Estimation of Wasserstein distances in the Spiked Transport Model

JONATHAN NILES-WEED∗ and PHILIPPE RIGOLLET†

Abstract. We propose a new statistical model, the spiked transport model, which formalizes the assumption that two probability distributions differ only on a low-dimensional subspace. We study the minimax rate of estimation for the Wasserstein distance under this model and show that this low-dimensional structure can be exploited to avoid the curse of dimensionality. As a byproduct of our minimax analysis, we establish a lower bound showing that, in the absence of such structure, the plug-in estimator is nearly rate-optimal for estimating the Wasserstein distance in high dimension. We also give evidence for a statistical-computational gap and conjecture that any computationally efficient estimator is bound to suffer from the curse of dimensionality.

Keywords: Wasserstein distance; optimal transport; high-dimensional statistics.

1. Introduction

Optimal transport is an increasingly useful toolbox in various data-driven disciplines such as machine learning [2, 3, 4, 17, 21, 35, 39, 45, 47, 70, 73, 79, 81, 86], computer graphics [34, 59, 84, 85], statistics [1, 7, 18, 19, 22, 46, 52, 55, 60, 71, 76, 77, 78, 82, 88, 96, 99] and the sciences [23, 63, 80, 98]. A core primitive of this toolbox is the computation of Wasserstein distances between probability measures, and a natural statistical question is the estimation of Wasserstein distances from data.

∗Part of this research was conducted while at the Institute for Advanced Study. JNW gratefully acknowledges its support.
†Supported by NSF awards IIS-BIGDATA-1838071, DMS-1712596 and CCF-TRIPODS-1740751; ONR grant N00014-17- 1-2147.
A key object in this endeavor is the empirical measure $\mu_n$ associated to $\mu$. It is the empirical measure defined by

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \quad X_i \sim \mu \text{ i.i.d.}$$

Owing to their flexibility, Wasserstein distances are notoriously hard to estimate in high dimension since in such cases, the empirical distribution is a poor proxy for the underlying distribution of interest. Indeed, for $d$-dimensional distributions, Wasserstein distances between empirical measures generally converge at the slow rate $n^{-1/d}$ \cite{11, 24, 30, 37, 95} and thus suffer from the curse of dimensionality. For example, the following behavior is typical.

**Proposition 1.** Let $\mu$ be a probability measure on $[-1, 1]^d$. If $\mu_n$ is an empirical measure comprising $n$ i.i.d. samples from $\mu$, then for any $p \in [1, \infty)$,

$$E_W^p(\mu_n, \mu) \leq \sqrt{d} \cdot r_{p,d}(n) := c_p \begin{cases} 
    n^{-1/2p} & \text{if } d < 2p \\
    n^{-1/2p} (\log n)^{1/p} & \text{if } d = 2p \\
    n^{-1/d} & \text{if } d > 2p 
\end{cases}$$

In many settings, this bound is known to be tight up to logarithmic factors. In fact, the rate in Proposition 1 has been shown to be essentially minimax optimal for the problem of estimating $\mu$ in Wasserstein distance \cite{83}, using any estimator, not necessarily the empirical distribution $\mu_n$.

Since $W_p$ satisfies the triangle inequality, Proposition 1 readily yields that $W_p(\mu^{(1)}, \mu^{(2)})$ for two probability measures $\mu^{(1)}, \mu^{(2)}$ on $[-1, 1]^d$ can be estimated at the rate $n^{-1/d}$ by the plug-in estimator $W_p(\mu^{(1)}_n, \mu^{(2)}_n)$ when $d > 2p$, and a lower bound of the same order can be shown for the plug-in estimator when $\mu^{(1)}$ and $\mu^{(2)}$ are, for example, the uniform measure on $[-1, 1]^d$. However, while the above results give a strong indication that the Wasserstein distance $W_p(\mu^{(1)}, \mu^{(2)})$ itself is also hard to estimate in high dimension, they do not preclude the existence of estimators that are better than $W_p(\mu^{(1)}_n, \mu^{(2)}_n)$. Indeed, until recently, the best known lower bound for the problem of estimating the distance itself was of order $n^{-3/2d}$ \cite{29}. A concurrent and independent result \cite{62} closes this gap for $p = 1$ and indicates that estimating the $W_1$ itself is essentially as hard as estimating the measure itself. Our results show that, in fact, estimating the distance $W_p(\mu^{(1)}, \mu^{(2)})$ is essentially as hard as estimating a measure $\mu$ in $W_p$-distance, for any $p \geq 1$. As a result, any estimator of the distance itself must suffer from the curse of dimensionality.

One goal of statistical optimal transport is to develop new models and methods that overcome this curse of dimensionality by leveraging plausible structure in the problem. Early contributions in this direction include assuming smoothness \cite{46, 62, 96} or sparsity \cite{36}. In this work, we propose a new model, called the spiked transport model, to formalize the assumption that two distributions differ only on a low-dimensional subspace of $\mathbb{R}^d$. Such an assumption forms the basis of several popular alternatives to Wasserstein distances such as the Sliced Wasserstein distance \cite{75} or random one-dimensional
projections [74]. More recently, several numerical methods that exploit a form of low-
dimensional structure were proposed together with illustrations of their good numerical
performance [25, 53, 72].

To exploit the low-dimensional structure of the spiked transport model, we consider a
standard method in statistics often called “projection pursuit” [38, 56, 57]. This general
method aims to alleviate the challenges of high-dimensional data analysis by considering
low-dimensional projections of the dataset that reveal “interesting” features of the data.
We show that a suitable instantiation of this method to the present problem, which we
call Wasserstein Projection Pursuit (WPP), leads to near optimal rates of estimation
of the Wasserstein distance in the spiked transport model and alleviates the curse of
dimensionality from which the plug-in estimator suffers.

While our results establish a clear statistical picture, it is unclear how to implement
WPP efficiently. An efficient relaxation of this estimator was recently proposed by Paty
and Cuturi [72], and a natural question is to analyze its performance in the spiked
transport model. Instead of pursuing this direction we bring strong evidence that, in fact,
no computationally efficient estimator is likely to be able to take advantage of the low-
dimensional structure inherent to the spiked transport model. Our computational lower
bounds come from the well-established statistical query framework [51]. In particular,
they indicate a fundamental tradeoff between statistical and computational efficiency [5,
6, 13, 16, 65]: computationally efficient methods to estimate Wasserstein distances are
bound to suffer the curse of dimensionality.

The rest of this paper is organized as follows. We introduce our model and main results
in Sections 2 and 3. In Section 4, we define a low-dimensional version of the Wasserstein
distance and establish its connection to the spiked transport model. Section 5 proves the
equivalence between transportation inequalities and subgaussian concentration properties
of the Wasserstein distance. In Section 6 we propose and analyze an estimator for the
Wasserstein distance under the spiked transport model. We establish a minimax lower
bound in Section 7. Finally, we prove a statistical query bound on the performance
of efficient estimators in Section 8. Supplementary proofs and lemmas appear in the
appendix and in Supplement A.

Notation.

We denote by ∥·∥ the Euclidean norm on \( \mathbb{R}^d \). The symbols \( ∥·∥_{op} \) and \( ∥·∥_F \) denote
the operator norm and Frobenius norm, respectively. If \( X \) is a random variable on \( \mathbb{R} \),
we let \( ∥X∥_p := (\mathbb{E}|X|^p)^{1/p} \). Throughout, we use \( c \) and \( C \) to denote positive constants
whose value may change from line to line, and we use subscripts to indicate when these
constants depend on other parameters. We write \( a \lesssim b \) if \( a \leq Cb \) holds for a universal
positive constant \( C \).

2. Model and methods

In this section, we describe the spiked transport model and Wasserstein projection pur-
suit.
2.1. Wasserstein distances

Given two probability measures $\mu$ and $\nu$ on $\mathbb{R}^d$, let $\Gamma_{\mu,\nu}$ denote the set of couplings between $\mu$ and $\nu$ so that $\gamma \in \Gamma_{\mu,\nu}$ iff $\gamma(U \times \mathbb{R}^d) = \mu(U)$ and $\gamma(\mathbb{R}^d \times V) = \nu(V)$.

For any $p \geq 1$, the $p$-Wasserstein distance $W_p$ between $\mu$ and $\nu$ is defined as

$$W_p(\mu, \nu) := \inf_{\gamma \in \Gamma_{\mu,\nu}} \left( \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, d\gamma(x, y) \right)^{1/p}.$$  \hspace{1cm} (1)

The definition of Wasserstein distances may be extended to measures defined on general metric spaces but such extensions are beyond the scope of this paper. We refer the reader to Villani [93] for a comprehensive treatment.

2.2. Spiked transport model

We introduce a new model that induces a low-dimensional structure on the optimal transport between two measures $\mu^{(1)}$ and $\mu^{(2)}$ over $\mathbb{R}^d$. To that end, fix a subspace $\mathcal{U} \subseteq \mathbb{R}^d$ of dimension $k \ll d$ and let $X^{(1)}, X^{(2)} \in \mathcal{U}$ be two random variables with arbitrary distributions. Next, let $Z$ be a third random variable, independent of $(X^{(1)}, X^{(2)})$ and such that $Z$ is supported on the orthogonal complement $\mathcal{U}^\perp$ of $\mathcal{U}$. Finally, let

$$\mu^{(1)} := \text{Law}(X^{(1)} + Z)$$
$$\mu^{(2)} := \text{Law}(X^{(2)} + Z).$$  \hspace{1cm} (2)

Though $\mu^{(1)}$ and $\mu^{(2)}$ are high-dimensional distributions, they differ only on the low-dimensional subspace $\mathcal{U}$. Borrowing terminology from principal component analysis [50], we say that the pair $(\mu^{(1)}, \mu^{(2)})$ satisfies the spiked transport model and we call $\mathcal{U}$ the spike.

This model is largely motivated by applications of optimal transport to high-dimensional problems. Indeed, in large-scale problems, the curse of dimensionality of optimal transport is intolerable and the need to develop practically relevant and tractable structured models is acute. For example, this situation arises in single-cell genomic data analysis where optimal transport may only be deployed after a projection to a low-dimensional structure [80], often using ad-hoc techniques such as low-dimensional embeddings or simply gene selection [58, 94]. Mathematical formulations of this problem [31, 69] analyze a model in which two groups of cells differ on a low-dimensional subspace and have independent, identical distributions orthogonal to this subspace—an instance of the model proposed in (2).

More generally, the domain adaptation problem refers to the setting common in a number of fields where estimators obtained from data generated in a source domain need to be adapted to apply in a different target domain [54]. One approach to domain adaptation based on optimal transport involves learning a transformation between the two domains [20]; in designing algorithms for this problem, it is often assumed that the relevant transformation acts on a low-dimensional subspace [33, 40]. The model proposed in (2) formalizes this assumption.
2.3. Concentration assumptions

In order to establish sharp statistical results for estimation of the Wasserstein distance, it is necessary to adopt smoothness and decay assumptions on the measures in question [see, e.g., 9, 37]. We focus on a family of such conditions known as transport inequalities, the study of which is a central object in the theory of concentration of measure [61].

A probability measure $\mu$ on $\mathbb{R}^d$ is said to satisfy the $T_p(\sigma^2)$ transport inequality if

$$W_p(\nu, \mu) \leq \sqrt{2\sigma^2 D(\nu \parallel \mu)}$$

for all probability measures $\nu$ on $\mathbb{R}^d$, where $D(\nu \parallel \mu)$ is the Kullback–Leibler divergence, defined by

$$D(\nu \parallel \mu) := \begin{cases} \int \log \frac{d\nu}{d\mu} \, d\nu & \text{if } \nu \ll \mu, \\ \infty & \text{otherwise}. \end{cases} \quad (3)$$

These inequalities interpolate between several well known assumptions in high-dimensional probability. For example inequality $T_1(\sigma^2)$ is essentially equivalent to the assertion that $\mu$ is subgaussian, and $T_2(\sigma^2)$ is implied by (and often equivalent to) a stronger log-Sobolev inequality [44]. In particular, we note that for $p > 1$, the transport inequality $T_p(\sigma^2)$ implies both that $\mu$ has light tails and that it has connected support. In this sense, $T_p(\sigma^2)$ is both a moment condition and a smoothness condition on the underlying measure.

Our main results on the estimation of $W_p(\mu^{(1)}, \mu^{(2)})$ are established under the assumption that both $\mu^{(1)}$ and $\mu^{(2)}$ satisfy a $T_p(\sigma^2)$ transport inequality. We show in Section 5 that this is precisely equivalent to requiring that the random variable $W_p(\mu_n, \mu)$ is subgaussian.

After the first version of our paper appeared, an analysis of our estimator under moment assumptions alone was performed by [64]. However, their assumptions are not sufficient to recover the rates we obtain in Theorem 1, even for compactly supported measures. We adopt the assumption that $\mu^{(1)}$ and $\mu^{(2)}$ satisfy a $T_p(\sigma^2)$ inequality because it encapsulates precisely the moment and smoothness conditions necessary to obtain optimal rates.

2.4. Wasserstein projection pursuit

To take advantage of the spiked transport model, we employ a natural estimation method that we call Wasserstein Projection Pursuit (WPP).

Let $\mu$ and $\nu$ be two probability distributions on $\mathbb{R}^d$. Given a $k \times d$ matrix $U$ with orthonormal rows, let $\mu_U$ (resp. $\nu_U$) denote the distribution of $UY$ where $Y \sim \mu$ (resp. $Y \sim \nu$). We define

$$\tilde{W}_{p,k}(\mu, \nu) := \max_{U \in \mathcal{V}_k(\mathbb{R}^d)} W_p(\mu_U, \nu_U),$$

where the maximization is taken over the Stiefel manifold $\mathcal{V}_k(\mathbb{R}^d)$ of $k \times d$ matrices with orthonormal rows.

Given empirical measures $\mu_n^{(1)}$ and $\mu_n^{(2)}$ associated to $\mu^{(1)}$ and $\mu^{(2)}$ that satisfy the spiked transport model, we propose the following WPP estimator of $W_p(\mu^{(1)}, \mu^{(2)})$:

$$\hat{W}_{p,k} = \tilde{W}_{p,k}(\mu_n^{(1)}, \mu_n^{(2)}).$$
In the next section, we show that this estimator is near-minimax-optimal.

3. Main results

As a theoretical justification for Wasserstein projection pursuit, we prove that our procedure successfully avoids the curse of dimensionality under the spiked transport model. Our results primarily focus on the estimation of the Wasserstein distance itself but we also obtain as a byproduct of Wasserstein projection pursuit an estimator for the spike \( \mathcal{U} \) using standard perturbation results.

3.1. Estimation of the Wasserstein distance

The following theorem shows that Wasserstein projection pursuit takes advantage of the low-dimensional structure of the spiked transport model when estimating the Wasserstein distance.

**Theorem 1.** Let \((\mu^{(1)}, \mu^{(2)})\) satisfy the spiked transport model (2). For any \( p \in [1, 2] \), if \( \mu^{(1)} \) and \( \mu^{(2)} \) satisfy the \( T_p(\sigma^2) \) transport inequality, then the WPP estimator \( \hat{W}_{p,k} \) satisfies

\[
E |\hat{W}_{p,k} - W_p(\mu^{(1)}, \mu^{(2)})| \leq c_k \cdot \sigma \left( r_{p,k}(n) + \sqrt{\frac{d \log n}{n}} \right),
\]

uniformly over all \((\mu^{(1)}, \mu^{(2)})\) satisfying the spiked transport model (2) and \( T_p(\sigma^2) \).

The constant \( c_k \) may be taken to be \( C\sqrt{k} \) for a universal constant \( C \). Strikingly, the rate \( r_{p,d}(n) \) achieved by the naïve plug-in estimator (see Proposition 1) has been replaced by \( r_{p,k}(n) \)—in other words, this estimator enjoys the rate typical for \( k \)-dimensional rather than \( d \)-dimensional measures. The only dependence on the ambient dimension is in the second term, which is of lower order than the first whenever \( p > 1 \) or \( k > 2 \). A more general version of this theorem appears in Section 6.

3.2. Estimation of the spike

We show that if the distance between \( \mu^{(1)} \) and \( \mu^{(2)} \) is large enough, Wasserstein projection pursuit recovers the subspace \( \mathcal{U} \). For simplicity, we state here the result when \( k = 1 \) and defer the full version to Section 6.

**Theorem 2.** Let \((\mu^{(1)}, \mu^{(2)})\) satisfy the spiked transport model with \( k = 1 \) and let \( \mathcal{U} \) be spanned by the unit vector \( u \in \mathbb{R}^d \). Fix \( p \in [1, 2] \) and assume that \( \mu^{(1)} \) and \( \mu^{(2)} \) satisfy the \( T_p(\sigma^2) \) transport inequality. Then the estimator

\[
\hat{u} := \arg\max_{v \in \mathbb{R}^d, \|v\| = 1} W_p(\mu^{(1)}_v, \mu^{(2)}_v)
\]
satisfies
\[ \mathbb{E} \sin^2 \left( \angle(\hat{u}, u) \right) \lesssim \frac{\sigma \cdot \left( n^{-1/2p} + \sqrt{\frac{d \log n}{n}} \right)}{W_p(\mu^{(1)}, \mu^{(2)})}. \]

### 3.3. Adaptive estimators

Implementing the estimator \( \hat{W}_{p,k} \) requires knowledge of the dimension \( k \) appearing in the definition of the spiked transport model. Since this quantity may not be known in practice, we also show that a variant of Wasserstein projection pursuit achieves the same rate without \textit{a priori} knowledge of \( k \).

Let
\[ \tilde{W}_p := \tilde{W}_{p,k}(\mu_n^{(1)}, \mu_n^{(2)}), \]
where
\[ \hat{k} := \arg\max_{k \in [d]} W_{p,k}(\mu_n^{(1)}, \mu_n^{(2)}) - p_k, \quad p_k := c' \sqrt{k} \cdot \sigma \left( \hat{r}_{p,k}(n) + \sqrt{\frac{d \log n}{n}} + \frac{d}{n} \right) \tag{4} \]
for a positive constant \( c' \).

The following result shows that \( \tilde{W}_p \) achieves comparable performance to \( \hat{W}_{p,k} \), without requiring knowledge of \( k \).

**Theorem 3.** Let \((\mu^{(1)}, \mu^{(2)})\) satisfy the spiked transport model (2) for some \( k \). For any \( p \in [1, 2] \), if \( \mu^{(1)} \) and \( \mu^{(2)} \) satisfy the \( T_p(\hat{\sigma}^2) \) transport inequality and \( c' \) is a sufficiently large universal constant, then \( \tilde{W}_p \) satisfies
\[ \mathbb{E} |\tilde{W}_p - W_p(\mu^{(1)}, \mu^{(2)})| \leq c'_k \cdot \sigma \left( \hat{r}_{p,k}(n) + \sqrt{\frac{d \log n}{n}} \right). \]

As in Theorem 1, we can take \( c'_k \leq C' \sqrt{k} \), where \( C' \) is a universal constant depending on \( c' \). The proof of Theorem 3 uses standard ideas from model selection, and is deferred to Supplement A.

### 3.4. Lower bounds

To show that Theorem 1 has the right dependence on \( n \) and \( d \), we exhibit two lower bounds, which imply that neither term in Theorem 1 can be avoided.

To show the optimality of the first term, we define
\[ r'_{p,d}(n) := c_{p,d} \begin{cases} n^{-1/2p} & \text{if } d < 2p \\ n^{-1/2p} & \text{if } d = 2p \\ (n \log n)^{-1/d} & \text{if } d > 2p \end{cases} \]
Theorem 4. Fix $p \geq 1$. For any estimator $\hat{W}$, there exists a pair of measures $\mu$ and $\nu$ supported on $[0, 1]^d$ such that

$$\mathbb{E}|\hat{W} - W_p(\mu^{(1)}, \mu^{(2)})| \geq r'_{p,d}(n).$$

This lower bound readily implies that the plug-in estimator for the Wasserstein distance is optimal up to logarithmic factors. By embedding $[0, 1]^k$ into $[0, 1]^d$, this result likewise implies that the term $r_{p,k}(n)$ in Theorem 1 is essentially optimal. A proof appears in Section 7.

Independently, Liang [62] recently obtained a similar result in the case $p = 1$. More specifically, he proved that when $d \geq 2$, for any estimator $\hat{W}$, there exist probability measures $\mu^{(1)}$ and $\mu^{(2)}$ such that the following lower bound holds:

$$\mathbb{E}|\hat{W} - W_1(\mu^{(1)}, \mu^{(2)})| \gtrsim \frac{\log \log n}{\log n} n^{-1/d}.$$

In particular, while our lower bound is slightly stronger and holds for all $p \geq 1$, both our result and that of Liang [62] fail to match the naive upper bound of order $n^{1/d}$ by logarithmic factors when $d$ is large. The presence of a logarithmic factor in our lower bound comes from a reduction to estimating the total variation distance. In that case, as in several other instances of functional estimation problems, the presence of this factor is, in fact, optimal, and has been dubbed sample size enlargement [48]. Closing this gap in the context of estimation of the Wasserstein distance is an interesting and fundamental question.

The only appearance of the ambient dimension $d$ is in the second term of Theorem 1. The following theorem shows that this dependence cannot be eliminated, even when $k = 1$.

Theorem 5. Let $p \in [1, 2]$ and $\sigma > 0$, and assume $k = 1$. For all estimators $\hat{W}$, there exists a pair of measures $\mu^{(1)}$ and $\mu^{(2)}$ satisfying the spiked transport model and $T_p(\sigma^2)$ such that

$$\mathbb{E}|\hat{W} - W_p(\mu^{(1)}, \mu^{(2)})| \gtrsim \frac{\sqrt{d}}{\sigma}.$$

The proof of Theorem 5 is deferred to the appendix. For problems where $d$ is large, dependence on $d$ may be a crippling limitation. In that case, we conjecture that assuming a sparse spike, in the same spirit as sparse PCA, can mitigate this effect and bring interpretability to the estimated spike.

3.5. A computational-statistical gap

The WPP estimator achieving the rate in Theorem 1 is computationally expensive to implement, which raises the question of whether an efficient estimator exists achieving the same rate. We give evidence in the form of a statistical query lower bound that
no such estimator exists. The statistical query model considers algorithms with access to an oracle $V_{\text{STAT}}(t)$, where $t > 0$ is a parameter which plays the role of sample size. We show that any such algorithm for estimating the Wasserstein distance needs an exponential number of queries to an oracle with exponential sample size parameter, even under the spiked transport model. By contrast, Theorem 1 implies that a non-efficient estimator needs a number of samples only polynomial in the dimension.

**Theorem 6.** Let $p \in [1, 2]$, and consider probability measures $\mu^{(1)}$ and $\mu^{(2)}$ satisfying the spiked transport model. There exists a positive constant $c$ such that any statistical query algorithm which estimates $W_p(\mu^{(1)}, \mu^{(2)})$ to accuracy $1/\text{poly}(d)$ with probability at least $2/3$ requires at least $2^{cd}$ queries to $V_{\text{STAT}}(2^{cd})$.

### 4. Low-dimensional Wasserstein distances

Motivated by projection pursuit, we define the following version of the Wasserstein distance which measures the discrepancy between low-dimensional projections of the measures.

**Definition 1.** For $k \in [d]$, the $k$-dimensional Wasserstein distance between $\mu^{(1)}$ and $\mu^{(2)}$ is

$$\tilde{W}_{p,k}(\mu^{(1)}, \mu^{(2)}) := \sup_{U \in \mathcal{V}_k(\mathbb{R}^d)} W_p(\mu^{(1)}_U, \mu^{(2)}_U).$$

This definition has been proposed independently and concurrently by a number of other recent works [25, 53, 72].

We will use throughout the following basic fact about the $k$-dimensional Wasserstein distance [72, Proposition 1].

**Proposition 2.** $\tilde{W}_{p,k}$ is a metric on the set of probability measures over $\mathbb{R}^d$ with finite $p$th moment.

The definition of $\tilde{W}_{p,k}$ is chosen so that, under the spiked transport model (2), the $k$-dimensional Wasserstein distance agrees with the standard Wasserstein distance.

**Proposition 3.** Under the spiked transport model (2),

$$\tilde{W}_{p,k}(\mu^{(1)}, \mu^{(2)}) = W_p(\mu^{(1)}_U, \mu^{(2)}_U).$$

Proposition 3 follows from the following statement, which pertains to distributions that are allowed to have a different component on the space orthogonal to the spike $U$.

Suppose $\nu^{(1)}$ and $\nu^{(2)}$ satisfy

$$\nu^{(1)} = \text{Law}(X^{(1)} + Z^{(1)})$$

$$\nu^{(2)} = \text{Law}(X^{(2)} + Z^{(2)}),$$

(5)
where as before $X^{(1)}$ and $X^{(2)}$ are supported on a subspace $\mathcal{U}$ and $Z^{(1)}$ and $Z^{(2)}$ are supported on its orthogonal complement $\mathcal{U}^\perp$, and where we assume that $X^{(i)}$ and $Z^{(i)}$ are independent for $i \in \{1, 2\}$. Note that unlike in the spiked transport model, the components $Z^{(1)}$ and $Z^{(2)}$ on the orthogonal complement of $\mathcal{U}$ need not be identical.

The following result shows that under this relaxed model, the $k$-dimensional Wasserstein distance between $\nu^{(1)}$ and $\nu^{(2)}$ still captures the true Wasserstein distance between the distributions as long as the distributions of $Z^{(1)}$ and $Z^{(2)}$ are sufficiently close.

**Proposition 4.** Under the relaxed spiked transport model (5),

$$|\tilde{W}_{p,k}(\nu^{(1)}, \nu^{(2)}) - W_p(\nu^{(1)}, \nu^{(2)})| \leq W_p(\text{Law}(Z^{(1)}), \text{Law}(Z^{(2)})).$$

5. Concentration

A key step to establish the upper bound of Section 6 consists in establishing good concentration properties for the Wasserstein distance between a measure and its empirical counterpart. The main assumption we adopt is that the measures in question satisfy a transport inequality. Since the pioneering work of Marton [66, 67] and Talagrand [87], transport inequalities have played a central role in the analysis of the concentration properties of high-dimensional measures.

We require two definitions.

**Definition 2.** Given a Polish space $\mathcal{X}$ equipped with a metric $\rho$, denote by $\mathcal{P}(\mathcal{X})$ the space of all Borel probability measures $\mathcal{X}$. Let $\mathcal{P}_p(\mathcal{X}) := \{\mu \in \mathcal{P}(\mathcal{X}) : \int \rho(x, \cdot)^p \, d\mu(x) < \infty\}$.

A measure $\mu \in \mathcal{P}_p(\mathcal{X})$ satisfies the $T_p(\sigma^2)$ inequality for some $\sigma > 0$ if

$$W_p(\nu, \mu) \leq \sqrt{2\sigma^2 D(\nu||\mu)} \quad \forall \nu \in \mathcal{P}(\mathcal{X}),$$

where $W_p$ is the Wasserstein-$p$ distance on $(\mathcal{X}, \rho)$ and $D$ is the Kullback-Leibler divergence (3).

**Definition 3.** A random variable $X$ on $\mathbb{R}$ is $\sigma^2$-subgaussian if $\mathbb{E}e^{\lambda (X - \mathbb{E}X)} \leq e^{\lambda^2 \sigma^2/2}$ for all $\lambda \in \mathbb{R}$.

In this section, we present a surprisingly simple equivalence between transport inequalities and subgaussian concentration for the Wasserstein distance. The essence of this result is present in the works of Gozlan [42] and Gozlan and Léonard [43, 44], and similar bounds have been obtained by Bolley et al. [12]. Nevertheless, we could not find this simple fact stated in a form suitable for our purposes in the literature. For any measure $\mu$, recall that the random measure $\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$, where $X_i \sim \mu$ i.i.d., denotes its associated empirical measure.
Theorem 7. Let $p \in [1, 2]$. A measure $\mu \in \mathcal{P}_p(\mathcal{X})$ satisfies $T_p(\sigma^2)$ if and only if the random variable $W_p(\mu_n, \mu)$ is $\sigma^2/n$-subgaussian for all $n$.

Because $T_{p'}(\sigma^2)$ implies $T_\mu(\sigma^2)$ when $p' \geq p$, Theorem 7 also implies that a measure satisfying $T_{p'}(\sigma^2)$ has good concentration for $W_p$ if $p \leq p'$. In the opposite direction, if $p > p'$, then satisfying $T_{p'}(\sigma^2)$ still yields a weaker concentration bound. A modification of the proof of Theorem 7 yields the following result.

Theorem 8. Let $p' \in [1, 2]$ and $p \geq 1$. If $\mu \in \mathcal{P}_{p'}(\mathcal{X})$ satisfies $T_{p'}(\sigma^2)$, then $W_{p'}(\mu_n, \mu)$ is $\sigma^2/n^{1 - \left(\frac{2}{p'} - \frac{2}{p}\right)}$ subgaussian.

The conclusion of Theorem 8 is interesting whenever $\left(\frac{2}{p'} - \frac{2}{p}\right) < 1$. For example, if we assume merely that $\mu$ satisfies $T_1(\sigma^2)$, Theorem 8 only yields a nontrivial concentration result for $W_p(\mu_n, \mu)$ when $p < 2$; by contrast, if $\mu$ satisfies $T_2(\sigma^2)$, then Theorem 8 implies a concentration result for $W_p(\mu_n, \mu)$ for all $p < \infty$.

In Section 6, we require concentration properties not of the Wasserstein distance itself but of the $k$-dimensional Wasserstein distance. The following result shows that low-dimensional projections inherit the concentration properties of the $d$-dimensional measure.

Proposition 5. Let $U \in V_k(\mathbb{R}^d)$. For any $p \in [1, 2]$ and $\sigma > 0$, if $\mu$ satisfies $T_p(\sigma^2)$, then so does $\mu_U$.

Proof. The projection $x \mapsto Ux$ is a contraction. The result then follows from Gozlan [41, Corollary 20] [see also 68].

We conclude this section by giving some simple conditions under which the $T_1(\sigma^2)$ inequality is satisfied. The following characterization is well known. Denote by Lip$(\mathcal{X})$ the space of all functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathbb{R}$.

Proposition 6 (10, Theorem 1.3). A measure $\mu \in \mathcal{P}_1(\mathcal{X})$ satisfies $T_1(\sigma^2)$ if and only if $f(X)$ is $\sigma^2$-subgaussian for all $f \in$ Lip$(\mathcal{X})$.

It is common to extend Definition 3 to random vectors as follows.

Definition 4. A random variable $X$ on $\mathbb{R}^d$ is $\sigma^2$-subgaussian if $u^\top X$ is $\sigma^2$-subgaussian for all $u \in \mathbb{R}^d$ satisfying $\|u\| = 1$.

Subgaussian random vectors yield a large collection of random variables satisfying a $T_1$ inequality.

Lemma 1. If $\mu$ on $\mathbb{R}^k$ satisfies $T_p(\sigma^2)$ for some $p \geq 1$, then $X \sim \mu$ is $\sigma^2$-subgaussian. Conversely, if $X \sim \mu$ on $\mathbb{R}^k$ is $\sigma^2$-subgaussian, then $\mu$ satisfies $T_1(Ck\sigma^2)$ for a universal constant $C > 0$. 


If the entries of $X$ are independent, then the result holds with $C = 1$ by a result of Marton \cite{marton1976convergence}. The presence of the factor $k$ in the converse statement is unavoidable; unlike $T_2$ inequalities, $T_1$ inequalities do not exhibit dimension-free concentration \cite{ledoux2001concentration}.

6. Upper bounds

In this section, we establish that under the spiked transport model, Wasserstein projection pursuit produces a significantly more accurate estimate of the Wasserstein distance than the plug-in estimator.

Let $\mu^{(1)}_n$ and $\mu^{(2)}_n$ be two measures generated according to the spiked transport model (2). For $i \in \{1, 2\}$, we let $\mu_n^{(i)} := \frac{1}{n} \sum_{j=1}^n \delta_{X_j^{(i)}}$, where $X_j^{(i)} \sim \mu^{(i)}$ are i.i.d. We define

$$\hat{W}_{p,k} := \tilde{W}_{p,k}(\mu^{(1)}_n, \mu^{(2)}_n).$$

Our main upper bound shows that $\hat{W}_{p,k}$ converges to the true Wasserstein distance $W_p(\mu, \nu)$ at a rate much faster than $n^{-1/d}$.

**Theorem 9.** Let $p' \in [1, 2]$ and $p \geq 1$. Under the spiked transport model, if $\mu^{(1)}$ and $\mu^{(2)}$ satisfy $T_{p'}(\sigma^2)$, then

$$\mathbb{E}|\hat{W}_{p,k} - W_p(\mu^{(1)}_n, \mu^{(2)}_n)| \lesssim \sigma \sqrt{k} \left( r_{p,k}(n) + c_p \cdot n^{\left(\frac{1}{p'} - \frac{1}{p}\right)} + \frac{\sqrt{d \log n}}{n} \right).$$

Theorem 9 can also be extended to the misspecified model proposed in (5).

**Theorem 10.** Let $p' \in [1, 2]$ and $p \geq 1$. Under the relaxed spiked transport model (5), if $\nu^{(1)}$ and $\nu^{(2)}$ satisfy $T_{p'}(\sigma^2)$, then

$$\mathbb{E}|\hat{W}_{p,k} - W_p(\nu^{(1)}_n, \nu^{(2)}_n)| \lesssim \sigma \sqrt{k} \left( r_{p,k}(n) + c_p \cdot n^{\left(\frac{1}{p'} - \frac{1}{p}\right)} + \frac{\sqrt{d \log n}}{n} \right) + \varepsilon,$$

where $\varepsilon = W_p(\text{Law}(Z^{(1)}), \text{Law}(Z^{(2)}))$.

Theorem 10, which follows almost immediately from the proof of Theorem 9, establishes that Wasserstein projection pursuit can bring statistical benefits even in situations where the spiked transport model holds only approximately. Our results leave open the question of whether a version of Theorem 9 can be shown to hold under more general misspecification, and we view finding further relaxations of the spiked transport model as an interesting question for future work.

Theorem 9 follows from the following two propositions. We first show that the quality of the proposed estimator $\hat{W}_{p,k}$ can be bounded by the sum of two terms depending only on $\mu^{(1)}_n$ and $\mu^{(2)}_n$ individually.
Proposition 7. If \((\mu^{(1)}, \mu^{(2)})\) satisfy the spiked transport model, then
\[
\mathbb{E}[\hat{W}_p - W_p(\mu^{(1)}, \mu^{(2)})] \leq \mathbb{E}\hat{W}_{p,k}(\mu^{(1)}, \mu^{(1)}_{n}) + \mathbb{E}\hat{W}_{p,k}(\mu^{(2)}, \mu^{(2)}_{n})
\]

**Proof.** Since \(\hat{W}_p = \hat{W}_{p,k}(\mu^{(1)}_{n}, \mu^{(2)}_{n})\) and \(W_p(\mu^{(1)}, \mu^{(2)}) = \hat{W}_{p,k}(\mu^{(1)}, \mu^{(2)})\), the claim is immediate from Proposition 2.

The following proposition allows us to bound both terms of Proposition 7 by the desired quantity.

Proposition 8. Let \(p' \in [1, 2]\) and \(p \geq 1\). If \(\mu\) satisfies \(T_{p'}(\sigma^2)\), then
\[
\mathbb{E}\hat{W}_{p,k}(\mu, \mu_{n}) \lesssim \sigma \sqrt{k} \left( r_{p,k}(n) + c_p \cdot n^{\left(\frac{1}{p'} - \frac{1}{p}\right) + \frac{d \log n}{n}} \right).
\]

**Proof.** Wasserstein distances are invariant under translating both measures by the same vector. Therefore, we can assume without loss of generality that \(\mu\) has mean 0. Likewise, by homogeneity, we assume \(\sigma = 1\).

Let \(Z_U := W_p(\mu_U, (\mu_{n})_U)\). We first show that the process \(Z_U\) is Lipschitz.

Lemma 2. There exists a random variable \(L\) such that for all \(U, V \in \mathcal{V}_k(\mathbb{R}^d)\),
\[
|Z_U - Z_V| \leq L \|U - V\|_{\text{op}}
\]
and \(\mathbb{E}L \lesssim \sqrt{dp}\).

**Proof.** Let \(X \sim \mu\). Then
\[
|Z_U - Z_V| \leq W_p(\mu_U, \mu_V) + W_p((\mu_{n})_U, (\mu_{n})_V)
\]
\[
\leq (\mathbb{E}(\|U - V\|X\|^p)^{1/p} + \frac{1}{n} \sum_{i=1}^{n} \|U - V\|X_i\|^p)^{1/p}
\]
\[
\leq \|U - V\|_{\text{op}} \left( (\mathbb{E}X\|^p)^{1/p} + \left( \frac{1}{n} \sum_{i=1}^{n} \|X_i\|^p \right)^{1/p} \right).
\]
We obtain that
\[
|Z_U - Z_V| \leq L \|U - V\|_{\text{op}}
\]
where \(L = (\mathbb{E}X\|^p)^{1/p} + \left( \frac{1}{n} \sum_{i=1}^{n} \|X_i\|^p \right)^{1/p}\). By Jensen’s inequality, we have \(\mathbb{E}L \leq 2(\mathbb{E}X\|^p)^{1/p}\). Together with Lemma S.2, it yields the claim. \(\square\)
Lemma S.1 shows that there exists a universal constant $c$ such that $\log N(V_k, \varepsilon, \|\cdot\|_{\text{op}}) \leq \varepsilon \log n$ for $\varepsilon \in (0, 1]$. Choosing $\varepsilon = \sqrt{k/n}$ yields

$$E \sup_{U \in V_k(\mathbb{R}^d)} (Z_U - EZ_U) \lesssim \sqrt{dkp/n} + n \left(\frac{1}{p} - \frac{1}{2}\right) + \sqrt{dk \log n/n}$$

Applying Proposition S.6 yields

$$E \sup_{U \in V_k(\mathbb{R}^d)} W_p(\mu_U, (\mu_n)_U) \lesssim \sup_{U \in V_k(\mathbb{R}^d)} EW_p(\mu_U, (\mu_n)_U) + E \sup_{U \in V_k(\mathbb{R}^d)} (Z_U - EZ_U) \lesssim \sqrt{kr_p,k(n)} + c_p \cdot n \left(\frac{1}{p} - \frac{1}{2}\right) + \sqrt{dk \log n/n}$$

as claimed.

We also obtain a Davis-Kahan-type theorem on subspace recovery. Given two subspaces $U_1$ and $U_2$, the minimal angle $[26, 28]$ between them is defined to be

$$\angle(U_1, U_2) := \arccos \left( \sup_{u_1 \in U_1, u_2 \in U_2} \frac{u_1^\top u_2}{\|u_1\|\|u_2\|} \right).$$

If $\angle(U_1, U_2) = 0$, then $U_1 \cap U_2 = \{0\}$, so that $U_1$ and $U_2$ are at least partially aligned. In the important special case that $U_1$ and $U_2$ are each one dimensional, this definition reduces to the angle between the subspaces.

The following result indicates that as long as $\mu^{(1)}$ and $\mu^{(2)}$ are well separated, Wasserstein projection pursuit also yields a subspace with at least partial alignment to $U$.

**Theorem 11.** Let $p' \in [1, 2]$ and $p \geq 1$. Assume that $\mu^{(1)}$ and $\mu^{(2)}$ satisfy the spiked transport model and $T_{p'}(\sigma^2)$ Let $\bar{U} := \text{span}(\bar{U})$, where

$$\bar{U} := \arg\max_{U \in V_k(\mathbb{R}^d)} W_p((\mu_n^{(1)})_U, (\mu_n^{(2)})_U).$$

Then

$$E \sin^2 (\angle(\bar{U}, U)) \lesssim \frac{\sigma \sqrt{k} \left( r_p,k(n) + c_p \cdot n \left(\frac{1}{p} - \frac{1}{2}\right) + \sqrt{dk \log n/n} \right)}{W_p(\mu^{(1)}, \mu^{(2)})}.$$
A proof of Theorem 11 appears in Supplement A. Note that the a bound on the minimal angle is a rather weak guarantee. Indeed, \( \angle(\hat{U}, V) \to 0 \) implies that the subspaces \( \hat{U} \) and \( U \) share at least a common line asymptotically but not more. When \( k = 1 \), this ensures recovery of the subspace, but this no longer holds true for higher dimensional spikes. In retrospect such a guarantee is all that we can hope for under the mere assumption that \( W_p(\mu^{(1)}, \mu^{(2)}) > 0 \). Indeed, it may be the case that these distributions differ only on a one dimensional space. Stronger guarantees may be achieved by assuming that \( W_p(\mu^{(1)}_V, \mu^{(2)}_V) > 0 \) for a large family of \( V \), but we leave them for future research.

7. A lower bound on estimating the Wasserstein distance

In this section, we prove that the rate \( r_{p,d}(n) \) is optimal for estimating the Wasserstein distance, up to logarithmic factors. The core idea of our lower bound is to relate estimating the Wasserstein distance to the problem of estimating total variation distance, sharp rates for which are known [48, 92]. To obtain sufficient control over the Wasserstein distance as a function of total variation, we prove a refined bound incorporating both total variation and the \( \chi^2 \) divergence (Proposition 9). We then show a modified lower bound (Proposition 10) for a testing problem involving the total variation distance over the class of distributions on \([m]\) close to the uniform measure in \( \chi^2 \) divergence.

In the interest of generality, we formulate our results for any compact metric space \( X \) whose covering numbers satisfy
\[
C \varepsilon^{-d} \leq N(X, \varepsilon) \leq C \varepsilon^{-d}
\]
for all \( \varepsilon \leq \text{diam}(X) \). This condition clearly holds for compact subsets of \( \mathbb{R}^d \) and more generally for metric spaces with Minkowski dimension \( d \). We adopt the assumption \( \text{diam}(X) = 1 \) without loss of generality.

Let \( \mathcal{P} \) be the set of distributions supported on \( X \) and let \( R(n, \mathcal{P}) \) denote the minimax risk over \( \mathcal{P} \),
\[
R(n, \mathcal{P}) := \inf_{W} \sup_{\mu, \nu \in \mathcal{P}} \mathbb{E}_{\mu, \nu} |\hat{W} - W_p(\mu, \nu)|.
\]

The bound \( R(n, \mathcal{P}) \gtrsim n^{-1/2p} \) is an almost trivial consequence of the fact that the distribution \( \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1 \) cannot be distinguished from \( \left( \frac{1}{2} + \varepsilon \right) \delta_{-1} + \left( \frac{1}{2} - \varepsilon \right) \delta_1 \) on the basis of \( n \) samples when \( \varepsilon \approx n^{-1/2} \). The interesting part of Theorem 4 is the rate when \( d > 2p \).

We prove the following.

**Theorem 12.** Let \( d > 2p \geq 2 \) and assume \( X \) satisfies (6). Then
\[
R(n, \mathcal{P}) \geq C_{d,p}(n \log n)^{-1/d}.
\]

Before proving Theorem 12, we establish the two propositions described above. Proposition 9 allows us to reduce Theorem 12 to an estimation problem involving total variation
distance, and Proposition 10 is a lower bound on the minimax rate for that total variation estimation problem.

**Proposition 9.** Assume \( d > 2p \geq 2 \), and let \( m \) be a positive integer. Let \( u \) be the uniform distribution on \([m] := \{1, \ldots, m\}\). There exists a random function \( F: [m] \to X \) such that for any distribution \( q \) on \([m]\),

\[
cm^{-1/d}d_{TV}(q, u)^{\frac{1}{2}} \leq W_p(F_q, F_u) \leq C_{d,p}m^{-1/d}(\chi^2(q, u))^{1/d}d_{TV}(q, u)^{\frac{1}{2} - \frac{1}{p}}
\]

with probability at least \( 0.9 \).

**Proof.** Lemma S.7 shows that the condition \( \mathcal{N}(\mathcal{Y}, \varepsilon) \geq c\varepsilon^{-d} \) implies the existence a set \( \mathcal{G}_m := \{x_1, \ldots, x_m\} \subseteq X \) such that \( d(x_i, x_j) \geq m^{-1/d} \) for all \( i \neq j \). We select \( F \) uniformly at random from the set of all bijections from \([m]\) to \( \mathcal{G}_m \).

To show the lower bound, we note that any points \( x, y \in \mathcal{G}_m \) satisfy

\[
d(x, y)^p \geq m^{-p/d} \mathbb{1}\{x \neq y\},
\]

which implies that for any coupling \( \pi \) between \( F_q \) and \( F_u \)

\[
\int d(x, y)^p d\pi(x, y) \geq m^{-p/d} \mathbb{P}_\pi[X \neq Y] \geq m^{-p/d}d_{TV}(F_q, F_u) = m^{-p/d}d_{TV}(q, u).
\]

The lower bound therefore holds with probability 1.

We now turn to the upper bound. We employ a dyadic covering bound [95, Proposition 1]. For any \( k^* \), there exists a dyadic partition \( \{Q^k_i\}_{1 \leq k \leq k^*} \) of \( X \) with parameter \( \delta = 1/3 \) such that \( |Q^k| \leq \mathcal{N}(\mathcal{X}, 3^{-(k+1)}) \). We obtain that for any \( k^* \geq 0 \),

\[
W_p^p(F_q, F_u) \leq 3^{-k*^p} + \sum_{k=1}^{k^*} 3^{-(k-1)p} \sum_{Q^k_i \in Q^k} |F_q(Q^k_i) - F_u(Q^k_i)|,
\]

By Lemma S.8, for any \( k \),

\[
\mathbb{E} \sum_{Q^k_i \in Q^k} |F_q(Q^k_i) - F_u(Q^k_i)| \leq 2d_{TV}(q, u) \wedge C_{d,p} \left( \frac{3^{kd}\chi^2(q, u)}{m} \right)^{1/2}.
\]

Let \( k_0 \) be a positive integer to be fixed later.

By applying the first bound, we obtain

\[
\mathbb{E} \sum_{k > k_0} 3^{-(k-1)p} \sum_{Q^k_i \in Q^k} |F_q(Q^k_i) - F_u(Q^k_i)| \leq 3^{-k_0p}d_{TV}(q, u).
\]

Applying the second bound and recalling that \( d/2 > p \) yields

\[
\mathbb{E} \sum_{k \leq k_0} 3^{-(k-1)p} \sum_{Q^k_i \in Q^k} |F_q(Q^k_i) - F_u(Q^k_i)| \leq C_{d,p} \left( \frac{\chi^2(q, u)}{m} \right)^{1/2} \sum_{k \leq k_0} 3^{k(d/2-p)}
\]

\[
\leq C_{d,p} \left( \frac{\chi^2(q, u)}{m} \right)^{1/2} 3^{k_0(d/2-p)},
\]
We obtain for any $k^* \geq 0$ that

\[
\mathbb{E} W_p^p(F_1^q, F_2^u) \leq C_{d,p} \left( \frac{\chi^2(q, u)}{m} \right)^{1/2} 3^{k_0(d/2-p)} + C \cdot 3^{-k_0} d_{TV}(q, u) + 3^{-k^*} \frac{d}{2} \cdot d_{TV}(q, u),
\]

and taking $k^* \to \infty$ it suffices to bound the first two terms.

Let $k_0$ to be the smallest positive integer such that

\[
3^{k_0} d_{TV}(q, u)^2 \leq \frac{m}{\chi^2(q, u)}.
\]

Then

\[
3^{k_0 d/2} \left( \frac{\chi^2(q, u)}{m} \right)^{1/2} \leq C_d \cdot d_{TV}(q, u),
\]

and hence

\[
\mathbb{E} W_p^p(F_1^q, F_2^u) \leq C_{d,p} \cdot 3^{-k_0} d_{TV}(q, u) \leq C_{d,p} m^{-p/d} (\chi^2(q, u))^{p/d} d_{TV}(q, u)^{1 - \frac{2p}{d}}.
\]

The claim follows from Markov’s inequality.

We now show that there are composite hypotheses that are well separated in total variation distance but nevertheless hard to distinguish on the basis of samples.

**Proposition 10.** Fix a positive integer $n$ and a constant $\delta \in [0, 1/10]$. Given a positive integer $m$, let $D_m$ be the set of probability distributions $q$ on $[m]$ satisfying $\chi^2(q, u) \leq 9$. Denote by $D^\ast_m, \delta$ the subset of $D_m$ of distributions satisfying $d_{TV}(q, u) \leq \delta$ and by $D_m^\ast$ the subset of $D_m$ satisfying $d_{TV}(q, u) \geq 1/4$. If $m = \lceil C \delta^{-1} n \log n \rceil$ for a sufficiently large universal constant $C$ and $n$ is sufficiently large, then

\[
\inf_{\psi} \left\{ \sup_{q \in D_m^\ast} \mathbb{P}_q[\psi = 1] + \sup_{q \in D_m^\ast, \delta} \mathbb{P}_q[\psi = 0] \right\} \geq .9,
\]

where the infimum is taken over all (possibly randomized) tests based on $n$ samples.

The proof of Proposition 10 follows a strategy due to Valiant and Valiant [91] and Wu and Yang [97], and our argument is a modification of theirs which permits simultaneous control of total variation and the $\chi^2$ divergence. We give the proof in Appendix A.2.

We now give a proof of the main theorem.

**Proof of Theorem 12.** Let $\hat{W}$ be any estimator for the Wasserstein distance between distributions on $X$ constructed on the basis of $n$ samples from each distribution.

Let $u$ be the uniform distribution on $[m]$, for some $m$ to be specified. Let $c^*$ be the constant appearing in the lower bound of Proposition 9 and define $\Delta_d = \frac{1}{16} c^* m^{-1/d}$.

Given $n$ samples $X_1, \ldots, X_n$ from an unknown distribution on $[m]$, define the randomized test

\[
\psi = \psi(X_1, \ldots, X_n) := 1 \{ \hat{W}(F(X_1), \ldots, F(X_n); F(Y_1), \ldots, F(Y_n)) \leq 2\Delta_d \},
\]
where \( F \) is the random function constructed in Proposition 9 and where \( Y_i \) are i.i.d. from \( u \).

By Proposition 9, if \( \delta \leq \delta_{d,p} := \left( \frac{c^{1/2}}{16c_{d,p}} \right)^{1/p-2/d} \), any \( q \in D_{m,\delta}^- \) satisfies the bound
\[
W_p(F_q, F_u) \leq \Delta_d
\]
with probability at least .9. Likewise, for \( q \in D_{m,\delta}^+ \), the bound
\[
W_p(F_q, F_u) \geq 3\Delta_d
\]
also holds with probability at least .9.

Define the event \( A = \{|\hat{W} - W_p(F_q, F_u)| \geq \Delta_d\} \). We obtain, for any \( q \in D_{m,\delta}^- \),
\[
\mathbb{E}_F \mathbb{P}_{F_q, F_u}[A] \geq \mathbb{E}_F \mathbb{P}_{F_q, F_u}[\hat{W} > 2\Delta_d \text{ and } W_p(F_q, F_u) \leq \Delta_d] = \mathbb{E}_F \mathbb{P}_{F_q, F_u}[W > 2\Delta_d] - \mathbb{P}[W_p(F_q, F_u) > \Delta_d] \geq \mathbb{P}_q[\psi = 0] - .1,
\]
and analogously for \( q \in D_{m,\delta}^+ \),
\[
\mathbb{E}_F \mathbb{P}_{F_q, F_u}[A] \geq \mathbb{P}_q[\psi = 1] - .1.
\]

For any estimator \( \hat{W} \), we have
\[
\sup_{\mu,\nu \in \mathcal{P}} \mathbb{P}_{\mu,\nu}[|\hat{W} - W_p(\mu, \nu)| \geq \Delta_d] \geq \frac{1}{2} \left( \sup_{q \in D_{m,\delta}^+} \mathbb{E}_F \mathbb{P}_{F_q, F_u}[A] + \sup_{q \in D_{m,\delta}^-} \mathbb{E}_F \mathbb{P}_{F_q, F_u}[A] \right) \geq \frac{1}{2} \left( \sup_{q \in D_{m,\delta}^+} \mathbb{P}_q[\psi = 1] + \sup_{q \in D_{m,\delta}^-} \mathbb{P}_q[\psi = 0] \right) - .1.
\]

Choosing \( m = \lceil C\delta^{-1}n \log n \rceil \) for a sufficiently large constant \( C \) and applying Proposition 10 yields that \( \sup_{\mu,\nu \in \mathcal{P}} \mathbb{P}_{\mu,\nu}[|\hat{W} - W_p(F_q, F_u)| \geq \Delta_d] \geq .8 \), and Markov’s inequality yields the claim.

8. Computational-statistical gaps for the spiked transport model

Sections 4 and 7 clarify the statistical price for estimating the Wasserstein distance for high-dimensional measures. Section 4 shows that the curse of dimensionality can be avoided under the spiked transport model. The WPP estimator exploits the low-dimensional structure in the spiked transport model, thereby beating the worst-case rate presented in Section 7. However, it is not clear how to make the estimator we propose computationally efficient. In this section, we give evidence that this obstruction is a fundamental obstacle, that is, that no computationally efficient estimator can beat the curse of dimensionality.

The statistical query model, first introduced in the context of PAC learning [51], is a well known computational framework for analyzing statistical algorithms. Instead of being given access to data points from a distribution, a statistical query (SQ) algorithm can approximately evaluate the expectation of arbitrary functions with respect to the
distribution. This model naturally captures the power of noise-tolerant algorithms [51] and is strong enough to implement nearly all common machine learning procedures [see, e.g. 8].

We recall the following definition.

**Definition 5.** Given a random variable $X$ on $\mathbb{R}^d$, for any sample size parameter $t > 0$ and function $f : \mathbb{R}^d \to [0, 1]$, the oracle $\text{VSTAT}(t)$ returns a value $v \in [p - \tau, p + \tau]$, where $p = \mathbb{E}f(X)$ and $\tau = \frac{1}{t} \vee \sqrt{\frac{p(1-p)}{t}}$.

A query to a $\text{VSTAT}(t)$ oracle can be simulated by using a data set of size approximately $t$. Our main result proves a lower bound against an oracle with sample size parameter $t = 2^{cd}$ for a positive constant $c$. Simulating such an oracle would require a number of samples exponential in the dimension. Nevertheless, we show that even under this strong assumption, at least $2^{cd}$ queries to the oracle are required. This result suggests that any computationally efficient procedure to estimate the Wasserstein distance under the spiked transport model requires an exponential number of samples. By contrast, Section 6 establishes that, information theoretically, only a polynomial number of samples are required.

We now state our main result.

**Theorem 13.** There exists a positive universal constant $c$ such that, for any $d$, estimating $W_1(\mu_1^{(1)}, \mu_2^{(2)})$ for distributions $\mu_1^{(1)}$ and $\mu_2^{(2)}$ on $\mathbb{R}^d$ satisfying the spiked transport assumption with $k = 1$ to accuracy $\Theta(1/\sqrt{d})$ with probability at least $2/3$ requires at least $2^{cd}$ queries to $\text{VSTAT}(2^{cd})$.

Our proof is based on a construction due to [27] [see also 14]. We defer the details to the appendix.

**Appendix A: Proofs of Lower Bounds**

**A.1. Proof of Theorem 5**

We reduce from the spiked covariance model. By homogeneity, we may assume that $\sigma = 1$. Let $\mu_1^{(1)}$ be the standard Gaussian measure on $\mathbb{R}^d$, and for $\mu_2^{(2)}$ we take either the standard Gaussian measure or the distribution of a centered Gaussian with covariance $I + \beta vv^T$, where $\|v\| = 1$ and $\beta > 0$ is to be specified. As long as $\beta \lesssim 1$, the measure $\mu_2^{(2)}$ is a $O(1)$-Lipschitz pushforward of the Gaussian measure. Hence, it satisfies $T_p(O(1))$ [41, Corollary 20].

Note that if $\mu_2^{(2)}$ has covariance $I + \beta vv^T$, then

$$W_p(\mu_1^{(1)}, \mu_2^{(2)}) \geq W_1(\mu_1^{(1)}, \mu_2^{(2)}) = W_1(\mathcal{N}(0, 1), \mathcal{N}(0, 1 + \beta)) \gtrsim \beta.$$
However, Cai et al. [15, Proposition 2] establish that the minimax testing error for
\[ H_0 : \mathcal{N}(0, I) \text{ vs. } H_1 : \mathcal{N}(0, I + \beta vv^T), \|v\| = 1 \]
is bounded below by a constant when \( \beta \lesssim \sqrt{d/n} \). A standard application of Le Cam’s two-point method [90] yields the claim.

A.2. Proof of Proposition 10

We require the existence of two distributions on \( \mathbb{R}_+ \), which will serve as the building blocks of our construction.

**Proposition 11.** For any integer \( L \geq 0 \) and \( \varepsilon \in [0, 1/6] \), there exists a pair of random variables \( U \) and \( V \) with the following properties:

- \( \mathbb{E} U^j = \mathbb{E} V^j \) for all \( j \leq L \)
- \( U, V \in [0, 16 \varepsilon - L^2] \) almost surely
- \( \mathbb{E} U = \mathbb{E} V = 1 \) and \( \mathbb{E} U^2 = \mathbb{E} V^2 \leq 6 \).
- \( \mathbb{E}|U - 1| \leq 12 \varepsilon \) but \( \mathbb{E}|V - 1| \geq 1 \).

The proof is deferred to Appendix A.3.

**Proof of Proposition 10.** The proof follows closely the approach of Wu and Yang [97, Proposition 1]. We first employ a standard argument showing that we can consider the Poissonalized setting. It is trivial to see that given samples \( X_1, \ldots, X_n \) from a distribution \( q \) on \([m]\), the counts \( N_i = N_i(X_1, \ldots, X_n) := |\{j \in [n] : X_j = i\}| \) are sufficient for \( q \). We therefore consider tests \( \psi \) based on count vectors. Note that, under \( q \), the count vector \((N_1, \ldots, N_m)\) has distribution Multinomial\((n, q)\).

Define
\[
R_n := \inf_{\psi} \left\{ \sup_{q \in D_m^+} \mathbb{P}_q[\psi = 1] + \sup_{q \in D_m^-} \mathbb{P}_q[\psi = 0] \right\}.
\]
We aim to prove a lower bound on \( R_n \).

Let \( \rho > 0 \), and let for \( n \geq 1 \) let \( \{\psi_n\} \) be a set of near optimal tests for a fixed sample size; i.e.
\[
\sup_{q \in D_m^+} \mathbb{P}_q[\psi_n = 1] + \sup_{q \in D_m^-} \mathbb{P}_q[\psi_n = 0] \leq R_n + \rho.
\]
Define set of approximate probability vectors
\[
\hat{D}_{m,\delta}^- := \left\{ q \in \mathbb{R}_+^m : \left| \sum_{i=1}^m q_i - 1 \right| \leq \delta, \frac{q}{\sum_{i=1}^m q_i} \in D_{m,\delta}^- \right\}
\]
\[
\hat{D}_{m,\delta}^+ := \left\{ q \in \mathbb{R}_+^m : \left| \sum_{i=1}^m q_i - 1 \right| \leq \delta, \frac{q}{\sum_{i=1}^m q_i} \in D_{m,\delta}^+ \right\}
\]
We let $\hat{D}_{m,\delta} := \hat{D}^-_{m,\delta} \cup \hat{D}^+_{m,\delta}$. Given $q \in \hat{D}_{m,\delta}$, define the renormalization $\bar{q} = \sum_{i=1}^m q_i/q$. We then define

$$\bar{R}_n := \inf \{ \sup_{q \in \hat{D}^+_{m,\delta}} \mathbb{P}_q[\psi = 1] + \sup_{q \in \hat{D}^-_{m,\delta}} \mathbb{P}_q[\psi = 0] \},$$

where the infimum is taken over all estimators based on the counts $N_1, \ldots, N_m$ and where $\mathbb{P}_q$ indicates the probability when $N_1, \ldots, N_m$ are independent and $N_i \sim \text{Pois}(n q_i)$ for all $i \in [m]$. We set $N = \sum_{i=1}^m N_i$, and note that, conditioned on $N = n'$, the count vector $(N_1, \ldots, N_m)$ has distribution Multinomial($n', \bar{q}$).

We define a test $\tilde{\psi}$ based on these Poissonalized counts by setting $\tilde{\psi} = \psi_N(N_1, \ldots, N_m) := \psi_N(N_1, \ldots, N_m)$. This definition along with the near optimality of $\tilde{\psi}_{n'}$ for $n' \geq 0$ implies

$$\sup_{q \in \hat{D}^+_{m,\delta}} \mathbb{P}_q[\tilde{\psi} = 1] + \sup_{q \in \hat{D}^-_{m,\delta}} \mathbb{P}_q[\tilde{\psi} = 0] \leq \sum_{n' \geq 0} R_{n'} \mathbb{P}_q[N = n'] + \rho \leq R_{n'/2} + \mathbb{P}_q[N < n/2] + \rho,$$

where the last inequality follows from the fact that $R_{n'} \leq 1$ for all $n' \geq 0$ and $R_{n'}$ is non-increasing in $n'$. Since $N = \text{Pois}(n \sum_{i=1}^m q_i)$ and $\sum_{i=1}^m q_i \geq 3/4$, a standard Chernoff bound implies $\mathbb{P}[N < n/2] \leq \exp(-Cn)$. Since $\rho$ was arbitrary, we obtain that

$$\bar{R}_n \leq R_{n'/2} + \exp(-Cn).$$

To prove a lower bound on $\bar{R}_n$, we consider random vectors

$$Q = \frac{1}{m}(U_1, \ldots, U_m), \quad Q' = \frac{1}{m}(V_1, \ldots, V_m),$$

where $U_i$ and $V_i$ for $i \in [m]$ are independent copies of $U$ and $V$ constructed in Proposition 11 with $\varepsilon = \frac{1}{2} \frac{n}{m}$. Conditioned on $Q$ and $Q'$, let $N$ and $N'$ be count vectors with independent entries generated by $N_i \sim \text{Pois}(n Q_i)$ and $N'_i \sim \text{Pois}(n Q'_i)$. Let us denote by $\mathbb{P}$ and $\mathbb{P}'$ the distributions of $N$ and $N'$ respectively. Under $\mathbb{P}$ and $\mathbb{P}'$, the entries of $N$ and $N'$ are i.i.d. Poisson mixtures, so applying Wu and Yang [97, Lemma 4] yields

$$d_{TV}(\mathbb{P}, \mathbb{P}') \leq m \left( \frac{8enL}{\varepsilon m} \right)^L.$$ 

Let $E = \{Q \in \hat{D}^-_{m,\delta}\}$ and $E' = \{Q' \in \hat{D}^+_{m}\}$. By Lemma S.11, $\mathbb{P}[E^C]$ and $\mathbb{P}'[E'^C]$ are each at most $C \frac{L^4}{m^2}$. Let $\pi_E$ be the law of $Q$ conditioned on $E$, and define $\pi'_E$ analogously, and let $\mathbb{P}_E$ and $\mathbb{P}'_E$ be the laws of $N$ and $N'$ under these priors. We obtain for any estimator $\bar{\psi}$ based on count vectors

$$\sup_{q \in \hat{D}^+_{m,\delta}} \mathbb{P}_q[\bar{\psi} = 1] + \sup_{q \in \hat{D}^-_{m,\delta}} \mathbb{P}_q[\bar{\psi} = 0] \geq \int \mathbb{P}_q'[\bar{\psi} = 1] d\pi'_E(q') + \int \mathbb{P}_q[\bar{\psi} = 0] d\pi_E(q) \geq 1 - d_{TV}(\mathbb{P}_E, \mathbb{P}'_E) \geq 1 - d_{TV}(\mathbb{P}, \mathbb{P}') - C \frac{L^4}{\delta^2 m}.$$
Choosing $L = c\frac{\delta m}{n}$ for a sufficiently small constant $c$ yields that
\[
\tilde{R}_n \geq 1 - m \exp(-C\frac{\delta m}{n}) - C\frac{\delta^2 m^3}{n^4}.
\]
Therefore
\[
R_n \geq 1 - m \exp(-C\frac{\delta m}{n}) - C\frac{\delta^2 m^3}{n^4} - \exp(-Cn),
\]
and choosing $m = [C\delta^{-1}n \log n]$ for $C$ a sufficiently large constant and $n$ sufficiently large yields the claim. 

A.3. Proof of Proposition 11

First, the reduction of Wu and Yang [97, Lemma 7] implies that it suffices to construct random variables $Y$ and $Y'$ such that
- $\mathbb{E}Y^j = \mathbb{E}Y'^j$ for $0 \leq j < L$
- $Y,Y' \in [1, 16\varepsilon^{-1}L^2]$ a.s.
- $\mathbb{E}Y = \mathbb{E}Y' \leq 6$
- $\mathbb{E}\frac{1}{Y} \geq 1 - 6\varepsilon$ but $\mathbb{E}\frac{1}{Y'} \leq \frac{1}{2}$.

Indeed, applying their construction yields $U$ and $V$ satisfying the first three requirements of Proposition 11 as well as $\mathbb{P}[U = 0] \leq 6\varepsilon$ and $\mathbb{P}[V = 0] \geq \frac{1}{4}$. Since the supports of $U$ and $V$ lie in $\{0\} \cup [1, +\infty)$, we have $\mathbb{E}[U - 1] = 2\mathbb{P}[U = 0] \leq 12\varepsilon$ and $\mathbb{E}[V - 1] = 2\mathbb{P}[V = 0] \geq 1$, as desired. We therefore focus on constructing such a $Y$ and $Y'$.

By Wu and Yang [97, Lemma 7] combined with Timan [89, Section 2.11.1], there exist random variables $X$ and $X'$ supported on $[1, 16L^2]$ such that $\mathbb{E}X^j = \mathbb{E}X'^j$ for $0 \leq j < L$ and $\mathbb{E}X \leq 1 - 6\varepsilon$ but $\mathbb{E}X' \leq \frac{1}{2}$. Let $P_\varepsilon$ and $P'_\varepsilon$ denote the distribution of $\varepsilon^{-1}X$ and $\varepsilon^{-1}X'$, respectively.

Let
\[
\Delta_\varepsilon := \int \frac{1}{(y-1)(y-2)} dP_\varepsilon(y) \\
\Delta'_\varepsilon := \int \frac{1}{(y'-1)(y'-2)} dP'_\varepsilon(y') \\
Z_\varepsilon := \int \frac{1}{y-2} dP_\varepsilon(y) - \int \frac{1}{y'-1} dP'_\varepsilon(y') .
\]

We define two new distributions $Q$ and $Q'$ by
\[
Q(dy) = \delta_1(dy) + \frac{1}{Z_\varepsilon} \left( \frac{1}{(y-1)(y-2)} P_\varepsilon(dy) - \Delta_\varepsilon \delta_1(dy) \right) 
\tag{7}
\]
\[
Q'(dy') = \delta_2(dy') + \frac{1}{Z_\varepsilon} \left( \frac{1}{(y'-1)(y'-2)} P'_\varepsilon(dy') - \Delta'_\varepsilon \delta_2(dy') \right) 
\tag{8}
\]
By Lemma S.9,
\[ \Delta_\varepsilon, \Delta_\varepsilon' \in [0, \frac{9}{5} \varepsilon^2], \quad Z_\varepsilon \geq \frac{3}{10} \varepsilon, \]
which implies in particular that both \( Q \) and \( Q' \) are probability distributions.

Let \( Y \sim Q \) and \( Y' \sim Q' \). We first check the last three conditions. Clearly \( Y \) and \( Y' \) are supported on \([1, 16\varepsilon^{-1}L^2]\), and Lemma S.10 implies that \( \mathbb{E} Y = \mathbb{E} Y' \leq 6 \). We have \( \mathbb{E} Y \geq 1 - \frac{5\varepsilon}{Z_\varepsilon} \geq 1 - 6\varepsilon \), and since \( Y' \geq 2 \) almost surely the bound \( \mathbb{E} Y' \leq \frac{1}{2} \) is immediate.

It remains to check the moment-matching condition. Any polynomial \( p(y) \) of degree at most \( L - 1 \) can be written
\[ p(y) = (y - 1)(y - 2)q(y) + \alpha y + \beta, \]
where \( q(y) \) has degree less than \( L - 1 \). Then
\[ \mathbb{E} p(Y) - \mathbb{E} p(Y') = \mathbb{E}(Y - 1)(Y - 2)q(Y) - \mathbb{E}(Y' - 1)(Y' - 2)q(Y') + \alpha(\mathbb{E} Y - \mathbb{E} Y'). \]
The last term vanishes because \( \mathbb{E} Y = \mathbb{E} Y' \), and
\[ \mathbb{E}(Y - 1)(Y - 2)q(Y) - \mathbb{E}(Y' - 1)(Y' - 2)q(Y') = \frac{1}{Z_\varepsilon} (\mathbb{E} q(\varepsilon^{-1} X) - \mathbb{E} q(\varepsilon^{-1} X')) = 0, \]

since \( \mathbb{E} X^j = \mathbb{E} X'^j \) for all \( j < L - 1 \).

Therefore \( \mathbb{E} p(Y) = \mathbb{E} p(Y') \) for all polynomials of degree at most \( L - 1 \).

\begin{proof}
\end{proof}

A.4. Proof of Theorem 13

We first establish the existence of a probability distribution on \( \mathbb{R} \) which agrees with \( N(0, 1) \) on many moments, but is far from \( N(0, 1) \) in Wasserstein distance.

**Proposition 12.** There exists a \( O(1) \)-subgaussian distribution \( A \) on \( \mathbb{R} \) that satisfies the following requirements.

- \( A \) agrees with \( N(0, 1) \) on the first \( 2m - 1 \) moments.
- \( W_1(A, N(0, 1)) = \Omega(1/\sqrt{m}) \).
- \( \chi^2(A, N(0, 1)) = \exp(O(m)) \).

A proof appears in Supplement A.

The separation \( W_1(A, N(0, 1)) = \Omega(1/\sqrt{m}) \) in Proposition 12 is easily seen to be tight. Indeed, Rigollet and Weed [78, Corollary 2] show that if \( \mu \) and \( \nu \) are \( O(1) \)-subgaussian and agree on their first \( O(m) \) moments, then \( W_1(\mu, \nu) = O(1/\sqrt{m}) \).

By planting the distribution constructed in Proposition 12 in a random direction, we obtain two high-dimensional measures satisfying the spiked transport model.
Lemma 3. Let $v$ be a unit vector in $\mathbb{R}^d$, and denote by $P_v$ the distribution on $\mathbb{R}^d$ of the random variable $Xv + Z$, where $X \sim A$ and $Z \sim \mathcal{N}(0, I_d - vv^\top)$ is independent of $X$. Then $\mu(1) = P_v$ and $\mu(2) = \mathcal{N}(0, I_d)$ satisfy the spiked transport model \((2), and W_1(P_v, \mathcal{N}(0, I_d)) = \Omega(1/\sqrt{m})$.

Proof. If we let $\xi \sim \mathcal{N}(0, 1)$, then $\mathcal{N}(0, I_d)$ is the law of $\xi v + Z$, where $\xi$ and $Z$ are independent and $Z \sim \mathcal{N}(0, I_d - vv^\top)$. Denoting by $\mathcal{U}$ the span of $v$, we see that $P_v$ and $\mathcal{N}(0, I_d)$ satisfy \((2) with $X(1) = Xv$ and $X(2) = \xi v$. By Propositions 3 and 12, $W_1(P_v, \mathcal{N}(0, I_d)) = W_1(A, \mathcal{N}(0, 1)) = \Omega(1/\sqrt{m})$.

The proof of Theorem 13 follows from a framework due to Feldman et al. [32], from which the following result is extracted. Given distributions $P_1$, $P_2$, and $Q$, define

$$\chi^2_Q(P_1, P_2) := \int \left( \frac{dP_1}{dQ} - 1 \right) \left( \frac{dP_2}{dQ} - 1 \right) dQ.$$ 

We call a set $\mathcal{P}$ of distributions $(\gamma, \beta)$ correlated with respect to $Q$ if for all $P_i, P_j \in \mathcal{P}$,

$$\chi^2_Q(P_i, P_j) \leq \begin{cases} \beta & \text{if } i = j, \\ \gamma & \text{if } i \neq j. \end{cases}$$

We then have the following.

Proposition 13. Let $Q$ be a set of distributions, and let $Q$ be a reference distribution. Suppose that there exists a set $\mathcal{P} \subseteq Q$ such that $\mathcal{P}$ is $(\gamma, \beta)$ correlated with respect to $Q$. Then any SQ algorithm that distinguishes queries from $P = Q$ and $P \in \mathcal{P}$ with success probability at least $2/3$ requires at least $|\mathcal{P}|\gamma/3\beta$ queries to VSTAT(1/2$\gamma$).

Proof. By choosing $\gamma' = \gamma$ in Feldman et al. [32, Lemma 3.10], we obtain that the set $\mathcal{P}$ satisfies $\text{SDA}(P, Q, 2\gamma) \geq |\mathcal{P}|\gamma/(\beta - \gamma) \geq |\mathcal{P}|\gamma/\beta$. Then Feldman et al. [32, Theorem 3.7] implies that distinguishing $Q$ from $Q$ requires at least $|\mathcal{P}|\gamma/3\beta$ queries to VSTAT(1/2$\gamma$).

We can now prove the lower bound.

Proof of Theorem 13. Let $Q$ be the set $\{P_v : v \in \mathbb{R}^d, \|v\| = 1\}$. The Johnson-Lindenstrauss lemma [49] implies that for any $\delta \in (0, 1)$, there exists a set of $2^{O(\delta^2d)}$ unit vectors in $\mathbb{R}^d$ with pairwise inner product at most $\delta$. Denote by $S$ a set of such vectors, and set $\mathcal{P} := \{P_v : v \in S\} \subseteq Q$.

Write $\Delta$ for $\chi^2(A, \mathcal{N}(0, 1))$, and recall that $\Delta = \exp(O(m))$. By Diakonikolas et al. [27, Lemma 3.4], $\chi^2_{\mathcal{N}(0,I)}(P_v, P_v) \leq \|v \cdot v\| 2^m \Delta$, and the set $\mathcal{P}$ is therefore $(\delta^2 m, \Delta, \Delta)$ correlated. By Proposition 13, any SQ algorithm that distinguishes between $\mu = \mathcal{N}(0, 1)$ and $\mu \in \mathcal{P}$ with probability at least $2/3$ requires $2^{O(\delta^2d)} \delta^2 m$ queries to VSTAT(1/2$\Delta$).

Let $\delta > 0$ be a constant small enough that $\delta^{-2m}/2\Delta = \exp(O(m))$. If $m = cd$ for a sufficiently small positive constant $c$, then $2^{O(\delta^2d)} \delta^2 m = \exp(O(d))$. An SQ algorithm
to distinguish queries from $\mathcal{N}(0,1)$ from those from a distribution in $\mathcal{Q}$ with probability at least $2/3$ therefore requires $\exp(\Omega(d))$ queries to $\text{VSTAT}(\exp(\Omega(d)))$. Therefore, by Lemma 3, any SQ algorithm which estimates $W_1$ under the spiked transport model to accuracy $\Theta(1/\sqrt{m}) = \Theta(1/\sqrt{d})$ requires $\exp(\Omega(d))$ queries to $\text{VSTAT}(\exp(\Omega(d)))$, as claimed.

Supplementary Material

Supplement A: Supplement to “Estimation of Wasserstein distances in the Spiked Transport Model”

(1). Additional lemmas and omitted proofs

Acknowledgements

The authors thank Oded Regev for helpful suggestions.

References


[27] Diakonikolas, I., Kane, D. M., and Stewart, A. Statistical query lower bounds for


[82] Seguy, V. and Cuturi, M. Principal geodesic analysis for probability measures under the optimal transport metric. NIPS, 2015.