ uniform samples, respectively.)
We first identify the convergence rate of $c_{N^*,N}$ and then use it to characterize that of $J_{N^*,N}(\theta^*, \theta)$. For

a standard location-scale model with $\theta = \theta^* = (0, 1)$, the explicit relationship $J_{N^*,N}(\theta, \theta) = 2(1 - c_{N^*,N})$ holds. Using the triangle inequality $W_2(\hat{f}_{\theta,N}, \hat{f}_{\theta,N^*}) \leq W_2(\hat{f}_{\theta,N}, f_\theta) + W_2(\hat{f}_{\theta,N^*}, f_\theta)$, squaring both sides via $(a + b)^2 \leq 2(a^2 + b^2)$, and taking expectations, we obtain:

$$2(1 - c_{N^*,N}) \leq 2\mathbb{E}[W_2^2(\hat{f}_{\theta,N}, f_\theta)] + 2\mathbb{E}[W_2^2(\hat{f}_{\theta,N^*}, f_\theta)]. \tag{32}$$

This shows that the error is controlled by the smaller of the two sample sizes, i.e., $\min(N, N^*)$. Thus, the convergence rates of $c_{N^*,N}$ and $J_{N^*,N}(\theta^*, \theta)$ are determined by the smaller sample size.

### A.4 Further Discussion of Bias Characterization and Correction

Consider $J_N(\theta^*, \theta)$ in (5) as a function over the joint parameter space. The loss is symmetric in its arguments, meaning that $J_N(\theta^*, \theta) = J_N(\theta, \theta^*)$. As a result, along $\theta = \theta^*$, the partial derivatives with respect to each argument must be equal, i.e., $\frac{\partial J_N(\theta^*,\theta)}{\partial \theta}\Big|_{\theta=\theta^*=\theta_0} = \frac{\partial J_N(\theta^*,\theta)}{\partial \theta^*}\Big|_{\theta^*=\theta=\theta_0}$ for any parameter value $\theta_0$.

In the one-dimensional parameter case, consider the directional derivative of $J_N(\theta^*, \theta)$ along the line $\theta^* = \theta$. It is given by

$$\frac{\partial J_N(\theta, \theta)}{\partial \theta}\bigg|_{\theta=\theta_0} = \frac{\partial J_N(\theta^*, \theta)}{\partial(\theta^*, \theta)}\bigg|_{\theta=\theta^*=\theta_0} \cdot (1, 1)^\top = 2 \cdot \frac{\partial J_N(\theta^*, \theta)}{\partial \theta}\bigg|_{\theta=\theta^*=\theta_0}, \tag{33}$$

which is nonzero if and only if the gradient at $\theta = \theta^*$ is nonzero.

By contrast, the directional derivative in the orthogonal direction $(1, -1)^\top$ vanishes due to symmetry, i.e., $\frac{\partial J_N(\theta^*,\theta)}{\partial(\theta^*,\theta)}\Big|_{\theta=\theta^*=\theta_0} \cdot (1, -1)^\top = 0$.

This implies that, locally, variation of the loss around $\theta = \theta^*$ occurs only along the diagonal. Hence, the condition $\frac{\partial J_N(\theta,\theta)}{\partial \theta} = 0$ fully characterizes whether $\theta = \theta^*$ is a stationary point of the expected loss for fixed $\theta^*$. If this condition fails, i.e., if the gradient is nonzero, then the minimizer must lie at some $\theta \neq \theta^*$.

Note that this reasoning extends naturally to the case of higher-dimensional data, where the loss $J_N(\theta^*, \theta) = \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})]$ remains symmetric in its two arguments. Furthermore, it is applicable to the case of higher-dimensional parameters, where $\theta = \theta^*$ defines the diagonal subspace along which local variations in the loss around $\theta = \theta^*$ are confined to occur.

**Gaussian Density Examples** We examine the expected Wasserstein loss $J_N(\theta^*, \theta)$ at $\theta = \theta^*$ in the context of Gaussian scale parameter estimation. Figure 2(c) shows that $J_N(\theta, \theta)$ increases with the scale parameter $\theta_2$, reflecting that sample transport distances grow with scale, with a steeper increase when the number of samples is small. By Remark 3.3 and (33), this implies that a scale parameter smaller than the true value $\theta^*$ can yield lower expected loss, thereby inducing the downward bias in the minimizer, as observed in Figure 2(b). As the sample size $N$ increases, the gradient of the expected loss along the diagonal diminishes, indicating that the bias vanishes asymptotically.

### A.5 Finite-Sample Bias and Its Asymptotic Convergence Behavior in Higher-Dimensional Settings

Leveraging established results on the convergence rate of the empirical Wasserstein distance in general dimensions (Boissard and Le Gouic, 2014; Fournier and Guillin, 2015; Weed and Bach, 2019), we expect the parameter estimate to converge at a rate governed by that of the empirical Wasserstein distance. This intuition is supported by the following perturbation analysis.

Let the convergence rate of the empirical Wasserstein distance be $O(s(N, d))$, where $s(N, d)$ is a function of the dimension $d$ and the sample size $N$. For sufficiently large $N$, we model the expected loss in (5) as

$$J_N(\theta^*, \theta) \approx W_2^2(f_{\theta^*}, f_\theta) + h(\theta)O(s(N, d)), \tag{34}$$

where $\theta^* \in \mathbb{R}^m$ denotes the fixed parameter, $\theta \in \mathbb{R}^m$ is the optimization variable, and $h(\theta)$ captures the coefficient of the fluctuation term. Let $\hat{\theta} \in \mathbb{R}^m$ denote a minimizer of $J_N(\theta^*, \theta)$, and define

$$g(\theta) = \nabla_\theta W_2^2(f_{\theta^*}, f_\theta) \in \mathbb{R}^m \quad \text{and} \quad H(\theta) = \nabla_\theta^2 W_2^2(f_{\theta^*}, f_\theta) \in \mathbb{R}^{m\times m}.$$

Then the first-order optimality condition suggests

$$g(\hat{\theta}) + \nabla_\theta h(\theta)|_{\theta=\hat{\theta}}\, O(s(N,d)) \approx 0. \tag{35}$$

Assuming that $\hat{\theta} - \theta^*$ is small, a first-order Taylor expansion of $g(\hat{\theta})$ around $\theta^*$ gives

$$g(\hat{\theta}) \approx g(\theta^*) + H(\theta^*)(\hat{\theta} - \theta^*). \tag{36}$$

Since $\theta^*$ minimizes the population loss, we have $g(\theta^*) = 0$. Hence, if $H(\theta^*)$ is invertible,

$$\hat{\theta} - \theta^* \approx -H(\theta^*)^{-1}\, \nabla_\theta h(\theta)|_{\theta=\hat{\theta}}\, O(s(N,d)). \tag{37}$$

This derivation suggests that the convergence rate of the parameter estimate $\hat{\theta}$ is governed by the order $O(s(N,d))$, namely, by the convergence rate of the empirical Wasserstein distance.

Although the above derivation is written around the well-specified optimum $\theta^*$, the same perturbative argument can be extended to misspecified settings by expanding around the population (or infinite-sample) minimizer $\theta^\dagger \in \mathbb{R}^m$ of the Wasserstein loss. We therefore expect the induced parameter error in misspecified settings to be governed by the same empirical convergence rate under suitable local regularity conditions.

### A.6 Finite-Sample Bias in the Semi-Discrete Case

In this section, we discuss the semi-discrete case in (4), where we fit a parametric model $f_\theta$ to a data set $x_1, \ldots, x_N \sim \mu$ by minimizing the Wasserstein loss. The expected squared $W_2$ loss between $\hat{\mu}_N$ and $f_\theta$ becomes (Bobkov and Ledoux, 2019)

$$J_N^{\mathrm{SD}}(\theta) = \mathbb{E}\big[W_2^2(\hat{\mu}_N, f_\theta)\big] = \sum_{i=1}^{N} \int_{(i-1)/N}^{i/N} \mathbb{E}\big[(x_{(i)} - F_\theta^{-1}(u))^2\big]\, \mathrm{d}u \tag{38}$$

$$= \mathbb{E}[x^2] + m_2(\theta) - \frac{2}{N} \sum_{i=1}^{N} \mathbb{E}[x_{(i)}]\, q_{1,i}(\theta), \tag{39}$$

where the expectations are taken with respect to the data distribution $\mu$ and $q_{1,i}(\theta) = N \int_{(i-1)/N}^{i/N} F_\theta^{-1}(u) du$. In (39), the last term depends on $N$ and can deviate from the infinite-sample limit, indicating the presence of finite-sample bias in the loss.

When $f_\theta$ belongs to the location-scale model discussed in Section 3.2, (39) becomes as follows:

$$J_N^{\mathrm{SD}}(\theta) = \mathbb{E}[x^2] + \theta_1^2 + \theta_2^2 - 2\mathbb{E}[x]\theta_1 - \frac{2}{N} \sum_{i=1}^{N} \mathbb{E}[x_{(i)}] q_i \theta_2, \tag{40}$$

where $q_i = N \int_{(i-1)/N}^{i/N} F_0^{-1}(u) du$. The optimal location and scale parameters minimizing (40) are $\hat{\theta}_{N,1} = \mathbb{E}[x]$ and $\hat{\theta}_{N,2} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[x_{(i)}] q_i$, respectively.[3] When $N$ is finite, $\hat{\theta}_{N,2}$ can deviate from its infinite sample limit (as verified in the numerical examples in Appendices B.4.2 and B.6.2), reconciling with the observation of Bellemare et al. (2017).

In the well-specified case of $\mu = f_{\theta^*}$ for a fixed $\theta^*$, the optimal parameters become $\hat{\theta}_{N,1} = \theta_1^*$ and $\hat{\theta}_{N,2} = \left(\frac{1}{N} \sum_{i=1}^{N} b_i q_i\right) \theta_2^*$, where $b_i = \int_0^1 F_0^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)}$ and we have used $\mathbb{E}[x_{(i)}] = \theta_1^* + \theta_2^* b_i$ and $\frac{1}{N} \sum_{i=1}^{N} q_i = 0$. Similarly to the results in Section 3.2, when $N$ is finite, we generally have $\hat{\theta}_{N,2} \neq \theta_2^*$, biased by a constant factor $\frac{1}{N} \sum_{i=1}^{N} b_i q_i$.

[3] When we minimize the empirical loss $W_2^2(\hat{\mu}_N, f_\theta)$, the optimal parameters are $\hat{\theta}_1 = \frac{1}{N} \sum_{i=1}^{N} x_i$ and $\hat{\theta}_2 = \frac{1}{N} \sum_{i=1}^{N} x_{(i)} q_i$, respectively. Their expectations become $\mathbb{E}[\hat{\theta}_1] = \mathbb{E}[x]$ and $\mathbb{E}[\hat{\theta}_2] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[x_{(i)}] q_i$, coinciding with the minimizer of the expected loss in (40) hence biased from the minimizer of $W_2^2(\mu, f_\theta)$.

**Bias Correction** For one-dimensional parameters, the bias can be corrected following Section 4.2 by considering the modified objective

$$\tilde{J}_N^{\mathrm{SD}}(\theta;\alpha) = J_N^{\mathrm{SD}}(\theta) - \alpha J_N(\theta,\theta). \tag{41}$$

Appropriately choosing $\alpha$ aligns the minimizer of (41) with the infinite-sample solution. We estimate such an $\alpha$ using the heuristic of Section 4.2, replacing the numerator in (16) with the gradient of $J_N^{\mathrm{SD}}(\theta)$. Experimental results for the semi-discrete case are discussed in Appendices B.4.2 (Gaussian) and B.6.2 (affine model).

The minimum expected Wasserstein estimator (MEWE) in Bernton et al. (2019b) also suffers from finite-sample bias, since the semi-discrete distance $W_2^2(\hat{\mu}_N, f_\theta)$ is approximated by replacing $f_\theta$ with an empirical version. The same heuristic discussed above can be applied to correct this bias, with details provided in Appendix A.7.

### A.7 Finite-Sample Bias in the Minimum Expected Wasserstein Estimator

The minimum expected Wasserstein estimator (MEWE) replaces $f_\theta$ in (4) with an empirical distribution $\hat{f}_{\theta,M(N)}$ constructed from it when the likelihood function is intractable but sampling is tractable. The consistency of this estimator is discussed in Bernton et al. (2019b):

$$\hat{\theta}_N = \arg\min_\theta \mathbb{E}_{\hat{f}}\left[W_p(\hat{\mu}_N, \hat{f}_{\theta,M(N)})\right] \to \theta^* = \arg\min_\theta W_p(\mu, f_\theta) \text{ as } N\to\infty, \tag{42}$$

where $M(N)$ is a function of $N$ satisfying $M(N)\to\infty$ as $N\to\infty$, and $\mathbb{E}_{\hat{f}}[\cdot]$ denotes expectation with respect to the randomness of $\hat{f}_{\theta,M(N)}$.

We analyze the finite-sample bias in the minimum expected Wasserstein estimator (MEWE). For simplicity, let $M(N) = rN$ with $r\in\mathbb{N}$. Using a derivation analogous to that in the proof of Lemma 3.1, the expected distance becomes

$$J_{rN}^{\mathrm{SD,approx}}(\theta) = \mathbb{E}_{\hat{f}}\left[W_2^2(\hat{\mu}_N, \hat{f}_{\theta,rN})\right], \tag{43}$$

$$= \frac{1}{N}\sum_{i=1}^N x_i^2 - \frac{2}{rN}\sum_{i=1}^{rN} x_{(\lceil i/r\rceil)} m_{1,i}(\theta) + m_2(\theta), \tag{44}$$

where $\lceil i/r\rceil$ is the ceiling of $i/r$ and $m_{1,i}(\theta) = \int_0^1 F_\theta^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)}$ with $p_{U_{(i)}}(u_{(i)}) = \frac{(rN)!}{(i-1)!(rN-i)!} u_{(i)}^{i-1}(1-u_{(i)})^{rN-i}$.

As $r\to\infty$, (43) converges to the semi-discrete loss $W_2^2(\hat{\mu}_N, f_\theta)$. For finite $r$, however, the loss and its minimizer can exhibit finite-sample bias.

In the location–scale model in Section 3.2, the approximate distance becomes

$$J_{rN}^{\mathrm{SD,approx}}(\theta) = \frac{1}{N}\sum_{i=1}^N x_i^2 - 2\bar{x}\theta_1 - \frac{2}{rN}\sum_{i=1}^{rN} x_{(\lceil i/r\rceil)} b_i \theta_2 + \theta_1^2 + \theta_2^2, \tag{45}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and $b_i = \int_0^1 F_0^{-1}(u_{(i)}) p_{U_{(i)}}(u_{(i)}) du_{(i)}$.

The corresponding minimizers are $\hat{\theta}_{rN,1} = \bar{x}$ and $\hat{\theta}_{rN,2} = \frac{1}{rN}\sum_{i=1}^{rN} b_i x_{(\lceil i/r\rceil)}$, respectively. As $r$ increases, $\hat{\theta}_{rN,2}$ converges to the semi-discrete solution $\hat{\theta}_2 = \frac{1}{N}\sum_{i=1}^N x_{(i)} q_i$ with $q_i = N\int_{(i-1)/N}^{i/N} F_0^{-1}(u)du$.

**Bias Correction** The heuristic from Section 4.2 also applies in this setting. For one-dimensional parameters, we can correct the finite-sample bias by minimizing

$$\tilde{J}_{rN}^{\mathrm{SD,approx}}(\theta;\alpha) = J_{rN}^{\mathrm{SD,approx}}(\theta) - \alpha J_{rN}(\theta,\theta), \tag{46}$$

with $\alpha$ chosen via a similar heuristic procedure. Specifically, we compute

$$\alpha_{rN}(\theta) = \frac{\partial J_{rN}^{\mathrm{SD,approx}}(\theta)/\partial\theta}{\partial J_{rN}(\theta,\theta)/\partial\theta} \tag{47}$$

for several values of $r$, and identify the $\theta$ at which $\alpha_{rN}(\theta)$ remains nearly constant across $r$. We can take this $\theta$ as a proxy for the semi-discrete solution $\hat{\theta} = \arg\min_\theta W_2^2(\hat{\mu}_N, f_\theta)$, and use the corresponding average of $\alpha_{rN}(\theta)$ across $r$ as $\alpha$ in the bias-corrected loss (46).

### A.8 Expected Sinkhorn Divergence for the Location-Scale Model as $\varepsilon \to \infty$

When $\varepsilon \to \infty$, the entropic regularization corresponds to pure entropy maximization. In this case, the optimal transport plan between two empirical distributions is the uniform coupling $\gamma(x_i, y_j) = \frac{1}{N^2}, i, j = 1, \ldots, N$.

Assuming a one-dimensional sample space, the expectation of (17) between two empirical distributions is

$$\mathbb{E}[\mathrm{OT}_\infty(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] = \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}[(x_i - y_j)^2] = m_2(\theta^*) + m_2(\theta) - 2m_1(\theta^*)m_1(\theta). \tag{48}$$

For the case of the same empirical distribution, we have

$$\mathbb{E}[\mathrm{OT}_\infty(\hat{f}_{\theta,N}, \hat{f}_{\theta,N})] = \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}[(x_i - x_j)^2] = \frac{1}{N^2}\sum_{i\neq j}\mathbb{E}[(x_i - x_j)^2] = \frac{2(N-1)}{N}(m_2(\theta) - m_1(\theta)^2). \tag{49}$$

Therefore, the expected Sinkhorn divergence in (19) is

$$S_{N,\infty}(\theta^*, \theta) = \frac{1}{N}(m_2(\theta^*) + m_2(\theta) - m_1(\theta^*)^2 - m_1(\theta)^2) + (m_1(\theta^*) - m_1(\theta))^2. \tag{50}$$

Substituting $m_1(\theta) = \theta_1$ and $m_2(\theta) = \theta_1^2 + \theta_2^2$ for the location–scale model, we obtain

$$S_{N,\infty}(\theta^*, \theta) = \frac{1}{N}({\theta_2^*}^2 + \theta_2^2) + (\theta_1^* - \theta_1)^2. \tag{51}$$

Thus, when $\varepsilon \to \infty$, the expected Sinkhorn divergence is minimized at a scale parameter of zero for any finite $N$. As $\varepsilon$ grows large, the downward bias on the scale parameter becomes stronger, which is consistent with the empirical observations reported in Section 5.

### A.9 Expected $W_2$ Loss for an Affine PDF Model

As a qualitatively different example to the location-scale models, we consider the following affine PDF model with a finite support:

$$f_a(x) = a(x - 0.5) + 1, \quad -2 \le a \le 2, \quad x \in [0, 1], \tag{52}$$

where $a$ is a slope parameter.

In what follows, we use the term *discrete—discrete case* to denote the setting where both arguments of the Wasserstein loss are empirical distributions, and *semi-discrete case* when one argument is an empirical distribution and the other is a population distribution.

#### A.9.1 Discrete–Discrete Case

We analyze the expected squared $W_2$ distance between empirical distributions $\hat{f}_{a^*,N}$ and $\hat{f}_{a,N}$, constructed from samples drawn from $f_{a^*}$ and $f_a$, respectively. To derive the expected distance, it suffices by Lemma 3.1 to compute $m_{1,i}(a)$ and $m_2(a)$ in (7)-(8). The second moment is

$$m_2(a) = \int_0^1 x^2 p(x; a)dx = \frac{a+4}{12}, \tag{53}$$

where $p(x; a)$ is the affine density defined in (52).

We next derive $m_{1,i}(a)$ in (7) for the affine PDF model. The inverse CDF is

$$F_a^{-1}(u) = \begin{cases} u, & a = 0, \\ \frac{1}{2} - \frac{1}{a} + \frac{1}{a}\sqrt{2au + \left(\frac{a}{2} - 1\right)^2}, & a \neq 0. \end{cases} \tag{54}$$

To compute $m_{1,i}(a)$, we make use of the following Gauss-hypergeometric integral identity (Gradshteyn and Ryzhik, 2014):

$$\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}(1-zt)^{-p}dt = B(\alpha,\beta)_2F_1(p,\alpha;\alpha+\beta;z), \tag{55}$$

where $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function, $\Gamma(\cdot)$ is the Gamma function, and ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the hypergeometric function.

Applying this identity, we obtain the following closed-form expression:

$$m_{1,i}(a) = \begin{cases} \frac{i}{N+1}, & a=0, \\ \frac{1}{2}-\frac{1}{a}+\frac{\sqrt{b}}{a}\ {}_2F_1\left(-\frac{1}{2},i;N+1;-\frac{2a}{b}\right), & a\neq 0, \end{cases} \tag{56}$$

where $b=\left(\frac{a}{2}-1\right)^2$.

Substituting (53) and (56) into (6) allows us to evaluate the expected loss. Minimization with respect to $a$ for a fixed $a^*$ can then be performed using a numerical solver.

### A.9.2 Semi-Discrete Case

We analyze the expected squared $W_2$ distance between an empirical distribution $\hat{\mu}_N$ and the density model $f_a$. To derive the distance in this semi-discrete case, it suffices by (39) to compute $q_{1,i}(a)=N\int_{(i-1)/N}^{i/N}F_a^{-1}(u)du$.

From the inverse CDF in (54), a straightforward computation yields

$$q_{1,i}(a) = \begin{cases} \frac{2k-1}{2N}, & a=0, \\ \frac{1}{2}-\frac{1}{a}+\frac{N}{3a^2}\left[\left(\frac{a^2}{4}-a+1+\frac{2ai}{N}\right)^{3/2}-\left(\frac{a^2}{4}-a+1+\frac{2a(i-1)}{N}\right)^{3/2}\right], & a\neq 0. \end{cases} \tag{57}$$

Substituting (53) and (57) into (39) allows us to evaluate the expected loss. Minimization with respect to $a$ can then be performed using a numerical solver, where the gradient of $q_{1,i}(a)$ in (57) can be obtained straightforwardly with $q'_{1,i}(0)=(6Nk-3N-6k^2+6k-2)/12N^3$.

## B ADDITIONAL NUMERICAL EXPERIMENTS

### B.1 Stochastic Optimization Settings: Gaussian, Tukey $g$-and-$h$, and Affine Models

We study stochastic optimization of expected Wasserstein losses using Gaussian (Appendix B.4), Tukey $g$-and-$h$ (Appendix B.5), and affine PDF models (Appendix B.6) as representative examples.

We approximate the expected losses by Monte Carlo estimation. In particular, for each Monte Carlo replicate, we draw the required independent samples, construct the corresponding empirical distributions, and compute the Wasserstein distances using them.

We consider both the discrete–discrete and semi-discrete settings, corresponding respectively to minimizing the Wasserstein loss between two empirical distributions and between an empirical distribution and a parametric model. The discrete-discrete setting is studied for all three models, whereas the semi-discrete setting is considered only for the Gaussian and affine models. Details are provided in Appendices B.1.1 and B.1.2, respectively.

#### B.1.1 Discrete–Discrete Case

At each iteration, we draw batches of size $N$ from both $f_{\theta^*}$ (or the data-generating distribution $\mu$ for misspecified cases) and $f_\theta$, compute the empirical Wasserstein loss, and update parameters using SGD. For gradient-based training, we construct models as a differentiable sampling map $z\mapsto g_\theta(z)$ from a simple latent variable $z$, enabling backpropagation through the sampling process. This allows stochastic gradients of the Wasserstein loss to be computed with respect to the parameter $\theta\in\mathbb{R}^m$.

We adopt a Robbins-Monro step-size schedule, $\eta_t = \frac{\eta_0}{1+\gamma \cdot t}$, for SGD (Robbins and Monro, 1951). In the Gaussian model experiments of Figures 3 (Section 3.3) and 8(a)-(b) (Appendix B.4.1), we set $(\eta_0, \gamma) = (0.01, 0.01)$, while for Figures 8 (c)-(d), we use $(0.01, 0.0025)$. For the Tukey $g$–$h$ model experiments in Figure 14 (Appendix B.5.2), we set $(\eta_0, \gamma) = (0.001, 0.001)$. For Figures 17(a), (b), and (c) (Appendix B.5.3), the settings are $(0.2, 0.001)$, $(1.5, 0.001)$, and $(0.001, 0.001)$, respectively. For the affine model experiments in Figure 19 (Appendix B.6.1), we set $(\eta_0, \gamma) = (1.0, 0.01)$.

### B.1.2 Semi-Discrete Case

We minimize the expected Wasserstein loss between an empirical distribution $\hat{\mu}_N$ constructed from $N$ samples of the data-generating distribution $\mu$ and the parametric model $f_\theta$. At each iteration, $N$ samples are drawn from $\mu$, the semi-discrete loss $W_2^2(\hat{\mu}_N, f_\theta)$ is evaluated, and parameters are updated using SGD with the same Robbins–Monro schedule as above. For the Gaussian model experiments in Figure 11 (Appendix B.4.2), we set $(\eta_0, \gamma) = (0.01, 0.01)$, and for the affine model experiments in Figure 22 (Appendix B.6.2), we set $(\eta_0, \gamma) = (1.0, 0.01)$.

## B.2 Stochastic Optimization Settings: Gaussian Mixture Models

We consider a well-specified setting in which both the true distribution $f_{\theta^*}$ and the model $f_\theta$ are multivariate Gaussian mixtures with the same known number of components and mixture weights. The model parameters are initialized using $k$-means clustering, and the goal is to recover the component means and covariances by minimizing the Wasserstein loss using SGD.

At each iteration of SGD, we draw batches of size $N$ from both $f_{\theta^*}$ and $f_\theta$, evaluate the empirical transport objective, and update the parameters using its gradient. Since analytic expressions for the Wasserstein loss, such as (2)–(3), are generally unavailable for higher-dimensional data, we compute an optimal transport plan between the empirical distributions using the `POT` library (Flamary et al., 2021) and use the resulting transport cost as the objective value. For the Sinkhorn divergence in Section 5, we compute the objective using the `GeomLoss` library (Feydy et al., 2019). In both cases, we compute the gradient by treating the current transport coupling as fixed and differentiating the corresponding transport cost with respect to the model parameters; this yields a gradient (or subgradient) of the loss, justified by an envelope-theorem argument (Milgrom and Segal, 2002). As noted in Appendix A.4, the bias correction scheme from Section 3.3 extends naturally to these higher-dimensional settings, and the bias correction term and its gradient are computed in the same way.

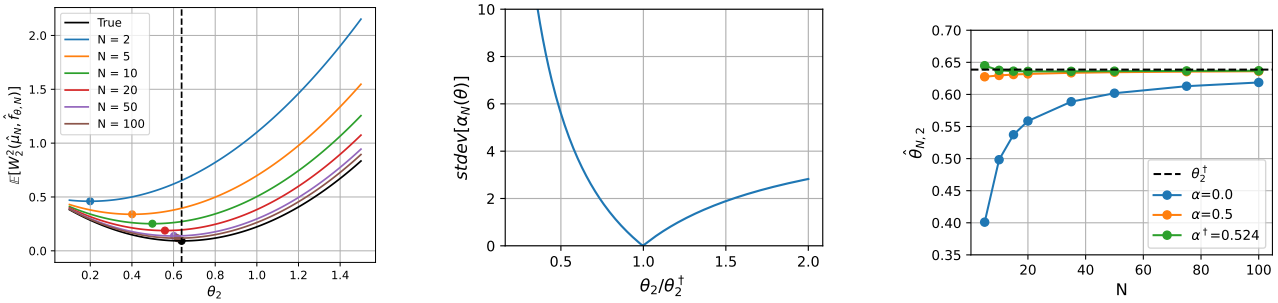
To quantitatively evaluate performance, we use the following fitting error:

$$\text{Fitting error} = \sum_{k=1}^{K} \pi_k W_2^2(\mathcal{N}(\mu_k^*, \Sigma_k^*), \mathcal{N}(\mu_k, \Sigma_k)), \tag{58}$$

where $\pi_k$ is the known weight of the $k$-th component, $\mu_k^* \in \mathbb{R}^d, \Sigma_k^* \in \mathbb{R}^{d\times d}$ are the true mean and covariance parameters, and $\mu_k \in \mathbb{R}^d, \Sigma_k \in \mathbb{R}^{d\times d}$ are the learned ones from optimization. We use the closed-form formula for the squared $W_2$ distance between two Gaussians (Peyré et al., 2019), applied component-wise. This weighted sum serves as an upper bound on the true distance and provides a practical accuracy metric.

We experiment with mixtures of $K = 4$ Gaussians in dimensions $d = 2, 5$, and 10. The mixture weights are fixed as $\pi = (0.4, 0.3, 0.2, 0.1)$. For $d = 2$, the means are $\mu_1^* = (3, 0)$, $\mu_2^* = (0, 3)$, $\mu_3^* = (-3, 0)$, and $\mu_4^* = (0, -3)$. For $d = 5$ and 10, they are $\mu_1^* = (3, 0, \ldots, 0)$, $\mu_2^* = (0, 3, 0, \ldots, 0)$, $\mu_3^* = (0, 0, 3, 0, \ldots, 0)$, and $\mu_4^* = (0, \ldots, 0)$. Covariance matrices are set randomly and kept fixed across runs.

To generate samples, we first draw a component index $k \in \{1, 2, 3, 4\}$ from the categorical distribution with weights $\pi$. A sample is then obtained as $y = L_k z + \mu_k$, where $z \sim \mathcal{N}(0, I)$, $\mu_k \in \mathbb{R}^d$, and $L_k \in \mathbb{R}^{d\times d}$. We train with SGD using batch sizes $N \in \{16, 32, 64, 128, 256, 512, 1024\}$, a learning rate of 0.01, and 20,000 iterations. When minimizing the Sinkhorn divergence, we use 50,000 iterations. Fitting performance is evaluated using the error metric in (58), with $\Sigma_k = L_k L_k^\top$. Figures 3(c) (Section 3.3) and 6(c) (Section 5) report fitting errors averaged over three random seeds.




(a) $J_N^{\mathrm{MS}}(\theta)$ (b) Standard deviation of $\alpha_N(\theta)$ (c) Bias reduction

Figure 7: The expected Wasserstein loss $J_N^{\mathrm{MS}}(\theta)$ and the application of the bias correction scheme in the misspecified case for Gaussian models across different sample sizes $N$. In (a), $J_N^{\mathrm{MS}}(\theta)$ is shown for the misspecified case (Exponential, scale $1/\sqrt{2}$). Colored solid curves depict $J_N^{\mathrm{MS}}(\theta)$, with minimizers $\hat{\theta}_{N,2}$ marked by dots. In (b), for each $\theta$, the standard deviation of $\{\alpha_N(\theta) \mid N \in \{5, 10, 15, 20, 35, 50, 75, 100\}\}$ is computed. In (c), solid curves represent the minimizers of the bias-corrected losses for $\alpha = 0, 0.5$, and $\alpha^\dagger$. The black dashed line indicates the infinite-sample solution $\theta^\dagger$.

### B.3 Stochastic Optimization Settings: A Neural Network Generator Example

We use the Gaussian mixture data configuration described in Appendix B.2 with $d = 2, 5$, and 10. The generator is a multilayer perceptron (MLP) with ReLU activations, mapping 10-dimensional Gaussian latent variables to the data space through linear layers of dimensions $10 \to 128 \to 32 \to d$.

We employ stochastic optimization. At each iteration, a batch of synthetic samples generated by the MLP is compared with an equal-sized batch drawn from the ground-truth Gaussian mixture. The model is trained by minimizing the Wasserstein loss or Sinkhorn divergence between the resulting empirical distributions, with the objectives and gradients computed as described in Appendix B.2.

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate 0.005 for 10,000 iterations. Experiments are conducted with batch sizes in $\{32, 64, 128, 256, 512, 1024\}$. For Sinkhorn divergences, we use $\varepsilon = 1$. Figures 5(c)–(d) (Section 4.3) and 23(c)–(d) (Appendix B.7) report averages over four random seeds.

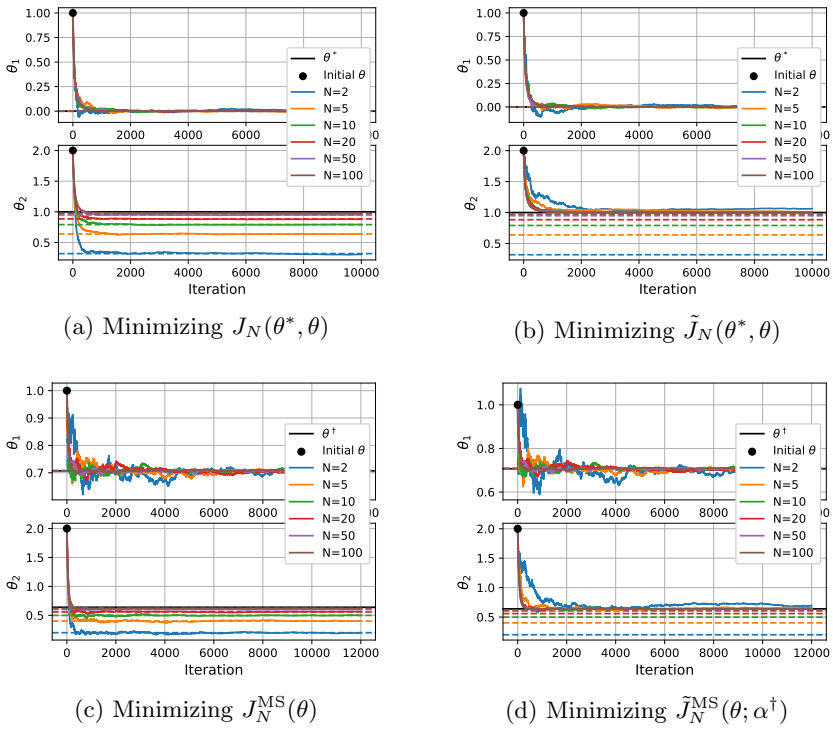
### B.4 Results for the Gaussian Model

#### B.4.1 Discrete–Discrete Case

In this section, we investigate the misspecified case using a Gaussian parametric model with data generated from an exponential distribution of scale $1/\sqrt{2}$. Since the location parameter is unbiased, we fix it at its optimal value and focus on the scale parameter. We denote the solutions in the infinite sample limit by $\theta^\dagger = (\theta_1^\dagger, \theta_2^\dagger) = (\hat{\theta}_{\infty,1}, \hat{\theta}_{\infty,2})$.

As shown in Figure 7(a), the expected Wasserstein loss $J_N^{\mathrm{MS}}(\theta)$ is strictly larger than its infinite-sample counterpart and yields biased minimizers, with $\hat{\theta}_{N,2}$ substantially underestimating $\theta_2^\dagger$ for small $N$ but converging as $N$ grows. Figure 7(b) further shows that the standard deviation of the heuristic coefficient $\alpha_N(\theta)$ is minimized near $\theta_2 = \theta_2^\dagger$. Finally, Figure 7(c) demonstrates that applying the proposed correction with $\alpha = \alpha^\dagger$ consistently reduces the finite-sample bias and yields solutions closer to the infinite-sample optimum, outperforming both the uncorrected case ($\alpha = 0$) and the fixed choice $\alpha = 0.5$.

Figures 8(a) and (b) extend the main-text results in Figure 3(b) on the stochastic optimization of $J_N(\theta^*, \theta)$ and $\tilde{J}_N(\theta^*, \theta)$ by incorporating both the location and scale parameters. In Figure 8(a), the location parameter remains unbiased even for small $N$, while the scale parameter shows a downward bias that diminishes as $N$ increases. In contrast, minimizing the bias-corrected loss $\tilde{J}_N(\theta^*, \theta)$ in Figure 8(b) eliminates this finite-sample bias, aligning the solutions with the true parameters even when $N$ is small. A similar pattern is observed in the misspecified case (Exponential, scale $1/\sqrt{2}$) in Figures 8(c) and (d), corresponding to minimization of $J_N^{\mathrm{MS}}(\theta)$ and $\tilde{J}_N^{\mathrm{MS}}(\theta; \alpha^\dagger)$, respectively.




(a) Minimizing $J_N(\theta^*, \theta)$

(b) Minimizing $\tilde{J}_N(\theta^*, \theta)$

(c) Minimizing $J_N^{\mathrm{MS}}(\theta)$

(d) Minimizing $\tilde{J}_N^{\mathrm{MS}}(\theta; \alpha^\dagger)$

Figure 8: Stochastic optimization of empirical Wasserstein losses for Gaussian distributions with different sample sizes $N$. In (a), SGD is applied to minimize $J_N(\theta^*, \theta)$ with respect to $\theta = (\theta_1, \theta_2)$ for fixed $\theta^*$. In (b), the same is shown for the bias-corrected loss $\tilde{J}_N(\theta^*, \theta)$. In (c) and (d), we report the misspecified case (Exponential, scale $1/\sqrt{2}$), minimizing $J_N^{\mathrm{MS}}(\theta)$ and $\tilde{J}_N^{\mathrm{MS}}(\theta; \alpha^\dagger)$, respectively. Colored solid lines represent parameter trajectories; dashed lines indicate the minimizers of $J_N(\theta^*, \theta)$ in (a)-(b) and $J_N^{\mathrm{MS}}(\theta)$ in (c)-(d); black solid lines denote the infinite-sample solutions.

### B.4.2 Semi-Discrete Case

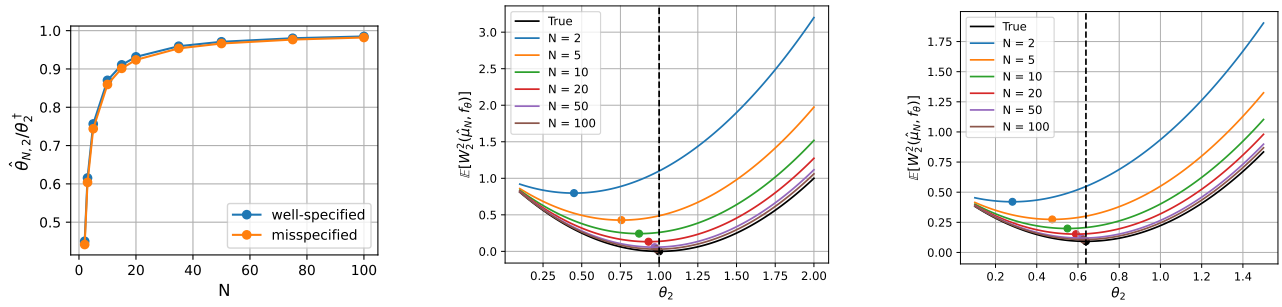
We illustrate finite-sample bias in the semi-discrete case using a Gaussian parametric model. We consider (i) a well-specified case where the data distribution $\mu$ is Gaussian with $\theta_1^* = 0$, $\theta_2^* = 1$, and (ii) a misspecified case where $\mu$ is exponential with scale $1/\sqrt{2}$. The infinite-sample solution is denoted $\theta^\dagger = (\theta_1^\dagger, \theta_2^\dagger)$. Since the location parameter is unbiased, we fix $\theta_1 = \theta_1^\dagger$ and focus on the scale parameter.

Figure 9(a) shows that $\hat{\theta}_{N,2}/\theta_2^\dagger$ is substantially below one for small $N$—indicating downward bias of the scale—but approaches one as $N$ increases, consistent with the asymptotic consistency of minimum Wasserstein estimation (Bassetti et al., 2006; Bernton et al., 2019b). Figures 9(b)–(c) show $J_N^{\mathrm{SD}}(\theta)$ in (40) for different $N$: the finite-sample objectives lie strictly above the population loss, are not minimized at $\theta = \theta^\dagger$, and yield biased minimizers.

Figure 10(a) plots the standard deviation of $\alpha_N(\theta)$ from (16). The variance is minimized near $\theta_2/\theta_2^\dagger = 1$, with similar shapes in both settings. This arises because, in location–scale models, $\partial J_\infty^{\mathrm{SD}}(\theta)/\partial\theta_2 = 2(\theta_2 - \theta_2^\dagger)$ and $d(\theta) = 4\theta_2$, implying that the standard deviation of $\alpha_N(\theta)$ can be approximated by $c|1 - \theta_2^\dagger/\theta_2|$ for some constant $c$ independent of the underlying distribution.

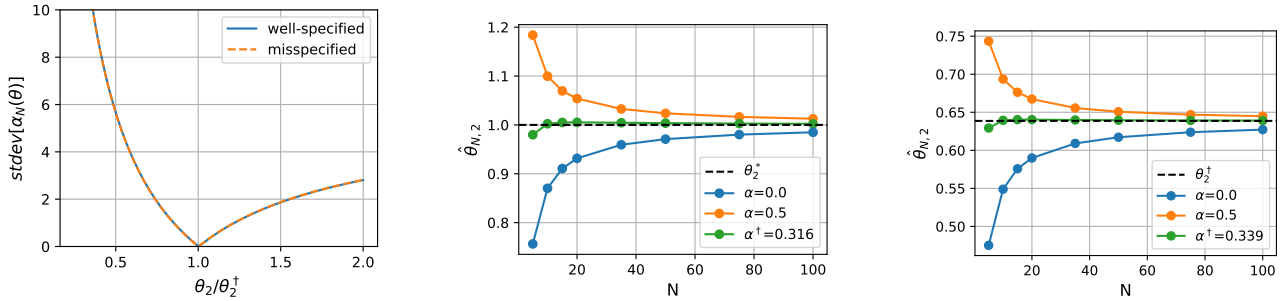
Figures 10(b)–(c) show bias reduction with different $\alpha$. Across $N \in [5, 100]$, setting $\alpha = \alpha^\dagger$ (as in (15), adapted to the semi-discrete case; see Appendix A.6) yields minimizers much closer to $\theta^\dagger$, outperforming both the uncorrected case ($\alpha = 0$) and the direct application of the scheme from Section 3.3 ($\alpha = 0.5$).

Finally, Figure 11 reports stochastic optimization results (see Appendix B.1.2 for details). For finite $N$, minimizing $J_N^{\mathrm{SD}}(\theta)$ in (40) with SGD converges to biased solutions in both settings (Figures 11 (a) and (c)), with bias diminishing as $N$ grows. In contrast, minimizing $\tilde{J}_N^{\mathrm{SD}}(\theta; \alpha^\dagger)$ in (41) leads to convergence near $\theta^\dagger$ even for small




(a) $\hat{\theta}_{N,2}/\theta_2^\dagger$ (b) $J_N^{\mathrm{SD}}(\theta)$ (well-specified) (c) $J_N^{\mathrm{SD}}(\theta)$ (misspecified)

Figure 9: Finite-sample bias in the semi-discrete case for Gaussian models with varying $N$. In (a), ratio $\hat{\theta}_{N,2}/\theta_2^\dagger$ is computed for $N \in [2, 100]$. In (b) and (c), the expected Wasserstein loss $J_N^{\mathrm{SD}}(\theta)$ is shown for the well-specified case (Gaussian, $\theta_1^* = 0$, $\theta_2^* = 1$) and the misspecified case (Exponential, scale $1/\sqrt{2}$), respectively. Colored solid curves depict $J_N^{\mathrm{SD}}(\theta)$ and dots mark the minimizers $\hat{\theta}_{N,2}$.




(a) Standard deviation of $\alpha_N(\theta)$ (b) Bias reduction (well-specified) (c) Bias reduction (misspecified)

Figure 10: Bias correction in the semi-discrete case for Gaussian models. In (a), the standard deviation of $\{\alpha_N(\theta) \mid N \in \{5, 10, 15, 20, 35, 50, 75, 100\}\}$ is computed as a function of $\theta$. In (b) and (c), we plot the bias-corrected minimizers in the well-specified case (Gaussian, $\theta_1^* = 0$, $\theta_2^* = 1$) and the misspecified case (Exponential, scale $1/\sqrt{2}$), respectively. Solid curves correspond to $\alpha = 0, 0.5$, and $\alpha^\dagger$; dashed lines indicate the infinite-sample solution $\theta^\dagger$.

$N$, confirming the effectiveness of the correction.

### B.4.3 Sinkhorn Divergence

We study the finite-sample bias of expected Sinkhorn divergence minimization under model misspecification. Let $\mu$ denote the data-generating distribution and $f_\theta$ the parametric model, with empirical distributions $\hat{\mu}_N$ and $\hat{f}_{\theta,N}$ formed from $N$ i.i.d. samples. The expected Sinkhorn divergence is defined as

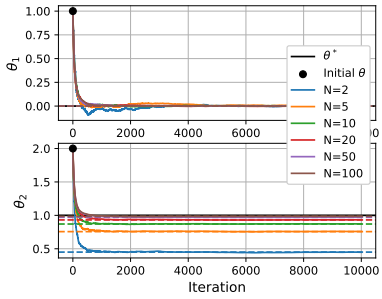
$$S_{N,\varepsilon}^{\mathrm{MS}}(\theta) = \mathbb{E}[S_\varepsilon(\hat{\mu}_N, \hat{f}_{\theta,N})], \tag{59}$$

where the expectation is taken over the sampling of $\hat{\mu}_N$ and $\hat{f}_{\theta,N}$.

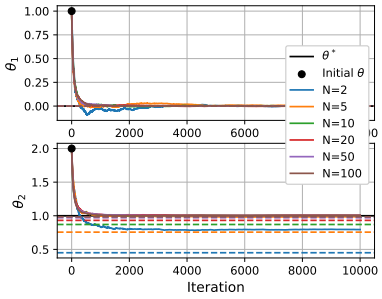
Figure 12(a) shows $S_{N,\varepsilon}^{\mathrm{MS}}(\theta)$ when $\mu$ is exponential (scale $1/\sqrt{2}$) and $f_\theta$ is Gaussian with $\theta_1 = 1/\sqrt{2}$. For finite $N$, the minimizer of $S_{N,\varepsilon}^{\mathrm{MS}}(\theta)$ deviates from the infinite-sample solution, denoted $\theta_\varepsilon^\dagger$. Since no analytic formula is available, we approximate $\theta_\varepsilon^\dagger$ by the minimizer at $N = 1{,}000$, which remains stable for larger $N$. Unlike the well-specified case, $\theta_\varepsilon^\dagger$ here can depend on $\varepsilon$.

For one-dimensional parameters, the bias can be corrected as done in Section 4.2 by considering the modified objective

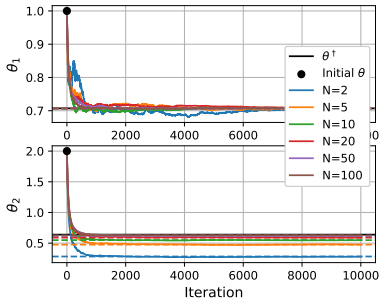
$$\tilde{S}_{N,\varepsilon}^{\mathrm{MS}}(\theta; \alpha) = S_{N,\varepsilon}^{\mathrm{MS}}(\theta) - \alpha S_{N,\varepsilon}(\theta, \theta). \tag{60}$$



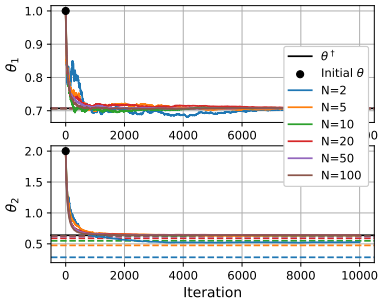

(a) Minimizing $J_N^{\mathrm{SD}}(\theta)$ (well-specified)




(b) Minimizing $\tilde{J}_N^{\mathrm{SD}}(\theta;\alpha^\dagger)$ (well-specified)




(c) Minimizing $J_N^{\mathrm{SD}}(\theta)$ (misspecified)




(d) Minimizing $\tilde{J}_N^{\mathrm{SD}}(\theta;\alpha^\dagger)$ (misspecified)

Figure 11: Stochastic optimization in the semi-discrete case for Gaussian models. In (a) and (b), SGD is applied to minimize $J_N^{\mathrm{SD}}(\theta)$ and $\tilde{J}_N^{\mathrm{SD}}(\theta;\alpha^\dagger)$ in the well-specified case (Gaussian, $\theta_1^* = 0$, $\theta_2^* = 1$), respectively. In (c) and (d), the same is shown for the misspecified case (Exponential, scale $1/\sqrt{2}$). Colored solid lines represent the parameter trajectories, and dashed lines indicate the corresponding minimizers of $J_N^{\mathrm{SD}}(\theta)$; black solid lines indicate infinite-sample solutions.

Choosing $\alpha$ appropriately aligns the minimizer of (60) with $\theta_\varepsilon^\dagger$ for each $\varepsilon$. We estimate such an $\alpha$ using the heuristic of Section 4.2, replacing the numerator and denominator in (16) with the gradients of $S_{N,\varepsilon}^{\mathrm{MS}}(\theta)$ and $S_{N,\varepsilon}(\theta,\theta)$, respectively. This relies on the assumption that both quantities converge at the same rate with $N$; see Genevay et al. (2019) and Mena and Niles-Weed (2019) for related convergence analyses.

Figure 12(b) shows the standard deviation of $\alpha_{N,\varepsilon}$ across $N$, computed from the adapted (16) using spline-smoothed gradients estimated from 10,000 trials. The minimum-variance location varies with $\varepsilon$ and serves as a proxy for $\theta_\varepsilon^\dagger$. We then define $\alpha_\varepsilon^\dagger$ in (15) as the mean of $\alpha_{N,\varepsilon}$ across $N$ at this parameter value.

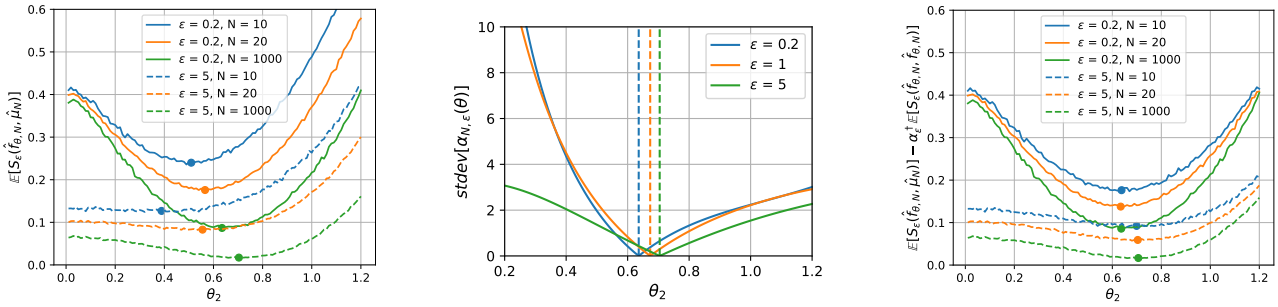
Finally, Figure 12(c) shows that applying the correction scheme with $\alpha = \alpha_\varepsilon^\dagger$ mitigates the finite-sample bias: the stationary point of the corrected objective shifts toward $\theta_\varepsilon^\dagger$ ($N = 1{,}000$ case). The resulting minimizers align with the vertical dashed lines in Figure 12(b), which indicate the minimum-variance parameter values for each $\varepsilon$.

## B.5 Results for the Tukey $g$-and-$h$ Model

### B.5.1 Tukey $g$-and-$h$ Distribution

The Tukey $g$-and-$h$ distribution provides a flexible model for data exhibiting skewness and heavy tails. A random variable $X$ from this family is defined as

$$X = \xi + \omega\, T_{g,h}(Z), \qquad T_{g,h}(Z) = \begin{cases} \dfrac{e^{gZ}-1}{g}\, e^{\frac{1}{2}hZ^2}, & g \neq 0, \\ Z\, e^{\frac{1}{2}hZ^2}, & g = 0, \end{cases} \tag{61}$$




(a) $S^{\mathrm{MS}}_{N,\varepsilon}(\theta)$ (b) Standard deviation of $\alpha_{N,\varepsilon}(\theta)$ (c) $\tilde{S}^{\mathrm{MS}}_{N,\varepsilon}(\theta;\alpha^{\dagger}_{\varepsilon})$

Figure 12: Expected Sinkhorn divergence and bias correction in the misspecified case. In (a), $S^{\mathrm{MS}}_{N,\varepsilon}(\theta)$ is shown when the data distribution is exponential (scale $1/\sqrt{2}$) and $f_\theta$ is Gaussian with $\theta_1 = 1/\sqrt{2}$; minimizers $\hat{\theta}_{N,\varepsilon,2}$ are marked by dots. In (b), the standard deviation of $\{\alpha_{N,\varepsilon}(\theta) \mid N \in \{2,5,10,20,50,100\}\}$ is plotted for each $\theta$. In (c), the bias-corrected objective $\tilde{S}^{\mathrm{MS}}_{N,\varepsilon}(\theta;\alpha^{\dagger}_{\varepsilon})$ is shown.

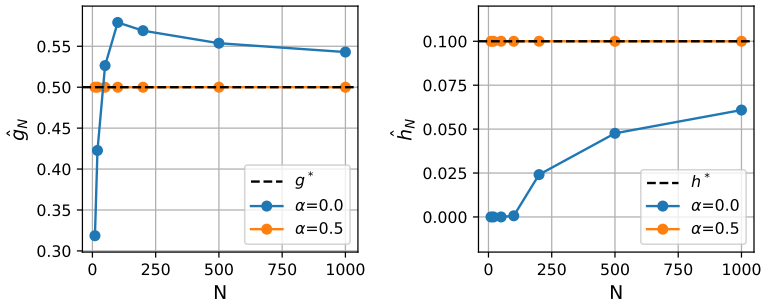



Figure 13: Finite-sample bias for Tukey $g$-and-$h$ models in well-specified case. Solid curves represent the minimizers of the bias-corrected losses for $\alpha = 0$ and 0.5. The black dashed lines indicate the fixed parameter $\theta^*$.

where $Z \sim \mathcal{N}(0,1)$, $\xi \in \mathbb{R}$ is a location parameter, $\omega > 0$ is a scale parameter, $g$ controls skewness, and $h \geq 0$ controls tail heaviness. The second-order moment exists when $h < \frac{1}{2}$, in which case the expected squared $W_2$ loss is well defined. Because the analytic form of the density is unavailable, maximum likelihood estimation is generally difficult, making Wasserstein-based objectives a practical alternative for parameter estimation. Since the inverse CDF is available, the expected Wasserstein loss can still be computed—based on Lemma 3.1 and Proposition 4.1—via numerical integration.

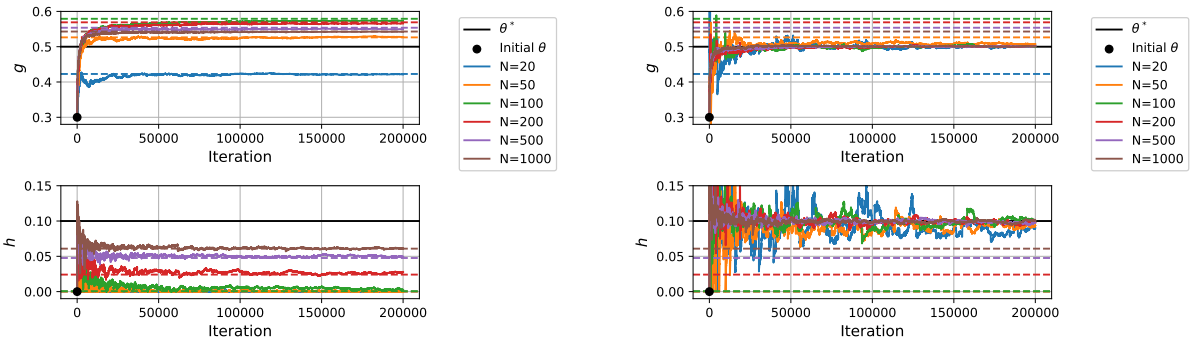
### B.5.2 Well-Specified Case

We first consider the well-specified case. Here, we fix $\xi = 0$ and $\omega = 1$ and estimate the two-dimensional parameter $(g,h)$ with $(g^*,h^*) = (0.5,0.1)$. As shown in Figure 13, for small $N$, the minimizers of the empirical Wasserstein loss (the curve for $\alpha = 0$) exhibit a clear finite-sample bias. Although the bias decreases as $N$ grows, achieving nearly unbiased estimation would require very large sample sizes. Applying the bias-correction scheme from Section 3.3 substantially improves estimation: the curve for $\alpha = 0.5$ shows that the corrected objective recovers the true parameters accurately even for small $N$. A similar effect is illustrated by the log-densities in Figure 1(b), plotted for $(g^*,h^*) = (0.5,0.2)$.

Figures 14(a) and (b) present the results of stochastic optimization of $J_N(\theta^*,\theta)$ and $\tilde{J}_N(\theta^*,\theta)$, respectively, confirming the convergence to biased solutions and their correction, as anticipated in Figure 13.

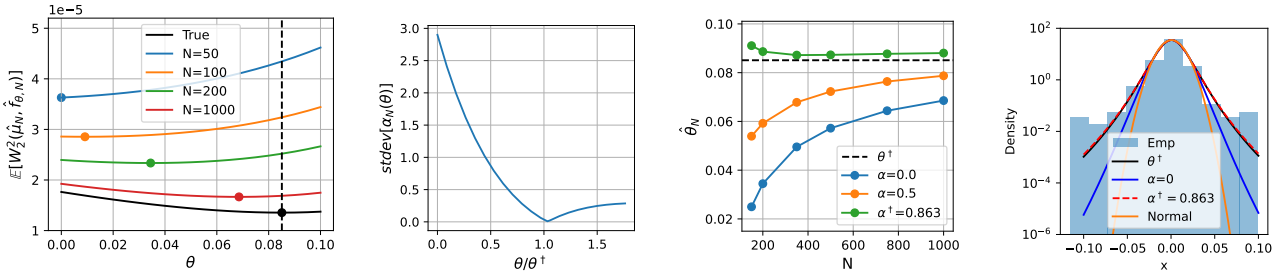
### B.5.3 Misspecified Case

We next extend the misspecified example from Section 4 using the SPY log-return dataset (obtained from Yahoo Finance via `yfinance` (Aroussi, 2019)) and the Diamonds price dataset (Wickham, 2016) (subsampled to




(a) Minimizing $J_N(\theta^*, \theta)$ (b) Minimizing $\tilde{J}_N(\theta^*, \theta)$

Figure 14: Stochastic optimization of empirical Wasserstein losses for Tukey $g$-and-$h$ models across different sample sizes $N$. In (a), SGD is used to minimize $J_N(\theta^*, \theta)$ with respect to $\theta = (g, h)$ for fixed $\theta^* = (0.5, 0.1)$. In (b), the same procedure is applied to the bias-corrected loss $\tilde{J}_N(\theta^*, \theta)$. Colored solid lines show parameter trajectories; dashed lines indicate the minimizers of $J_N(\theta^*, \theta)$; black solid lines denote the fixed parameter values.




(a) $J_N^{\mathrm{MS}}(\theta)$ (b) Stdev. of $\alpha_N(\theta)$ (c) Bias reduction (d) PDFs

Figure 15: The expected Wasserstein loss $J_N^{\mathrm{MS}}(\theta)$ and the application of the bias correction scheme in the misspecified case for Tukey $g$-and-$h$ models with $\theta = h$, using the SPY dataset, across different sample sizes $N$. In (a), $J_N^{\mathrm{MS}}(\theta)$ is shown. Colored solid curves depict $J_N^{\mathrm{MS}}(\theta)$, with minimizers $\hat{\theta}_N$ marked by dots. In (b), for each $\theta$, the standard deviation (Stdev.) of $\{\alpha_N(\theta) \mid N \in \{200, 500, 1000\}\}$ is computed. In (c), solid curves represent the minimizers of the bias-corrected losses for $\alpha = 0, 0.5$, and $\alpha^\dagger$. The black dashed line indicates the infinite-sample solution $\theta^\dagger$. In (d), we show the empirical histogram of the data and the PDFs at $\theta^\dagger$, the biased ($\alpha = 0$) and bias-corrected ($\alpha = \alpha^\dagger$) estimates with $N = 200$, and the normal distribution, all on a log scale.

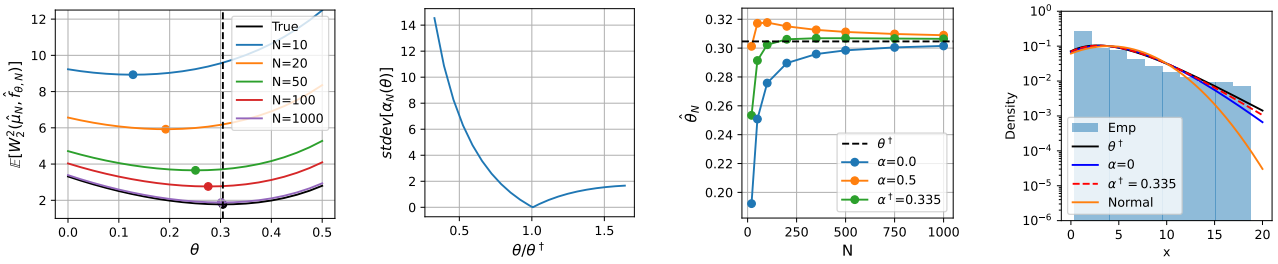
$M = 4000$). For SPY, we estimate the tail-heaviness parameter $h$, and for Diamonds the skewness parameter $g$, while fixing $(\xi, \omega)$ to the sample mean and standard deviation and setting the remaining parameter to zero.

Figures 15 and 16 show the same patterns as in Figure 4, namely the finite-sample bias of the loss and its minimizer, the minimization of $\alpha_N(\theta)$'s standard deviation near $\theta^\dagger$, and effective bias correction with $\alpha = \alpha^\dagger$ from (15).

Figures 17(a)–(c) show the results of stochastic optimization of $J_N^{\mathrm{MS}}(\theta)$ and $\tilde{J}_N^{\mathrm{MS}}(\theta; \alpha^\dagger)$ for the BTC–USD ($\theta = h$), SPY ($\theta = h$), and Diamonds ($\theta = g$) datasets, respectively. The results confirm convergence to biased solutions and their correction, consistent with the behaviors observed in Figures 4(c), 15(c), and 16(c).

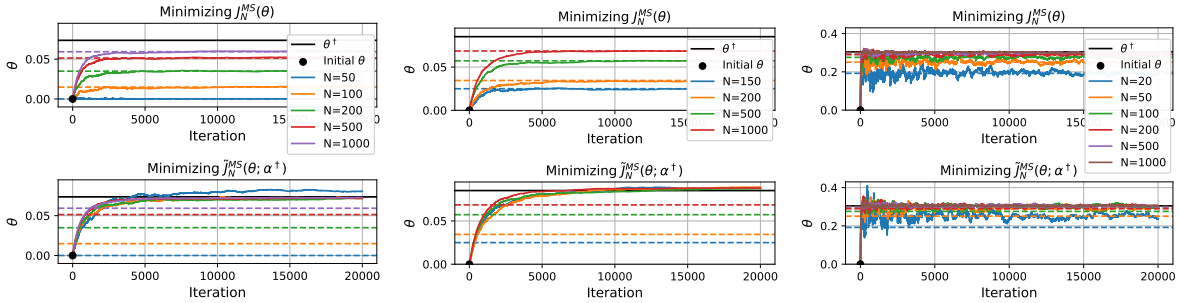
### B.6 Results for the Affine PDF Model

We use the affine PDF family $f_a(x) = a(x - 0.5) + 1$, $-2 \le a \le 2$, $x \in [0, 1]$, as a qualitatively different non-location-scale example. For this model, the expected Wasserstein losses in both the discrete–discrete and semi-discrete settings can be evaluated in closed form using the formulas derived in Appendix A.9. We now examine the resulting finite-sample bias and its correction empirically.




(a) $J_N^{\mathrm{MS}}(\theta)$ (b) Stdev. of $\alpha_N(\theta)$ (c) Bias reduction (d) PDFs

Figure 16: The expected Wasserstein loss $J_N^{\mathrm{MS}}(\theta)$ and the application of the bias correction scheme in the misspecified case for Tukey $g$-and-$h$ models with $\theta = g$, using the Diamonds dataset, across different sample sizes $N$. In (a), $J_N^{\mathrm{MS}}(\theta)$ is shown. Colored solid curves depict $J_N^{\mathrm{MS}}(\theta)$, with minimizers $\hat{\theta}_N$ marked by dots. In (b), for each $\theta$, the standard deviation (Stdev.) of $\{\alpha_N(\theta) \mid N \in \{200, 500, 1000\}\}$ is computed. In (c), solid curves represent the minimizers of the bias-corrected losses for $\alpha = 0, 0.5$, and $\alpha^\dagger$. The black dashed line indicates the infinite-sample solution $\theta^\dagger$. In (d), we show the empirical histogram of the data and the PDFs at $\theta^\dagger$, the biased ($\alpha = 0$) and bias-corrected ($\alpha = \alpha^\dagger$) estimates with $N = 20$, and the normal distribution, all on a log scale.




(a) BTC-USD (b) SPY (c) Diamonds

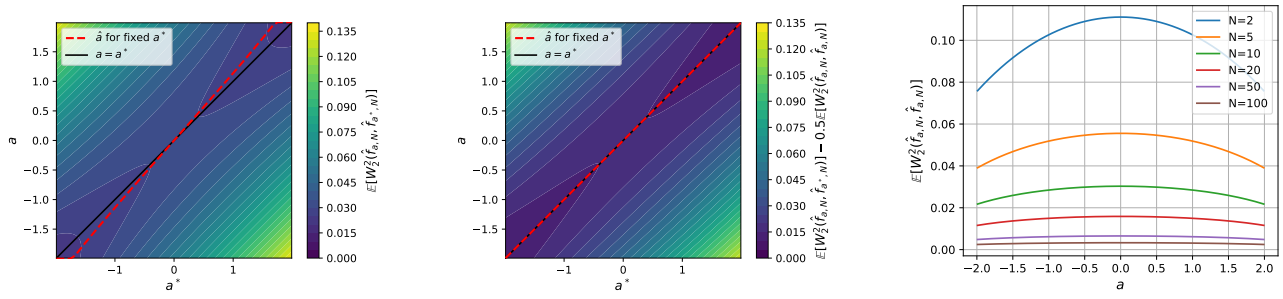
Figure 17: Stochastic optimization of empirical Wasserstein losses in the misspecified setting for Tukey $g$-and-$h$ models across different sample sizes $N$. SGD is used to minimize $J_N^{\mathrm{MS}}(\theta)$ and $\tilde{J}_N^{\mathrm{MS}}(\theta;\alpha^\dagger)$ for (a) the BTC–USD dataset with respect to $\theta = h$, (b) the SPY dataset with respect to $\theta = h$, and (c) the Diamonds dataset with respect to $\theta = g$. Colored solid lines represent parameter trajectories; dashed lines indicate the minimizers of $J_N^{\mathrm{MS}}(\theta)$; black solid lines denote the infinite-sample solutions.

### B.6.1 Discrete–Discrete Case

We first consider the bias in minimizing the expected Wasserstein loss between two empirical distributions. Figure 18(a) shows that the minimizer $\hat{a} = \arg\min_a J_N(a^*, a)$ generally differs from $a^*$, except in the uniform case $a^* = 0$. The estimates exhibit an outward bias, tending to lie farther from zero than $a^*$. In contrast, the modified loss $\tilde{J}_N(a^*, a)$ is minimized exactly at $a^*$ for all fixed parameters as shown in Figure 18(b).

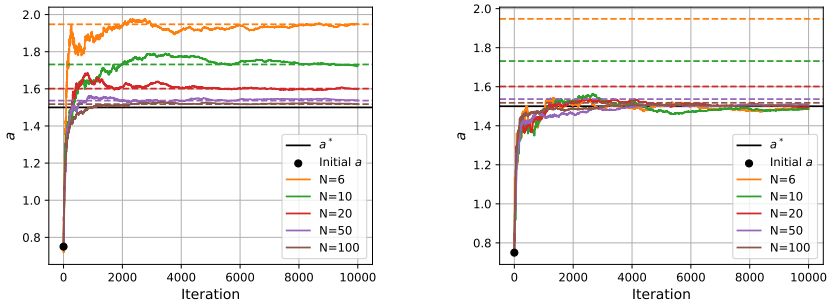
Along the diagonal $a = a^*$ (Figure 18(c)), the expected loss decreases as $|a|$ grows, since larger $|a|$ concentrates mass near the boundaries and reduces transport among dense regions. This decrease is steeper when $N$ is small. By Remark 3.3 and (33), this explains the outward bias in Figure 18(a). As $N$ increases, the gradient along the diagonal diminishes, indicating that the bias vanishes asymptotically.

Figure 19 shows stochastic optimization results. When $N$ is finite, minimizing $J_N(a^*, a)$ via SGD converges to a biased solution (Figure 19(a)), and the bias diminishes as $N$ grows. In contrast, minimizing the modified loss $\tilde{J}_N(a^*, a)$ yields convergence to the fixed parameter even for small $N$ (Figure 19(b)), confirming the effectiveness of the bias correction scheme.




(a) $J_N(a^*, a)$ (b) $\tilde{J}_N(a^*, a)$ (c) $J_N(a, a)$

Figure 18: Expected Wasserstein loss in the affine PDF model. In (a) and (b), we depict the expected loss $J_N(a^*, a)$ in (6) and the modified loss $\tilde{J}_N(a^*, a)$ in (10) over $(a^*, a) \in (-2, 2)^2$ with $N = 10$, respectively. Red dashed curves trace the minimizer for each fixed $a^*$; the black solid line denotes the diagonal $a = a^*$. In (c), $J_N(a^*, a)$ along the diagonal $a^* = a$ for various $N$ is shown.




(a) Minimizing $J_N(a^*, a)$ (b) Minimizing $\tilde{J}_N(a^*, a)$

Figure 19: Stochastic optimization of empirical Wasserstein losses for affine PDF models with different sample sizes $N$. SGD is applied to minimize (a) $J_N(a^*, a)$ and (b) $\tilde{J}_N(a^*, a)$ with respect to slope parameter $a$ for fixed $a^*$. Colored solid lines show parameter trajectories; dashed lines mark the minimizers of $J_N(a^*, a)$; black solid lines indicate the fixed parameter value.

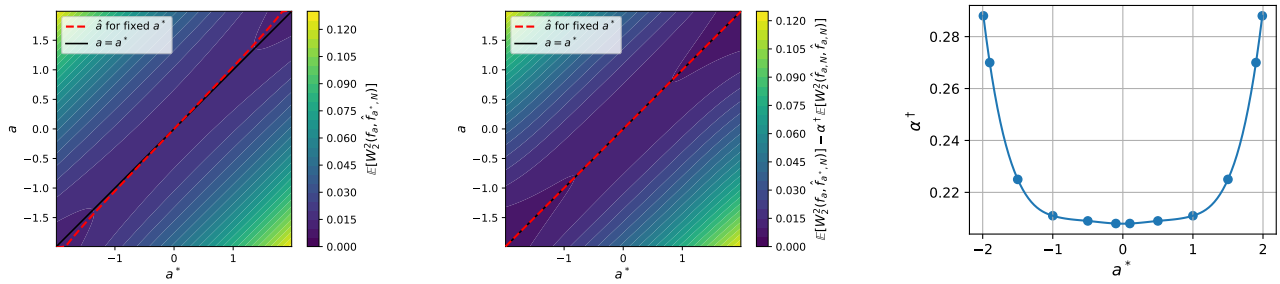
### B.6.2 Semi-Discrete Case

We now examine finite-sample bias in the semi-discrete case, focusing on the well-specified setting where the data-generating distribution is the affine PDF with slope $a^*$. As in Figure 18, Figure 20(a) shows that the minimizer $\hat{a} = \arg\min_a J_N^{\mathrm{SD}}(a)$ generally exhibits an outward bias, lying farther from zero than $a^*$, except in the uniform case $a^* = 0$. By contrast, the modified loss $\tilde{J}_N^{\mathrm{SD}}(a; \alpha^\dagger)$ with $\alpha = \alpha^\dagger$, chosen as in Section 4.2, successfully corrects this bias for all $a^*$, as shown in Figure 20(b).

Figure 20(c) reports $\alpha^\dagger$, obtained by minimizing the standard deviation of $\{\alpha_N(a) \mid N \in \{5, 10, 15, 20, 35, 50, 75, 100\}\}$ in (16) for each $a^*$. When $a^* = 0$, where finite-sample bias is absent, the gradient of the self-distance is zero for all $N$, causing $\alpha_N$ in (16) to diverge. In this case, we interpolate smoothly from $\alpha^\dagger$ values at nearby $a^*$.

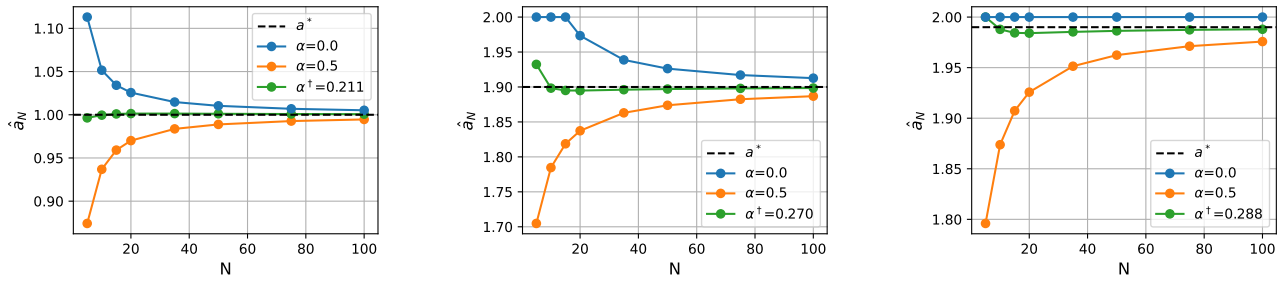
Figures 21(a)–(c) show minimization of the bias-corrected loss in (14) for $a^* \in \{1, 1.9, 1.99\}$. In all cases, setting $\alpha = \alpha^\dagger$ as proposed in Section 4.2 effectively reduces finite-sample bias and yields solutions closer to the infinite-sample optimum, even for small $N$. This approach consistently outperforms both the case without correction ($\alpha = 0$) and the direct application of the correction from Section 3.3 ($\alpha = 0.5$).

Finally, Figure 22 shows stochastic optimization results. When $N$ is finite, minimizing $J_N^{\mathrm{SD}}(a)$ via SGD converges to a biased solution (Figure 22(a)), with the bias diminishing as $N$ increases. In contrast, minimizing $\tilde{J}_N^{\mathrm{SD}}(a; \alpha^\dagger)$ with $\alpha^\dagger$ selected as above leads to convergence to the correct parameter even for small $N$ (Figure 22(b)), demonstrating




(a) $J_N^{\mathrm{SD}}(a)$ for different $a^*$ (b) $\tilde{J}_N^{\mathrm{SD}}(a;\alpha^\dagger)$ for different $a^*$ (c) $\alpha^\dagger$ for different $a^*$

Figure 20: Expected Wasserstein loss in the semi-discrete case for the affine PDF model with data-generating distribution $\mu = f_{a^*}$ (well-specified). In (a) and (b), we depict expected losses $J_N^{\mathrm{SD}}(a)$ in (39) and modified losses $\tilde{J}_N^{\mathrm{SD}}(a;\alpha^\dagger)$ in (14) with $\alpha = \alpha^\dagger$ over $(a^*, a) \in (-2, 2)^2$ for $N = 10$, respectively. Red dashed curves trace the minimizer for each fixed $a^*$; the black solid line indicates the diagonal $a = a^*$. In (c), values of $\alpha^\dagger$ across different $a^*$ are shown.



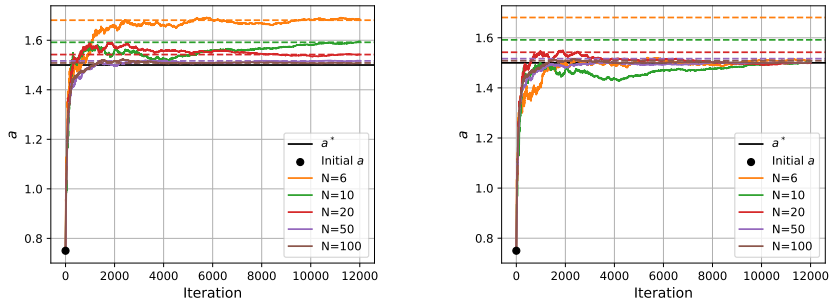

(a) Bias reduction ($a^* = 1$) (b) Bias reduction ($a^* = 1.9$) (c) Bias reduction ($a^* = 1.99$)

Figure 21: Bias reduction results in the semi-discrete case with data-generating distribution $\mu = f_{a^*}$ (well-specified). Solid curves represent the minimizers of the bias-corrected losses for $\alpha = 0, 0.5$, and $\alpha^\dagger$; dashed lines indicate the infinite-sample solution $a^*$.

the effectiveness of the bias correction scheme.

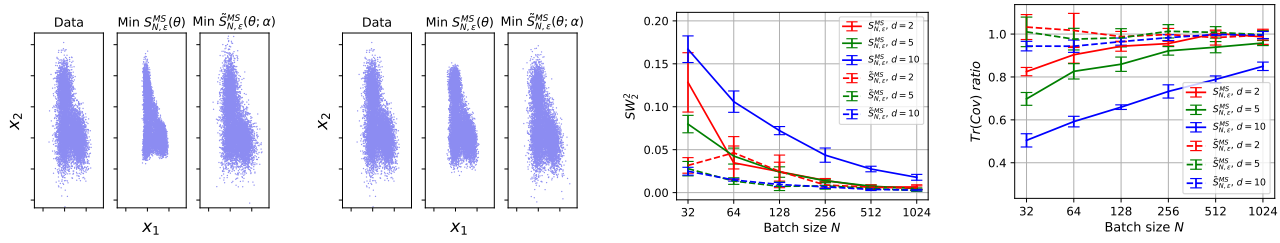
### B.7 Results for the Neural Network Generator Minimizing the Sinkhorn Divergence

Similar to the results in Section 4.3, Figure 23 shows that finite-sample bias persists for NN generators minimizing the expected Sinkhorn divergence, becomes more pronounced in higher dimensions, and decreases as the batch size grows. Bias correction yields samples that are qualitatively closer to the data and substantially reduces the test $SW_2^2$ in most cases. The covariance trace ratio further confirms that training with bias correction better recovers the true scale of the distribution (ratio close to one), whereas training without correction leads to severe shrinkage, which becomes more evident in higher dimensions.




(a) Minimizing $J_N^{\mathrm{SD}}(a)$ (b) Minimizing $\tilde{J}_N^{\mathrm{SD}}(a;\alpha^\dagger)$

Figure 22: Stochastic optimization of semi-discrete losses for affine PDF models with different sample sizes $N$. SGD is applied to minimize (a) $J_N^{\mathrm{SD}}(a)$ and (b) $\tilde{J}_N^{\mathrm{SD}}(a;\alpha^\dagger)$ with respect to slope parameter $a$ for fixed $a^*$. Colored solid lines show parameter trajectories; dashed lines mark the minimizers of $J_N^{\mathrm{SD}}(a)$; black solid lines indicate the fixed parameter value.




(a) Batch size = 64 (b) Batch size = 1024 (c) $SW_2^2$ (d) Tr(Cov)

Figure 23: Results for neural network (NN) generators fitted to Gaussian mixture data by stochastic optimization of $S_{N,\varepsilon}^{\mathrm{MS}}$ and $\tilde{S}_{N,\varepsilon}^{\mathrm{MS}}$ (defined in Appendix B.4.3) with $\alpha = 0.5$ and regularization weight $\varepsilon = 1$ for different batch sizes. In (a) and (b), we show two-dimensional projections of the data and of samples generated by NNs trained without and with bias correction in ten-dimensional space, for batch sizes 64 and 1024, respectively. In (c) and (d), performance is evaluated using the sliced Wasserstein distance (lower is better) and the covariance trace ratio (closer to 1.0 is better), respectively, for $d = 2, 5$, and 10.