Wasserstein Projection Pursuit of Non-Gaussian Signals

Satyaki Mukherjee *
National University of Singapore

Soumendu Sundar Mukherjee *
Indian Statistical Institute, Kolkata

Debarghya Ghoshdastidar
Technical University of Munich

August 27, 2024

Abstract

We consider the general dimensionality reduction problem of locating in a high-dimensional data cloud, a k-dimensional non-Gaussian subspace of interesting features. We use a projection pursuit approach—we search for mutually orthogonal unit directions which maximise the q-Wasserstein distance of the empirical distribution of data-projections along these directions from a standard Gaussian. Under a generative model, where there is a underlying (unknown) low-dimensional non-Gaussian subspace, we prove rigorous statistical guarantees on the accuracy of approximating this unknown subspace by the directions found by our projection pursuit approach. Our results operate in the regime where the data dimensionality is comparable to the sample size, and thus supplement the recent literature on the non-feasibility of locating interesting directions via projection pursuit in the complementary regime where the data dimensionality is much larger than the sample size.

1 Introduction

A central question in statistics and machine learning concerns the recovery of useful or interesting features from data. A huge body of literature exists that focuses on such feature extraction tasks. Often the statistician encounters high-dimensional data of which only a relatively low-dimensional subspace is of interest. A family of algorithms, often described by the umbrella term *projection pursuit* (Friedman and Tukey (1974); Huber (1985)), are particularly well-suited for such tasks. By restricting attention to low-dimensional subspaces, projection pursuit allows the statistician to evade the so-called "curse-of-dimensionality", which plagues most classical learning algorithms in high-dimensional settings. Furthermore, projection pursuit helps the statistician to discard noisy and information-poor features. Some prominent members of this family of techniques include Principal Component Analysis (PCA), Independent Component Analysis (ICA), matching pursuit, etc.

Perhaps the simplest projection pursuit algorithm is PCA (see, e.g., Jolliffe (2002), Jolliffe and Cadima (2016)), which considers the subspace generated by the top k eigenvectors of the sample covariance matrix. In effect, PCA tries to find linear combinations of the original features which explain the most variability. While very useful in its own regard, PCA is limited by the fact that it only considers variances. Thus it works very well when the superfluous features have considerably

*Equal contribution

lower variance than the signal, e.g., in noise reduction problems. On the other hand, consider a setup where the interesting components are non-Gaussian, while the rest are Gaussian of comparable variability. The Gaussian components cannot be treated as noise (in the sense of having smaller variance than the signal component) but are simply superfluous or "uninteresting". PCA has difficulty separating the interesting non-Gaussian components in such scenarios.

To overcome this limitation, various methods conceptually similar to PCA have been proposed. A broad class of such methods goes by the name of ICA (Lee and Lee (1998)). Broadly speaking, there are two families of ICA algorithms. One class of algorithms focuses more on ensuring that the signal directions are statistically independent, thus they minimize mutual information. The other focuses on finding directions in which the data is the "least Gaussian" (i.e. most interesting). In this paper, we are also interested in the latter objective.

Many approaches are possible for finding non-Gaussian directions, depending upon our definition of "non-Gaussian". A natural way to quantify non-Gaussianity would be to measure the deviation of some aspect of a probability measure of interest from that of a standard Gaussian. For instance, one could use measures such as kurtosis (Girolami and Fyfe (1996)) or negentropy (Cao et al. (2003); Novey and Adali (2008)). Alternatively, one could measure the deviation from a Gaussian using suitable probability metrics such as the Kolmogorov-Smirnov (KS) distance, the Wasserstein distance, etc.

We now state the general projection pursuit approach in the context of the problem of identifying non-Gaussian components with the following simple example. Consider a p-dimensional random vector X which satisfies the following: there is a unknown direction u_* such that $u_*^\top X$ is non-Gaussian, while $(I - u_* u_*^\top)X$, the distribution of X in the orthogonal complement of u_*, is $(p-1)$-dimensional standard Gaussian, and, further, the non-Gaussian component $u_*^\top X$ and the Gaussian part $(I - u_* u_*^\top)X$ are statistically independent. Suppose $\varrho(\nu, \nu')$ is some measure of quantifying the distance between two probability measures ν and ν'. ϱ could be a divergence between probability measures (e.g., the Kullback-Liebler divergence) or a proper metric (e.g., KS distance). Suppose we observe a sample $X_1, \ldots, X_n$ from ν, the distribution of X. Our goal is to recover the unknown direction u_*. The main idea of projection pursuit is then to find a unit direction $\widehat{u}$ such that the empirical distribution of the data projected on $\widehat{u}$ (i.e. $\frac{1}{n}\sum_{i=1}^n \delta_{\widehat{u}^\top X_i}$) is the farthest from the standard Gaussian distribution with respect to ϱ, i.e.

$$\widehat{u} = \operatorname*{argmax}_{v \in \mathbb{S}_{p-1}} \varrho\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \Phi\right), \tag{1}$$

where Φ is the standard Gaussian measure (here $\mathbb{S}_{p-1}$ denotes the unit sphere in $\mathbb{R}^p$). We will later formalise a version of this for general k. Our paper is interested in the case when ϱ is the q-Wasserstein distance (for $1 \le q \le 2$) between probability measures. Specifically, we analyse the question of whether the recovered directions can be guaranteed (with high probability) to be from the signal space or not.

Some work in this regard has been done in Bickel et al. (2018) and Montanari and Zhou (2022). Bickel et al. (2018) use the KS distance for ϱ. They show that if the data is purely Gaussian (i.e. in a null model with no interesting directions), then two completely different phenomena occur according as whether the data-dimensionality-to-sample-size ratio $\frac{p}{n}$ goes to zero or infinity. In the former regime, all projections are Gaussians (in fact, this is known from the earlier work of Diaconis and Freedman (1984)). On the other hand, in the latter regime, given any arbitrary probability distribution $\widetilde{\nu}$, with high probability, one can find a (data-dependent) direction along which the data set is distributed as $\widetilde{\nu}$. In other words, one can find directions along which the data is as far

from Gaussianity (in the KS metric) as one desires. This means that projection-pursuit can spot fake signal amidst complete noise. When applied to actual data in this regime, there is no way of knowing if the found direction came from some underlying signal space, or if it is a mirage of signal in a Gaussian desert!

Montanari and Zhou (2022) prove a similar result for the 2-Wasserstein distance. Both the above papers argue that under the null model of $N(0, \mathbb{I}_p)$, when p/n converges to a sufficiently small constant, the empirical distributions of projections of the data points in every direction are close to the standard Gaussian distribution. This obviously begs the question if, under a spiked alternative, one could find directions along which the data projections are non-Gaussian. (Montanari and Zhou, 2022, Theorem 4.6) study the question of obtaining such a signal direction under a specific model of supervised learning.

With ϱ as the q-Wasserstein distance ($q \geq 1$), the recent work Niles-Weed and Rigollet (2022) has considered a spike estimation problem. Given two samples from measures μ_1, μ_2, they consider the problem of finding a k-dimensional subspace U which maximizes the q-Wasserstein distance (for $1 \leq q \leq 2$) between the empirical measures of $\widehat{\mu}_1$ and $\widehat{\mu}_2$ restricted to the subspace U. For instance, their main result shows for $q = 2$ that their estimator achieves an error of

$$O\left(\sqrt{k}\max\{n^{-1/k}, n^{-1/2q}(\log n)^{1/q}\} + \sqrt{\frac{p\log n}{n}}\right). \tag{2}$$

In comparison, we assume that μ_2 is Gaussian and μ_1 is mostly non-Gaussian in the spiked directions, with its (signal) orthogonal directions also being Gaussian. Using a more careful analysis, we are able to remove the $\sqrt{\log n}$ factor from the second term in (2) (see Theorem 3.1). While the setup of Niles-Weed and Rigollet (2022) does not require the non-signal direction to be Gaussian, our concentration bounds, i.e. Proposition 3.1 along with Theorem 10 in Niles-Weed and Rigollet (2022) can be used to extend our results to their setting thereby removing the $\log n$ factor in general. On the other hand, in the special case that the non-signal direction of the data is not exactly Gaussian, but close enough, we can still use Theorem 3.2 to prove error bounds on the estimator (in terms of a measure of closeness to Guassian for the non-signal direction).

Our contributions. Our analysis is done in the context of an alternative model of unsupervised learning. We suppose that the sample comes from a spiked Gaussian model, i.e. there is a k-dimensional subspace in which the distribution is decidedly not Gaussian. We first show that under sub-Gaussian tail assumptions, in every direction, the empirical distribution formed by the data-projections and the true marginal distribution in the same direction are uniformly close. This result is a substantial extension of similar results in Bickel et al. (2018) and Montanari and Zhou (2022) to a more general setting. Further, we also show that one can recover an orthonormal set of vectors which form an approximate basis of the signal space. (This can be thought of as an instance of the general strategy of matching pursuit.) In particular, each recovered vector's component in the independent Gaussian space is inversely proportional to the signal-to-noise ratio. Finally, if the signal-to-noise ratio is sufficiently large, then we give a methodology to accurately estimate k, i.e. the dimension of the signal space. This allows one to use our sequential procedure even in cases where very little is known about the signal space.

2 Set-up

Given a probability measure μ in $\mathbb{R}^p$, and a vector $v \in \mathbb{R}^p$, we define the action of v on μ, $v\sharp\mu$ to be the marginal density of μ in the direction v. In particular,

Definition 2.1. If X is a random variable in $\mathbb{R}^p$ having distribution μ and $v^\top X$ is the dot product of v and X, then $v\sharp\mu$ is defined to be the density function of the real-valued random variable $v^\top X$.

For $q > 1$, the q-Wasserstein distance $d_{W_q}(\mu, \nu)$ between two probability measures μ and ν is defined as

$$d_{W_q}(\mu,\nu) \coloneqq \inf_{\substack{\text{all couplings } \pi \text{ of } X,Y \\ X\sim\mu, Y\sim\nu}} \left[\mathbb{E}_\pi \|X-Y\|^q\right]^{1/q}.$$

We now introduce a formal set-up for the non-Gaussian component recovery problem which we will analyse.

Assumption 1. Suppose that we have data $X_1, \ldots, X_n$ i.i.d. from a σ-sub-Gaussian distribution ν on $\mathbb{R}^p$,[1] where Ψ has the following structure: Suppose $X \sim \Psi$.

1. $\mathbb{E}X = 0; \mathrm{var}(X) = I_p$[2].
2. There is a (unknown) k-dimensional subspace U such that $\Pi_U X$ has a sufficiently non-Gaussian distribution, and $\Pi_{U^\perp} X$ has a close-to-Gaussian distribution, in the sense that there exists constants κ_1, κ_2 with,

$$\inf_{v\in U\cap\mathbb{S}_{p-1}} d_{W_q}(v\sharp\Psi,\Phi) > \kappa_1 > \kappa_2 > \sup_{v\in U^\perp\cap\mathbb{S}_{p-1}} d_{W_q}(v\sharp\Psi,\Phi).$$

3. $\Pi_U X$ and $\Pi_{U^\perp} X$ are independent.

Given the sample $X_1, \cdots, X_n$, our goal is to recover the space U. The quantities κ_1 and κ_2, as we will see below, dictate a separation condition necessary to distinguish the non-Gaussian signal components from the Gaussian part.

To further motivate our setup let us quickly look at a simple distribution satisfying the assumptions above:

Example 2.1. Let the data X be generated from a mixture of Gaussians i.e. $\sum_{i=1}^k \frac{N(u_i,\mathbb{I}_p)}{k}$, where u_i's are some vectors in $\mathbb{R}^p$. Note quickly that if v is any norm 1 vector orthogonal to all u_i, then $v^\top X$ follows $N(0,1)$ and is independent of $u_i^\top X$. Clearly then this is a specific example of our model, with $U = \mathrm{Span}\{u_1, \ldots, u_k\}$. In fact the projection of Ψ to the subspace $U^\perp$ is the distribution $N(0, I_{p-k})$.

Before moving onto the technical results we also quickly define the following notations we will be using throughout.

[1] A random variable $X \in \mathbb{R}^p$ with mean μ is sub-Gaussian with parameter σ or the distribution is in $SG_p(\sigma)$ iff $P\Big(\|X-\mu\| \geq t\Big) \leq Ce^{-\frac{t^2}{2\sigma^2}}$.

[2] While for the purposes of our proof, we have assumed that the covariance matrix is identity throughout, this assumption is heuristically not much different from working with whitened data.

Definition 2.2. Given a p-dimensional distribution ν and a 1-dimensional distribution μ

$$d_{(q)}(\nu,\mu) \coloneqq \sup_{v\in\mathbb{S}_{p-1}} d_{W_q}(v\sharp\nu,\mu).$$

Note that when $p=1$, this is simply the q-Wasserstein distance between the ν and μ. For larger p, when $\mu=\Phi$, our distance $d_{(q)}$ captures how non-Gaussian the distribution ν can become in a particular direction.

Definition 2.3. Given a p-dimensional distribution ν and a 1-dimensional distribution μ

$$\mathrm{sep}_{(q)}(\nu,\mu) \coloneqq \inf_{v\in\mathbb{S}_{p-1}} d_{W_q}(u\sharp\nu,\mu).$$

In essence, when $\mu=\Phi$, $\mathrm{sep}_{(q)}$ gives a measure of separation from the 1-dimensional Gaussian.

Definition 2.4. For a subspace W of $\mathbb{R}^p$ and let Q be a matrix whose columns form an orthonormal basis of W. For a random variable $X\in\mathbb{R}^p$ with distribution μ, we define $\mu_{|Q}$ to be the distribution of $Q^\top X$.

We remark that as the distance $d_{(q)}(\mu,\Phi)$ is rotationally invariant, i.e. given a fixed subspace W, $d_{(q)}(\mu_{|Q},\Phi)$ is the same regardless of what orthonormal basis one chooses for W. Thus one can consider the quantity $d_{(q)}(\mu_{|W},\Phi)$ unambiguously.

Finally for the sake of clarity of our conclusion we define a signal to noise ratio for the distribution Ψ (under the assumption that $d_{(2)}(\Psi,\Phi)\neq \mathrm{sep}_{(2)}(\Psi_{|U},\Phi)$)[3], SNR_2 as

$$\mathrm{SNR}_2 = \sqrt{\frac{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}{d_{(2)}(\Psi,\Phi)^2 - \mathrm{sep}_{(2)}(\Psi_{|U},\Phi)^2}}.$$

We can also define a signal to noise ratio for a general q. Unfortunately because of technical reasons this ratio is slightly weaker than the specific case of $q=2$. In particular let us define for $q\geq 1, q\neq 2$, the signal to noise ratio SNR_q (under the assumption that $d_{(q)}(\Psi,\Phi)\neq \mathrm{sep}_{(q)}(\Psi_{|U},\Phi)$) as

$$\mathrm{SNR}_q = \sqrt{\frac{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2}{d_{(q)}(\Psi,\Phi)^2 - \mathrm{sep}_{(q)}(\Psi_{|U},\Phi)^2 + d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)/4}}.$$

Note that based on our definition of κ_1 and κ_2, our SNR_q^2 is always positive.

We will show that for p/n smaller than a constant and n large enough we can construct with high probability k orthonormal vectors $\widehat{u}_1,\dots,\widehat{u}_k$ such that

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\| \leq \frac{2}{\mathrm{SNR}_q}.$$

We also show that we can estimate k if $k\leq \frac{\mathrm{SNR}_q^2}{4}$.

[3] If $d_{(2)}(\Psi,\Phi)^2 = \mathrm{sep}_{(2)}(\Psi_{|U},\Phi)^2$, then we can choose any arbitrary $\delta>0$, and define $SNR_2 = \sqrt{\frac{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}{(1-\delta^2)d_{(2)}(\Psi,\Phi)^2}}$. Then for p/n smaller than a constant depending on δ and n large enough we will again get $\left\|\mathrm{Proj}_{U^\perp}(\widehat{v}_j)\right\| \leq \frac{2}{\mathrm{SNR}}$, as in the normal case. A similar equation is true for the general q

3 Main results

The following proposition is the central pivot granting us leverage to most of our results.

Proposition 3.1. *Let $X_1, \ldots, X_n$ be n data points from Ψ. Let n,p go to infinity in a way such that $p/n \to \gamma$. Then given a positive constant ϵ, there exists a positive constant $\gamma_{\sigma,\epsilon}$ depending on σ and ϵ such that when $\gamma \leq \gamma_{\sigma,\epsilon}$, we have for $1 \leq q \leq 2$*

$$\mathbb{P}\left(\sup_{v \in \mathbb{S}_{p-1}} \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, v\sharp\Psi \right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, v\sharp\Psi \right) \right] \right| > \epsilon \right) < D\exp(-nc_{\sigma,\gamma,\epsilon}),$$

where $c_{\sigma,\gamma,\epsilon}$ is some positive constant dependent on σ, γ, and ϵ.

Proposition 3.1 uniformly bounds the difference between the data dependent (and thus random) quantity, $d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, v\sharp\Psi \right)$, and the deterministic quantity $\mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, v\sharp\Psi \right) \right]$, dependent only on v.

Then given our assumptions 1 on the distribution Ψ above, we will state the following theorem (proved in Section 5):

Theorem 3.1 (Empirical non-Gaussianity implies true non-Gaussianity)**.** *Let $X_1, \ldots, X_n$ be n data points from Ψ. Let n,p go to infinity in a way such that $p/n \to \gamma$ and let $1 \leq q \leq 2$. Then given an $\epsilon > 0$, there exists a constants $\gamma_{\sigma,\epsilon}$ dependent on σ and ϵ and C_σ depending on σ such that if $\gamma \leq \gamma_{\sigma,\epsilon}$, the following statement is true with high probability for all unit vectors v in $\mathbb{R}^p$ simultaneously.*

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, \Phi \right) - d_{W_q}\left(v\sharp\Psi, \Phi \right) \right| \leq \epsilon + \frac{C_\sigma}{\sqrt[2q]{n}}.$$

We note that Theorem 3.1 needs very little assumptions on the distribution Ψ. We only need Ψ to be σ-sub-Gaussian. The main upshot of the theorem is that it implies with uniform high probability that in every direction the empirical distribution of the projection is as far away from Gaussian, as the true marginal distribution in that direction. Thus heuristically if we want to find directions in which Ψ is not Gaussian it makes sense to maximise the quantity $d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, \Phi \right)$. We can also derive a version of the above theorem which shows that in the proportional asymptotic setting (i.e. $p/n \to \gamma$), the error of the estimate is relatively small. A proof of Theorem 3.1 (and Corollary 3.1) is written in Section 5.

Corollary 3.1. *Let $X_1, \ldots, X_n$ be n data points from Ψ. Let n, p go to infinity in a way such that $p/n \to \gamma < 1$ and let $1 \leq q \leq 2$. Then there exists a constants γ_σ dependent on σ and C_σ depending on σ such that if $\gamma \leq \gamma_\sigma$, then the following is true with high probability simultaneously for all unit vectors v in $\mathbb{R}^p$:*

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, \Phi \right) - d_{W_q}\left(v\sharp\Psi, \Phi \right) \right| \leq \sqrt{\gamma}\log\frac{1}{\gamma} + \frac{C_\sigma}{\sqrt[2q]{n}}.$$

We can now proceed to state conditions under which the recovered directions have a very small component in the Gaussian subspace $U^\perp$.

Theorem 3.2 (Recovered direction is almost orthogonal to the Gaussian subspace)**.** *Let $U^\perp$ be the Gaussian subspace of Ψ. Let $X_1, \ldots, X_n$ be n data points from Ψ. Let n,p go to infinity in a way such that $p/n \to \gamma$. Given $\epsilon > 0$, there exists a constant $\gamma_{\sigma,\epsilon}$ dependent on σ and ϵ such that if $\gamma \leq \gamma_{\sigma,\epsilon}$, then with asymptotic high probability for any $v \in \mathbb{S}_{p-1}$ such that $d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \Phi\right) \geq \sqrt{1-\delta^2}d(\Psi,\Phi) + \epsilon + \frac{C_\sigma}{\sqrt[2q]{n}}$, we have that for $1 \leq q \leq 2$*

$$\left\|\mathrm{Proj}_{U^\perp}(v)\right\|_2^2 \leq \delta^2 \frac{d_{(q)}(\Psi,\Phi)^2}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2} + \frac{d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2}.$$

For $q = 2$ we have the sharper bound:

$$\left\|\mathrm{Proj}_{U^\perp}(v)\right\|_2^2 \leq \delta^2 \frac{d_{(2)}(\Psi,\Phi)^2}{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}.$$

Now that we have introduced most of our bulky technology, we can use it to prove the following simple Corollary. In the interest of space we have moved a detailed proof of the Corollary to the Appendix in Section A. This in turn allows us to argue the validity of a procedure in the vein of the general idea of matching pursuit.

Corollary 3.2 (Guarantee that recovery is possible)**.** *Let U be a k-dimensional sub-space of $\mathbb{R}^p$, where k is a constant. Let $X_1, \ldots, X_n$ be n data points from Ψ. Let $l < k$ be some integer. Let $v_1, \ldots, v_l$ be some vectors in $\mathbb{R}^p$. Then there exists some constant C_σ, depending on σ such that given $\epsilon > 0$ there exists with high probability a unit vector $\widehat{u}$ which is orthonormal to all v_i such that for $1 \leq q \leq 2$*

$$d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{\widehat{u}^\top X_i}, \Phi\right) \geq \mathrm{sep}_{(q)}(\Psi_{|U}, \Phi) - \epsilon - \frac{C_\sigma}{\sqrt[2q]{n}}.$$

Suppose now that the distribution Ψ is such that there is a k dimensional subspace U such that there exists a δ with $\mathrm{sep}_{(2)}(\Psi_{|U}, \Phi) \geq \sqrt{1-\delta^2}d_{(2)}(\Psi,\Phi)$. (If $\mathrm{sep}_{(2)}(\Psi_{|U}, \Phi) = d_{(2)}(\Psi,\Phi)$, then we can choose any $\delta > 0$. We discuss this case in Remark 3.1.) That is the "top k directions" are a constant factor far from Gaussian as the maximum possible. Then using Corollary 3.2 we can with high probability sequentially construct vectors $\widehat{u}_1, \ldots, \widehat{u}_k$ such that for every $1 \leq j \leq k$,

$$d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{\widehat{u}_j^\top X_i}, \Phi\right) \geq \sqrt{1-\delta^2}d_{(q)}(\Psi,\Phi) - \epsilon - \frac{C_\sigma}{\sqrt[2q]{n}}.$$

Then setting $\epsilon = \frac{\delta^2 d_{(q)}(\Psi,\Phi)}{2}$, and n large enough such that $\frac{4C_\sigma}{\sqrt[2q]{n}} \leq \delta^2 d(\Psi,\Phi)$, we have that

$$d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{\widehat{u}_j^\top X_i}, \Phi\right) \geq \sqrt{1-\delta^2}d_{(q)}(\Psi,\Phi) - \epsilon - \frac{C_\sigma}{\sqrt[2q]{n}} \geq \sqrt{1-4\delta^2}d_{(2)}(\Psi,\Phi) + \epsilon + \frac{C_\sigma}{\sqrt[2q]{n}}.$$

Now we can use Theorem 3.2 with ϵ set as above. Thus if p/n converges to a sufficiently small constant γ depending on δ and σ, then for large enough n with high probability we have that for every j,

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\| \leq 2\delta \frac{d_{(2)}(\Psi,\Phi)}{\sqrt{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}}.$$

By a similar argument for a general q we get,

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\|^2 \leq 4\delta^2 \frac{d_{(q)}(\Psi,\Phi)^2}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2} + \frac{d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2}.$$

In other words the k-space that we found (i.e. the one spanned by $\widehat{u}_1,\ldots,\widehat{u}_k$) is approximately orthogonal to $U^\perp$, the subspace where the distribution is close to Gaussian. Thus we have the following natural method to estimate k vectors which are almost orthogonal to $U^\perp$. For $1 \leq j \leq k$, let

$$\widehat{u}_j = \underset{\substack{v\in\mathbb{S}_{p-1} \text{ and} \\ v^\top \widehat{u}_t = 0 \ \forall t<j}}{\mathrm{argmax}} \ d_{W_2}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \Phi\right).$$

By invoking Theorem 3.2 with δ such that $\mathrm{sep}_{(2)}(\Psi_{|U},\Phi) = \sqrt{1-\delta^2}\, d_{(2)}(\Psi,\Phi)$, (Assuming $\mathrm{sep}_{(q)}(\Psi_{|U},\Phi) \neq d_{(q)}(\Psi,\Phi)$. We discuss the equality case in Remark 3.1) the above discussion then implies that, with high probability for large enough n we have that:

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\| \leq 2\sqrt{\frac{d_{(2)}(\Psi,\Phi)^2 - \mathrm{sep}_{(2)}(\Psi_{|U},\Phi)^2}{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}} = \frac{2}{\mathrm{SNR}_2}.$$

Similarly for a general q we have that

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\| \leq \sqrt{4\frac{d_{(q)}(\Psi,\Phi)^2 - \mathrm{sep}_{(q)}(\Psi_{|U},\Phi)^2}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2} + \frac{d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2}} = \frac{2}{\mathrm{SNR}_q}.$$

Remark 3.1. When $\mathrm{sep}_{(2)}(\Psi_{|U},\Phi) = \sqrt{1-\delta^2}\, d_{(2)}(\Psi,\Phi)$, we can in fact choose an arbitrary $\delta > 0$. We choose an apropriate ϵ as before and apply Theorem 3.2 along with Corollary 3.2. Thus, if p/n is smaller than a constant depending upon δ, with high probability for a large enough n we have that for every j,

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\| \leq 2\delta \frac{d_{(2)}(\Psi,\Phi)}{\sqrt{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}}.$$

Similarly for $q \geq 3$ we have, under similar conditions, that:

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\|^2 \leq 4\delta^2 \frac{d_{(q)}(\Psi,\Phi)^2}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2} + \frac{d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2}.$$

A common problem that often occurs in such problems is that k is unknown. To give some answer to this question we first consider the following corollary which is proved in detail in the Appendix in Section B.

Corollary 3.3. *Given integers $m > k+1$, let δ be a positive real number such that $4\delta^2 < \frac{1}{m}\left(1 - \frac{d_{(2)}(\Psi_{|U^\perp},\Phi)^2}{d_{(2)}(\Psi,\Phi)^2}\right)$. Let $X_1,\ldots,X_n$ be n data points from Ψ. Let n,p go to infinity in a way such that $p/n \to \gamma$ Given $\epsilon > 0$ there is a $\gamma_{\sigma,\epsilon}$, where $\gamma_{\sigma,\epsilon}$ is a constant depending on σ, ϵ, such that if $\gamma \leq \gamma_{\sigma,\epsilon}$ then with high probability there **does not exist** a set of $k+1$ orthonormal unit vectors $\widehat{u}_1,\ldots,\widehat{u}_{k+1}$ such that*

$$d_{W_2}\left(\frac{1}{n}\sum_{i=1}^n \delta_{\widehat{u}_j^\top X_i}, \Phi\right) \geq \sqrt{1-4\delta^2}\, d_{(2)}(\Psi,\Phi) + \epsilon + \frac{C_\sigma}{\sqrt[4]{n}},$$

for all $1 \leq j \leq k+1$.

Remark 3.2. For a general $q \neq 2$, we can get a similar result but with δ such that

$$4\delta^2 < \frac{1}{m} \frac{d_{(q)}(\Psi, \Phi)^2 - d_{(q)}(\Psi_{|U^\perp}, \Phi)^2 - d_{(q)}(\Psi, \Phi) d_{(q)}(\Psi_{|U^\perp}, \Phi)}{d_{(q)}(\Psi, \Phi)^2}.$$

Continuing the discussion prior to the corollary, we consider δ such that $\mathrm{sep}_{(2)}(\Psi_{|U}, \Phi) \geq \sqrt{1-\delta^2}\, d_{(2)}(\Psi, \Phi)$. Note that when $k+1 < \frac{1}{4} \frac{d_{(2)}(\Psi,\Phi)^2 - d_{(2)}(\Psi_{|U^\perp},\Phi)^2}{d_{(2)}(\Psi,\Phi)^2 - \mathrm{sep}_{(2)}(\Psi_{|U},\Phi)^2} = \frac{\mathrm{SNR}_2^2}{4}$ the hypothesis of Corollary 3.3 is true. This gives us a natural cutoff point for our sequential algorithm. We can stop at $\widehat{k}$, if for $\epsilon = \frac{\delta^2 d_{(2)}(\Psi,\Phi)}{2}$, and n large enough such that $\frac{4C_\sigma}{\sqrt[4]{n}} \leq \delta^2 d_{(2)}(\Psi, \Phi)$ we have that

$$d_{W_2}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}_{\widehat{k}+1}^\top X_i}, \Phi\right) < \sqrt{1-4\delta^2} d_{(2)}(\Psi, \Phi) + \epsilon + \frac{C_\sigma}{\sqrt[4]{n}}.$$

Corollary 3.3 then implies that this stopping rule ensures with high probability that $\widehat{k} \leq k$. On the other hand, the discussion following Corollary 3.2 means that the same stopping rule ensures that $\widehat{k} \geq k$. In effect we have that if $k+1 < \frac{\mathrm{SNR}_2^2}{4}$, then with high probability $\widehat{k} = k$.

4 Conclusion

In this article, we have considered the problem of isolating a non-Gaussian independent component from a Gaussian counterpart under certain separability assumptions. We have theoretically analysed the approximation accuracy of a projection pursuit procedure. In contrast to more traditional procedures like PCA, we do not need the variances of the superfluous feature directions to be small. We only need a distributional gap between directions which are Gaussian and those which are not.

Since the proposed method involves optimisation of the objective function $d_{W_2}\big(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \Phi\big)$ as v varies over the unit sphere, two natural questions immediately come to mind. First of all, since our objective function is markedly non-convex, designing an efficient algorithm that can find a global minimum (or even good local minima) would be a significant addition to present work.

Secondly, it needs to be investigated if similar results are true for distances other than the 2-Wasserstein distance. It is plausible that some distances would be more suitable both from a theoretical perspective and also the practical optimisation aspect. We leave the investigation of these questions for future work.

5 Proofs

5.1 Proof of Proposition 3.1

Quickly noting that for any vector $v \in \mathbb{R}^p$, with $\|v\| = 1$ we have,

$$\mathbb{P}\big(\big|v^\top (X-\mu)\big| \geq t\big) \leq \mathbb{P}\big(\|X-\mu\| \geq t\big),$$

we obtain the following simple proposition.

Proposition 5.1. *If $X \in \mathbb{R}^p$ is in $SG_p(\sigma)$ and $v \in \mathbb{R}^p$ be any unit norm vector (i.e. $\|v\| = 1$), then $v^\top X \in SG_1(\sigma)$.*

The following simple proposition bounds the operator norm of the sample covariance matrix of i.i.d. sub-Gaussian random vectors. It is a slightly reworded version of Theorem 6.5 of Wainwright (2019).

Proposition 5.2. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from a σ-sub-Gaussian distribution in $\mathbb{R}^p$ with covariance matrix $\mathbb{I}$. Then there exist universal constants c_1, c_2, c_3 such that we have for all $\delta > 0$ that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum X_i X_i^\top\right\|_2 \geq 1 + \sigma^2\left(c_1\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right) + \delta\right)\right) \leq c_2 e^{-nc_3 \min(\delta, \delta^2)}.$$

We will also use the following result on Wasserstein distances and sample convergences found in Bobkov and Ledoux (2019) as Corollary 7.17.

Proposition 5.3. *Let $q \geq 1$. Let μ be some distribution such that for some $s > 2q$ its s-th moment exists and is bounded. Then given v such that $\|v\| = 1$, if $X_1, \ldots, X_n$ are iid random variables sampled from μ, we have*

$$\mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \mu\right)^q\right] \leq \frac{C}{\sqrt{n}},$$

where C is some absolute constant dependent upon the upper bound of the s-th moment.

To prove Proposition 3.1, we will finally be needing the following lemma (proved in the Appendix in Section C,) on the concentration of the q-Wasserstein distance between a sub-Gaussian measure μ and the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ of an i.i.d. sample $X_1, \ldots, X_n$ from μ (a similar result with a log-Sobolev assumption on μ appears as Theorem 7.1 in Bobkov and Ledoux (2019)).

Lemma 5.1. *Let μ be a σ-sub-Gaussian measure. Let μ_n be the empirical measure formed from an i.i.d. sample of size n from μ. Then*

$$\mathbb{P}\big(|d_{W_q}(\mu_n, \mu) - \mathbb{E}d_{W_q}(\mu_n, \mu)| \geq t\big) \leq C \exp\left(-\frac{c\, n^{2/\max\{q,2\}} t^2}{\sigma^2}\right).$$

for some absolute constants $C, c > 0$.

Now armed with the above preliminaries we can move onto proving our central results:

Proof of Proposition 3.1. As mentioned before we begin by using Lemma 5.1 and the hypothesis that $v^\top X$ is σ-sub-Gaussian to get that for any fixed v (with $\|v\| = 1$) we have

$$\mathbb{P}\left(\left|d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)\right]\right| \geq t\right) \leq A' \exp\left(-\frac{c\, n^{2/\max\{q,2\}}\epsilon^2}{\sigma^2}\right),$$

where the constants A', c are absolute constants

Let $E_{p,\delta}$ be a smallest δ-net on the unit sphere in $\mathbb{R}^p$, i.e. given any $v \in \mathbb{S}_{p-1}$, $\exists$ a $w \in E_{p,\delta}$ such that $\|v - w\| \leq \delta$. It is known that there exists such a net for any p such that $|E_{p,\delta}| \leq A''\frac{1}{\delta^p}$. Then we have that

$$\mathbb{P}\left(\sup_{w \in E_{p,\delta}}\left|d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, w\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, w\sharp\Psi\right)\right]\right| \geq \epsilon\right)$$
$$\leq \left(\frac{1}{\delta}\right)^p A \exp\left(-\frac{c\, n^{2/\max\{q,2\}}\epsilon^2}{\sigma^2}\right).$$

To go from taking the supremum over the net to that on the entire sphere then we would have to control how small changes in v affect the quantity of interest. To that end let $v \in \mathbb{S}_{p-1}$. Let $w \in E_{p,\delta}$ such that $\|w - v\| \leq \delta$. Then we have

$$\begin{aligned}
&\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, w\sharp\Psi\right)\right| \\
&\qquad \leq \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, v\sharp\Psi\right)\right| \\
&\qquad\qquad + \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, v\sharp\Psi\right) - d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, w\sharp\Psi\right)\right| \\
&\qquad \leq \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}\right)\right| + \left| d_{W_q}\big(v\sharp\Psi, w\sharp\Psi\big)\right|.
\end{aligned}$$

We now upper bound each of the two terms above. The second term can be upper bounded for $1 \leq q \leq 2$ as follows:

$$\begin{aligned}
\left| d_{W_q}\big(v\sharp\Psi, w\sharp\Psi\big)\right|^q &= \inf_{\substack{(X,X')\text{: the marginals} \\ X \text{ and } X' \text{ are distributed as } \Psi}} \mathbb{E}[|v^\top X - w^\top X'|^q] \\
&\leq \mathbb{E}_{X\sim\Psi}[|(v-w)^\top X|^q] && \text{(considering the coupling } X = X'\text{)} \\
&= \mathbb{E}_{X\sim\Psi}[(v^\top X - w^\top X)^2]^{q/2} \\
&\leq \big((v-w)^\top \mathbb{E}_{X\sim\Psi}[XX^\top](v-w)\big)^{q/2} && \text{(by concavity of } x \mapsto x^{q/2} \text{ for } 0 \leq q/2 \leq 1\text{)} \\
&\leq \delta^q. && \text{(as } \mathbb{E}[XX^\top] = \mathbb{I}_p\text{)}
\end{aligned}$$

Similarly, for the first term we note that

$$\begin{aligned}
\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}\right)\right|^q &= \left((v-w)^\top\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)(v-w)\right)^{q/2} \\
&\leq \big(\delta \left\|\widehat{\Sigma}_n\right\|_2\big)^q,
\end{aligned}$$

where $\widehat{\Sigma}_n$ is the sample covariance matrix and $\|\cdot\|_2$ denotes the operator norm. We will now use a result that the operator norm of the sample covariance matrix is with high probability smaller than $2 + C'\sigma^2\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right)$, for some universal constant C' (see Proposition 5.2).

Then we can condition on this event as this is true with high probability $(1 - e^{-\theta n})$. Thus we have w.h.p.

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, \frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}\right)\right| \leq \delta\left(3 + C'\sigma^2\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right)\right).$$

Combining everything then we have w.h.p.

$$\left| d_{W_q}\left(\frac{\sum_{i=1}^n \delta_{v^\top X_i}}{n}, v\sharp\Psi\right) - d_{W_q}\left(\frac{\sum_{i=1}^n \delta_{w^\top X_i}}{n}, w\sharp\Psi\right)\right| \leq \delta\left(3 + C'\sigma^2\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right)\right).$$

We then have that whenever there exists an $v \in \mathbb{S}_{p-1}$ with

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)\right]\right| \geq \epsilon + 2\delta\left(3 + C'\sigma^2\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right)\right),$$

there exists a $v \in E_{p,\delta}$ whp (with the property that $\|v - w\| \le \delta$) such that

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, v\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)\right]\right| \ge \epsilon.$$

Thus using the probability bound on the δ-net gives us

$$\begin{aligned}
&\mathbb{P}\left(\sup_{v\in\mathbb{S}_{p-1}} \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)\right]\right| \ge \epsilon + 2\delta\left(3 + C'\sigma^2\sqrt{\frac{p}{n}} + C'\sigma^2\frac{p}{n}\right)\right) \\
&\le \mathbb{P}\left(\sup_{w\in E_{p,\delta}} \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, w\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{w^\top X_i}, w\sharp\Psi\right)\right]\right| \ge \epsilon\right) + e^{-\theta n} \\
&\le \left(\frac{1}{\delta}\right)^p A e^{-\frac{cn\epsilon^2}{\sigma^2}} + e^{-\theta n} = A e^{-\frac{cn\epsilon^2}{\sigma^2} - p\log\delta} + e^{-\theta n},
\end{aligned}$$

where we get the extra $\frac{p}{n}$ term as it is no longer true that $\frac{p}{n} \ll \sqrt{\frac{p}{n}}$ (and from invoking Proposition 5.2). Using the hypothesis that $\frac{p}{n} \to \gamma$ and choosing a δ such that

$$\delta = \frac{\epsilon}{2\left(3 + C'\sigma^2\sqrt{\gamma} + C'\sigma^2\gamma\right)},$$

we get

$$\begin{aligned}
&\mathbb{P}\left(\sup_{v\in\mathbb{S}_{p-1}} \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)\right]\right| \ge 2\epsilon\right) \\
&\le A\exp\left(-\frac{cn\epsilon^2}{\sigma^2} - p\log\frac{\epsilon}{6 + 2C'\sigma^2\sqrt{\gamma} + 2C'\sigma^2\gamma}\right) + e^{-\theta n} \\
&= A\exp\left(-n\left(\frac{c\epsilon^2}{\sigma^2} + (1+o(1))\gamma\log\frac{\epsilon}{6 + 2C'\sigma^2\sqrt{\gamma} + 2C'\sigma^2\gamma}\right)\right) + e^{-\theta n}. \qquad (3)
\end{aligned}$$

As there is some constant $\gamma_{\sigma,\epsilon}$ such that for all $\gamma \le \gamma_{\sigma,\epsilon}$, $\frac{c\epsilon^2}{\sigma^2} + \gamma\log\frac{\epsilon}{6+2C'\sigma^2\sqrt{\gamma}+2C'\sigma^2\gamma}$ is positive and lower bounded, the proof follows. □

5.2 Proof of Theorem 3.1

Proof of Theorem 3.1. Note that Proposition 3.1 probabilistically bounds the difference between $d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)$ and the its expectation as v varies over all possible unit vectors. Define $\gamma_{\sigma,\epsilon}$ as in Proposition 3.1. Thus when $\gamma \le \gamma_{\sigma,\epsilon}$ we have with high probability for all unit vectors $u \in \mathbb{R}^p$ simultaneously that

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{v^\top X_i}, v\sharp\Psi\right)\right]\right| \le \epsilon. \qquad (4)$$

Since $v\sharp\Psi$ is σ-sub-Gaussian we can conclude that all its moments are upper bounded (by a suitable function of σ). Choose any $s > 4$ and combine the upper bound on s'th moment of $v^\top X$ with

Proposition 5.3 (Corollary 7.17 of Bobkov and Ledoux (2019)). We then get that there is a constant C_σ, dependent on σ, such that

$$\mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{v^\top X_i}, v\sharp\Psi\right)\right]^q \leq \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{v^\top X_i}, v\sharp\Psi\right)^q\right] \leq \frac{C_\sigma^2}{\sqrt{n}}, \tag{5}$$

using the convexity of $x \mapsto x^q$ for $q \geq 1$.

Combining Equations (4) and (5) with the triangle inequality, we have with high probability that

$$\begin{aligned}\left|d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{v^\top X_i}, \Phi\right) - d_{W_q}\big(v\sharp\Psi, \Phi\big)\right| &\leq d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{v^\top X_i}, v\sharp\Psi\right)\\ &\leq \epsilon + \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{v^\top X_i}, v\sharp\Psi\right)\right] \leq \epsilon + \frac{C_\sigma^2}{n^{1/2q}}.\end{aligned}$$

This completes the proof. □

5.3 Proof of Corollary 3.1

We note from proof of Theorem 3.1 and equation 3 of the proof of Proposition 3.1 that for the upperbound to hold we can choose ϵ, γ (for $\gamma < 1$) such that $\frac{c\epsilon^2}{\sigma^2} + \gamma \log \frac{\epsilon}{6+2C'\sigma^2\sqrt{\gamma}+2C'\sigma^2\gamma}$ is positive.

In particular then if we choose $\epsilon = -\sqrt{\gamma}\log\gamma$, our quantity of interest becomes

$$\begin{aligned}\frac{c\epsilon^2}{\sigma^2} + \gamma\log\frac{\epsilon}{6+2C'\sigma^2\sqrt{\gamma}+2C'\sigma^2\gamma} &= \frac{c\gamma|\log\gamma|^2}{\sigma^2} + \gamma\log\frac{\sqrt{\gamma}|\log\gamma|}{6+2C'\sigma^2\sqrt{\gamma}+2C'\sigma^2\gamma}\\ &= \gamma\left[\frac{c|\log\gamma|^2}{\sigma^2} + \log\frac{\sqrt{\gamma}|\log\gamma|}{6+2C'\sigma^2\sqrt{\gamma}+2C'\sigma^2\gamma}\right].\end{aligned}$$

Thus for small enough γ (depending on σ) this quantity is always positive and our upper bound holds. Plugging in the value of this ϵ in Theorem 3.1 gives us with high probability that

$$\left|d_{W_2}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{v^\top X_i}, \Phi\right) - d_{W_2}\big(v\sharp\Psi, \Phi\big)\right| \leq \sqrt{\gamma}\log\frac{1}{\gamma} + \frac{C_\sigma}{\sqrt[4]{n}}.$$

5.4 Proof of Theorem 3.2

We first prove a lemma.

Lemma 5.2. *Let $\beta = \big\|\mathrm{Proj}_W(v)\big\|$, then for any $q \geq 1$,*

$$d_{W_q}\big(v\sharp\Psi, \Phi\big) \leq \sqrt{1-\beta^2}\, d(\Psi, \Phi) + \beta\, d(\Psi_{|U^\perp}, \Phi).$$

Further, for $q = 2$, we have the sharper inequality

$$d_{W_2}\big(v\sharp\Psi, \Phi\big)^2 \leq (1-\beta^2)\, d(\Psi, \Phi)^2 + \beta^2\, d(\Psi_{|U^\perp}, \Phi)^2.$$

Proof. Let $v = \sqrt{1-\beta^2}u + \beta w$, where $u \in U$, $w \in U^\perp$, and $\|w\| = \|u\| = 1$. Let X be a random variable distributed as Ψ. Thus $v^\top X = \sqrt{1-\beta^2}u^\top X + \beta w^\top X$, where $u^\top X$ and $w^\top X$ are independent random variables following the distributions $u\sharp\Psi$ and $w\sharp\Psi$ respectively. In other words if Y_1 and Y_2 are two independent random variables from the distributions $u\sharp\Psi$ and $w\sharp\Psi$ respectively we can write

$$u^\top X \overset{d}{=} \sqrt{1-\beta^2}Y_1 + \beta Y_2, \tag{6}$$

where $\overset{d}{=}$ means equal in distribution. Note then that if Z is a random variable distributed as Φ(i.e. $N(0,1)$), and Z_2, Z_3 are two iid copies distributed as Φ, we can also write

$$Z \overset{d}{=} \sqrt{1-\beta^2}Z_1 + \beta Z_2. \tag{7}$$

Let Ω, be the set of all possible couplings of the distributions of $v\sharp\Psi$ and Φ. Similarly, let Ω_1 (resp. Ω_2) be the set of all possible couplings of the distributions of $u\sharp\Psi$ and Φ (resp. $w\sharp\Psi$ and Φ). Then there is a natural way to construct a coupling in Ω given a coupling in Ω_1 and another in Ω_2. That given any joint distribution μ in Ω_1 whose marginals are $u\sharp\Psi$ and Φ respectively, define Y_1 and Z_1 be the corresponding marginal random variables i.e. $(Y_1, Z_1) \sim \mu$. Similarly given any joint distribution ν in Ω_2, we can define random variables Y_2, Z_2 where $(Y_2, Z_2) \sim \nu$. Note that by construction we can keep the pair (Y_1, Z_1) independent of (Y_2, Z_2). Then equations 6 and 7 can be used to define a joint distribution in Ω. We then derive the following inequality,

$$\begin{aligned}
d_{W_q}\big(v\sharp\Psi, \Phi\big) &= \inf_{(v^\top X, Z)\in\Omega} \mathbb{E}\big[|v^\top X - Z|^q\big]^{1/q} \\
&\leq \inf_{\substack{(Y_1,Z_1)\in\Omega_1 \\ (Y_2,Z_2)\in\Omega_2}} \mathbb{E}\big[\big|\sqrt{1-\beta^2}Y_1 + \beta Y_2 - \sqrt{1-\beta^2}Z_1 - \beta Z_2\big|^q\big]^{1/q} \\
&= \inf_{\substack{(Y_1,Z_1)\in\Omega_1 \\ (Y_2,Z_2)\in\Omega_2}} \mathbb{E}\big[\big|\sqrt{1-\beta^2}(Y_1 - Z_1) + \beta(Y_2 - Z_2)\big|^q\big]^{1/q} \\
&\leq \inf_{\substack{(Y_1,Z_1)\in\Omega_1 \\ (Y_2,Z_2)\in\Omega_2}} \sqrt{1-\beta^2}\mathbb{E}\big[|Y_1 - Z_1|^q\big]^{1/q} + \beta\mathbb{E}\big[|Y_2 - Z_2|^q\big]^{1/q} \text{ (by Minkowski's inequality)} \\
&= \sqrt{1-\beta^2}\, d_{W_q}(u\sharp\Psi, \Phi) + \beta\, d_{W_q}(w\sharp\Psi, \Phi). \\
&\leq \sqrt{1-\beta^2}\, d_{(q)}(\Psi, \Phi) + \beta\, d_{(q)}(\Psi_{|W}, \Phi).
\end{aligned}$$

For the case $q = 2$, we get a stronger inequality by carefully expanding the square instead of using

Minkowski's inequality. In particular,

$$\begin{aligned}
d_{W_2}\big(v\sharp\Psi,\Phi\big)^2 &= \inf_{(v^\top X,Z)\in\Omega}\mathbb{E}\Big[(v^\top X-Z)^2\Big]\\
&\le \inf_{\substack{(Y_1,Z_1)\in\Omega_1\\(Y_2,Z_2)\in\Omega_2}}\mathbb{E}\Big[(\sqrt{1-\beta^2}Y_1+\beta Y_2-\sqrt{1-\beta^2}Z_1-\beta Z_2)^2\Big]\\
&= \inf_{\substack{(Y_1,Z_1)\in\Omega_1\\(Y_2,Z_2)\in\Omega_2}}(1-\beta^2)\mathbb{E}\Big[(Y_1-Z_1)^2\Big]+\beta^2\mathbb{E}\Big[(Y_2-Z_2)^2\Big]\\
&\qquad\qquad +2\beta\sqrt{1-\beta^2}\mathbb{E}\Big[(Y_1-Z_1)(Y_2-Z_2)\Big]\\
&= \inf_{\substack{(Y_1,Z_1)\in\Omega_1\\(Y_2,Z_2)\in\Omega_2}}(1-\beta^2)\mathbb{E}\Big[(Y_1-Z_1)^2\Big]+\beta^2\mathbb{E}\Big[(Y_2-Z_2)^2\Big]\\
&= (1-\beta^2)\inf_{(Y_1,Z_1)\in\Omega_1}\mathbb{E}\Big[(Y_1-Z_1)^2\Big]+\beta^2\inf\,(Y_2,Z_2)\in\Omega_2\mathbb{E}\Big[(Y_2-Z_2)^2\Big]\\
&= (1-\beta^2)d_{W_2}(u\sharp\Psi,\Phi)^2+\beta^2 d_{W_2}(w\sharp\Psi,\Phi)^2\\
&\le (1-\beta^2)d_{(2)}(\Psi,\Phi)^2+\beta^2 d_{(2)}(\Psi_{|U^\perp},\Phi)^2.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.2. We begin by using Lemma 5.2 to get

$$d_{W_q}\big(v\sharp\Psi,\Phi\big)\le\sqrt{1-\beta^2}\,d_{(q)}(\Psi,\Phi)+\beta\,d_{(q)}(\Psi_{|U^\perp},\Phi).$$

Squaring both sides,

$$\begin{aligned}
d_{W_q}\big(v\sharp\Psi,\Phi\big)^2 &\le (1-\beta^2)\,d_{(q)}(\Psi,\Phi)^2+\beta^2\,d_{(q)}(\Psi_{|U^\perp},\Phi)^2+2\beta\sqrt{(1-\beta^2)}d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)\\
&\le (1-\beta^2)\,d_{(q)}(\Psi,\Phi)^2+\beta^2\,d_{(q)}(\Psi_{|U^\perp},\Phi)^2+d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi),
\end{aligned}$$

where the last line follows by using the AM-GM inequality to get $\sqrt{\beta^2(1-\beta^2)}\le 1/2$. Now using Theorem 3.1 and the hypothesis, we have for an appropriate $\gamma_{\sigma,\epsilon}$ and C_σ that $d_{W_q}(v\sharp\Psi,\Phi)\ge\sqrt{1-\delta^2}d_{(q)}(\Psi,\Phi)$. Thus

$$\begin{aligned}
&(1-\delta^2)d_{(q)}(\Psi,\Phi)^2\le(1-\beta^2)\,d_{(q)}(\Psi,\Phi)^2+\beta^2\,d_{(q)}(\Psi_{|U^\perp},\Phi)^2+d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)\\
\implies &\big\|\mathrm{Proj}_{U^\perp}(v)\big\|^2=\beta^2\le\delta^2\frac{d_{(q)}(\Psi,\Phi)^2}{d_{(q)}(\Psi,\Phi)^2-d_{(q)}(\Psi_{|U^\perp},\Phi)^2}+\frac{d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi)}{d_{(q)}(\Psi,\Phi)^2-d_{(q)}(\Psi_{|U^\perp},\Phi)^2}
\end{aligned}$$

For $q=2$, we can use the sharper upper bound in Lemma 5.2, i.e.

$$d_{W_2}\big(v\sharp\Psi,\Phi\big)^2\le(1-\beta^2)\,d_{(2)}(\Psi,\Phi)^2+\beta^2\,d_{(2)}(\Psi_{|U^\perp},\Phi)^2.$$

As before using Theorem 3.1 and the hypothesis, we have for an appropriate $\gamma_{\sigma,\epsilon}$ and C_σ that $d_{W_2}(v\sharp\Psi,\Phi)\ge\sqrt{1-\delta^2}d(\Psi,\Phi)$. Using this and rewriting the inequality we get

$$\big\|\mathrm{Proj}_{U^\perp}(v)\big\|^2=\beta^2\le\delta^2\frac{d_{(2)}(\Psi,\Phi)^2}{d_{(2)}(\Psi,\Phi)^2-d_{(2)}(\Psi_{|U^\perp},\Phi)^2}.$$

Remark 5.1. Instead of bounding $\sqrt{\beta^2(1-\beta^2)}$ by $1/2$, we could get a sharper inequality by instead doing the following. We note that if $(1-\delta^2)d_{(q)}(\Psi,\Phi)^2 - (1-\beta^2)\, d_{(q)}(\Psi,\Phi)^2 - \beta^2\, d_{(q)}(\Psi_{|U^\perp},\Phi)^2$ is negative then we already get the bound $\beta^2 \leq \delta^2 \frac{d_{(q)}(\Psi,\Phi)^2}{d_{(q)}(\Psi,\Phi)^2 - d_{(q)}(\Psi_{|U^\perp},\Phi)^2}$. Otherwise we can square both sides to get

$$\left((1-\delta^2)d_{(q)}(\Psi,\Phi)^2 - (1-\beta^2)\, d_{(q)}(\Psi,\Phi)^2 - \beta^2\, d_{(q)}(\Psi_{|U^\perp},\Phi)^2\right)^2 \leq 4\beta^2(1-\beta^2)d_{(q)}(\Psi,\Phi)d_{(q)}(\Psi_{|U^\perp},\Phi).$$

On expansion we note that this is a quadratic in β^2. Thus if Δ is its discriminant and we write $x = d_{(q)}(\Psi,\Phi)$, $y = d_{(q)}(\Psi_{|U^\perp},\Phi)$, we get the bound

$$\beta^2 \leq \frac{2\delta^2x^2(x^2-y^2) + 4x^2y^2 + \sqrt{\Delta}}{2(x^2+y^2)^2} \leq \frac{2\delta^2x^2(x^2-y^2) + 4x^2y^2}{(x^2+y^2)^2},$$

where the last inequality follows from the fact that the quadratics leading and constant terms are positive and hence $\sqrt{\Delta} \leq 2\delta^2x^2(x^2-y^2) + 4x^2y^2$.

□

Acknowledgements

SSM was partially supported by an INSPIRE research grant (DST/INSPIRE/04/2018/002193) from the Dept. of Science and Technology, Govt. of India and a Start-Up Grant from Indian Statistical Institute, Kolkata.

This work was carried out when Satyaki was at TUM and was supported by the German Research Foundation (DFG) through DFG-ANR PRCI "ASCAI" (GH 257/3-1).

References

Bickel, P. J., Kur, G., and Nadler, B. (2018). Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, 115(37):9151–9156.

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

Cao, L., Chua, K. S., Chong, W., Lee, H., and Gu, Q. (2003). A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, 55(1-2):321–336.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The Annals of statistics*, pages 793–815.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890.

Girolami, M. and Fyfe, C. (1996). Negentropy and kurtosis as projection pursuit indices provide generalised ica algorithms. In *Advances in Neural Information Processing Systems Workshop*, volume 9. Denver, CO.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, pages 435–475.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Lee, T.-W. and Lee, T.-W. (1998). *Independent component analysis*. Springer.

Montanari, A. and Zhou, K. (2022). Overparametrized linear dimensionality reductions: From projection pursuit to two-layer neural networks. *arXiv preprint arXiv:2206.06526*.

Niles-Weed, J. and Rigollet, P. (2022). Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688.

Novey, M. and Adali, T. (2008). Complex ica by negentropy maximization. *IEEE Transactions on Neural Networks*, 19(4):596–609.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

A Proof of Corollary 3.2

Proof of Corollary 3.2. As $l < k$ and $\dim(U) = k$, by rank nullity theorem, there exists a unit vector $\widehat{u}$ in U which is orthogonal to all the vectors $\{v_1, \ldots, v_l\}$. As k is constant. $k \ll n$ thus we can invoke Proposition 3.1 with $p = k$ to get that with high probability

$$\sup_{v \in U \cap \mathbb{S}_{p-1}} \left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, v\sharp\Psi_\Phi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{v^\top X_i}, v\sharp\Psi_\Phi\right)\right]\right| < \epsilon.$$

In particular, then

$$\left| d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}^\top X_i}, \widehat{u}\sharp\Psi_\Phi\right) - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}^\top X_i}, \widehat{u}\sharp\Psi_\Phi\right)\right]\right| < \epsilon.$$

Finally similar to the proof of Theorem 3.1, invoking Proposition 5.3 gives us

$$\mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}^\top X_i}, \widehat{u}\sharp\Psi_\Phi\right)\right] \leq \frac{C_\sigma}{\sqrt[2q]{n}}.$$

Therefore we have

$$\begin{aligned} d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}^\top X_i}, \Phi\right) &\geq d_{W_q}\big(\widehat{u}\sharp\Psi, \Phi\big) - d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}^\top X_i}, \widehat{u}\sharp\Psi\right) \\ &\geq \mathrm{sep}_q(\Psi_{|U}, \Phi) - \epsilon - \mathbb{E}\left[d_{W_q}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}^\top X_i}, \widehat{u}\sharp\Psi\right)\right] \\ &\geq \mathrm{sep}_q(\Psi_{|U}, \Phi) - \epsilon - \frac{C_\sigma}{\sqrt[2q]{n}}. \end{aligned}$$

□

B Proof of Corollary 3.3

We will need the following simple linear algebraic lemma.

Lemma B.1. *Let $v_1, \ldots, v_k$ be a set of orthonormal vectors in a vector space $\mathbb{R}^p$. Let G be the subspace spanned by $v_1, \ldots, v_k$. Let H be a subspace of $\mathbb{R}^p$. Then $\forall g \in G$ and $h \in H$ such that $\|g\| = \|h\| = 1$, we have that*

$$\left(g^\top h\right)^2 \leq \sum_{j=1}^{k} \left\|\mathrm{Proj}_H(v_j)\right\|^2.$$

Proof. We first remember from basic linear algebra that for any unit vectors $v \in \mathbb{R}^p$ and $h \in H$, we have $|v^\top h| \leq |\mathrm{Proj}_H(v)|$. Then we can write $g = \sum_{j=1}^{k} \alpha_j v_j$, where $\sum_j \alpha_j^2 = 1$ as $\|g\| = 1$ and $v_1, \ldots, v_k$ form an orthonormal basis of G. Combining these we get

$$\begin{aligned}
(g^\top h)^2 &= \left(\sum_{j=1}^{k} \alpha_j v_j^\top h\right)^2 \\
&\leq \left(\sum_{j=1}^{k} \alpha_i^2\right)\left(\sum_{j=1}^{k} (v_j^\top h)^2\right) && \text{(by Cauchy-Schwartz)} \\
&\leq \sum_{j=1}^{k} \left\|\mathrm{Proj}_H(v_j)\right\|^2.
\end{aligned}$$

This completes the proof. □

We can now prove Corollary 3.3.

Proof of Corollary 3.3. To prove this we will use Theorem 3.2 along with the trivial linear algebraic Lemma B.1. We prove by contradiction. Suppose a orthonormal set $\widehat{u}_1, \ldots, \widehat{u}_{k+1}$ exists satisfying the hypothesis:

$$d_{W_2}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{u}_j^\top X_i}, \Phi\right) \geq \sqrt{1 - 4\delta^2}\, d_{(2)}(\Psi, \Phi) + \epsilon + \frac{C_\sigma}{\sqrt[4]{n}}.$$

We can invoke Theorem 3.2 to get with high probability,

$$\left\|\mathrm{Proj}_{U^\perp}(\widehat{u}_j)\right\| \leq 2\delta \frac{d_{(2)}(\Psi, \Phi)}{\sqrt{d_{(2)}(\Psi, \Phi)^2 - d_{(2)}(\Psi_{|U^\perp}, \Phi)^2}}.$$

Then let $G = \mathrm{Span}\{\widehat{u}_1, \ldots, \widehat{u}_{k+1}\}$. As As $\dim(G) + \dim(U^\perp) = p + 1$, there exists a non-zero vector $s \in G \cap H = U^\perp$ such that $\|s\| = 1$. Invoking Lemma B.1 with $H = U^\perp$ and $g = h = s$, we get the contradiction

$$1 = \left(s^\top s\right)^2 \leq \sum_{j=1}^{k+1} \left\|\mathrm{Proj}_H(\widehat{u}_j)\right\|^2 \leq \frac{4(k+1)\delta^2 d_{(2)}(\Psi, \Phi)^2}{d_{(2)}(\Psi, \Phi)^2 - d_{(2)}(\Psi_{|W}, \Phi)^2} < 1.$$

This completes the proof. □

C Proof of Lemma 5.1

To prove Lemma 5.1, we need a bound on the concentration function of sub-Gaussian random variables. For a Borel set A, let A^r denote the r-fattening of A:

$$A_r = \{x : d(x, A) < r\}.$$

Let μ be a probability measure on $\mathbb{R}$. Let

$$\alpha_\mu(r) = \sup_{A:\mu(A)\geq 1/2} 1 - \mu(A^r), r > 0,$$

denote the concentration function of μ.

Lemma C.1. *Let μ be a σ-sub-Gaussian probability measure. Then there exist absolute constants $C, c > 0$ such that $\alpha_\mu(r) \leq Ce^{-\frac{cr^2}{\sigma^2}}$ for all $r > 0$.*

Proof. Without loss of generality, we may assume that $\sigma = 1$. Choose r_0 such that $\mu((-r_0, r_0)^c) < \frac{1}{2}$. Then any A such that $\mu(A) > \frac{1}{2}$ must intersect $(-r_0, r_0)$, for otherwise one would get $\mu(A) \leq \mu((-r_0, r_0)^c) < \frac{1}{2}$. Take $x_0 \in A \cap (-r_0, r_0)$. Then one must have

$$(-r, r) \subseteq x_0 + (-(r_0 + r), r_0 + r) \subseteq A^{r_0+r}.$$

Now, by sub-Gaussianity, there exist constants $C_1, c_1 > 0$ such that $\mu((-r, r)^c) \leq C_1 e^{-c_1 r^2}$ for all $r > 0$. Therefore

$$1 - \mu(A^{r+r_0}) \leq \mu((-r, r)^c) \leq C_1 e^{-c_1 r^2} \leq Ce^{-c(r+r_0)^2},$$

where the last inequality holds for some constants $C, c > 0$ for all large enough r, say $r > r_1$. (For example, one can take $c = \frac{1}{2}c_1, C = C_1 e^{\frac{1}{2}Br_0^2}$ and $r_1 = 2r_0$.) Thus for all $r > r_0 + r_1$, we have that $\alpha_\mu(r) \leq Ce^{-cr^2}$.

We can always increase the constant C so that one has $\sup_{r\in(0,r_0+r_1]} \alpha_\mu(r) \leq Ce^{-c(r_0+r_1)^2}$. Then for any $r \leq r_0 + r_1$,

$$\alpha_\mu(r) \leq \sup_{r\in(0,r_0+r_1]} \alpha_\mu(r) \leq Ce^{-c(r_0+r_1)^2} \leq Ce^{-cr^2}.$$

We conclude that there exist absolute constants $C, c > 0$ such that $\alpha_\mu(r) \leq Ce^{-cr^2}$ for all $r > 0$. □

We are now ready to prove Lemma 5.1.

Proof of Lemma 5.1. The proof is the same as the proof of Theorem 7.1 in Bobkov and Ledoux (2019), except that we replace their log-Sobolev assumption on μ with a sub-Gaussianity assumption, which yields a stronger bound on the concentration function as in Lemma C.1, which in turns gives us a tail bound of the form

$$\mathbb{P}\big(|d_{W_q}(\mu_n, \mu) - \mathbb{E}d_{W_q}(\mu_n, \mu)| \geq t\big) \leq C \exp\left(-\frac{c\, n^{2/\max\{q,2\}} t^2}{\sigma^2}\right)$$

for some absolute constants $C, c > 0$. □