SHARP BOUNDS FOR MAX-SLICED WASSERSTEIN DISTANCES

MARCH T. BOEDIHARDJO

Abstract. We obtain essentially matching upper and lower bounds for the expected max-sliced 1-Wasserstein distance between a probability measure on a separable Hilbert space and its empirical distribution from n samples. By proving a Banach space version of this result, we also obtain an upper bound, that is sharp up to a log factor, for the expected max-sliced 2-Wasserstein distance between a symmetric probability measure μ on a Euclidean space and its symmetrized empirical distribution in terms of the operator norm of the covariance matrix of μ and the diameter of the support of μ.

1. Introduction

Suppose that μ is a probability measure on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|x\|_2^2\,dx < \infty$, where $\|\,\|_2$ is the Euclidean norm on $\mathbb{R}^d$. Let $X_1, \ldots, X_n$ be i.i.d. samples of μ. How many samples are needed so that the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is "close" to μ? Obviously the answer depends on the notion of "close" we use. If we want the covariance matrix of $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ to be close, in the operator norm, to the covariance matrix of μ, it is already a very deep question of how many samples are needed, though by now, in some aspects, this question has been settled after a series of work [32, 2, 3, 39, 33, 21, 15, 35, 44, 1]. In general, after certain rescaling, $O(d \log d)$ samples suffice to accurately approximate the covariance matrix of μ. On the other hand, if we want $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and μ to be close in the Wasserstein distance, we need n to be exponentially large in d (see, e.g., [12]).

To circumvent this curse of dimensionality issue, in recent years, the notions of sliced, max sliced and projection robust Wasserstein distances have been introduced and used in applications [31, 7, 9, 10, 13, 11, 14, 22, 28, 43, 18, 23, 24]. They were further studied in [26, 42, 19, 4, 25, 27]. The max sliced p-Wasserstein distance between two probability measures μ_1 and μ_2 on $\mathbb{R}^d$ is

$$W_{p,1}(\mu_1, \mu_2) := \sup_{v\in\mathbb{R}^d,\, \|v\|_2=1} W_p(v_{\#}\mu_1, v_{\#}\mu_2), \tag{1.1}$$

where $v_{\#}\mu_i$ is the pushforward probability measure of μ_i by the map $\langle \cdot, v\rangle$, i.e., if μ_i is the distribution of a random vector X_i in $\mathbb{R}^d$, then $v_{\#}\mu_i$ is the distribution of the random variable $\langle X_i, v\rangle$. The quantity $W_p(v_{\#}\mu_1, v_{\#}\mu_2)$ denotes the p-Wasserstein distance between the measures $v_{\#}\mu_1$ and $v_{\#}\mu_2$ on $\mathbb{R}$. The sliced Wasserstein distance (which we do not study in this paper) is the notion where in (1.1), we replace the supremum over v by the integral of $W_p(v_{\#}\mu_1, v_{\#}\mu_2)^p$ over v on the unit sphere and then take the pth root. The projection robust Wasserstein distance $W_{p,s}$ (which we also study in this paper) is the notion where in (1.1), we take the p-Wasserstein distance between the pushforward measures of μ_1 and μ_2 by a projection onto a subspace of a

2020 *Mathematics Subject Classification.* 60B11, 62G05, 62R20.
Key words and phrases. Max-sliced Wasserstein, Projection robust Wasserstein.

fixed dimension s and then take supremum over all such subspaces. When $s = 1$, this is the max-sliced Wasserstein distance $W_{p,1}$.

1.1. **Max-sliced 1-Wasserstein distance.** When $p = 1$, by the Kantorovich-Rubinstein theorem, the max-sliced 1-Wasserstein distance between two probability measures μ_1 and μ_2 on $\mathbb{R}^d$ coincides with the following quantity:

$$W_{1,1}(\mu_1, \mu_2) = \sup_{\substack{v\in\mathbb{R}^d,\, \|v\|_2=1 \\ f \text{ is 1-Lipschitz}}} \left| \int_{\mathbb{R}^d} f(\langle x, v\rangle)\, d\mu_1(x) - \int_{\mathbb{R}^d} f(\langle x, v\rangle)\, d\mu_2(x) \right|, \tag{1.2}$$

where the supremum is over all the v on the unit sphere and over all the 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$ (i.e., $|f(x) - f(y)| \le |x - y|$ for all $x, y \in \mathbb{R}$). Consider the following problem:

Problem 1. Suppose that μ is a probability measure on $\mathbb{R}^d$. Let $X_1, \dots, X_n$ be i.i.d. samples of μ. Estimate $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$.

There are known estimates (some of which are sharp) of $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ under certain regularity assumptions on the measure μ, e.g., log-concavity of μ [25, Theorem 1] and [4, Theorem 1.6], or μ satisfying the spiked transport model and the transport inequality [26, Theorem 1], or μ satisfying the projection Bernstein tail condition or the projection Poincaré inequality [19, Theorem 3.5 and Theorem 3.6], or μ being isotropic with its marginal distributions having uniformly bounded 4th moments [4, Proposition 4.1] (see also [4, Remark 4.2]).

As for the most general setting, under the only assumption of μ being supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, it was shown in [25, Proposition 1] that $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C \cdot \frac{rd}{\sqrt{n}}$, where $C \ge 1$ is a universal constant. In [27, Theorem 2], this was improved to $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C \cdot \frac{r\sqrt{d}}{\sqrt{n}}$. In these two bounds, the rate of convergence $\frac{1}{\sqrt{n}}$ is optimal in n, but both bounds involve the dimension d.

There is a dimension-free bound for $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ that holds with the same generality. More precisely, if μ is supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, then $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C \cdot r \cdot n^{-1/3}$, where $C \ge 1$ is a universal constant. This follows by taking $k = 1$ and optimizing the $\epsilon > 0$ in the term $\mathcal{J}_n$ in [42, Theorem 1]. This estimate is dimension-free but comes at the cost of slower convergence rate in n.

In short, the literature concerning Problem 1 can be summarized as follows.

(1) If μ is supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, then $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C(d)\cdot \frac{r}{\sqrt{n}}$, where $C(d) \ge 1$ is a constant that depends only on d.
(2) If μ is supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, then $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C \cdot r \cdot n^{-1/3}$, where $C \ge 1$ is a universal constant.
(3) If in addition, μ satisfies certain regularity assumptions, then $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C \cdot \frac{r}{\sqrt{n}}$, where $C \ge 1$ is a universal constant.

These results together suggest the following question. Does the dimension-free bound $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \le C \cdot \frac{r}{\sqrt{n}}$, where $C \ge 1$ is a universal constant, actually hold for every μ supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le r\}$ even without any regularity assumptions?

In the first main result of this paper, we answer this question affirmatively. We obtain essentially matching dimension-free upper and lower bounds for $\mathbb{E}W_{1,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ in the most general setting. This essentially settles Problem 1.

Theorem 1.1. *Suppose that μ is a probability measure on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|x\|_2 \, d\mu(x) < \infty$ and $\int_{\mathbb{R}^d} x \, d\mu(x) = 0$. Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ sampled according to μ. Then*

$$\frac{1}{2\sqrt{2n}} \int_{\mathbb{R}^d} \|x\|_2 \, d\mu(x) \leq \mathbb{E} W_{1,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \leq \frac{C}{\sqrt{n}} \cdot \inf_{0<\delta\leq 1} \frac{1}{\sqrt{\delta}} \left(\int_{\mathbb{R}^d} \|x\|_2^{2+\delta} \, d\mu(x)\right)^{\frac{1}{2+\delta}},$$

where $C \geq 1$ is a universal constant.

We also obtain a version of Theorem 1.1 for probability measures on Banach spaces. Beside being a result of intrinsic interest in the study of probability in Banach spaces (see [17]), this result is essential for proving the second main result Theorem 1.4 of this paper on the max-sliced 2-Wasserstein distance for probability measures on Euclidean spaces. Indeed, in proving the latter result, we will take the Banach space E to be the space of all $d \times d$ matrices equipped with the operator norm. In the Banach space setting, to define the metric $W_{1,1}$, in (1.2), instead of taking supremum over v on the unit sphere, we take supremum over all linear functionals $v^* \in B_{E^*}$, where B_{E^*} is the unit ball of the dual space E^* centered at the origin. See Section 1.3 for the precise definition.

Theorem 1.2. *Suppose that μ is a probability measure on a Banach space $(E, \| \|)$ with separable dual E^* and that $\int_E \|x\| \, d\mu(x) < \infty$ and $\int_E x \, d\mu(x) = 0$. Let $X_1, \ldots, X_n$ be i.i.d. random elements of E sampled according to μ. Then*

$$\begin{aligned} \frac{1}{2n}\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i X_i\right\| &\leq \mathbb{E} W_{1,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \\ &\leq \frac{C}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i X_i\right\| + \frac{C\sqrt{\ln n}}{n} \cdot \mathbb{E} \sup_{v^* \in B_{E^*}} \left(\sum_{i=1}^n |v^*(X_i)|^2\right)^{\frac{1}{2}}, \end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. uniform ± 1 random variables and $g_1, \ldots, g_n$ are i.i.d. standard Gaussian random variables that are independent from $X_1, \ldots, X_n$, and $C \geq 1$ is a universal constant.

Remark. If we fix $X_1, \ldots, X_n$, the quantity in the last term $\sup_{v^* \in B_{E^*}} (\sum_{i=1}^n |v^*(X_i)|^2)^{\frac{1}{2}}$ is exactly the Lipschitz constant of the function $(g_1, \ldots, g_n) \mapsto \|\sum_{i=1}^n g_i X_i\|$ with respect to the Euclidean norm on $\mathbb{R}^n$. Moreover, by Khintchine's inequality, if we take the expectation $\mathbb{E}_\epsilon$ on $\epsilon_1, \ldots, \epsilon_n$, we have for $v^* \in B_{E^*}$,

$$\mathbb{E}_\epsilon \left\|\sum_{i=1}^n \epsilon_i X_i\right\| \geq \mathbb{E}_\epsilon \left|\sum_{i=1}^n \epsilon_i v^*(X_i)\right| \geq c \left(\mathbb{E}_\epsilon \left|\sum_{i=1}^n \epsilon_i v^*(X_i)\right|^2\right)^{\frac{1}{2}} = c\left(\sum_{i=1}^n |v^*(X_i)|^2\right)^{\frac{1}{2}},$$

where $c > 0$ is a universal constant. So if we take supremum over $v^* \in B_{E^*}$ and then take the full expectation, we obtain

$$\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i X_i\right\| \geq c \cdot \mathbb{E} \sup_{v^* \in B_{E^*}} \left(\sum_{i=1}^n |v^*(X_i)|^2\right)^{\frac{1}{2}}.$$

Therefore, since $\mathbb{E}\|\sum_{i=1}^n g_i X_i\| \leq C\sqrt{\ln n} \cdot \mathbb{E}\|\sum_{i=1}^n \epsilon_i X_i\|$ [37, Exercise 7.1], the upper and lower bounds in Theorem 1.2 differ by at most a $\sqrt{\ln n}$ factor.

1.2. **Max-sliced 2-Wasserstein distance.** We now turn to the problem of estimating the expected max-sliced 2-Wasserstein distance $\mathbb{E}W_{2,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$.

Unlike in Theorem 1.1, for the max-sliced 2-Wasserstein distance, the convergence rate is not always the same. Even in dimension one, for certain log-concave measures μ on $\mathbb{R}$, for $p \geq 1$, the quantity $\mathbb{E}W_p(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ is of order $\frac{1}{\sqrt{n}}$ [6]. However, if μ is uniformly distributed on two points $1, -1 \in \mathbb{R}$, one can easily see that $\mathbb{E}W_p(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ is of order $n^{-1/(2p)}$, which is much slower than $\frac{1}{\sqrt{n}}$ when $p > 1$.

Similarly, for the max-sliced 2-Wasserstein distance, if we assume certain regularity assumptions on μ (e.g., μ is log-concave [4, 25]), then $\mathbb{E}W_2(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) = O(\frac{1}{\sqrt{n}})$ or $O(\frac{\log n}{\sqrt{n}})$. (Let's ignore the dimension factors for a short moment.) On the other hand, even if μ is isotropic and its marginal distributions have uniformly bounded 4th moments, the quantity $\mathbb{E}W_2(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ could already be as large as $c \cdot (d/n)^{\frac{1}{4}}$ for some universal constant $c > 0$ [4, Example 3.3].

Thus, in the most general setting (i.e., no regularity assumptions on μ), the best convergence rate in n for the max-sliced 2-Wasserstein distance we can hope for is $n^{-1/4}$.

Corollary 1.3. *Let $r > 0$. Suppose that μ is a probability measure on $\{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$. Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ sampled according to μ. Then for all $p \geq 1$,*

$$\mathbb{E}W_{p,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \leq C \cdot r \cdot n^{-1/(2p)},$$

where $C \geq 1$ is a universal constant.

Proof. For two probability measures μ_1, μ_2 on $\{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$, it is easy to see that $W_{p,1}(\mu_1, \mu_2)^p \leq (2r)^{p-1} \cdot W_{1,1}(\mu_1, \mu_2)$. Thus by Theorem 1.1, the result follows. $\square$

Corollary 1.3 removes the dimension factor in the estimate of $\mathbb{E}W_{p,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right)$ in [27, Theorem 2].

The upper bound $C \cdot r \cdot n^{-1/(2p)}$ in Corollary 1.3 is attained, up to the constant C, when $\mu = \frac{1}{2}\delta_{y_0} + \frac{1}{2}\delta_{y_0}$ is uniformly distributed on two points $y_0, -y_0 \in \mathbb{R}^d$ with y_0 being any vector with $\|y_0\|_2 = r$.

While the bound $C \cdot r \cdot n^{-1/(2p)}$ in Corollary 1.3 is sharp in n, r, p, if one also has information on the covariance matrix of μ, then perhaps, one can obtain a better bound that can depend on the covariance matrix of μ. Before we go into further discussions on this, we mention some simple connections between the max-sliced 2-Wasserstein distance and sample covariance matrices. The literature on sample covariance matrices gives us important intuition regarding the convergence in the max-sliced 2-Wasserstein distance.

If μ is a probability measure on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|x\|_2^2 \, d\mu(x) < \infty$, then the max-sliced 2-Wasserstein distance between μ and δ_0 (the probability measure with an atom of mass 1 at the origin) is equal to

$$W_{2,1}(\mu, \delta_0) = \sup_{\|v\|_2=1} \left(\int_{\mathbb{R}^d} |\langle x, v\rangle|^2 \, d\mu(x)\right)^{\frac{1}{2}} = \sup_{\|v\|_2=1} \langle \Sigma v, v\rangle^{\frac{1}{2}} = \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}},$$

where $\Sigma = \int_{\mathbb{R}^d} xx^T\, d\mu(x)$ is a $d \times d$ matrix and $\|\ \|_{\mathrm{op}}$ denotes the operator norm. Thus, for $X_1, \ldots, X_n$ in $\mathbb{R}^d$, we have

$$W_{2,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \geq W_{2,1}\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \delta_0\right) - W_{2,1}(\mu, \delta_0) = \left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}. \tag{1.3}$$

So in order for $W_{2,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right)$ to be small, it is necessary that $\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right\|_{\mathrm{op}}$ cannot be too much larger than $\|\Sigma\|_{\mathrm{op}}$.

Given that $W_{2,1}(\mu, \delta_0) = \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}$, the quantity $W_{2,1}(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i})$ should be assessed relative to $\|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}$.

Problem 2. Suppose that μ is a probability measure on $\mathbb{R}^d$. Let $\Sigma = \int_{\mathbb{R}^d} xx^T\, d\mu(x)$. How many i.i.d. samples $X_1, \ldots, X_n$ of μ are needed to make $\|\Sigma\|_{\mathrm{op}}^{-\frac{1}{2}} \cdot \mathbb{E} W_{2,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right)$ small?

In [4, Theorem 1.3], it was shown that if μ is centered and isotropic (i.e., $\Sigma = I$) with $\sup_{v\in\mathbb{R}^d,\, \|v\|_2=1}(\mathbb{E}|\langle X, v\rangle|^q)^{\frac{1}{q}} \leq L$ where $q > 4$, then with high probability,

$$W_{2,1}\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \leq C(q, L)\left[\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T - I\right\|_{\mathrm{op}}^{\frac{1}{2}} + \left(\frac{d}{n}\right)^{\frac{1}{4}}\right], \tag{1.4}$$

where $C(q, L) \geq 1$ is a constant that depends only on q and L. By [35], the sample covariance error term $\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T - I\right\|_{\mathrm{op}}^{\frac{1}{2}}$ is of order $\left(\frac{d}{n}\right)^{\frac{1}{4}}$ with high probability. Thus, under the assumptions mentioned above, $n = O(d)$ suffices in Problem 2.

The literature on sample covariance matrices (see e.g., [32, 38, 36]) suggests that for a general isotropic probability measure μ supported on $\{x \in \mathbb{R}^d : \|x\|_2 \leq C\sqrt{d}\}$ but without the assumption $\sup_{v\in\mathbb{R}^d,\, \|v\|_2=1}(\mathbb{E}|\langle X, v\rangle|^q)^{\frac{1}{q}} \leq L$, the number of samples $n = O(d \log d)$ should suffice in Problem 2. More generally, if μ is supported on $\{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ but not necessarily isotropic, $n = O(\frac{r^2}{\|\Sigma\|_{\mathrm{op}}} \log \frac{r^2}{\|\Sigma\|_{\mathrm{op}}})$ should suffice in Problem 2.

In this paper, we show that these are indeed true for symmetric μ and its symmetrized empirical distribution. A probability measure μ on $\mathbb{R}^d$ is symmetric if $\mu(A) = \mu(-A)$ for all measurable $A \subset \mathbb{R}^d$.

Theorem 1.4. *Let $r > 0$. Suppose that μ is a symmetric probability measure on $\mathbb{R}^d$ supported on $\{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$. Let $X_1, \ldots, X_n$ be i.i.d. random vectors sampled according to μ. Then*

$$\mathbb{E}\left[W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{X_i} + \delta_{-X_i})\right)^2\right] \leq C\|\Sigma\|_{\mathrm{op}}\left(\frac{r^2 \ln n}{n\|\Sigma\|_{\mathrm{op}}} + \sqrt{\frac{r^2 \ln n}{n\|\Sigma\|_{\mathrm{op}}}}\right),$$

where $\Sigma = \int_{\mathbb{R}^d} xx^T\,d\mu(x)$ *and* $C \geq 1$ *is a universal constant.*

The $\ln n$ factors in Theorem 1.4 cannot always be removed. Indeed, consider the probability measure μ uniformly distributed on the $2d$ points $\pm\sqrt{d}\,e_1,\ldots,\pm\sqrt{d}\,e_d$, where $\{e_1,\ldots,e_d\}$ is the unit vector basis for $\mathbb{R}^d$. Then by (1.3), we have

$$W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n(\delta_{X_i}+\delta_{-X_i})\right) \geq \left\|\frac{1}{n}\sum_{i=1}^n X_iX_i^T\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|^{\frac{1}{2}},$$

where $\Sigma = \int_{\mathbb{R}^d} xx^T\,d\mu(x) = I$. If we view $e_1,\ldots,e_d$ as d bins and each $X_iX_i^T$ as a ball into a bin, then $\frac{1}{d}\left\|\sum_{i=1}^n X_iX_i^T\right\|$ is the maximum number of balls in a bin after n balls are thrown into d bins. So by [30, Theorem 1], when $\frac{d}{\mathrm{polylog}(d)} \leq n \ll d\log d$,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n X_iX_i^T\right\|_{\mathrm{op}}^{\frac{1}{2}} \geq c\left(\frac{d}{n}\cdot\frac{\log d}{\log\frac{d\log d}{n}}\right)^{\frac{1}{2}},$$

where $c > 0$ is a universal constant. Thus, in this example, the $\ln n$ factors in Theorem 1.4 cannot be removed.

The following lower bound result shows that the upper bound in Theorem 1.4 is sharp for every covariance matrix Σ up to the $\ln n$ factor.

Proposition 1.5. *Let* Σ *be a* $d\times d$ *positive semidefinite matrix such that* $\|\Sigma\|_{\mathrm{op}} \leq \frac{1}{2}\mathrm{Tr}(\Sigma)$. *Then there exists a symmetric probability measure* μ *on* $\mathbb{R}^d$ *supported on* $\{x\in\mathbb{R}^d : \|x\|_2^2 = \mathrm{Tr}(\Sigma)\}$ *such that* $\int_{\mathbb{R}^d} xx^T\,d\mu(x) = \Sigma$ *and for every* $n \in \mathbb{N}$,

$$\mathbb{E}\left[W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n(\delta_{X_i}+\delta_{-X_i})\right)^2\right] \geq \frac{1}{16}\|\Sigma\|_{\mathrm{op}}\left(\frac{\mathrm{Tr}(\Sigma)}{n\|\Sigma\|_{\mathrm{op}}} + \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n\|\Sigma\|_{\mathrm{op}}}}\right),$$

where $X_1,\ldots,X_n$ *are i.i.d. random vectors sampled according to* μ.

1.3. **Some definitions.** Throughout this paper, unless specified otherwise, we always use the Euclidean metric $\|\ \|_2$ on $\mathbb{R}^d$. If $f:\Lambda\to\mathbb{R}$ is a bounded function, then $\|f\|_\infty := \sup_{x\in\Lambda}|f(x)|$. A function $f:\mathbb{R}^s\to\mathbb{R}$ is 1-Lipschitz function if $|f(x)-f(y)| \leq \|x-y\|_2$ for all $x,y\in\mathbb{R}^s$. The operator norm (or equivalently the largest singular value) of a matrix A is denoted by $\|A\|_{\mathrm{op}}$

If (T,ρ) is a metric space and $\epsilon > 0$, then the covering number $N(T,\rho,\epsilon)$ is the smallest size of $S\subset T$ for which every element of T has distance at most ϵ from an element of S. The packing number $N_{\mathrm{pack}}(T,\rho,\epsilon)$ is the largest size of $S\subset T$ for which all elements of S have distance more than ϵ away from each other. We always have $N(T,\rho,\epsilon) \leq N_{\mathrm{pack}}(T,\rho,\epsilon) \leq N(T,\rho,\frac{\epsilon}{2})$.

If E is a Banach space, then the unit ball $\{x\in E : \|x\|\leq 1\}$ of E is denoted by B_E. The dual space of all bounded linear functionals $v^*: E\to\mathbb{R}$ is denoted by E^*.

Pushforward measure: If μ is a probability measure on a separable Banach space E and $Q: E\to\mathbb{R}^s$ is a map, then $Q_\#\mu$ is the pushforward measure of μ by Q, i.e., if X is a random element of E with distribution μ, then $Q(X)$ has distribution $Q_\#\mu$. In particular, $Q_\#\mu$ is a probability measure on $\mathbb{R}^s$.

Classical Wasserstein distance: If μ_1 and μ_2 are probability measures on E and $p \geq 1$, then the p-Wasserstein distance between μ_1 and μ_2 is

$$W_p(\mu_1, \mu_2) = \inf_{\gamma} \left(\int_{E \times E} \|x - y\|^p \, d\gamma(x, y) \right)^{\frac{1}{p}},$$

where the infimum is over all distributions γ on $E \times E$ with μ_1 and μ_2 being its marginal distributions for its first and second components.

Max-sliced and projection robust Wasserstein distances: If μ and ν are probability measures on E and $p \geq 1$, $s \in \mathbb{N}$, then

$$W_{p,s}(\mu_1, \mu_2) = \sup_{Q} W_p(Q_{\#}\mu_1, Q_{\#}\mu_2),$$

where the supremum is over all $Q : E \to \mathbb{R}^s$ of the form $Qx = (v_1^*(x), \ldots, v_s^*(x))$, for $x \in E$, with $v_1^*, \ldots, v_s^*$ in the unit ball B_{E^*} of E^*. Here we use the Euclidean distance $\| \, \|_2$ on $\mathbb{R}^s$ to define the Wasserstein distance W_p on the right hand side.

When $p = 1$, we have

$$\begin{aligned} &W_{1,s}(\mu_1, \mu_2) \\ &= \sup_{\substack{v_1^*, \ldots, v_s^* \in B_{E^*} \\ f \text{ is 1-Lipschitz}}} \left| \int_E f(v_1^*(x), \ldots, v_s^*(x)) \, d\mu_1(x) - \int_E f(v_1^*(x), \ldots, v_s^*(x)) \, d\mu(x) \right|, \end{aligned}$$

where the supremum is over all $v_1^*, \ldots, v_s^* \in B_{E^*}$ and all the 1-Lipschitz functions $f : \mathbb{R}^s \to \mathbb{R}$.

1.4. **Organization of this paper.** In the rest of this paper, we prove the results stated in this introduction section.

In Section 2, we prove Theorem 1.1 and Theorem 1.2. The upper bound parts of Theorem 1.1 and Theorem 1.2 are contained in Corollary 2.8 and Corollary 2.9, respectively. The lower bound parts of Theorem 1.1 and Theorem 1.2 are stated as Corollary 2.11 and Proposition 2.10, respectively.

In Section 3, we prove Theorem 1.4 and Proposition 1.5. Theorem 1.4 is restated as Theorem 3.3. Proposition 1.5 is restated as Proposition 3.4.

2. Max-sliced 1-Wasserstein distance

In this section, we first derive a general upper bound result Theorem 2.7 (which we obtain at a greater generality of $W_{1,s}$) for the expected max-sliced 1-Wasserstein distance between a probability measure on a Banach space and its empirical distribution. From this result, Corollary 2.8 and Corollary 2.9 follow as consequences. These give the upper bound parts of Theorem 1.1 and 1.2, respectively. Lower bound results are proved at the end of this section.

To prove Theorem 2.7, we use Gaussian symmetrization to reduce the problem of bounding the expected max-sliced 1-Wasserstein distance to bounding the expected supremum of a Gaussian process. To bound this expected supremum, we use Talagrand's majorizing measure theorem. We bound the metric induced by the Gaussian process by the product metric of (1) a metric on some function space (which is locally an $\| \, \|_\infty$ metric) and (2) a Hilbert space metric. Since Talagrand's γ_2 quantity of the product metric space is bounded by 3 times the sum of the γ_2 for each metric space, it suffices to bound the γ_2 for each of these two metric spaces. To bound the γ_2 for the first metric space, we use the Dudley's entropy integral. As for the second metric space, since it is a Hilbert space metric, the γ_2 for that metric space is equivalent to the supremum of some Gaussian process which, in fact, coincides with the norm of a Gaussian sum.

Throughout this section,

$$N_k = \begin{cases} 2^{2^k}, & k \geq 1 \\ 1, & k = 0 \end{cases}.$$

The following notion was introduced by Talagrand [34] (see also [40, Chapter 8] and [37, Chapter 6]). For a given metric space (T, ρ), define

$$\gamma_2(T, \rho) = \inf_{\text{admissible}\, T_0, T_1, \ldots} \sup_{t \in T} \sum_{k=0}^{\infty} 2^{\frac{k}{2}} \rho(t, T_k), \tag{2.1}$$

where admissible means that $T_0, T_1, \ldots \subset T$ with $|T_k| \leq N_k$ for all $k \geq 0$. Also $\rho(t, T_k) = \inf_{t_k \in T_k} \rho(t, t_k)$.

Talagrand's majorizing measure theorem states that if $(X_t)_{t \in T}$ is a mean zero Gaussian process, then letting $\rho(t, s) = (\mathbb{E}|X_t - X_s|^2)^{\frac{1}{2}}$, we have

$$c\gamma_2(T, \rho) \leq \mathbb{E} \sup_{t \in T} X_t \leq C\gamma_2(T, \rho), \tag{2.2}$$

where $C, c > 0$ are universal constants.

Lemma 2.1. *Let (T, ρ_T) and (Z, ρ_Z) be metric spaces. Define the metric $\rho_T \times \rho_Z$ on $T \times Z$ by*

$$(\rho_T \times \rho_Z)((t_1, z_1), (t_2, z_2)) = \rho_T(t_1, t_2) + \rho_Z(z_1, z_2).$$

Then

$$\gamma_2(T \times Z, \rho_T \times \rho_Z) \leq 3\gamma_2(T, \rho_T) + 3\gamma_2(Z, \rho_Z).$$

Proof. Fix $\epsilon > 0$. Let $T_0, T_1, \ldots \subset T$ be an admissible sequence that almost attains the infimum in (2.1), i.e.,

$$\sup_{t \in T} \sum_{k=0}^{\infty} 2^{\frac{k}{2}} \rho_T(t, T_k) \leq \gamma_2(T, \rho_T) + \epsilon.$$

Similarly, let $Z_0, Z_1, \ldots \subset Z$ be an admissible sequence such that

$$\sup_{z \in Z} \sum_{k=0}^{\infty} 2^{\frac{k}{2}} \rho_Z(z, Z_k) \leq \gamma_2(Z, \rho_Z) + \epsilon.$$

For notational convenience, let $T_{-1} = T_0$ and $Z_{-1} = Z_0$.

Observe that the sequence $(T_{k-1} \times Z_{k-1})_{k \geq 0}$ is admissible. For all $t \in T$ and $z \in Z$, we have

$$\begin{aligned} \sum_{k=0}^{\infty} 2^{\frac{k}{2}} (\rho_T \times \rho_Z)((t, z), T_{k-1} \times Z_{k-1}) &= \sum_{k=0}^{\infty} 2^{\frac{k}{2}} [\rho_T(t, T_{k-1}) + \rho_Z(z, Z_{k-1})] \\ &= \sum_{k=-1}^{\infty} 2^{\frac{k+1}{2}} \rho_T(t, T_k) + \sum_{k=-1}^{\infty} 2^{\frac{k+1}{2}} \rho_Z(z, Z_k) \\ &\leq 3 \sum_{k=0}^{\infty} 2^{\frac{k}{2}} \rho_T(t, T_k) + 3 \sum_{k=-1}^{\infty} 2^{\frac{k}{2}} \rho_Z(z, Z_k) \end{aligned}$$

$$\leq \quad 3[\gamma_2(T,\rho_T)+\epsilon]+3[\gamma_2(Z,\rho_Z)+\epsilon].$$

So

$$\gamma_2((T,Z),\rho_T\times\rho_Z)\leq 3\gamma_2(T,\rho_T)+3\gamma_2(Z,\rho_Z)+6\epsilon.$$

Since this holds for all $\epsilon>0$, the result follows. □

Lemma 2.2 ([34], page 12-13)**.** *Let (T,ρ_T) be a metric space. Then*

$$\gamma_2(T,\rho_T)\leq C\int_0^\infty \sqrt{\log N(T,\rho_T,\epsilon)}\,d\epsilon,$$

where $C\geq 1$ is a universal constant.

Next we bound the covering number of a set of 1-Lipschitz functions with respect to a certain norm (see Lemma 2.4 below). This will be needed when we apply Lemma 2.2 to bound the γ_2 quantity for that metric space of 1-Lipschitz functions. Before we do that, we need a basic result.

In the sequel, the readers who are interested in the max-sliced Wasserstein distances but not the general projection robust Wasserstein distances may take $s=1$ in the rest of this paper. This will be enough to prove the main results mentioned in the introduction section.

Lemma 2.3. *Let $a>0$. Let*

$$D=\{h:[-a,a]^s\to\mathbb{R}|\,h \textit{ is 1-Lipschitz and } h(0)=0\}.$$

Then

$$N(D,\|\,\|_\infty,\epsilon)\leq \exp\left(\left(\frac{Ca\sqrt{s}}{\epsilon}\right)^s\right),$$

for all $\epsilon>0$, where $C\geq 1$ is a universal constant.

Proof. The map $h\to(x\mapsto h(ax))$ defines an isometry from the metric space $(D,\|\,\|_\infty)$ to the metric space $(\widetilde{D},\|\,\|_\infty)$, where

$$\widetilde{D}=\{h:[-1,1]^s\to\mathbb{R}|\,h \text{ is } a\text{-Lipschitz and } h(0)=0\}.$$

So

$$N(D,\|\,\|_\infty,\epsilon)=N(\widetilde{D},\|\,\|_\infty,\epsilon).$$

Since

$$\widetilde{D}\subset\widehat{D}:=\{h:[-1,1]^s\to\mathbb{R}:h(0)=0 \text{ and } |h(x)-h(y)|\leq a\sqrt{s}\max_i|x_i-y_i|\;\forall x,y\in[-1,1]^s\}$$

and it is well known (see, e.g., [41, page 129]) that $N(\widehat{D},\|\,\|_\infty,\epsilon)\leq\exp((\frac{Ca\sqrt{s}}{\epsilon})^s)$, it follows that

$$N(\widetilde{D},\|\,\|_\infty,\epsilon)\leq N(\widehat{D},\|\,\|_\infty,\frac{\epsilon}{2})\leq\exp\left(\left(\frac{Ca\sqrt{s}}{\epsilon}\right)^s\right).$$

So the result follows. □

Lemma 2.4. *Let T be the set of all 1-Lipschitz functions $f : \mathbb{R}^s \to \mathbb{R}$ with $f(0) = 0$. For $0 < \delta \leq 1$, define the norm $\| \, \|_{(\delta)}$ on T by*

$$\|f\|_{(\delta)} = \sup_{x \in \mathbb{R}^s} \frac{|f(x)|}{\|x\|_2^{1+\delta} + 1}. \tag{2.3}$$

Then

$$\log N(T, \| \, \|_{(\delta)}, \epsilon) \leq \left(\frac{C\sqrt{s}}{\epsilon} \right)^s \frac{1}{\delta},$$

for all $\epsilon > 0$ and $0 < \delta \leq 1$, where $C \geq 1$ is a universal constant.

Proof. Set $\Omega_0 = \{x \in \mathbb{R}^s : \|x\|_2 \leq 1\}$, and for $j \in \mathbb{N}$, set

$$\Omega_j = \{x \in \mathbb{R}^s : 2^{j-1} \leq \|x\|_2 \leq 2^j\} \cup \{0\}.$$

Let

$$A_j = \{h : \Omega_j \to \mathbb{R} | \, h \text{ is 1-Lipschitz and } h(0) = 0\}.$$

Define the following norm $\| \, \|_{(\delta),j}$ on A_j:

$$\|h\|_{(\delta),j} = \sup_{x \in \Omega_j} \frac{|h(x)|}{\|x\|_2^{1+\delta} + 1} \quad \text{for } h \in A_j.$$

For every $f \in T$, observe that the restriction $f|_{\Omega_j} \in A_j$ and

$$\|f\|_{(\delta)} = \sup_{j \geq 0} \left\| f|_{\Omega_j} \right\|_{(\delta),j}.$$

Thus, $(T, \| \, \|_{(\delta)})$ can be identified as a metric subspace of the product metric space $\prod_{j=0}^{\infty} (A_j, \| \, \|_{(\delta)})$. So the ϵ-covering number of T is bounded by the $\frac{\epsilon}{2}$-covering number of $\prod_{j \in \mathbb{N}} A_j$. So

$$N(T, \| \, \|_{(\delta)}, \epsilon) \leq \prod_{j=0}^{\infty} N(A_j, \| \, \|_{(\delta),j}, \frac{\epsilon}{2}). \tag{2.4}$$

Note that for all $j \geq 1 + \frac{1}{\delta} \log_2 \frac{1}{\epsilon}$ and $h \in A_j$, we have

$$\|h\|_{(\delta),j} = \sup_{x \in \Omega_j} \frac{|h(x)|}{\|x\|_2^{1+\delta} + 1} \leq \sup_{x \in \Omega_j \setminus \{0\}} \frac{\|x\|_2}{\|x\|_2^{1+\delta} + 1} \leq \sup_{x \in \Omega_j \setminus \{0\}} \|x\|_2^{-\delta} \leq 2^{-\delta(j-1)} \leq \epsilon.$$

So $N(A_j, \| \, \|_{*,j}, \epsilon) = 1$ for all $j \geq 1 + \frac{1}{\delta} \log_2 \frac{1}{\epsilon}$.

For $j \geq 0$, let

$$D_j = \{h : [-2^j, 2^j]^s \to \mathbb{R} | \, h \text{ is 1-Lipschitz and } h(0) = 0\}. \tag{2.5}$$

Note that $\Omega_j \subset [-2^j, 2^j]^s$. Every function $h \in A_j$ can be extended to a function $\tau(h) \in D_j$ (by Kirszbraun extension), where

$$[\tau(h)](x) = \inf_{y \in \Omega_j} (h(y) + \|x - y\|_2) \quad \text{for } x \in [-2^j, 2^j]^s.$$

(Note that $\tau(0)$ is not the zero function, but $[\tau(h)](0) = 0$.) For all $h_1, h_2 \in A_j$ with $j \geq 0$,

$$\|h_1 - h_2\|_{(\delta),j} \quad = \quad \sup_{x \in \Omega_j \setminus \{0\}} \frac{|h_1(x) - h_2(x)|}{\|x\|_2^{1+\delta} + 1}$$

$$
\begin{aligned}
&\leq \sup_{x\in\Omega_j\setminus\{0\}} \frac{|h_1(x)-h_2(x)|}{2^{(j-1)(1+\delta)}}\\
&\leq 2^{(1-j)(1+\delta)} \sup_{x\in[-2^j,2^j]^s} |[\tau(h_1)](x)-[\tau(h_2)](x)|\\
&= 2^{(1-j)(1+\delta)}\|\tau(h_1)-\tau(h_2)\|_\infty,
\end{aligned}
$$

where $\|h\|_\infty = \sup_{x\in[-2^j,2^j]^s} |h(x)|$ for $h\in D_j$. So for all $j\geq 0$,

$$
\begin{aligned}
N(A_j,\|\,\|_{(\delta),j},\epsilon) &\leq N_{\text{pack}}(A_j,\|\,\|_{(\delta),j},\epsilon)\\
&\leq N_{\text{pack}}(D_j,\|\,\|_\infty,2^{(j-1)(1+\delta)}\epsilon)\\
&\leq \exp\left(\left(\frac{C\cdot 2^j\sqrt{s}}{2^{(j-1)(1+\delta)}\epsilon}\right)^s\right) = \exp\left(\left(\frac{C\sqrt{s}}{\epsilon}\right)^s 2^{s(1+\delta-j\delta)}\right),
\end{aligned}
$$

where the last inequality follows from Lemma 2.3. Therefore, by (2.4),

$$
\log N(T,\|\,\|_{(\delta)},\epsilon) \leq \sum_{j=0}^{\infty}\left(\frac{C\sqrt{s}}{\epsilon}\right)^s 2^{s(1+\delta-j\delta)}.
$$

But

$$
\sum_{j=0}^{\infty} 2^{s(1+\delta-j\delta)} = \frac{2^{s(1+\delta)}}{1-2^{-s\delta}} \leq \frac{2^{2s}}{1-2^{-\delta}} \leq C\cdot\frac{2^{2s}}{\delta},
$$

since $0<\delta\leq 1$. So the result follows. □

The following result is the main lemma of this section. We bound the expected supremum of the Gaussian process that arises when we use Gaussian symmetrization to prove Theorem 2.7. The key ingredient in proving this lemma is Talagrand's majorizing measure theorem.

Lemma 2.5. *Let $0<\delta\leq 1$. Suppose that E is a Banach space with separable dual E^* and $x_1,\dots,x_n\in E$. Let $g_1,\dots,g_n$ be i.i.d. standard Gaussian random variables. Let T be the set of all 1-Lipschitz functions $f:\mathbb{R}^s\to\mathbb{R}$ with $f(0)=0$. Then*

$$
\begin{aligned}
&\mathbb{E}\sup_{\substack{v_1^*,\dots,v_s^*\in B_{E^*}\\ f\in T}}\left|\frac{1}{n}\sum_{i=1}^n g_i f(v_1^*(x_i),\dots,v_s^*(x_i))\right|\\
\leq &\frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\| + \frac{CM\sqrt{s}}{\sqrt{n}}\cdot\begin{cases}(\delta n)^{-\frac{1}{2}}, & s=1\\ (\ln(\delta n+2))\cdot(\delta n)^{-\frac{1}{2}}, & s=2\\ (\delta n)^{-\frac{1}{s}}, & s\geq 3\end{cases},
\end{aligned}
$$

where

$$
M = \sqrt{2}\left(n + s^{1+\delta}\sup_{v^*\in B_{E^*}}\sum_{i=1}^n |v^*(x_i)|^{2+2\delta}\right)^{\frac{1}{2}}, \tag{2.6}
$$

and $B_{E^*}=\{v^*\in E^*:\|v^*\|\leq 1\}$.

Proof. Let $Z = \{(v_1^*, \ldots, v_s^*) : v_1^*, \ldots, v_s^* \in B_{E^*}\}$. Define the Gaussian process $(X_{f,z})_{(f,z)\in T\times Z}$ as follows. If $f \in T$ and $z = (v_1^*, \ldots, v_s^*) \in Z$, then

$$X_{f,z} = \sum_{i=1}^{n} g_i f(v_1^*(x_i), \ldots, v_s^*(x_i)).$$

Recall that $\|\ \|_{(\delta)}$ is defined in (2.3). For $f, h \in T$ and $(v_1^*, \ldots, v_s^*) \in Z$, we have

$$\begin{aligned}
&\|\{f(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1\le i\le n} - \{h(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1\le i\le n}\|_2 \qquad (2.7)\\
=&\left(\sum_{i=1}^{n} |f(v_1^*(x_i), \ldots, v_s^*(x_i)) - h(v_1^*(x_i), \ldots, v_s^*(x_i))|^2\right)^{\frac{1}{2}}\\
\le&\|f-h\|_{(\delta)}\left(\sum_{i=1}^{n}\left[1 + \|(v_1^*(x_i), \ldots, v_s^*(x_i))\|_2^{1+\delta}\right]^2\right)^{\frac{1}{2}}\\
\le&\|f-h\|_{(\delta)}\left(\sum_{i=1}^{n} 2\left[1 + \|(v_1^*(x_i), \ldots, v_s^*(x_i))\|_2^{2+2\delta}\right]\right)^{\frac{1}{2}}\\
\le&\|f-h\|_{(\delta)}\sqrt{2}\left(\sum_{i=1}^{n}\left[1 + s^{\delta}(|v_1^*(x_i)|^{2+2\delta} + \ldots + |v_s^*(x_i)|^{2+2\delta})\right]\right)^{\frac{1}{2}}\\
\le&\|f-h\|_{(\delta)}\sqrt{2}\left(n + s^{1+\delta}\sup_{v^*\in B_{E^*}}\sum_{i=1}^{n}|v^*(x_i)|^{2+2\delta}\right)^{\frac{1}{2}}\\
=&M\|f-h\|_{(\delta)},
\end{aligned}$$

where $M > 0$ is defined in (2.6).

Fix $b > 0$. Let $T^{(b)} \subset T$ be a b-covering of T with respect to $\|\ \|_{(\delta)}$ that has the smallest size, i.e., $|T^{(b)}| = N(T, \|\ \|_{(\delta)}, b)$. For every $f \in T$, there exists $h \in T^{(b)}$ such that $\|f - h\|_{(\delta)} \le b$ so by (2.7),

$$\|\{f(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1\le i\le n} - \{h(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1\le i\le n}\|_2 \le bM,$$

for all $v_1^*, \ldots, v_s^* \in B_{E^*}$. So

$$\begin{aligned}
&\sup_{v_1^*,\ldots,v_s^*\in B_{E^*}} \frac{1}{n}\sum_{i=1}^{n} g_i f(v_1^*(x_i), \ldots, v_s^*(x_i))\\
\le&\sup_{v_1^*,\ldots,v_s^*\in B_{E^*}} \frac{1}{n}\sum_{i=1}^{n} g_i h(v_1^*(x_i), \ldots, v_s^*(x_i)) + \frac{1}{n}\|(g_1, \ldots, g_n)\|_2 \cdot bM.
\end{aligned}$$

So since $T = -T$, we have

$$\begin{aligned}
&\mathbb{E}\sup_{\substack{v_1^*,\ldots,v_s^*\in B_{E^*}\\ f\in T}}\left|\frac{1}{n}\sum_{i=1}^{n} g_i f(v_1^*(x_i), \ldots, v_s^*(x_i))\right| \qquad (2.8)\\
=&\mathbb{E}\sup_{\substack{v_1^*,\ldots,v_s^*\in B_{E^*}\\ f\in T}}\frac{1}{n}\sum_{i=1}^{n} g_i f(v_1^*(x_i), \ldots, v_s^*(x_i))
\end{aligned}$$

$$\leq \mathbb{E} \sup_{\substack{v_1^*,\ldots,v_s^* \in B_{E^*} \\ h \in T^{(b)}}} \frac{1}{n} \sum_{i=1}^n g_i h(v_1^*(x_i), \ldots, v_s^*(x_i)) + \mathbb{E}\frac{1}{n} \|(g_1, \ldots, g_n)\|_2 \cdot bM$$
$$\leq \mathbb{E} \sup_{\substack{v_1^*,\ldots,v_s^* \in B_{E^*} \\ h \in T^{(b)}}} \frac{1}{n} \sum_{i=1}^n g_i h(v_1^*(x_i), \ldots, v_s^*(x_i)) + \frac{bM}{\sqrt{n}}$$
$$= \frac{1}{n} \mathbb{E} \sup_{(h,z) \in T^{(b)} \times Z} X_{h,z} + \frac{bM}{\sqrt{n}},$$

where $X_{h,z}$ is defined at the beginning of this proof.

For $f, h \in T$ and $z_1 = (v_1^*, \ldots, v_s^*) \in Z$, $z_2 = (w_1^*, \ldots, w_s^*) \in Z$, we have

$$\begin{aligned}
&\left(\mathbb{E}|X_{f,z_1} - X_{h,z_2}|^2\right)^{\frac{1}{2}} \\
=&\|\{f(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1 \leq i \leq n} - \{h(w_1^*(x_i), \ldots, w_s^*(x_i))\}_{1 \leq i \leq n}\|_2 \\
\leq&\|\{f(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1 \leq i \leq n} - \{h(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1 \leq i \leq n}\|_2 \\
&+ \|\{h(v_1^*(x_i), \ldots, v_s^*(x_i))\}_{1 \leq i \leq n} - \{h(w_1^*(x_i), \ldots, w_s^*(x_i))\}_{1 \leq i \leq n}\|_2 \\
\leq&M\|f-h\|_{(\delta)} + \left(\sum_{i=1}^n |h(v_1^*(x_i), \ldots, v_s^*(x_i)) - h(w_1^*(x_i), \ldots, w_s^*(x_i))|^2\right)^{\frac{1}{2}} \\
\leq&M\|f-h\|_{(\delta)} + \left(\sum_{i=1}^n \sum_{j=1}^s |v_j^*(x_i) - w_j^*(x_i)|^2\right)^{\frac{1}{2}},
\end{aligned} \tag{2.9}$$

where the second inequality follows from (2.7) and the last inequality follows from h being 1-Lipschitz. Recall that $M > 0$ is defined in (2.6) and $\| \ \|_{(\delta)}$ is defined in (2.3). Consider the metric $\rho_T(f, h) = M\|f - h\|_{(\delta)}$ on T. Also, define the metric ρ_Z on Z by

$$\rho_Z((v_1^*, \ldots, v_s^*), (w_1^*, \ldots, w_s^*)) = \left(\sum_{i=1}^n \sum_{j=1}^s |v_j^*(x_i) - w_j^*(x_i)|^2\right)^{\frac{1}{2}}.$$

Then by (2.9), we have

$$\left(\mathbb{E}|X_{f,z_1} - X_{h,z_2}|^2\right)^{\frac{1}{2}} \leq \rho_T(f, h) + \rho_Z(z_1, z_2),$$

for all $(f, z_1), (h, z_2) \in T \times Z$. So by (2.2) and Lemma 2.1,

$$\mathbb{E} \sup_{(f,z) \in T^{(b)} \times Z} X_{f,z} \leq C\gamma_2(T^{(b)} \times Z, \rho_T \times \rho_Z) \leq C\gamma_2(T^{(b)}, \rho_T) + C\gamma_2(Z, \rho_Z). \tag{2.10}$$

Let's bound each of these two terms. For the first term, by Lemma 2.2,

$$\begin{aligned}
&\gamma_2(T^{(b)}, \rho_T) \\
\leq&C \int_0^\infty \sqrt{\log N(T^{(b)}, \rho_T, \epsilon)}\, d\epsilon \\
=&C \int_0^\infty \sqrt{\log N(T^{(b)}, \| \ \|_{(\delta)}, \frac{\epsilon}{M})}\, d\epsilon \\
=&CM \int_0^\infty \sqrt{\log N(T^{(b)}, \| \ \|_{(\delta)}, \epsilon)}\, d\epsilon
\end{aligned} \tag{2.11}$$

$$\begin{aligned}
&\leq CM\left(\int_b^\infty \sqrt{\log N(T^{(b)}, \|\ \|_{(\delta)}, \epsilon)}\, d\epsilon + b\sqrt{\log|T^{(b)}|}\right)\\
&\leq CM\left(\int_b^\infty \sqrt{\log N(T, \|\ \|_{(\delta)}, \frac{\epsilon}{2})}\, d\epsilon + b\sqrt{\log|T^{(b)}|}\right)\\
&\leq CM\int_{\frac{b}{2}}^\infty \sqrt{\log N(T, \|\ \|_{(\delta)}, \epsilon)}\, d\epsilon,
\end{aligned}$$

where the second last inequality follows from $T^{(b)} \subset T$ and the last inequality follows from $|T^{(b)}| = N(T, \|\ \|_{(\delta)}, b)$ (by definition of $T^{(b)}$) and $b = 2\int_{\frac{b}{2}}^{b} 1\, d\epsilon$.

We now bound the other term in (2.10). Let $(g_{i,j})_{1\leq i\leq n,\, 1\leq j\leq s}$ be i.i.d. standard Gaussian random variables. Then for $(v_1^*, \ldots, v_s^*), (w_1^*, \ldots, w_s^*) \in Z$, we have

$$\begin{aligned}
\left(\mathbb{E}\left|\sum_{i=1}^n\sum_{j=1}^s g_{i,j}v_j^*(x_i) - \sum_{i=1}^n\sum_{j=1}^s g_{i,j}w_j^*(x_i)\right|^2\right)^{\frac{1}{2}} &= \left(\sum_{i=1}^n\sum_{j=1}^s |v_j^*(x_i) - w_j^*(x_i)|^2\right)^{\frac{1}{2}}\\
&= \rho_Z((v_1^*, \ldots, v_s^*), (w_1^*, \ldots, w_s^*)).
\end{aligned}$$

So by (2.2),

$$\begin{aligned}
\gamma_2(Z, \rho_Z) &\leq C\cdot\mathbb{E}\sup_{(v_1^*,\ldots,v_s^*)\in Z}\sum_{i=1}^n\sum_{j=1}^s g_{i,j}v_j^*(x_i)\\
&= C\cdot\mathbb{E}\sum_{j=1}^s\sup_{v^*\in B_{E^*}}\sum_{i=1}^n g_{i,j}v^*(x_i)\\
&= C\cdot\mathbb{E}\sum_{j=1}^s\left\|\sum_{i=1}^n g_{i,j}x_i\right\| = Cs\cdot\mathbb{E}\left\|\sum_{i=1}^n g_ix_i\right\|.
\end{aligned}$$

So we have bounded the second term in (2.10). Together with the bound (2.11) for the first term, we obtain the following from (2.10).

$$\mathbb{E}\sup_{(f,z)\in T^{(b)}\times Z} X_{f,z} \leq CM\int_{\frac{b}{2}}^\infty \sqrt{\log N(T, \|\ \|_{(\delta)}, \epsilon)}\, d\epsilon + Cs\cdot\mathbb{E}\left\|\sum_{i=1}^n g_ix_i\right\|.$$

Combining this with (2.8), we obtain

$$\begin{aligned}
&\mathbb{E}\sup_{\substack{v_1^*,\ldots,v_s^*\in B_{E^*}\\ f\in T}}\left|\frac{1}{n}\sum_{i=1}^n g_if(v_1^*(x_i), \ldots, v_s^*(x_i))\right|\\
&\leq C\inf_{b>0}\left(\frac{bM}{\sqrt{n}} + \frac{M}{n}\int_b^\infty \sqrt{\log N(T, \|\ \|_{(\delta)}, \epsilon)}\, d\epsilon\right) + \frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_ix_i\right\|\\
&\leq C\inf_{0<b\leq 1}\left(\frac{bM}{\sqrt{n}} + \frac{M}{n}\int_b^1 \sqrt{\left(\frac{C\sqrt{s}}{\epsilon}\right)^s\frac{1}{\delta}}\, d\epsilon\right) + \frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_ix_i\right\|
\end{aligned}$$

by Lemma 2.4

$$
\begin{aligned}
&= \frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\| + \frac{CM}{\sqrt{n}} \cdot \inf_{0<b\leq 1}\left(b + \frac{1}{\sqrt{\delta n}}\int_b^1 \sqrt{\left(\frac{C\sqrt{s}}{\epsilon}\right)^s}\, d\epsilon\right) \\
&\leq \frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\| + \frac{CM}{\sqrt{n}} \cdot \begin{cases} (\delta n)^{-\frac{1}{2}}, & s=1 \\ (\ln(\delta n+2))\cdot(\delta n)^{-\frac{1}{2}}, & s=2 \\ \sqrt{s}\cdot(\delta n)^{-\frac{1}{s}}, & s\geq 3 \end{cases},
\end{aligned}
$$

where we take $b = \begin{cases} 0, & s=1 \\ \min((\delta n)^{-\frac{1}{2}}, 1), & s=2 \\ \min(C\sqrt{s}\cdot(\delta n)^{-\frac{1}{s}},\, 1), & s\geq 3 \end{cases}$ $\square$

In the sequel, we define

$$
\Phi(n,s,\delta) = \begin{cases} (\delta n)^{-\frac{1}{2}}, & s=1 \\ (\ln(\delta n+2))\cdot(\delta n)^{-\frac{1}{2}}, & s=2 \\ (\delta n)^{-\frac{1}{s}}, & s\geq 3 \end{cases}. \tag{2.12}
$$

Next we adjust the scale in Lemma 2.5.

Lemma 2.6. *Let $0<\delta\leq 1$. Suppose that E is a Banach space with separable dual E^* and $x_1,\dots,x_n\in E$. Let $g_1,\dots,g_n$ be i.i.d. standard Gaussian random variables. Let T be the set of all 1-Lipschitz functions $f:\mathbb{R}^s\to\mathbb{R}$ with $f(0)=0$. Then*

$$
\begin{aligned}
&\mathbb{E}\sup_{\substack{v_1^*,\dots,v_s^*\in B_{E^*}\\ f\in T}}\left|\frac{1}{n}\sum_{i=1}^n g_i f(v_1^*(x_i),\dots,v_s^*(x_i))\right| \\
\leq& \frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\| + Cs\cdot\left(\frac{1}{n}\sup_{v^*\in B_{E^*}}\sum_{i=1}^n |v^*(x_i)|^{2+2\delta}\right)^{\frac{1}{2+2\delta}}\cdot\Phi(n,s,\delta).
\end{aligned}
$$

Proof. Observe that if $f\in T$ and $a>0$, then the map $y\mapsto \frac{1}{a}f(ay)$ from $\mathbb{R}^s$ to $\mathbb{R}$ is also in T. Thus, without loss of generality, by rescaling $x_1,\dots,x_n$, we may assume that

$$
\sup_{v^*\in B_{E^*}}\sum_{i=1}^n |v^*(x_i)|^{2+2\delta} = n\cdot s^{-(1+\delta)}.
$$

Then in Lemma 2.5,

$$
M\leq\sqrt{2}\left[\sqrt{n}+s^{\frac{1+\delta}{2}}\left(\sup_{v^*\in B_{E^*}}\sum_{i=1}^n |v^*(x_i)|^{2+2\delta}\right)^{\frac{1}{2}}\right] = 2\sqrt{2}\cdot\sqrt{n}.
$$

So by Lemma 2.5,

$$
\begin{aligned}
&\mathbb{E}\sup_{\substack{v_1^*,\dots,v_s^*\in B_{E^*}\\ f\in T}}\left|\frac{1}{n}\sum_{i=1}^n g_i f(v_1^*(x_i),\dots,v_s^*(x_i))\right| \\
\leq& \frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\| + C\sqrt{s}\cdot\Phi(n,s,\delta)
\end{aligned}
$$

$$=\frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^{n}g_ix_i\right\|+Cs\cdot\left(\frac{1}{n}\sup_{v^*\in B_{E^*}}\sum_{i=1}^{n}|v^*(x_i)|^{2+2\delta}\right)^{\frac{1}{2+2\delta}}\cdot\Phi(n,s,\delta),$$

since we assume that $\sup_{v^*\in B_{E^*}}\sum_{i=1}^{n}|v^*(x_i)|^{2+2\delta}=n\cdot s^{-(1+\delta)}$. So the result follows. □

Theorem 2.7. *Let $0<\delta\le 1$. Suppose that μ is a probability measure on a Banach space E with separable dual E^* and $\int_E\|x\|\,d\mu(x)<\infty$. Let $X_1,\ldots,X_n$ be i.i.d. random elements of E sampled according to μ. Then*

$$\begin{aligned}&\mathbb{E}W_{1,s}\left(\mu,\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)\\ \le&\frac{Cs}{n}\mathbb{E}\left\|\sum_{i=1}^{n}g_iX_i\right\|+Cs\cdot\mathbb{E}\left[\left(\frac{1}{n}\sup_{v^*\in B_{E^*}}\sum_{i=1}^{n}|v^*(X_i)|^{2+2\delta}\right)^{\frac{1}{2+2\delta}}\right]\cdot\Phi(n,s,\delta),\end{aligned}$$

where $g_1,\ldots,g_n$ are i.i.d. standard Gaussian random variables that are independent from $X_1,\ldots,X_n$, and $\Phi(n,s,\delta)$ is defined in (2.12).

Proof. By the definition of $W_{1,s}$ in Section 1.3,

$$\begin{aligned}&W_{1,s}\left(\mu,\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)\\ =&\sup_{\substack{v_1^*,\ldots,v_s^*\in B_{E^*}\\ f\text{ is 1-Lipschitz}}}\left|\frac{1}{n}\sum_{i=1}^{n}f(v_1^*(X_i),\ldots,v_s^*(X_i))-\int_E f(v_1^*(x),\ldots,v_s^*(x))\,d\mu(x)\right|,\end{aligned}$$

where the supremum is over all $v_1^*,\ldots,v_s^*\in B_{E^*}$ and all 1-Lipschitz functions $f:\mathbb{R}^s\to\mathbb{R}$ with $f(0)=0$. By symmetrization,

$$\mathbb{E}W_{1,s}\left(\mu,\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)\le C\cdot\mathbb{E}\sup_{\substack{v_1^*,\ldots,v_s^*\in B_{E^*}\\ f\text{ is 1-Lipschitz}}}\left|\frac{1}{n}\sum_{i=1}^{n}g_if(v_1^*(X_i),\ldots,v_s^*(X_i))\right|.$$

So by Lemma 2.6, the result follows. □

Corollary 2.8. *Let $0<\delta\le 1$. Suppose that μ is a probability measure on a separable Hilbert space E with $\int_E\|x\|\,d\mu(x)<\infty$. Let $X_1,\ldots,X_n$ be i.i.d. random elements of E sampled according to μ. Then*

$$\mathbb{E}W_{1,s}\left(\mu,\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)\le Cs\cdot\left(\int_E\|x\|^{2+2\delta}\,d\mu(x)\right)^{\frac{1}{2+2\delta}}\cdot\Phi(n,s,\delta),$$

where $\Phi(n,s,\delta)$ is defined in (2.12).

Proof. In Theorem 2.7,

$$\mathbb{E}\left\|\sum_{i=1}^{n}g_iX_i\right\|\le\left(\mathbb{E}\left\|\sum_{i=1}^{n}g_iX_i\right\|^2\right)^{\frac{1}{2}}=\left(\sum_{i=1}^{n}\mathbb{E}\|X_i\|^2\right)^{\frac{1}{2}}=\sqrt{n}\left(\int_E\|x\|^2\,d\mu(x)\right)^{\frac{1}{2}}.$$

We also have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sup_{v^*\in B_{E^*}}\sum_{i=1}^n|v^*(X_i)|^{2+2\delta}\right)^{\frac{1}{2+2\delta}}\right] &\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\|X_i\|^{2+2\delta}\right)^{\frac{1}{2+2\delta}}\right]\\ &\leq \left(\int_E\|x\|^{2+2\delta}\,d\mu(x)\right)^{\frac{1}{2+2\delta}}.\end{aligned}$$

Since $\frac{1}{\sqrt{n}}\leq\Phi(s,\delta,n)$, by Theorem 2.7, the result follows. ☐

Corollary 2.9. *Suppose that μ is a probability measure on a Banach space E with separable dual E^* and $\int_E\|x\|\,d\mu(x)<\infty$. Let $X_1,\ldots,X_n$ be i.i.d. random elements of E sampled according to μ. Then*

$$\mathbb{E}W_{1,1}\left(\mu,\frac{1}{n}\sum_{i=1}^n\delta_{X_i}\right)\leq\frac{C}{n}\mathbb{E}\left\|\sum_{i=1}^n g_iX_i\right\|+\frac{C\sqrt{\ln n}}{n}\cdot\mathbb{E}\sup_{v^*\in B_{E^*}}\left(\sum_{i=1}^n|v^*(X_i)|^2\right)^{\frac{1}{2}},$$

where $g_1,\ldots,g_n$ are i.i.d. standard Gaussian random variables that are independent from $X_1,\ldots,X_n$.

Proof. By Theorem 2.7 with $s=1$,

$$\begin{aligned}&\mathbb{E}W_{1,1}\left(\mu,\frac{1}{n}\sum_{i=1}^n\delta_{X_i}\right)\\ \leq&\frac{C}{n}\mathbb{E}\left\|\sum_{i=1}^n g_iX_i\right\|+\frac{C}{\sqrt{n}}\cdot\inf_{0<\delta\leq1}\frac{1}{\sqrt{\delta}}\mathbb{E}\left[\left(\frac{1}{n}\sup_{v^*\in B_{E^*}}\sum_{i=1}^n|v^*(X_i)|^{2+2\delta}\right)^{\frac{1}{2+2\delta}}\right]\\ \leq&\frac{C}{n}\mathbb{E}\left\|\sum_{i=1}^n g_iX_i\right\|+\frac{C}{\sqrt{n}}\cdot\inf_{0<\delta\leq1}\frac{1}{\sqrt{\delta}}\mathbb{E}\left[n^{-\frac{1}{2+2\delta}}\left(\sup_{v^*\in B_{E^*}}\sum_{i=1}^n|v^*(X_i)|^2\right)^{\frac{1}{2}}\right].\end{aligned}$$

Take $\delta=1/\lceil\ln n\rceil$. Then $n^{-\frac{1}{2+2\delta}}\leq\frac{C}{\sqrt{n}}$. The result follows. ☐

In the rest of this section, we prove some lower bound results. These results are quite standard.

Proposition 2.10. *Suppose that μ is a probability measure on a Banach space E with separable dual E^* and that $\int_E\|x\|\,d\mu(x)<\infty$ and $\int_E x\,d\mu(x)=0$. Let $X_1,\ldots,X_n$ be i.i.d. random elements of E sampled according to μ. Then*

$$\mathbb{E}W_{1,1}\left(\mu,\frac{1}{n}\sum_{i=1}^n\delta_{X_i}\right)\geq\frac{1}{2n}\mathbb{E}\left\|\sum_{i=1}^n\epsilon_iX_i\right\|,$$

where $\epsilon_1,\ldots,\epsilon_n$ are i.i.d. uniform ±1 random variables that are independent from $X_1,\ldots,X_n$.

Proof. For fixed $x_1,\ldots,x_n\in E$, by considering the 1-Lipschitz function $f(t)=t$, we have

$$W_{1,1}\left(\mu,\frac{1}{n}\sum_{i=1}^n\delta_{x_i}\right)\geq\sup_{v^*\in B_{E^*}}\left|\int_E v^*(x)\,d\mu(x)-\frac{1}{n}\sum_{i=1}^n v^*(x_i)\right|=\left\|\frac{1}{n}\sum_{i=1}^n x_i\right\|.$$

So

$$\mathbb{E}W_{1,1}\left(\mu, \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \geq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\right\|.$$

Let $Y_1, \ldots, Y_n$ be i.i.d. random elements of E sampled according to μ that are independent from $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$. Then

$$\mathbb{E}\left\|\sum_{i=1}^{n}X_i\right\| \geq \frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^{n}(X_i - Y_i)\right\| = \frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^{n}\epsilon_i(X_i - Y_i)\right\| \geq \frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^{n}\epsilon_i X_i\right\|,$$

where the last inequality follows from Jensen's inequality and taking expectation on $Y_1, \ldots, Y_n$. The result follows. □

Corollary 2.11. *Suppose that μ is a probability measure on a separable Hilbert space E with $\int_E \|x\|\,d\mu(x) < \infty$ and $\int_E x\,d\mu(x) = 0$. Let $X_1, \ldots, X_n$ be i.i.d. random elements of E sampled according to μ. Then*

$$\mathbb{E}W_{1,1}\left(\mu, \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \geq \frac{1}{2\sqrt{2n}}\int_E \|x\|\,d\mu(x).$$

Proof. By Proposition 2.10, it suffices to show that

$$\mathbb{E}\left\|\sum_{i=1}^{n}\epsilon_i X_i\right\| \geq \sqrt{\frac{n}{2}}\cdot\mathbb{E}\|X_1\|.$$

If we first take expectation on $\epsilon_1, \ldots, \epsilon_n$, then by the Kahane-Khintchine inequality [16], we have

$$\mathbb{E}_\epsilon\left\|\sum_{i=1}^{n}\epsilon_i X_i\right\| \geq \frac{1}{\sqrt{2}}\left(\mathbb{E}_\epsilon\left\|\sum_{i=1}^{n}\epsilon_i X_i\right\|^2\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2}}\left(\sum_{i=1}^{n}\|X_i\|^2\right)^{\frac{1}{2}} \geq \frac{1}{\sqrt{2n}}\sum_{i=1}^{n}\|X_i\|.$$

So

$$\mathbb{E}\left\|\sum_{i=1}^{n}\epsilon_i X_i\right\| \geq \frac{1}{\sqrt{2n}}\sum_{i=1}^{n}\mathbb{E}\|X_i\| = \sqrt{\frac{n}{2}}\cdot\mathbb{E}\|X_1\|.$$

□

3. Max-sliced 2-Wasserstein distance

The following lemma is known. See e.g., [32].

Lemma 3.1. *Let $r > 0$. Suppose that μ is a probability measure on $\mathbb{R}^d$ supported on $\{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$. Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ sampled according to μ. Let $g_1, \ldots, g_n$ be i.i.d. standard Gaussian random variables that are independent from $X_1, \ldots, X_n$. Then*

$$\mathbb{E}\left\|\sum_{i=1}^{n}X_i X_i^T\right\|_{\mathrm{op}} \leq 2n\|\mathbb{E}X_1 X_1^T\|_{\mathrm{op}} + Cr^2\ln n,$$

and

$$\mathbb{E}\left\|\sum_{i=1}^{n}g_i X_i X_i^T\right\|_{\mathrm{op}} \leq Cr\sqrt{n\ln n}\,\|\mathbb{E}X_1 X_1^T\|_{\mathrm{op}}^{\frac{1}{2}} + Cr^2\ln n.$$

Proof. Fix $x_1, \dots, x_n \in \mathbb{R}^d$ with $\|x_i\|_2 \le r$ for all i. By the noncommutative Khintchine inequality (see [20, 29, 8]), for $p \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\,\mathrm{Tr}\left(\sum_{i=1}^n g_i x_i x_i^T\right)^{2p} &\le (C\sqrt{p})^{2p}\,\mathrm{Tr}\left[\left(\sum_{i=1}^n (x_i x_i^T)^2\right)^p\right] \\
&= (C\sqrt{p})^{2p}\,\mathrm{Tr}\left[\left(\sum_{i=1}^n \|x_i\|_2^2\, x_i x_i^T\right)^p\right] \\
&\le (C\sqrt{p})^{2p} n \left\|\left(\sum_{i=1}^n \|x_i\|_2^2\, x_i x_i^T\right)^p\right\|_{\mathrm{op}} \\
&\le (Cr\sqrt{p})^{2p} n \left\|\sum_{i=1}^n x_i x_i^T\right\|_{\mathrm{op}}^p,
\end{aligned}$$

where the second last inequality follows from the fact that $\sum_{i=1}^n \|x_i\|_2^2\, x_i x_i^T$ has rank at most n. Taking $p = \lceil \ln n \rceil$, we obtain

$$\mathbb{E}\left\|\sum_{i=1}^n g_i x_i x_i^T\right\|_{\mathrm{op}} \le Cr\sqrt{\ln n}\left\|\sum_{i=1}^n x_i x_i^T\right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

Now we randomize $x_1, \dots, x_n$. We get

$$\mathbb{E}\left\|\sum_{i=1}^n g_i X_i X_i^T\right\|_{\mathrm{op}} \le Cr\sqrt{\ln n}\left(\mathbb{E}\left\|\sum_{i=1}^n X_i X_i^T\right\|_{\mathrm{op}}\right)^{\frac{1}{2}}. \tag{3.1}$$

By symmetrization,

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^n X_i X_i^T\right\|_{\mathrm{op}} &\le \|n\mathbb{E}X_1X_1^T\|_{\mathrm{op}} + C\cdot\mathbb{E}\left\|\sum_{i=1}^n g_i X_i X_i^T\right\|_{\mathrm{op}} \\
&\le n\|\mathbb{E}X_1X_1^T\|_{\mathrm{op}} + Cr\sqrt{\ln n}\left(\mathbb{E}\left\|\sum_{i=1}^n X_i X_i^T\right\|_{\mathrm{op}}\right)^{\frac{1}{2}}.
\end{aligned}$$

So

$$\mathbb{E}\left\|\sum_{i=1}^n X_i X_i^T\right\|_{\mathrm{op}} \le 2n\|\mathbb{E}X_1X_1^T\|_{\mathrm{op}} + Cr^2\ln n.$$

This proves the first inequality. Combining this with (3.1), we obtain the second inequality. □

Lemma 3.2. *Suppose that μ_1, μ_2 are symmetric probability measures on $(\mathbb{R}^d, \|\,\|_2)$ supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le r\}$. Consider the map $\eta(x) := xx^T$ from the Hilbert space $(\mathbb{R}^d, \|\,\|_2)$ to the Banach space $(\mathbb{R}^{d\times d}, \|\,\|_{\mathrm{op}})$. Let $\eta_\#\mu_1$ and $\eta_\#\mu_2$ be the pushforward measures of μ_1 and μ_2 by η, respectively. Then*

$$W_{2,1}(\mu_1, \mu_2)^2 \le W_{1,1}(\eta_\#\mu_1, \eta_\#\mu_2).$$

Proof. Define abs : $\mathbb{R} \to \mathbb{R}$ and sq : $\mathbb{R} \to \mathbb{R}$ by $\mathrm{abs}(t) = |t|$ and $\mathrm{sq}(t) = t^2$. Observe that if ν_1 and ν_2 are symmetric probability measures on the interval $[-r, r]$, then

$$\begin{aligned}
W_2(\nu_1, \nu_2)^2 &= W_2(\mathrm{abs}_\#\nu_1, \mathrm{abs}_\#\nu_2)^2 \\
&= \inf_\gamma \int_{[0,r]\times[0,r]} |t-s|^2 \, d\gamma(t,s) \\
&\leq \inf_\gamma \int_{[0,r]\times[0,r]} |t^2-s^2| \, d\gamma(t,s) \\
&= W_1(\mathrm{sq}_\#\mathrm{abs}_\#\nu_1, \mathrm{sq}_\#\mathrm{abs}_\#\nu_2) \\
&= W_1(\mathrm{sq}_\#\nu_1, \mathrm{sq}_\#\nu_2),
\end{aligned}$$

where the infimum is over all coupling γ of the pushforward measures $\mathrm{abs}_\#\nu_1$ and $\mathrm{abs}_\#\nu_2$ on $[0, r]$.

For $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$, let $u_\#\mu_i$ be the pushforward measure of μ_i by the map $\langle \cdot, u\rangle$. Taking $\nu_i = u_\#\mu_i$ in the above, we obtain

$$\begin{aligned}
W_{2,1}(\mu_1, \mu_2)^2 &= \sup_{u\in\mathbb{R}^d,\, \|u\|_2=1} W_2(u_\#\mu_1,\, u_\#\mu_2)^2 \\
&\leq \sup_{u\in\mathbb{R}^d,\, \|u\|_2=1} W_1(\mathrm{sq}_\#u_\#\mu_1,\, \mathrm{sq}_\#u_\#\mu_2)
\end{aligned}$$

Observe that $\mathrm{sq}_\#u_\#\mu_i$ is the pushforward measure of μ_i by the map

$$x \mapsto \langle x, u\rangle^2 = \mathrm{Tr}(uu^T xx^T) = \mathrm{Tr}(uu^T\eta(x)).$$

Moreover, since the trace class norm of uu^T is equal to 1, we can identify uu^T as an element in the unit ball of the dual of the Banach space $(\mathbb{R}^d, \|\,\|_{\mathrm{op}})$. Thus the result follows. $\square$

Below we restate and prove Theorem 1.4.

Theorem 3.3. *Let $r > 0$. Suppose that μ is a symmetric probability measure on $\mathbb{R}^d$ supported on $\{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$. Let $X_1, \ldots, X_n$ be i.i.d. random vectors sampled according to μ. Then*

$$\mathbb{E}\left[W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n(\delta_{X_i} + \delta_{-X_i})\right)^2\right] \leq C\|\Sigma\|_{\mathrm{op}}\left(\frac{r^2\ln n}{n\|\Sigma\|_{\mathrm{op}}} + \sqrt{\frac{r^2\ln n}{n\|\Sigma\|_{\mathrm{op}}}}\right),$$

where $\Sigma = \int_{\mathbb{R}^d} xx^T\, d\mu(x)$.

Proof. Since μ and $\frac{1}{2n}\sum_{i=1}^n(\delta_{X_i} + \delta_{-X_i})$ are symmetric, by Lemma 3.2,

$$\begin{aligned}
(3.2)\quad W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n(\delta_{X_i} + \delta_{-X_i})\right)^2 &\leq W_{1,1}\left(\eta_\#\mu,\, \eta_\#\left[\frac{1}{2n}\sum_{i=1}^n(\delta_{X_i} + \delta_{-X_i})\right]\right) \\
&= W_{1,1}\left(\eta_\#\mu,\, \frac{1}{n}\sum_{i=1}^n \delta_{\eta(X_i)}\right).
\end{aligned}$$

Note that $\eta(X_i) = X_i X_i^T$ are i.i.d. random matrices with distribution $\eta_\# \mu$. Taking $E = (\mathbb{R}^{d\times d}, \|\ \|_{\mathrm{op}})$ in Corollary 2.9, we obtain

$$\begin{aligned}
&\mathbb{E}W_{1,1}\left(\eta_\#\mu, \frac{1}{n}\sum_{i=1}^n \delta_{\eta(X_i)}\right) \\
\leq &\frac{C}{n}\mathbb{E}\left\|\sum_{i=1}^n g_i X_i X_i^T\right\| + \frac{C\sqrt{\ln n}}{n}\cdot \mathbb{E}\sup_{V^*\in B_{E^*}}\left(\sum_{i=1}^n |V^*(X_iX_i^T)|^2\right)^{\frac{1}{2}}
\end{aligned}$$

Since B_{E^*} coincides with the convex hull of $\{\pm vv^T : v\in\mathbb{R}^d, \|v\|_2\leq 1\}$,

$$\begin{aligned}
\sup_{V^*\in B_{E^*}}\left(\sum_{i=1}^n |V^*(X_iX_i^T)|^2\right)^{\frac{1}{2}} &= \sup_{v\in\mathbb{R}^d,\, \|v\|_2\leq 1}\left(\sum_{i=1}^n |\mathrm{Tr}(vv^TX_iX_i^T)|^2\right)^{\frac{1}{2}} \\
&= \sup_{v\in\mathbb{R}^d,\, \|v\|_2\leq 1}\left(\sum_{i=1}^n \langle X_i, v\rangle^4\right)^{\frac{1}{2}} \\
&\leq r \sup_{v\in\mathbb{R}^d,\, \|v\|_2 = 1}\left(\sum_{i=1}^n \langle X_i, v\rangle^2\right)^{\frac{1}{2}} = r\left\|\sum_{i=1}^n X_iX_i^T\right\|_{\mathrm{op}}^{\frac{1}{2}}.
\end{aligned}$$

Therefore,

$$\mathbb{E}W_{1,1}\left(\eta_\#\mu, \frac{1}{n}\sum_{i=1}^n \delta_{\eta(X_i)}\right) \leq \frac{C}{n}\mathbb{E}\left\|\sum_{i=1}^n g_iX_iX_i^T\right\| + \frac{Cr\sqrt{\ln n}}{n}\cdot\mathbb{E}\left\|\sum_{i=1}^n X_iX_i^T\right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

So by Lemma 3.1,

$$\begin{aligned}
&\mathbb{E}W_{1,1}\left(\eta_\#\mu, \frac{1}{n}\sum_{i=1}^n \delta_{\eta(X_i)}\right) \\
\leq &\frac{Cr\sqrt{\ln n}}{\sqrt{n}}\|\mathbb{E}X_1X_1^T\|_{\mathrm{op}}^{\frac{1}{2}} + \frac{Cr^2\ln n}{n} + \frac{Cr\sqrt{\ln n}}{\sqrt{n}}\|\mathbb{E}X_1X_1^T\|_{\mathrm{op}}^{\frac{1}{2}} + \frac{Cr^2\ln n}{n} \\
= &\frac{Cr\sqrt{\ln n}}{\sqrt{n}}\|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} + \frac{Cr^2\ln n}{n}.
\end{aligned}$$

So by (3.2), the result follows. □

Below we restate and prove Proposition 1.5.

Proposition 3.4. *Let Σ be a $d\times d$ positive semidefinite matrix such that $\|\Sigma\|_{\mathrm{op}} \leq \frac{1}{2}\mathrm{Tr}(\Sigma)$. Then there exists a symmetric probability measure μ on $\mathbb{R}^d$ supported on $\{x\in\mathbb{R}^d : \|x\|_2^2 = \mathrm{Tr}(\Sigma)\}$ such that $\int_{\mathbb{R}^d} xx^T\, d\mu(x) = \Sigma$ and for every $n\in\mathbb{N}$,*

$$\mathbb{E}\left[W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n(\delta_{X_i}+\delta_{-X_i})\right)^2\right] \geq \frac{1}{16}\|\Sigma\|_{\mathrm{op}}\left(\frac{\mathrm{Tr}(\Sigma)}{n\|\Sigma\|_{\mathrm{op}}} + \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n\|\Sigma\|_{\mathrm{op}}}}\right), \tag{3.3}$$

where $X_1, \ldots, X_n$ are i.i.d. random vectors sampled according to μ.

Proof. Without loss of generality, we may assume that Σ is a diagonal matrix with diagonal entries $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$. We may also assume that $\mathrm{Tr}(\Sigma) = \lambda_1 + \ldots + \lambda_d = 1$. Let $\{e_1, \ldots, e_d\}$ be the unit vector basis for $\mathbb{R}^d$. Take

$$\mu(\{e_j\}) = \mu(\{-e_j\}) = \frac{1}{2}\lambda_j \quad \text{for } j = 1, \ldots, d.$$

Then μ is symmetric.

We need to show that the left hand side of (3.3) is at least each of the two terms on the right hand side. So the proof has two parts. The first part of the proof is similar to the proofs of Proposition 2.10 and Corollary 2.11. Let $\mathbb{R}^d_+ = \{(v_1, \ldots, v_d) : v_1, \ldots, v_d \geq 0\}$. By considering the 1-Lipschitz function $f(t) = |t|$, we have

$$\begin{aligned}
&W_{1,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{X_i} + \delta_{-X_i})\right) \\
&\geq \sup_{v \in \mathbb{R}^d_+,\, \|v\|_2 = 1} \left| \int_{\mathbb{R}^d} |\langle x, v\rangle|\, d\mu(x) - \frac{1}{2n}\sum_{i=1}^n (|\langle X_i, v\rangle| + |\langle -X_i, v\rangle|)\right| \\
&= \sup_{v \in \mathbb{R}^d_+,\, \|v\|_2 = 1} \left| \sum_{i=1}^d \lambda_i v_i - \frac{1}{n}\sum_{i=1}^n |\langle X_i, v\rangle| \right| \\
&= \sup_{v \in \mathbb{R}^d_+,\, \|v\|_2 = 1} \left| \sum_{i=1}^d \lambda_i v_i - \frac{1}{n}\sum_{i=1}^n \langle \mathrm{abs}(X_i), v\rangle \right|,
\end{aligned}$$

where $\mathrm{abs}(X_i)$ is the vector for which we take absolute value on each entry of X_i. (Since X_i is distributed according to μ, the vector X_i actually has only one nonzero entry.) So

$$W_{1,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{X_i} + \delta_{-X_i})\right) \geq \frac{1}{2}\left\| \mathrm{diag}(\Sigma) - \frac{1}{n}\sum_{i=1}^n \mathrm{abs}(X_i)\right\|_2,$$

where $\mathrm{diag}(\Sigma) = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$.

Since $X_1, \ldots, X_n$ are i.i.d. with distribution μ, the random vectors $\mathrm{abs}(X_1), \ldots, \mathrm{abs}(X_n)$ are i.i.d. with the following distribution

$$\mathrm{abs}_{\#}\mu(\{e_j\}) = \lambda_j \quad \text{for } j = 1, \ldots, d.$$

In particular, $\mathbb{E}[\mathrm{abs}(X_1)] = \sum_{j=1}^d \lambda_j e_j = \mathrm{diag}(\Sigma)$. So

$$\begin{aligned}
\mathbb{E}\left[W_{1,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{X_i} + \delta_{-X_i})\right)^2 \right] &\geq \frac{1}{4}\mathbb{E}\left\| \mathrm{diag}(\Sigma) - \frac{1}{n}\sum_{i=1}^n \mathrm{abs}(X_i)\right\|_2^2 \\
&= \frac{1}{4n}\left[\mathbb{E}\|\mathrm{abs}(X_1)\|_2^2 - \|\mathrm{diag}(\Sigma)\|_2^2\right] \\
&= \frac{1}{4n}(1 - (\lambda_1^2 + \ldots + \lambda_d^2)).
\end{aligned}$$

Since by assumption $\|\Sigma\|_{\mathrm{op}} \le \frac{1}{2}\mathrm{Tr}(\Sigma) = \frac{1}{2}$, we have $\lambda_j \le \frac{1}{2}$ for all j. So $\lambda_1^2 + \ldots + \lambda_d^2 \le \frac{1}{2}(\lambda_1 + \ldots + \lambda_d) = \frac{1}{2}$. So

$$\mathbb{E}\left[W_{1,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{X_i} + \delta_{-X_i})\right)^2\right] \ge \frac{1}{8n}. \tag{3.4}$$

This proves that the left hand side of (3.3) is at least twice the first term on the right hand side. We now move to the second part of the proof. The second term on the right hand side of (3.3) is larger than the first term precisely when $\frac{\mathrm{Tr}(\Sigma)}{n\|\Sigma\|_{\mathrm{op}}} < 1$, or equivalently, $\frac{1}{n} < \lambda_1$. So we may assume this in the rest of the proof.

Consider the pushforward measure $(e_1)_\# \mu$ of μ by the map $\langle \cdot, e_1\rangle$. Note that

$$(e_1)_\#\mu(\{-1\}) = \frac{1}{2}\lambda_1, \quad (e_1)_\#\mu(\{1\}) = \frac{1}{2}\lambda_1, \quad (e_1)_\#\mu(\{0\}) = 1 - \lambda_1.$$

We have

$$\begin{aligned} W_{2,1}\left(\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{X_i} + \delta_{-X_i})\right) &\ge W_2\left((e_1)_\#\mu, \frac{1}{2n}\sum_{i=1}^n (\delta_{\langle X_i, e_1\rangle} + \delta_{-\langle X_i, e_1\rangle})\right) \\ &= W_2\left(\mathrm{abs}_\#(e_1)_\#\mu, \frac{1}{n}\sum_{i=1}^n \delta_{|\langle X_i, e_1\rangle|}\right), \end{aligned} \tag{3.5}$$

where $\mathrm{abs}_\#(e_1)_\#\mu(\{1\}) = \lambda_1$ and $\mathrm{abs}_\#(e_1)_\#\mu(\{0\}) = 1 - \lambda_1$. (See the beginning of the proof of Lemma 3.2.) Moreover, the random variables $|\langle X_i, e_1\rangle|$, for $i = 1, \ldots, d$, are i.i.d. with this distribution. Thus, the probability measure $\frac{1}{n}\sum_{i=1}^n \delta_{|\langle X_i, e_1\rangle|}$ is supported on only two points 0 and 1 with the mass at 1 being $\frac{1}{n}$ times a $\mathrm{binom}(n, \lambda_1)$ random variable, which we denote by Y. So we have

$$W_2\left(\mathrm{abs}_\#(e_1)_\#\mu, \frac{1}{n}\sum_{i=1}^n \delta_{|\langle X_i, e_1\rangle|}\right)^2 = \left|\frac{1}{n}Y - \lambda_1\right|.$$

As explained above, we may assume that $\frac{1}{n} \le \lambda_1$. Also by assumption, $\|\Sigma\|_{\mathrm{op}} \le \frac{1}{2}\mathrm{Tr}(\Sigma)$ so $\lambda_1 \le \frac{1}{2} \le 1 - \frac{1}{n}$. Therefore, $\frac{1}{n} \le \lambda_1 \le 1 - \frac{1}{n}$. With λ_1 in this range, by [5, Theorem 1],

$$\mathbb{E}\left|\frac{1}{n}Y - \lambda_1\right| \ge \frac{1}{\sqrt{2}}\left(\mathbb{E}\left|\frac{1}{n}Y - \lambda_1\right|^2\right)^{\frac{1}{2}} = \frac{\sqrt{n\lambda_1(1-\lambda_1)}}{n\sqrt{2}} = \sqrt{\frac{\lambda_1(1-\lambda_1)}{2n}} \ge \frac{1}{2}\sqrt{\frac{\lambda_1}{n}}.$$

This proves that the left hand side of (3.3) is at least twice the second term on the right hand side. Together with (3.4), this completes the proof. □

Acknowledgement: The author is grateful to Ramon van Handel, Nikita Zhivotovskiy and Sloan Nietert for some useful discussions.

References

[1] P. Abdalla and N. Zhivotovskiy, Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails, Journal of the European Mathematical Society, to appear.

[2] R. Adamczak, A. E. Litvak, A. Pajor and N. Tomczak-Jaegermann, Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles, Journal of the American Mathematical Society, 23:535-561, 2009

[3] R. Adamczak, A. E. Litvak, A. Pajor and N. Tomczak-Jaegermann, Sharp bounds on the rate of convergence of the empirical covariance matrix, Comptes Rendus. Mathématique 349.3-4 (2011): 195-200.

[4] D. Bartl and S. Mendelson, Structure preservation via the Wasserstein distance, arXiv preprint arXiv:2209.07058 (2022).

[5] D. Berend and A. Kontorovich, A sharp estimate of the binomial mean absolute deviation with applications, Statistics & Probability Letters 83.4 (2013): 1254-1259.

[6] S. Bobkov and M. Ledoux, One-dimensional empirical measures, order statistics, and Kantorovich transport distances, Vol. 261. No. 1259. American Mathematical Society, 2019.

[7] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, Sliced and Radon Wasserstein barycenters of measures, Journal of Mathematical Imaging and Vision, 1(51):22-45, 2015.

[8] A. Buchholz, Operator Khintchine inequality in non-commutative probability, Math. Ann., 319(1):1-16, 2001.

[9] M. Carrière, M. Cuturi, and S. Oudot, Sliced Wasserstein kernel for persistence diagrams, In International Conference on Machine Learning (ICML), 2017.

[10] I. Deshpande, Z. Zhang, and A. G. Schwing, Generative modeling using the sliced Wasserstein distance, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[11] I. Deshpande, Y-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing Max-sliced Wasserstein distance and its use for GANs, In CVPR, pages 10648-10656, 2019.

[12] N. Fournier and A. Guillin, On the rate of convergence in Wasserstein distance of the empirical measure, Probability theory and related fields 162.3-4 (2015): 707-738.

[13] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced Wasserstein autoencoders. In International Conference on Learning Representations, 2018.

[14] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau and G. Rohde, Generalized sliced wasserstein distances, Advances in neural information processing systems, 32 (2019).

[15] V. Koltchinskii and K. Lounici, Concentration inequalities and moment bounds for sample covariance operators. Bernoulli, 23:110-133, 2014.

[16] R. Latała and K. Oleszkiewicz, On the best constant in the Khinchin-Kahane inequality, Studia Mathematica 109.1 (1994): 101-104.

[17] M. Ledoux and M. Talagrand. Probability in Banach Spaces: isoperimetry and processes, Springer Science & Business Media, 2013.

[18] T. Lin, C. Fan, N. Ho, M. Cuturi and M. Jordan, Projection robust Wasserstein distance and Riemannian optimization. Advances in neural information processing systems, 33 (2020): 9383-9397.

[19] T. Lin, Z. Zheng, E. Chen, M. Cuturi, M. I. Jordan, On projection robust optimal transport: Sample complexity and model misspecification, In International Conference on Artificial Intelligence and Statistics PMLR 2021.

[20] F. Lust-Piquard, Inégalités de Khintchine dans C_p $(1 < p < \infty)$, C. R. Acad. Sci. Paris Sér. I Math., 303(7):289-292, 1986.

[21] S. Mendelson and G. Paouris, On the singular values of random matrices, Journal of the European Mathematical Society, 16:823-834, 2014.

[22] K. Nadjahi, A. Durmus, U. Simsekli and R. Badeau, Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. Advances in Neural Information Processing Systems, 32 (2019).

[23] K. Nguyen and N. Ho, Amortized projection optimization for sliced Wasserstein generative models, Advances in Neural Information Processing Systems 35 (2022): 36985-36998.

[24] K. Nguyen and N. Ho, Energy-based sliced Wasserstein distance, Advances in Neural Information Processing Systems 36 (2024).

[25] S. Nietert, Z. Goldfeld, R. Sadhu and K. Kato, Statistical, robustness, and computational guarantees for sliced Wasserstein distances, Advances in Neural Information Processing Systems, 35, 28179-28193 (2022).

[26] J. Niles-Weed and P. Rigollet, Estimation of Wasserstein distances in the spiked transport model, Bernoulli 28.4 (2022): 2663-2688.

[27] J. Olea, C. Rush, A. Velez, and J. Wiesel, On the generalization error of norm penalty linear regression models, arXiv preprint arXiv:2211.07608, 2022.

[28] F.-P. Paty and M. Cuturi, Subspace robust Wasserstein distances, In ICML, pages 5072-5081, 2019.

[29] G. Pisier, Introduction to operator space theory, volume 294 of London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, 2003.

[30] M. Raab and A. Steger, "Balls into bins"-A simple and tight analysis, International Workshop on Randomization and Approximation Techniques in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[31] J. Rabin, G. Peyré, J. Delon, and M. Bernot, Wasserstein barycenter and its application to texture mixing, In International Conference on Scale Space and Variational Methods in Computer Vision, pages 435-446. Springer, 2011.

[32] M. Rudelson, Random vectors in the isotropic position, Journal of Functional Analysis 164.1 (1999): 60-72.

[33] N. Srivastava and R. Vershynin, Covariance estimation for distributions with $2 + \epsilon$ moments, Annals of Probability 41 (2013), 3081-3111.

[34] M. Talagrand, The generic chaining: upper and lower bounds of stochastic processes. Springer Science & Business Media, 2005.

[35] K. Tikhomirov, Sample covariance matrices of heavy-tailed distributions, International Mathematics Research Notices 2018.20 (2018): 6254-6289.

[36] J. A. Tropp, An introduction to matrix concentration inequalities, Foundations and Trends in Machine Learning 8.1-2 (2015): 1-230.

[37] R. van Handel, Probability in high dimension, Lecture Notes (Princeton University) (2014).

[38] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices. Compressed sensing, 210–268, Cambridge Univ. Press, Cambridge, 2012.

[39] R. Vershynin, How close is the sample covariance matrix to the actual covariance matrix?, Journal of Theoretical Probability 25.3 (2012): 655-686.

[40] R. Vershynin, High-dimensional probability: An introduction with applications in data science, Vol. 47. Cambridge university press, 2018.

[41] M. J. Wainwright, High-dimensional statistics: A non-asymptotic viewpoint. Vol. 48. Cambridge university press, 2019.

[42] J. Wang, R. Gao, and Y. Xie, Two-sample test using projected Wasserstein distance, IEEE International Symposium on Information Theory (ISIT), 2021.
Updated version: arxiv: 2010.11970

[43] J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. P. Paudel, and L. V. Gool, Sliced Wasserstein generative models, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3713-3722, 2019.

[44] N. Zhivotovskiy, Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle, Electronic Journal of Probability 29 (2024): 1-28.

Department of Mathematics, Michigan State University, East Lansing, MI 48824
Email address: boedihar@msu.edu