OPT2025: 17th Annual Workshop on Optimization for Machine Learning

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

Cheongjae Jang CJJANG@HANYANG.AC.KR
Hanyang University, Seoul, Korea

Yung-Kyun Noh NOHYUNG@HANYANG.AC.KR
Hanyang University, Seoul, Korea
Korea Institute for Advanced Study, Seoul, Korea

Abstract

We show that minimizing the expected Wasserstein loss between empirical distributions can lead to biased parameter estimates in finite-sample regimes. Specifically, when two empirical distributions are sampled from the same parametric family—one at a fixed parameter value and the other at a variable—we find that minimizing the expected loss with respect to the variable parameter generally fails to recover the fixed one. We analytically verify this bias in simple one-dimensional settings, including location-scale models, by deriving closed-form expressions for the expected empirical Wasserstein loss. The analysis reveals that when the expected loss varies along the diagonal (where the two parameter values coincide), the gradient at the fixed parameter value is nonzero, shifting the minimizer away from it. To address this, we propose a simple correction scheme that eliminates the bias in well-specified cases. Numerical experiments confirm that stochastic gradient descent on the empirical Wasserstein loss converges to biased solutions and demonstrate the effectiveness of the proposed bias 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 [20]. In particular, parameter estimation methods based on minimizing the Wasserstein distance have drawn increasing attention as an alternative to classical likelihood-based approaches [2, 5]. Unlike maximum likelihood estimation, 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 [4, 8, 18, 19] and generative modeling frameworks [1, 11, 15, 17].

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 [2, 5]. In finite-sample regimes, however, the expected empirical Wasserstein loss between an empirical distribution and a parametric model can differ from its infinite-sample counterpart, leading to biased gradients and minimizers [3]. 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. Although prior work has examined the properties of minibatch Wasserstein loss and noted that it differs from the true distance between the underlying distributions [12], and non-asymptotic bounds on this gap have been proposed [23], its impact 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. Specifically, when two empirical distributions are independently generated from the same parametric family—one at a fixed parameter value and the other at a variable parameter value—we find that minimizing the expected loss with respect to the variable parameter generally fails to recover the fixed one.

To make this phenomenon analytically tractable, we focus on one-dimensional settings where closed-form expressions for optimal transport are available [6, 20, 22, 24]. We derive the expected empirical Wasserstein loss in closed form for representative models such as location-scale families, and 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 correction scheme that eliminates the bias in well-specified cases. Numerical experiments confirm that stochastic gradient descent on the empirical Wasserstein loss converges to biased solutions and demonstrate the effectiveness of the proposed bias correction scheme.

2. Background

The p-Wasserstein distance (denoted W_p) between two probability density functions (PDFs) μ and ν over $\mathbb{R}^d$ is defined as $W_p^p(\mu,\nu) = \inf_{\gamma\in\Gamma}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|x-y\|^p d\gamma(x,y)$, where Γ denotes the set of all joint distributions on $\mathbb{R}^d\times\mathbb{R}^d$ that have respective marginals μ and ν.

In the one-dimensional case ($d=1$), this admits a closed-form expression using the cumulative distribution functions (CDFs) P and Q of μ and ν, respectively [9, 20]:

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

We focus on the one-dimensional case, which allows more precise analytical characterization of the Wasserstein distance and is sufficient to reveal the core phenomena under study.

Let $\hat{\mu}_N$ and $\hat{\nu}_N$ denote empirical distributions constructed from i.i.d. samples $x_1,\dots,x_N\sim\mu$ and $y_1,\dots,y_N\sim\nu$, 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 δ_x is a Dirac measure at $x\in\mathbb{R}$. Assuming the samples are ordered without loss of generality, i.e., $x_i\le x_{i+1}$ and $y_i\le y_{i+1}$ for $i=1,\dots,N-1$, the W_p distance between $\hat{\mu}_N$ and $\hat{\nu}_N$ becomes

$$W_p^p(\hat{\mu}_N,\hat{\nu}_N) = \frac{1}{N}\sum_{k=1}^N |x_k - y_k|^p. \tag{2}$$

It is well known that under mild conditions, $W_p(\hat{\mu}_N,\mu)\to 0$ in expectation as $N\to\infty$ [6, 7, 10, 14, 25], and hence, $\mathbb{E}\left[W(\hat{\mu}_N,\hat{\nu}_N)\right]\to W(\mu,\nu)$ [23].

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, i.e., $\hat{\theta}_N = \arg\min_\theta W_p(\hat{\mu}_N, f_\theta)$. Under suitable conditions, this estimator is known to be consistent [2, 5]: $\hat{\theta}_N \to \theta^* = \arg\min_\theta W_p(\mu, f_\theta)$ as $N \to \infty$.

However, this consistency holds only asymptotically, and the estimator can behave quite differently in finite-sample settings. For instance, in a Bernoulli model, minimizing the expected empirical Wasserstein loss yields a biased estimate [3], i.e., $\hat{\theta}_N = \arg\min_\theta \mathbb{E}[W_p^p(\hat{\mu}_N, f_\theta)] \neq \arg\min_\theta W_p^p(\mu, f_\theta)$ in general.

In practice, Wasserstein-based objectives are often used to compare *two empirical distributions*, each constructed from finite samples [4, 11, 17, 19]. Although a consistency result has been established for this setting under certain asymptotic conditions [5], the behavior of such estimators 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 θ^* (fixed) and θ (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., $\hat{\theta}_N = \arg\min_\theta J_N(\theta^*, \theta) \stackrel{?}{=} \theta^*$.

3. Finite-sample bias of minimizing expected Wasserstein loss between empirical distributions

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, by focusing on location-scale models. We then propose a simple method to correct the bias in well-specified settings.

3.1. Bias in minimizing expected squared W_2 for location-scale models

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 $y_1, \ldots, y_N \sim f_\theta$ are ordered samples, and $\theta^* = (\theta_1^*, \theta_2^*)$ is fixed while θ is variable. Setting $p = 2$ in (2), our loss is

$$J_N(\theta^*, \theta) = \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[(x_i - y_i)^2]. \tag{3}$$

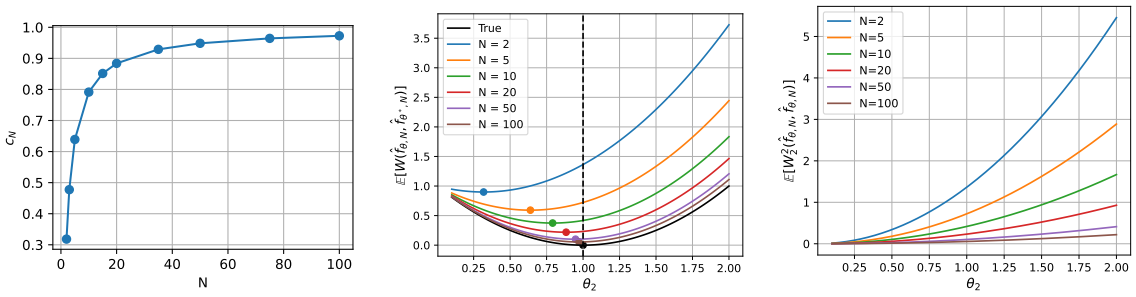
Each term is $\mathbb{E}[(x_i - y_i)^2] = \int_{\mathbb{R}}\int_{\mathbb{R}} (x_i - y_i)^2 \; p_{X(i)}(x_i) \; p_{Y(i)}(y_i) \; dx_i \; dy_i$, where $p_{X(i)}(x_i)$ and $p_{Y(i)}(y_i)$ are the densities of the i-th order statistics from f_{θ^*} and f_θ, respectively [6].

We analyze the bias in the location-scale family $f_\theta(y) = \frac{1}{\theta_2} f_0\left(\frac{y-\theta_1}{\theta_2}\right)$, with location parameter θ_1 and scale parameter $\theta_2 > 0$. The reference density f_0 is normalized to have zero mean and unit variance without loss of generality.

The following proposition shows that, both the expected Wasserstein loss in (3) and its minimizer can be derived analytically, revealing that the optimal parameters $\hat{\theta}_N = (\hat{\theta}_{N,1}, \hat{\theta}_{N,2}) = \arg\min_\theta J_N(\theta^*, \theta)$ generally differ from the fixed parameter θ^* when N is finite.

Proposition 1 *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*

$$J_N(\theta^*, \theta) = \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] = (\theta_1 - \theta_1^*)^2 + \theta_2^2 - 2c_N \cdot \theta_2\theta_2^* + \theta_2^{*2}, \tag{4}$$



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

Figure 1: 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 $N \in [2, 100]$ via numerical integration. 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 $\theta^* = \theta$.

where $c_N = \left(\frac{1}{N} \sum_{i=1}^{N} \left(\int_0^1 F_0^{-1}(u)\, p_{U(i)}(u)\, du \right)^2 \right)$, $p_{U(i)}(u)$ *denotes the probability density function of the* i*-th order statistic of a uniform distribution, and* F_0 *is the cumulative distribution function 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* (4) *are given by* $\hat{\theta}_{N,1} = \theta_1^*$ *and* $\hat{\theta}_{N,2} = c_N \cdot \theta_2^*$.

Proof. *The 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.

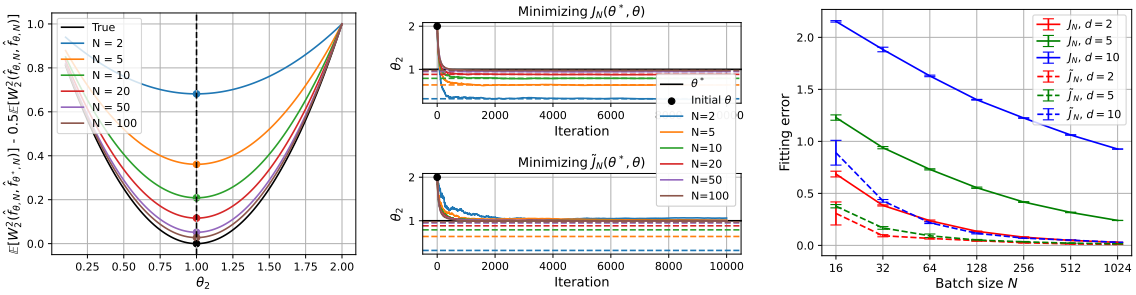
Numerical examples. We illustrate the finite-sample bias using the Gaussian distribution, where the mean and standard deviation serve as the location and scale parameters, respectively. Since Proposition 1 shows no bias in the location parameter, we focus on the scale parameter.

Figure 1(a) shows that c_N in (4) 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 [2, 5].

In Figure 1(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.

3.2. Conditions for finite-sample bias and a simple bias-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.3. This identity leads directly to the following remark on the condition under which minimizing the expected Wasserstein loss exhibits finite-sample bias:



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

Figure 2: 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 $d = 2, 5, 10$.

Remark 2 *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 1(c)). The fact that $J_N(\theta, \theta)$ increases with θ_2 aligns with the downward bias of $\hat{\theta}_{N,2}$ observed in Figure 1(b), as further explained in Appendix A.3.

Accordingly, we define a bias-corrected loss by subtracting a self-distance term:

$$\tilde{J}_N(\theta^*, \theta) = J_N(\theta^*, \theta) - \frac{1}{2} J_N(\theta, \theta). \tag{5}$$

Differentiating (5) with respect to θ gives $\frac{\partial \tilde{J}_N(\theta^*,\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = 0$, since $\frac{\partial J_N(\theta,\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = 2 \cdot \frac{\partial J_N(\theta^*,\theta)}{\partial \theta}\Big|_{\theta=\theta^*}$. Therefore, $\theta = \theta^*$ is always a stationary point of $\tilde{J}_N(\theta^*, \theta)$, for any θ^* and N.

Numerical examples. We first evaluate the modified loss $\tilde{J}_N(\theta^*, \theta)$ for Gaussian density examples in Figure 2(a). Unlike the biased minimizers of $J_N(\theta^*, \theta)$ in Figure 1(b), the minimizers of $\tilde{J}_N(\theta^*, \theta)$ coincide with those of the true Wasserstein loss between the underlying densities.

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

In Figure 2(b), for finite N, the scale parameter minimizing $J_N(\theta^*, \theta)$ converges to a biased solution, which matches the minimizers characterized in Proposion 1. 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 2(c) reports fitting errors for high-dimensional GMMs 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. Additional results are provided in Appendix B.3 for an affine PDF model defined in Appendix A.4.

4. Conclusion

In this paper, we have investigated the finite-sample bias that arises when minimizing the expected Wasserstein distance between two empirical distributions. We analytically characterized this bias in well-specified settings, with detailed results for one-dimensional location–scale models. The fact that such bias arises even in these simple cases suggests that similar or more pronounced effects may occur in practical scenarios, where models are often misspecified and batch sizes are limited. Future work will further extend the analysis and bias correction scheme to broader settings, including misspecified models and Sinkhorn divergences [15], and evaluate their effectiveness on real-world data.

Acknowledgements

C. Jang was partly supported by NRF/ME (RS-2023-00249714, RS-2025-25427337). Y.-K. Noh was partly supported by NRF/MSIT (RS-2024-00421203) and IITP/MSIT (RS-2023-00220628).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

[3] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[4] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019.

[5] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4): 657–676, 2019.

[6] Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

[7] Emmanuel Boissard and Thibaut Le Gouic. 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, 2014.

[8] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

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

[10] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 1183–1203, 2013.

[11] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491, 2018.

[12] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *International Conference on Artificial Intelligence and Statistics*, pages 2131–2141. PMLR, 2020.

[13] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

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

[15] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

[16] Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

[17] Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

[18] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180, 2012.

[19] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli. 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, 2020.

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

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

[23] Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105): 1–23, 2019.

[24] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[25] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.

Appendix A. Derivations and proofs

A.1. Derivations for expected squared W_2 between empirical measures

To prove Proposition 1, we first establish the following lemma.

Lemma 3 *Let $\{f_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^m\}$ be a parametric family of distributions, and let $\hat{f}_{\theta^*,N}$ and $\hat{f}_{\theta,N}$ denote empirical distributions based on N i.i.d. samples drawn from the parametric models f_{θ^*} and f_θ, respectively. Then the expected squared W_2 loss between $\hat{f}_{\theta^*,N}$ and $\hat{f}_{\theta,N}$ is*

$$\begin{aligned} J_N(\theta^*, \theta) &= \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})] \\ &= m_2(\theta^*) + m_2(\theta) - \frac{2}{N}\sum_{i=1}^{N} m_{1,i}(\theta^*) m_{1,i}(\theta), \end{aligned} \tag{6}$$

where

$$m_{1,i}(\theta) \equiv \int_0^1 F_\theta^{-1}(u_i) p_{U(i)}(u_i) du_i, \tag{7}$$

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

Here $p_{U(i)}(u) = \frac{N!}{(i-1)!(N-i)!} u^{i-1}(1-u)^{N-i}$ is the density of the i-th order statistic of a $\mathrm{Uniform}(0,1)$ distribution, and F_θ denotes the CDF associated with f_θ.

Proof. *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 $y_1, \ldots, y_N \sim f_\theta$ are ordered samples. Then the expected squared W_2 loss is expressed as*

$$J_N(\theta^*, \theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(x_i - y_i)^2] \tag{9}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}}\int_{\mathbb{R}} (x_i - y_i)^2 \, p_{X(i)}(x_i) \, p_{Y(i)}(y_i) \, dx_i \, dy_i, \tag{10}$$

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_{θ^*} [6]. An analogous expression holds for $p_{Y(i)}(y_i)$, the i-th order statistic from f_θ.

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 (10) can be written as

$$J_N(\theta^*, \theta) = \frac{1}{N}\sum_{i=1}^{N} \int_0^1 \int_0^1 (F_{\theta^*}^{-1}(u_i) - F_\theta^{-1}(v_i))^2 \, p_{U_{(i)}}(u_i) p_{U_{(i)}}(v_i) \, dv_i \, du_i. \tag{11}$$

Using (7) and $m_{2,i}(\theta) = \int_0^1 (F_\theta^{-1}(u_i))^2 p_{U(i)}(u_i) du_i$, the expected loss in (11) simplifies to:

$$J_N(\theta^*, \theta) = \frac{1}{N}\sum_{i=1}^{N} m_{2,i}(\theta^*) - 2m_{1,i}(\theta^*)m_{1,i}(\theta) + m_{2,i}(\theta) \tag{12}$$

$$= m_2(\theta^*) + m_2(\theta) - \frac{2}{N}\sum_{i=1}^{N} m_{1,i}(\theta^*) m_{1,i}(\theta), \tag{13}$$

where we have used $\frac{1}{N}\sum_{i=1}^{N} m_{2,i}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 (F_\theta^{-1}(u_i))^2 p_{U(i)}(u_i)du_i = \int_0^1 (F_\theta^{-1}(u))^2 du = \int_{\mathbb{R}} y^2 f_\theta(y)dy$ *to derive* (13)*, which follows from the permutation-invariance of the sum and the change of variable* $u \mapsto y = F_\theta^{-1}(u)$. □

A.2. Proof of Proposition 1

By Lemma 3, it remains to compute the terms $m_{1,i}(\theta)$ in (7) and $m_2(\theta)$ in (8) for the location-scale model in order to obtain (4).

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

$$m_{1,i}(\theta) = \theta_1 + \theta_2 \cdot b_i, \tag{14}$$

$$m_2(\theta) = \theta_1^2 + \theta_2^2, \tag{15}$$

with $b_i = \int_0^1 F_0^{-1}(u_i) p_{U(i)}(u_i)du_i$.

Substituting (14) and (15) into (6) and simplifying, we obtain

$$J_N(\theta^*, \theta) = \theta_1^2 + \theta_2^2 + \theta_1^{*2} + \theta_2^{*2} - \frac{2}{N}\sum_{i=1}^{N}(\theta_1^* + \theta_2^* \cdot b_i)(\theta_1 + \theta_2 \cdot b_i) \tag{16}$$

$$= (\theta_1 - \theta_1^*)^2 + \theta_2^2 - 2c_N \cdot \theta_2\theta_2^* + \theta_2^{*2}, \tag{17}$$

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

The resulting objective is quadratic in (θ_1, θ_2). Solving the first-order necessary condition $\frac{\partial J_N(\theta^*,\theta)}{\partial \theta} = 0$ yields $\hat{\theta}_{1,N} = \theta_1^*$ and $\hat{\theta}_{2,N} = c_N \cdot \theta_2^*$. □

A.3. Additional discussion of finite-sample bias and the correction scheme

Consider $J_N(\theta^*, \theta)$ in (3) 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., $\left.\frac{\partial J_N(\theta^*,\theta)}{\partial \theta}\right|_{\theta=\theta^*=\theta_0} = \left.\frac{\partial J_N(\theta^*,\theta)}{\partial \theta^*}\right|_{\theta^*=\theta=\theta_0}$ for any parameter value θ_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

$$\left.\frac{\partial J_N(\theta,\theta)}{\partial \theta}\right|_{\theta=\theta_0} = \left.\frac{\partial J_N(\theta^*,\theta)}{\partial(\theta^*,\theta)}\right|_{\theta=\theta^*=\theta_0} \cdot (1,1)^\top = 2 \cdot \left.\frac{\partial J_N(\theta^*,\theta)}{\partial \theta}\right|_{\theta=\theta^*=\theta_0}, \tag{18}$$

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., $\left.\frac{\partial J_N(\theta^*,\theta)}{\partial(\theta^*,\theta)}\right|_{\theta=\theta^*=\theta_0} \cdot (1,-1)^\top = 0$.

This implies that, locally, variation of the loss around $\theta = \theta^*$ occurs only along the diagonal. Hence, the condition $\frac{\partial J_N(\theta,\theta)}{\partial \theta} \neq 0$ fully characterizes whether $\theta = \theta^*$ is a stationary point of the

expected loss for fixed θ^*. 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 $J_N(\theta^*, \theta) = \mathbb{E}[W_2^2(\hat{f}_{\theta^*,N}, \hat{f}_{\theta,N})]$ is symmetric in its two arguments. Furthermore, it is applicable to the case of high-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 1(c) shows that $J_N(\theta, \theta)$ increases with the scale parameter θ_2, reflecting that sample transport distances grow with scale, particularly when the number of samples is small. By Remark 2 and (18), this implies that a scale parameter smaller than the true value θ^* can yield lower expected loss, thereby inducing the downward bias in the minimizer, as observed in Figure 1(b). As the sample size N increases, the gradient of the expected loss along the diagonal diminishes, indicating that the bias vanishes asymptotically.

Bias correction in location-scale models. Applying the bias correction scheme from Section 3.2 to the location-scale model in Proposition 1, where $J_N(\theta^*, \theta) = (\theta_1 - \theta_1^*)^2 + \theta_2^2 - 2c_N \cdot \theta_2\theta_2^* + \theta_2^{*2}$ as in (4), the modified loss becomes $\tilde{J}_N(\theta^*, \theta) = (\theta_1 - \theta_1^*)^2 + c_N \cdot \theta_2^2 - 2c_N \cdot \theta_2\theta_2^* + \theta_2^{*2}$, which is minimized at $(\theta_1, \theta_2) = (\theta_1^*, \theta_2^*)$.

A.4. Expected squared W_2 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:

$$f_a(x) = a(x - 0.5) + 1, \quad -2 \leq a \leq 2, \quad x \in [0, 1], \tag{19}$$

where a is a slope parameter. 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 to compute $m_{1,i}(a)$ in (7) and $m_2(a)$ in (8). The second moment is

$$m_2(a) = \int_0^1 x^2 p(x; a) dx = \frac{a + 4}{12}, \tag{20}$$

where $p(x; a)$ is the affine density defined in (19).

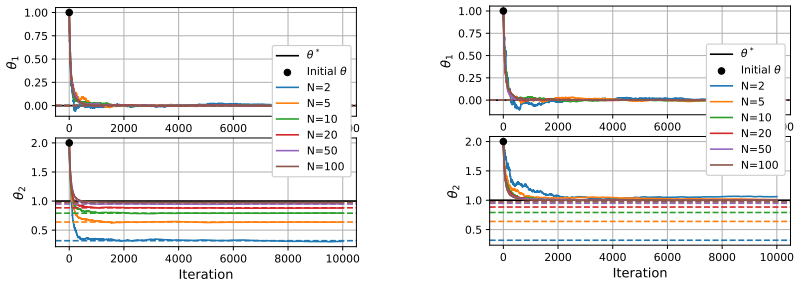
We next derive $m_{1,i}(a)$ in (7) for the affine PDF model. The inverse CDF is

$$F_a^{-1}(u) = \begin{cases} u, & a = 0, \\ \frac{1}{2} - \frac{1}{a} + \frac{1}{a}\sqrt{2au + \left(\frac{a}{2} - 1\right)^2}, & a \neq 0. \end{cases} \tag{21}$$

To compute (7), we make use of the following Gauss-hypergeometric integral identity [16]:

$$\int_0^1 t^{\alpha-1}(1 - t)^{\beta-1}(1 - zt)^{-p} dt = B(\alpha, \beta)_2F_1(p, \alpha; \alpha + \beta; z), \tag{22}$$

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.



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

Figure 3: Stochastic optimization of empirical Wasserstein losses for Gaussian distributions with different sample sizes N. SGD is applied to minimize (a) $J_N(\theta^*, \theta)$ and (b) $\tilde{J}_N(\theta^*, \theta)$ with respect to $\theta = (\theta_1, \theta_2)$ for fixed θ^*. Colored solid lines represent the parameter trajectories, and dashed lines indicate the corresponding minimizers of $J_N(\theta^*, \theta)$; black solid lines indicate the fixed parameter values.

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

$$m_{1,i}(a) = \begin{cases} \frac{i}{N+1}, & a = 0, \\ \frac{1}{2} - \frac{1}{a} + \frac{\sqrt{b}}{a}\, {}_2F_1\left(-\frac{1}{2}, i; N+1; -\frac{2a}{b}\right), & a \neq 0, \end{cases} \tag{23}$$

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

Substituting (20) and (23) 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.

Appendix B. Details for numerical experiments

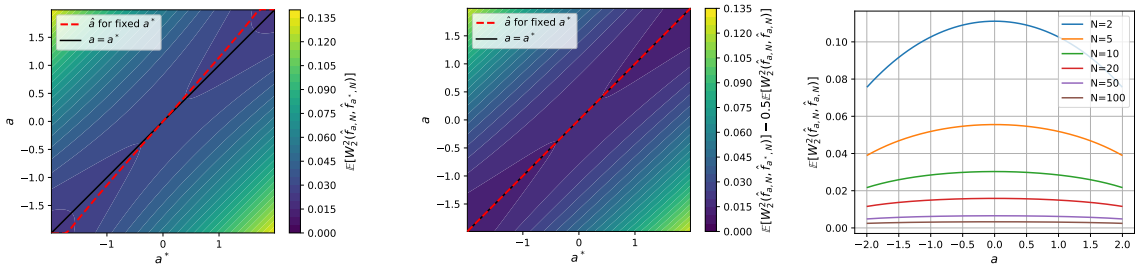
B.1. Stochastic optimization settings: Gaussian and affine models

We study the well-specified case where both the data-generating distribution and the model belong to the same parametric family $\{f_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^m\}$. The data distribution is f_{θ^*} and the model is f_θ, trained to minimize the Wasserstein distance between their empirical distributions.

For gradient-based training, we consider models that admit a differentiable sampling map $z \mapsto g_\theta(z)$ from a simple latent variable z, enabling backpropagation through the sampling process. This structure allows computation of stochastic gradients of the Wasserstein loss with respect to θ.

At each iteration, we draw batches of size N from both f_{θ^*} and f_θ, compute the empirical Wasserstein loss, and update parameters using SGD with a Robbins-Monro step-size schedule [21], $\eta_t = \frac{\eta_0}{1+\gamma \cdot t}$. In the Gaussian model experiment in Figure 2 (b) and (c), we use $\eta_0 = 0.01$ and $\gamma = 0.01$, while for the affine model in Figure 5, we set $\eta_0 = 1.0$ and $\gamma = 0.01$.

Figure 3 illustrates the stochastic optimization of $J_N(\theta^*, \theta)$ and $\tilde{J}_N(\theta^*, \theta)$ for Gaussian distributions with both location and scale parameters. Results for the affine model are provided in Appendix B.3.



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

Figure 4: 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 (5) 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.

B.2. Stochastic optimization settings: Gaussian mixture models

We fit Gaussian mixtures in higher dimensions by minimizing the Wasserstein loss. Both the data (or true) distribution f_{θ^*} and the model f_θ are mixtures with the same known number of components and weights (well-specified setting), and the goal is to recover the component means and covariances. At each iteration, batches of size N are drawn from both f_{θ^*} and f_θ, the empirical Wasserstein loss is computed, and parameters are updated via SGD. Model parameters are initialized using k-means clustering.

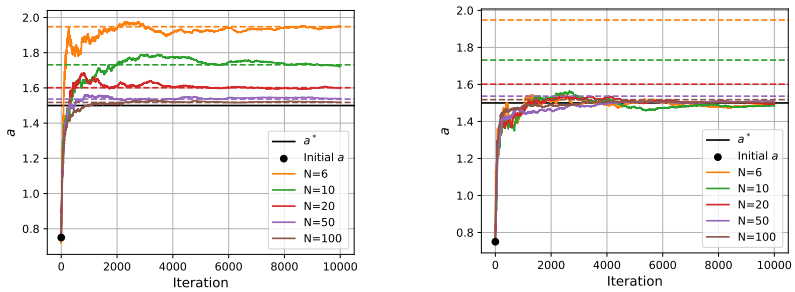
For high-dimensional data, analytic expressions for the Wasserstein distance such as (1)–(2) are not available. Therefore, we compute the optimal transport plan between empirical distributions using the `POT` library [13] and use the resulting Wasserstein distance as the objective. As noted in Appendix A.3, the bias correction scheme from Section 3.2 also applies in these higher-dimensional data and parameter settings.

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

$$\text{Fitting error} = \sum_{k=1}^{K} \pi_k W_2^2(\mathcal{N}(\mu_k^*, \Sigma_k^*), \mathcal{N}(\mu_k, \Sigma_k)), \tag{24}$$

where π_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 2-Wasserstein distance between two Gaussians [20], applied component-wise. This weighted sum serves as an upper bound on the mixture–mixture distance and provides a practical accuracy metric.

We experiment with $K = 4$ components in dimensions $d = 2, 5, 10$, batch sizes $N \in \{16, \ldots, 1024\}$, learning rate 0.01, and 20,000 SGD iterations. Figure 2(c) reports the average fitting errors over three random seeds.



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

Figure 5: 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 values.

B.3. Results for the affine PDF model

We use the affine PDF family (Appendix A.4) to further examine finite-sample bias in minimizing the expected empirical Wasserstein loss. Let f_{a^*} be the data distribution with fixed slope a^* and f_a the model with variable slope a.

Figure 4(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 4(b).

Along the diagonal $a = a^*$ (Figure 4(c)), the expected loss decreases as $|a|$ grows, since larger $|a|$ concentrates mass near the boundaries and reduces transport among dense regions, especially for small N. By Remark 2 and (18), this explains the outward bias in Figure 4(a). As N increases, the gradient along the diagonal diminishes, indicating that the bias vanishes asymptotically.

Figure 5 shows stochastic optimization results. When N is finite, minimizing $J_N(a^*, a)$ via SGD converges to a biased solution (Figure 5(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 5(b)), confirming the effectiveness of the bias correction scheme.

Also note that the observations in Figures 3 and 5 are consistent with [12], which shows that the empirical Wasserstein loss using minibatches is unbiased with respect to its expectation over empirical distributions of the same size.