Aligning Embeddings and Geometric Random Graphs: Informational Results and Computational Approaches for the Procrustes-Wasserstein Problem

Mathieu Even, Luca Ganassali, Jakob Maier, Laurent Massoulié

Aligning Embeddings and Geometric Random Graphs: Informational Results and Computational Approaches for the Procrustes-Wasserstein Problem

Mathieu Even
DI ENS, CRNS, PSL University, INRIA Paris

Luca Ganassali
Université Paris-Saclay, LMO

Jakob Maier
D.I ENS, CRNS, PSL University, INRIA Paris

Laurent Massoulié
D.I ENS, CRNS, PSL University, INRIA, MSR-INRIA Joint Centre, Paris

Abstract

The Procrustes-Wasserstein problem consists in matching two high-dimensional point clouds in an unsupervised setting, and has many applications in natural language processing and computer vision. We consider a planted model with two datasets X, Y that consist of n datapoints in $\mathbb{R}^d$, where Y is a noisy version of X, up to an orthogonal transformation and a relabeling of the data points. This setting is related to the graph alignment problem in geometric models. In this work, we focus on the euclidean transport cost between the point clouds as a measure of performance for the alignment. We first establish information-theoretic results, in the high ($d \gg \log n$) and low ($d \ll \log n$) dimensional regimes. We then study computational aspects and propose the 'Ping-Pong algorithm', alternatively estimating the orthogonal transformation and the relabeling, initialized via a Franke-Wolfe convex relaxation. We give sufficient conditions for the method to retrieve the planted signal after one single step. We provide experimental results to compare the proposed approach with the state-of-the-art method of Grave et al. [2019].

1 Introduction

Finding an alignment between high dimensional vectors or across two point clouds of embeddings has been the focus of recent threads of research and has a variety of applications in computer vision, such as inferring scene geometry and camera motion from a stream of images [Tomasi and Kanade, 1992], as well as in natural language processing such as automatic unsupervised translation [Rapp, 1995, Fung, 1995].

Many practical algorithms proposed for this task view this problem as minimizing the distance across distributions in $\mathbb{R}^d$. Some approaches are based e.g. on optimal transport and Gromov-Wasserstein distance Alvarez-Melis and Jaakkola [2018] or adversarial learning [Zhang et al., 2017, Conneau et al., 2018]. Another line of methods adapt the iterative closest points procedure (ICP) – originally introduced in Besl and McKay [1992] for 3-D shapes – to higher dimensions Hoshen and Wolf [2018]. Another recent contribution is that of Grave et al. [2019], where a method is proposed to jointly learn an orthogonal transformation and an alignment between two point clouds by alternating the objectives in the corresponding minimization problem.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To formalize this problem, we consider a Gaussian model in which both datasets $X, Y \in \mathbb{R}^{d\times n}$ (or two point clouds of n datapoints in $\mathbb{R}^d$) are sampled as follows. First, $X = (x_1, \dots, x_n)$ is a collection of i.i.d. $\mathcal{N}(0, I_d)$ Gaussian vectors, and $Y = (y_1, \dots, y_n)$ is a noisy version of $X = (x_1, \dots, x_n)$, up to an orthogonal transformation $Q^\star$ and a relabeling $\pi^\star : [n] \to [n]$ of the data points, that is:

$$\forall i \in [n]\,, \quad y_i = Q^\star x_{\pi^\star(i)} + \sigma z_i\,, \tag{1}$$

or, in matrix form:

$$Y = Q^\star X (P^\star)^\top + \sigma Z\,,$$

where $Z = (z_1, \dots, z_n) \in \mathbb{R}^{d\times n}$ is also made of i.i.d. $\mathcal{N}(0, I_d)$ Gaussian vectors, $P^\star$ is the permutation matrix associated with some permutation $\pi^\star$, and $\sigma > 0$ is the noise parameter. Recovering (in some sense that will be made precise in the sequel) the (unknown) permutation $\pi^\star$ and orthogonal transformation $Q^\star$ defines the *Procrustes-Wasserstein problem* (sometimes abbreviated as PW in the sequel), which will be the focus of this study.

The practical approaches previously mentioned have shown good empirical results and are often scalable to large datasets. However, they suffer from a lack of theoretical results to guarantee their performance or to exhibit regimes where they fail. Model (1) described here above appears to be the simplest one to obtain such guarantees. We are interested in pinning down the *fundamental* limits of the Procrustes-Wasserstein problem, hence providing an ideal baseline for any computational method to be compared to, before delving into computational aspects. Our contributions are as follows:

- (i) We define a planted model for the Procrustes-Wassertein problem and discuss the appropriate choice of metrics to measure the performance of any estimator. Based on these metrics, we establish[1]:
  - $(i.a)$ information-theoretic results in the high-dimensional $d \gg \log n$ regime which was not explored before for this problem;
  - $(i.b)$ new information-theoretic results in the low-dimensional regime ($d \ll \log n$) for our metric of performance (the L^2 transport cost), which substantially differ from those obtained in Wang et al. [2022] for the overlap.
- (ii) We study computational aspects and propose the 'Ping-Pong algorithm', alternatively estimating the orthogonal transformation and the relabeling, initialized via a Franke-Wolfe convex relaxation. This method is quite close to that proposed in Grave et al. [2019] although the alternating part differs. We give sufficient conditions for the method to retrieve the planted signal after one single step.
- (iii) Finally, we provide experimental results to compare the proposed approach with the state-of-the-art method of Grave et al. [2019].

1.1 Discussion and related work

One can check that under the above model (1), the maximum likelihood (ML) estimators of $(P^\star, Q^\star)$ given (X, Y) is given by:

$$(\hat{P}, \hat{Q}) \in \operatorname*{arg\,min}_{(P,Q)\in\mathcal{S}_n\times\mathcal{O}(d)} \frac{1}{n}\|XP^\top - Q^\top Y\|_F^2 = \operatorname*{arg\,min}_{(P,Q)\in\mathcal{S}_n\times\mathcal{O}(d)} \|QX - YP\|_F^2\,, \tag{2}$$

which is strictly equivalent[2] to the formulation of the non-planted problem of Grave et al. [2019]. Exactly solving the joint optimization problem (2) is non convex and difficult in general. However, if $P^\star$ is known then (2) boils down to the following *orthogonal Procrustes problem*:

$$\hat{Q} \in \operatorname*{arg\,min}_{Q\in\mathcal{O}(d)} \frac{1}{n}\|XP^\star - Q^\top Y\|_F^2\,, \tag{3}$$

which has a simple closed form solution given by $\hat{Q} = UV^\top$ where $USV^\top$ is the singular value decomposition (SVD) of $Y(XP^\star)^\top$ (see Schönemann [1966]). Conversely, when $Q^\star$ is known, (2) amounts to the following *linear assignment problem* (LAP in the sequel):

$$\operatorname*{arg\,min}_{P\in\mathcal{S}_n} \frac{1}{n}\|XP^\top - Q^\star Y\|_F^2 = \operatorname*{arg\,max}_{P\in\mathcal{S}_n} \frac{1}{n}\langle XP^\top, Q^\star Y\rangle, \tag{4}$$

[1] this very dichotomy ($d \gg \log n$ versus $d \ll \log n$) is fundamental in high dimensional statistics and not a mere artifact of our rationale; see Remark 1.

[2] their matrices X and Y are the transposed versions of ours.

Reference	Setting	Metrics	Regime	Condition
Kunisky and Niles-Weed [2022]	$Q^\star = I_d$	ov for $P^\star$	$d \ll \log n$	$\sigma \ll n^{-2/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$ $\sigma \ll n^{-1/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = o(1)$
			$d \sim a \log n$	$\sigma < (e^{4/a} - 1)^{-1/2}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$ $\sigma < ((2e^{1/a} - 1)^2 - 1)^{-1/2}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = o(1)$
			$d \gg \log n$	$\sigma < (1/2 - \varepsilon)(d/\log n)^{1/2}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$
Wang et al. [2022]	$A = X^\top X, B = Y^\top Y$	ov for $P^\star$	$d \ll \log n$	$\sigma \ll n^{-2/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$ $\sigma \ll n^{-1/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = o(1)$
This paper	X, Y from (1)	c^2 for $P^\star$, ℓ^2 for $Q^\star$	$d \ll \log n$ $d \gg \log n$	$\sigma \ll d^{-1/2}$ for $c^2(\hat{\pi}, \pi^\star) = \ell^2(\hat{Q}, Q^\star) = o(d)$ $\sigma \ll 1$ for $c^2(\hat{\pi}, \pi^\star) = \ell^2(\hat{Q}, Q^\star) = o(d)$

Table 1: Summary of previous informational results, together with the ones in this paper

which can be solved in polynomial time, e.g. in cubic time by the celebrated Hungarian algorithm [Kuhn, 1955], or more efficiently at the price of regularizing the objective and using the celebrated Sinkhorn algorithm [Cuturi, 2013].

Previous results when $Q^\star$ is known. As seen above, when $Q^\star$ is known (assume e.g. $Q^\star = I_d$), the Procrustes-Wasserstein problem reduces to a simpler objective, that of aligning Gaussian databases. This problem has been studied by Dai et al. [2019, 2023] in the context of feature matching. Kunisky and Niles-Weed [2022] study the same problem as a geometric extension of planted matching and establish state-of-the-art statistical bounds in the Gaussian model in the low-dimensional ($d \ll \log n$), logarithmic ($d \sim a \log n$) and high-dimensional ($d \gg \log n$) regimes. In particular, they show that exact recovery is feasible in the logarithmic regime $d \sim a \log n$ if $\sigma^2 < \frac{1}{e^{4/a}-1}$, and in the high-dimensional regime if $\sigma^2 < (1/4 - \varepsilon)\frac{d}{\log n}$. Note that in this problem, there is no computational/statistical gap since the LAP is always solvable in polynomial time.

Geometric graph alignment. Strongly connected to the Procrustes-Wassertein problem is the topic of graph alignment where the instances come from a geometric model. Wang et al. [2022] investigate this problem for complete weighted graphs. In their setting, given a permutation $\pi^\star$ on $[n]$ and n i.i.d. pairs of correlated Gaussian vectors $(X_{\pi^\star(i)}, Y_i)$ in $\mathbb{R}^d$ with noise parameter σ, they observe matrices $A = X^\top X$ and $B = Y^\top Y$ (i.e all inner products $\langle X_i, X_j \rangle$ and $\langle Y_i, Y_j \rangle$) and are interested in recovering the hidden vertex correspondence $\pi^\star$. The maximum likelihood estimator in this setting writes

$$\operatorname*{arg\,min}_{P \in \mathcal{S}_n} \frac{1}{n} \|X^\top X P - P Y^\top Y\|_F^2 = \operatorname*{arg\,max}_{P \in \mathcal{S}_n} \frac{1}{n} \langle P^\top X^\top X P, Y^\top Y \rangle \,, \tag{5}$$

which is an instance of the *quadratic assignment problem* (QAP in the sequel), known to be NP-hard in general, as well as some of its approximations [Makarychev et al., 2014]. In fact, we have the following informal equivalence (see Appendix A for a proof):

Lemma 1 (Informal)**.** *PW and geometric graph alignement are equivalent, that is, one knows how to (approximately) solve the former iff they know how to (approximately) solve the latter.*

Wang et al. [2022] focus on the low-dimensional regime $d = o(\log n)$, where geometry plays the most important role (see Remark 1). They prove that exact (resp. almost exact) recovery of $\pi^\star$ is information-theoretically possible as soon as $\sigma = o(n^{-2/d})$ (resp. $\sigma = o(n^{-1/d})$). They conduct numerical experiments which suggest good performance of the celebrated Umeyama algorithm [Umeyama, 1988], which is confirmed by a follow-up work by Gong and Li [2024] analyzing the Umeyama algorithm (which is polynomial time in the low dimensional regime $d = o(\log n)$) in the same setting and shows that it achieves exact (resp. almost exact) recovery of $\pi^\star$ if $\sigma = o(d^{-3} n^{-2/d})$ (resp. $\sigma = o(d^{-3} n^{-1/d})$), hence coinciding with the information thresholds up to a $\mathrm{poly}(d)$ factor. However, their algorithm is of time complexity at least $\Omega(2^d n^3)$, which is not polynomial in d. This is why we do not include this method in our baselines.

We emphasize that our results clearly depart from those obtained in Wang et al. [2022] and Gong and Li [2024], because (i) we are also interested in the high dimensional case $d \gg \log n$, and (ii) we work with a different performance metric which provides less stringent conditions for the recovery to be feasible, see Section 1.2. A summary of previous informational results together with ours (see also Section 2) is given in Table 1.

On the orthogonal transformation $Q^\star$. Generalizing the standard linear assignment problem, our model described above in (1) introduces an additional orthogonal transformation $Q^\star$ across the datasets. This orthogonal transformation can be motivated in the context of aligning embeddings in a

high-dimensional space: indeed, the task of learning embeddings is often agnostic to orientation in the latent space. In other words, two point clouds may represent the same data points while having different global orientations. Hence, across different data sets, learning this orientation shift is crucial in order to compare (or align) the point clouds. As an illustration of this fact, Xing et al. [2015] provides empirical evidence that orthogonal transformations are particularly adapted for bilingual word translation.

Discussing the method proposed in Grave et al. [2019]. We conclude this introduction by discussing the work of Grave et al. [2019]. Their proposed algorithm is as follows. At each iteration t, given a current estimate Q_t of the orthogonal transformation, we sample mini-batches X_t, Y_t of same size b and find the optimal matching P_t between $Y_t Q_t^\top$ and X_t, via solving a linear assignment problem of size b. This matching P_t in turn helps to refine the estimation of the orthogonal transformation via a projected gradient descent step, and the procedure repeats. This method has the main advantage to be scalable to very large datasets and to perform well in practice ; however, no guarantees are given for this method, and in particular the mini-batch step which can justifiably raise some concerns. Indeed, since $X_t = (x_{t,j})_{j\in[b]}$ and $Y_t = (y_{t,j})_{j\in[b]}$ are chosen independently, if $b \ll \sqrt{n}$ it is likely that for any matching π_t the pairs $(x_{t,j}, y_{t,\pi_t(j)})$ always correspond to disjoint pairs, and thus aligning $Y_t Q_t^\top$ and X_t does not reveal any useful information about the true $P^\star$ – this is even more striking when the data is non-isotropic.

1.2 Problem setting and metrics of performance

Notations. We denote by $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ the fact that X follows a Gaussian distribution in $\mathbb{R}^d$ of mean μ and covariance matrix Σ. If $\mu = 0$ and $\Sigma = I_d$, variable X is called *standard Gaussian*. We denote by $\mathcal{O}(d)$ the orthogonal group in dimension d, and by $\mathcal{S}_n$ the group of permutations on $[n]$. Throughout, $\|\cdot\|$ and $\langle\cdot\rangle$ are is the standard euclidean norm and scalar product on $\mathbb{R}^d$, and $\|\cdot\|_F$ and $\|\cdot\|_{op}$ are respectively the Frobenius matrix norm and the operator matrix norm. The spectral radius of a matrix A is denoted $\rho(A)$. In all the proofs, quantities c_i where i is an integer are unspecified constants which are universal, that is independent from the parameters. Finally, all considered asymptotics are when $n \to \infty$. Note that d also depends on n. An event is said to hold *with high probability (w.h.p.)* if its probability tends to 1 when n goes to ∞.

Problem setting and performance metrics. We work with the planted model as introduced in (1) and recall that our goal is to recover the permutation $\pi^\star$ and the orthogonal matrix $Q^\star$ from the observation of X and Y.

Performance metrics. Previous works measure the performance of an estimator $\hat{\pi}$ of a planted relabeling $\pi^\star$ via the overlap:

$$\mathrm{ov}(\pi, \pi') := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\hat{\pi}(i)=\pi'(i)\}}\,, \tag{6}$$

defined for any two permutations π, π'. This is an interesting metric when we have no hierarchy in the errors, that is when only the true match is valuable, and all wrong matches cost the same. However, this discrete measure does not take into account the underlying geometry of the model. A performance metric which is more adapted to our setting is the L^2 transport cost between the point clouds. The natural intuition is that a mismatch is less costy if it corresponds to embeddings which are in fact close in the underlying space. We define

$$c^2(\pi, \pi') = \frac{1}{n}\sum_{i=1}^{n} \left\|x_{\pi(i)} - x_{\pi'(i)}\right\|^2,$$

for any two permutations π, π'. Note that this cost can also be written in matrix form as $c^2(P, P') = \left\|(P - P')X^\top\right\|_F^2$. From this form it is clear that, as stated before, $c^2(P, P')$ is nothing but the euclidean transport cost for aligning $XP^\top$ onto $X(P')^\top$. Note that these two measures, ov and c^2, are also well-defined[3] when P, P' are more general (and in particular when they are bistochastic matrices). Finally, we measure the performance for the estimation of $Q^\star$ via the Frobenius norm:

$$\ell^2(Q, Q') = \|Q - Q'\|_F^2\,,$$

[3]for the overlap, one could extend its definition using that $\mathrm{ov}(P, P') = \langle P, P'\rangle$.

defined for any two orthogonal matrices Q, Q'.

Comparison between metrics. For a Haar-distributed matrix Q on $\mathcal{O}(d)$, we have that $\mathbb{E}\left[\ell^2(Q, Q^\star)\right] = 2d$, while for π sampled uniformly from the set of all permutations, we have $\mathbb{E}\left[c^2(\hat{\pi}, \pi^\star)\right] = 2d(1 - 1/n)$ and $\mathbb{E}\left[\mathrm{ov}(\hat{\pi}, \pi^\star)\right] = 1/n$. Hence, some estimators $\hat{\pi}, \hat{Q}$ of $\pi^\star, Q^\star$ will perform well in our metrics if they can achieve $\ell^2(Q, Q^\star) \leqslant \varepsilon d$, and $c^2(\pi, \pi^\star) \leqslant \varepsilon d$ for some small (possibly vanishing) $\varepsilon > 0$.

Depending on dimension d, similarity measures given by c^2 and the overlap can behave differently or coincide. In the case where d is small, and thus plays a very important role, ov and c^2 have very different behaviors, and lead to very different results. In particular, there is a wide regime in which inferring $\pi^\star$ for the overlap sense is impossible, but reachable in the transport cost sense, see Section 2.

For any fixed permutation π, we have that $\mathbb{E}\left[c^2(\pi, \pi^\star)\right] = 2d(1 - \mathrm{ov}(\pi, \pi^\star))$, where the mean is taken with respect to the randomness of X. We also have the basic deterministic inequality

$$c^2(\pi, \pi^\star) \leqslant (1 - \mathrm{ov}(\pi, \pi^\star)) \times \sup_{(i,j)\in[n]^2} \|x_i - x_j\|^2 .$$

Thus, as long as $\sup_{(i,j)\in[n]^2} \|x_i - x_j\|^2 = O(d)$, an estimator $\hat{\pi}$ with good overlap ($1 - \mathrm{ov}(\pi, \pi^\star) \leqslant \varepsilon$) also has a good c^2 cost ($c^2(\pi, \pi^\star) = O(\varepsilon d)$). However, this required control $\sup_{(i,j)\in[n]^2} \|x_i - x_j\|^2 = O(d)$ only holds as long as $d \gg \log(n)$.

The blessing of large dimensions lead to an equivalence between the discrete metric ov, and the continuous transport metric c^2. We gather several important points highlighting the dichotomy between small and large dimensions for our problem in the following remark.

Remark 1. *On the blessings of large dimensions for our problem:*

1. *For any open ball $\mathcal{B}$ of radius $\varepsilon > 0$, denoting $\mathcal{X} = \{x_i, i \in [n]\}$, we have that $\mathbb{P}(\mathcal{B} \cap \mathcal{X} = \varnothing) \to 1$ if $d \gg \log(n)$, while if $d \ll \log(n)$ then for all $M > 0$, $\mathbb{P}(|\mathcal{B} \cap \mathcal{X}| \geqslant M) \to 1$. In small dimensions, any fixed non-empty ball will contain infinitely many points of $\mathcal{X}$ as n increases, while in large dimensions these points are separated and any fixed ball will contain no such points w.h.p.*
2. *For $d \gg \log(n)$, matrix $X/\sqrt{d}$ satisfies the* restricted isometry property *[Candès, 2008].*
3. *For $d \gg \log(n)$, the overlap and the transport cost metrics are equivalent: there exist numerical constants $\alpha, \beta > 0$ such that w.h.p., for all permutation matrices π, π', $\alpha c^2(\pi, \pi') \leqslant 2d(1 - \mathrm{ov}(\pi, \pi')) \leqslant \beta c^2(\pi, \pi')$.*

Organization of the rest of the paper. Section 2 is dedicated to our informational results, giving their essential content as well as the main ideas on the proofs. We next discuss in Section 3 some computational results, introducing the Ping-Pong algorithm, and presenting our numerical experiments.

2 Informational results

The substantial theoretic part of the paper stands in the informational results obtained for the Procrustes-Wasserstein problem which we describe hereafter.

2.1 High dimensions

In the high-dimensional case when $\log n \ll d$ (and $d \log d \ll n$), our results – Theorem 1 below – imply that if $\sigma \to 0$ then the ML estimators defined in (2) satisfy w.h.p.

$$\mathrm{ov}(\pi^\star, \hat{\pi}) = 1 - o(1),\ c^2(\hat{P}, P^\star) = o(d),\ \text{and } \ell^2(\hat{Q}, Q^\star) = o(d),$$

that is one can infer $\pi^\star$ and $Q^\star$ almost exactly, for all introduced metrics, as soon as $\sigma \to 0$.

Note that this is the first result in the high-dimensional regime for the Procrustes Wassertein problem: Kunisky and Niles-Weed [2022] also considered this regime but only for the LAP problem (that

is recovering $\pi^\star$ when $Q^\star$ in known), and the only existing results for geometric graph alignment Wang et al. [2022], Gong and Li [2024] do not consider this high dimensional case. Our result thus complements the existing picture and shows that almost exact recovery is feasible under the loose assumption $\sigma \to 0$, in the c^2 and the overlap sense, since these metrics are equivalent in large dimensions (see Remark 1). Our result is in fact more specific and only requires $d \geqslant 2\log n$. We prove the following Theorem:

Theorem 1. *Assume that $d \geqslant 2\log n$. There exists universal constants $c_1, c_2, c_3 > 0$ so that for n large enough, with probability $1 - o(1)$, the ML estimators defined in* (2) *satisfy*

$$\mathrm{ov}(\pi^\star, \hat{\pi}) \geqslant 1 - \max\left(60\sigma^2, c_1\frac{d}{n}, c_2\frac{\log n}{d\log d}\right), \tag{7}$$

and

$$\frac{\ell^2(Q^\star, \hat{Q})}{2d} \leqslant c_1\frac{d}{n} + c_2\sigma^2 + c_3\max\left(\frac{d\log n}{n}, \sqrt{\frac{\log n}{n}}\right). \tag{8}$$

The proof of Theorem 1 is detailed in Appendix C and builds upon controlling the probability of existence of a certain subset of indices $\mathcal{K}(\hat{Q}, \hat{\pi}, Q^\star)$ of vectors with prescribed properties in order to show that $\pi^\star$ can be recovered. We apply standard concentration inequalities to control the previous probability. The $d \geqslant 2\log(n)$ assumption is crucial here since it allows the union bound over $\mathcal{S}_n$ to work.

2.2 Low dimensions

In the low-dimensional case when $d \ll \log n$, Theorem 2 below implies that if $\sigma = o(d^{-1/2})$ then there exist estimators $\hat{\pi}, \hat{Q}$ that satisfy w.h.p.

$$c^2(\hat{P}, P^\star) = o(d), \text{ and } \ell^2(\hat{Q}, Q^\star) = o(d)\,,$$

that is, one can approximate $\pi^\star$ (in the c^2 sense *only*) and $Q^\star$ as soon as $\sigma = o(d^{-1/2})$. This is of course to be put in contrast with the previous results on geometric graph alignment in this low-dimensional regime: for almost exact recovery in Wang et al. [2022] *in the overlap sense*, we need $\sigma = o(n^{-1/d})$, which is far more restrictive than $\sigma = o(d^{-1/2})$ as soon as $d\log(d) < \log n$, that is nearly in the whole low dimensional regime when $d \ll \log n$. In particular, since the rates of Wang et al. [2022] are sharp when d is of constant order, in order to approximate $\pi^\star$ in the overlap sense it is necessary to have σ to decreasing polynomially (at rate $1/n^{1/d}$) to 0, whereas approximating $\pi^\star$ in the transport cost sense requires only $\sigma = o(1)$.

There is no contradiction here, since we recall that the c^2 metric and the overlap are not equivalent in small dimensions: let us give a few more insights on this. This scaling $n^{-1/d}$ comes from the fact that in small dimensions, points of the dataset are close to each other, and the order of magnitude between some x_i and its closest point in the dataset scales exactly as $n^{-1/d}$: if the noise is smaller than this quantity, one should be able to recover the planted permutation. However, when it comes to considering the c^2 metric, matching i with j such that $\|x_i - x_j\|^2 \ll d$ is sufficient, thus suggesting that recovering a permutation with small c^2 cost and recovering $Q^\star$ with small Frobenius norm error should be achievable even with large σ (*i.e.*, that does no tend to 0 as n increases).

Our main theorem for low dimensions is as follows.

Theorem 2. *Let $\delta_0 \in (0,1)$. There exist estimators $\hat{\pi}, \hat{Q}$ of $\pi^\star, Q^\star$ such that if for some numerical constants $C_1, C_2 > 0$ we have $\sigma \leqslant C_1\delta_0^2 d^{-1/2}$ and $\log(n) \geqslant C_2 d\log(1/\delta_0)$, then:*

$$\frac{c^2(\hat{\pi}, \pi^\star)}{2d} \leqslant \delta_0 \quad \textit{and} \quad \frac{\ell^2(\hat{Q}, Q^\star)}{2d} \leqslant \delta_0\,.$$

A refined version of Theorem 2, namely Theorem 3, is proved in Appendix D. We emphasize that the estimators considered in Theorem 2 are *not* the ML estimators: recall that the strategy to analyse the former as rolled out for Theorem 1 required the union bound over $\mathcal{S}_n$ to work. This drastically fails when $d \ll \log(n)$. Hence, we will instead focus on an estimator that takes advantage of the fact that d is small, and show that even in small dimensions, the signal-to-noise ratio σ does not need to decrease with n.

Let us first describe the intuition behind the estimators $\hat{\pi}, \hat{Q}$. When $d = 1$, $Q^\star = \pm 1$ and a simple strategy to recover $Q^\star$ is to count the number $N_+(\mathcal{X}), N_-(\mathcal{X})$ (resp. $N_+(\mathcal{Y}), N_-(\mathcal{Y})$) of positive and negative x_i (resp. positive and negative y_j): if $N_+(\mathcal{X})$ and $N_+(\mathcal{Y})$ are close, then we output $\hat{Q} = +1$, whereas if $N_+(\mathcal{X})$ and $N_-(\mathcal{Y})$ are close, then $\hat{Q} = -1$. In dimension d, an analog strategy can be applied at the cost of looking in all relevant directions, and the number of such directions is exponentially big in d. Our strategy is thus as follows. We compute the number of points that lie in a given cone $\mathcal{C}(u, \delta)$ of given angle δ and direction u. Then, we estimate $Q^\star$ by the orthogonal transformation $\hat{Q}$ which makes the number of y_j in $\mathcal{C}(u, \delta)$ closest to the number of x_j in $\mathcal{C}(\hat{Q}u, \delta)$, for any direction u. Note that this approach heavily relies on the small dimension assumption $d \ll \log n$: in this case, for any constant δ, all theses cones contain w.h.p. a large number of points (tending to ∞ with n), which does not hold anymore when $d \gg \log n$.

For $\delta > 0$ and $u \in \mathcal{S}^{d-1}$, let $\mathcal{C}(u, \delta) := \left\{ v \in \mathbb{R}^d \,|\, \langle u, v \rangle \geqslant (1 - \delta)\|v\| \right\}$ be the cone of angle δ centered around u. Let $\mathcal{X} := \{x_i, i \in [n]\}$, $\mathcal{Y} := \{y_i, i \in [n]\}$. We now introduce the following sets, for some $\kappa > 0$:

$$\mathcal{C}_\mathcal{X}(u, \delta) := \mathcal{X} \cap \mathcal{C}(u, \delta) \cap \mathcal{B}(0, 1/\kappa)^C \quad \text{and} \quad \mathcal{C}_\mathcal{Y}(u, \delta) := \mathcal{Y} \cap \mathcal{C}(u, \delta) \cap \mathcal{B}(0, \sqrt{1 + \sigma^2}/\kappa)^C \,,$$

where $\mathcal{B}(0, r)^C$ contains all vectors in $\mathbb{R}^d$ of norm larger than or equal to r. The role of $\kappa > 0$ is to prevent side effects: indeed, since the cones are centered at the origin, points that are too close to 0 fall into cones with arbitrary directions and are not informative for the statistics we want to compute.

Now, for some $p \geqslant 1$ and directions $u_1, \dots, u_p \in \mathcal{S}^{d-1}$ to be set later, we define the following *conical alignment loss*:

$$\forall Q \in \mathcal{O}(d)\,, \quad F(Q) = \frac{1}{p} \sum_{k=1}^{p} \left(|\mathcal{C}_\mathcal{X}(Qu_k, \delta)| - |\mathcal{C}_\mathcal{Y}(u_k, \delta)| \right)^2 . \tag{9}$$

The estimator $\hat{Q}$ in Theorem 2 is then defined as a minimizer of the conical alignment loss over a finite set $\mathcal{N} \subseteq \mathcal{O}(d)$:

$$\hat{Q} \in \operatorname*{arg\,min}_{Q \in \mathcal{N}} F(Q)\,,$$

where $\mathcal{N}$ will further be some ε-net of $\mathcal{O}(d)$, while $\hat{\pi}$ is then obtained by a LAP as in (10).

2.3 From $P^\star$ to $Q^\star$ and vice versa

In our proofs, we often prove that one of the estimators $\hat{P}$ or $\hat{Q}$ performs well in order to deduce that both perform well. This is thanks to the following two results, proved in Appendix B.1 and B.2.

Lemma 2 (From $\hat{Q}$ to $\hat{P}$)**.** *Let $\delta \in (0, 1/2)$. Assume that there exists $\hat{Q}$ that is $\sigma(\{x_1, \dots, x_n, y_1, \dots, y_n\})$-measurable such that $\ell^2_{\mathrm{ortho}}(\hat{Q}, Q^*) := \|\hat{Q} - Q^*\|^2 \leqslant \delta d$. There exist constants $C_1, C_2, C_3 > 0$ such that with probability at least $1 - 2e^{-nd} - 2e^{-(d^2+\sqrt{n})}$,*

$$\hat{\pi} \in \operatorname{argmin}_{\pi \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} \left\| x_{\pi(i)} - \hat{Q}^\top y_i \right\|^2 , \tag{10}$$

that can be computed in polynomial time (complexity $O(n^3)$) as the solution of a LAP, satisfies:

$$\frac{c^2(\hat{\pi}, \pi^\star)}{d} \leqslant C_1 \delta + C_2 \sigma^2 + C_3 \max\left(\frac{d \ln(1/\delta)}{n}, \sqrt{\frac{\ln(1/\delta)}{n}} \right) .$$

Lemma 3 (From $\hat{P}$ to $\hat{Q}$)**.** *Let $\delta \in (0, 1/2)$. Assume that there exists $\hat{\pi}$ that is $\sigma(\{x_1, \dots, x_n, y_1, \dots, y_n\})$-measurable such that $c^2(\hat{\pi}, \pi^\star) \leqslant \delta d$. Let $\hat{Q}$ be the solution to the following optimization problem: There exist constants $C_1, C_2, C_3 > 0$ such that with probability at least $1 - 2e^{-nd} - 2e^{-(d^2+\sqrt{n})}$,*

$$\hat{Q} \in \operatorname*{arg\,min}_{Q \in \mathcal{O}(d)} \frac{1}{n} \sum_{i=1}^{n} \left\| x_{\hat{\pi}(i)} - Q^\top y_i \right\|^2 , \tag{11}$$

that can be computed in closed form with an SVD of $XY^\top$, satisfies:

$$\frac{\ell^2_{\mathrm{ortho}}(\hat{Q}, Q^\star)}{d} \leqslant C_1 \delta + C_2 \sigma^2 + C_3 \max\left(\frac{d \ln(1/\delta)}{n}, \sqrt{\frac{\ln(1/\delta)}{n}} \right) .$$

3 Computational aspects

The estimators provided this far in Section 2, namely the joint minimization in P and Q in (2) and the minimizer of the conical alignment loss in (9) are of course not poly-time in general. In this section, we are interested in computational aspects of the problem.

3.1 Convex relaxation and Ping-Pong algorithm

Estimating $P^\star$ can be made via solving the QAP (5), that can be convexified into the *relaxed quadratic assignment problem* (relaxed QAP):

$$\hat{P}_{\mathrm{relaxed}} \in \operatorname*{arg\,min}_{P \in \mathcal{D}_n} \frac{1}{n} \|X^\top X P - P Y^\top Y\|_F^2 \,, \tag{12}$$

where $\mathcal{D}_n$ is the polytope of *bistochastic* matrices, which is the convex envelope of the set of permutation matrices. Note that unlike in (5), this argmin is not necessarily equal to $\arg\max_{P \in \mathcal{D}_n} \langle P^\top X^\top X P, Y^T Y' \rangle$ since $\mathcal{D}_n$ contains non-orthogonal matrices.

The estimate $\hat{P}_{\mathrm{relaxed}}$ gives a first estimate to then perform alternate minimizations in Q through an SVD – see (11) – and P through a LAP – see (10). Combining an initialization with convex relaxation, computed via Frank-Wolfe algorithm [Jaggi, 2013] and the alternate minimizations in P and Q yields the *Ping-Pong algorithm*.

Algorithm 1: PING-PONG ALGORITHM

Input: Number of Frank-Wolfe steps T, number of alternate-minimization steps K, $\tilde{P}_0 = \frac{1 1^\top}{n}$

1 **for** $k = 0$ *to* $T - 1$ **do**

2 Compute $S_k = \arg\min_{P \in \mathcal{S}_n} \langle P, \nabla f(\tilde{P}_k) \rangle$ (LAP), where $f(P) = \left\| X^\top X P - P Y^\top Y \right\|_F^2$

3 $\tilde{P}_{k+1} = (1 - \gamma_k)\tilde{P}_k + \gamma_k S_k$ for $\gamma_k = \frac{1}{2+k}$

4 $P_0 = \tilde{P}_T$ and $Q_0 = I_d$

5 **for** $k = 0$ *to* $K - 1$ **do**

6 $Q_{k+1} = U_k V_k^\top$ for $Y P_k X^\top = U_k D_k V_k$ the SVD of $Y P_k X^\top$ *(Ping)*

7 $P_{k+1} \in \arg\max_{P \in \mathcal{S}_n} \langle P, Y^\top Q_{k+1} X \rangle$ (LAP) *(Pong)*

Output: P_K, Q_K

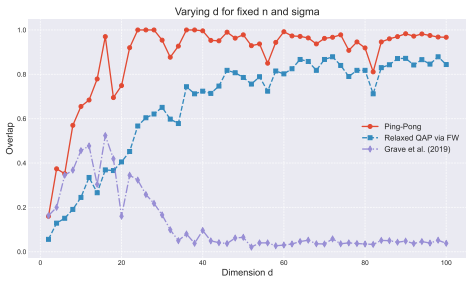
Algorithm 1 is structurally similar to Grave et al. [2019]'s algorithm, as explained in the introduction. The difference lies in the steps in Lines 6-7 of Algorithm 1: while Grave et al. [2019] perform projected gradient steps, our approach is more greedy and directly minimizes in each variable. Both approaches are experimentally compared in Section 3.3.

3.2 Guarantees for one step of Ping-Pong algorithm

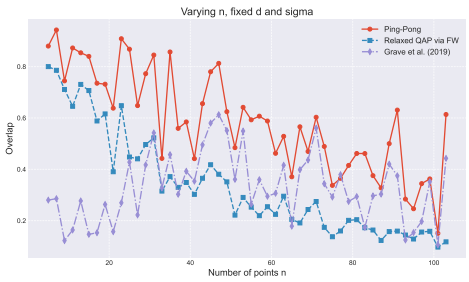
Providing statistical rates for the outputs of Algorithm 1 is a challenging problem for two reasons. First, relaxed QAP is not a well-understood problem: the only existing guarantees in the literature are for correlated Gaussian Wigner models in the noiseless case (i.e., $\sigma = 0$ in our model) [Valdivia and Tyagi, 2023], while for correlated Erdös-Rényi graphs, the relaxation is known to behave badly in general [Lyzinski et al., 2016]. Secondly, studying the iterates in lines 6 and 7 of the algorithm is challenging, since these are projections on non-convex sets. While Lemmas 2 and 3 show that if P_k (resp. Q_k) has small c^2 loss, then Q_{k+1} has small ℓ^2 loss (resp. P_{k+1} has small c^2 loss), showing that there is a contraction at each iteration 'à la Picard's fix-point Theorem' remains out of reach for this paper. We thus resort to proving that *one single step* of Algorithm 1 ($K = T = 1$) can recover the planted signal, provided that the noise σ is small enough.

Proposition 1. *There exists $C > 0$ such that for any $\delta \in (0, 1)$, if $\sigma \leqslant n^{-\frac{13}{\delta}}$, then the permutation $\hat{\pi}$ associated to the outputs $\hat{\pi}, \hat{Q}$ of Algorithm 1 for $K = T = 1$ satisfies, with probability $1 - 1/n$:*

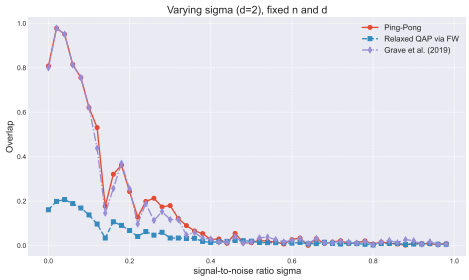
$$\mathrm{ov}(\pi^\star, \hat{\pi}) \geqslant 1 - \delta \,.$$



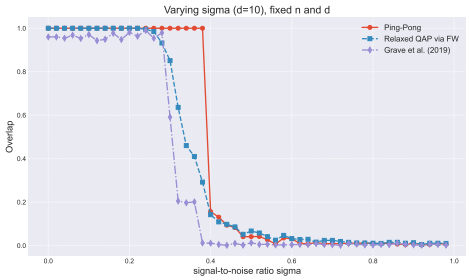
(a) varying d



(b) varying n



(c) varying σ, small d



(d) varying σ, large d

Figure 1: Influence of the parameters (dimensions d, number of points n, and noise level σ) on the accuracy (in terms of overlap) of three different estimators: the relaxed QAP estimator (12) projected on the set of permutation matrices (blue curve), the output of Alg. 1 (red curve), and the output of Grave et al. [2019]'s algorithm (purple curve). Each dot corresponds to averaging scores over 10 experiments. Figure 1a: $\sigma = 0.34, n = 100$. Figure 1b: $\sigma = 0.34, d = 5$. Figures 1c and 1d: $n = 200$, $d = 2$ and $d = 60$ respectively.

In the high-dimensional setting ($d \gg \log(n)$), there exist some constants c_1, c_2 such that if $\sigma \leqslant n^{-c_1}$, then $\hat{\pi}$ satisfies w.h.p.

$$\mathrm{ov}(\pi^\star, \hat{\pi}) \geqslant 1 - c_2 \max\left(\sqrt{\frac{d\log(d)}{n} + \frac{\log(n)}{d}}, \frac{d\log(d)}{n} + \frac{\log(n)}{d}\right).$$

Thus, for σ polynomially small in n and exponentially small in $1/\delta$, one step of Alg. 1 recovers $\pi^\star$ in the overlap sense with error δ. In large dimensions, this is improved, since σ is no longer required to be exponentially small as the target error decreases to zero. Proof of Proposition 3 is given in Appendix E.

3.3 Numerical experiments

We compare in Figure 1 our Alg. 1 with (i) the naive initialization of the relaxed QAP estimator (12), and (ii) the method in Grave et al. [2019]. The curve 'relaxed QAP via FW' is obtained by computing the relaxed QAP estimator with Frank-Wolfe algorithm with $T = 1000$ steps, enough for convergence. This estimator is then taken as initialization for Alg. 1 and Grave et al. [2019]'s algorithm, that are both taken with the same large number of steps ($K = 100$, empirically leading to convergence to stationary points of the algorithms). For fair comparison, we take full batches in Grave et al. [2019] (smaller batches lead to even worse performances).

Conclusion

We establish new informational results for the Procrustes-Wassertein problem, both in the high ($d \gg \log n$) and low ($d \ll \log n$) dimensional regimes. We propose the 'Ping-Pong algorithm', alternatively estimating the orthogonal transformation and the relabeling, initialized via a Franke-Wolfe convex relaxation. Our experimental results show that our method most globally outperforms the algorithm proposed in Grave et al. [2019].

References

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL https://aclanthology.org/D18-1214.

P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008. ISSN 1631-073X. doi: https://doi.org/10.1016/j.crma.2008.03.014. URL https://www.sciencedirect.com/science/article/pii/S1631073X08000964.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Osman E. Dai, Daniel Cullina, and Negar Kiyavash. Database alignment with gaussian features. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3225–3233. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/dai19b.html.

Osman Emre Dai, Daniel Cullina, and Negar Kiyavash. Gaussian database alignment and gaussian planted matching, 2023.

Pascale Fung. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Third Workshop on Very Large Corpora*, 1995. URL https://aclanthology.org/W95-0114.

Shuyang Gong and Zhangsong Li. The umeyama algorithm for matching correlated gaussian geometric models in the low-dimensional regime, 2024.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation, 2018.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/jaggi13.html.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Dmitriy Kunisky and Jonathan Niles-Weed. *Strong recovery of geometric planted matchings*, pages 834–876. 2022. doi: 10.1137/1.9781611977073.36. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977073.36.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Vince Lyzinski, Donniell E. Fishkind, Marcelo Fiori, Joshua T. Vogelstein, Carey E. Priebe, and Guillermo Sapiro. Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):60–73, 2016. doi: 10.1109/TPAMI.2015.2424894.

Konstantin Makarychev, Rajsekar Manokaran, and Maxim Sviridenko. Maximum quadratic assignment problem: Reduction from maximum label cover and lp-based approximation algorithm. *ACM Trans. Algorithms*, 10(4), aug 2014. ISSN 1549-6325. doi: 10.1145/2629672. URL https://doi.org/10.1145/2629672.

Reinhard Rapp. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA, June 1995. Association for Computational Linguistics. doi: 10.3115/981658.981709. URL https://aclanthology.org/P95-1050.

C. A. Rogers. Covering a sphere with spheres. *Mathematika*, 10(2):157–164, 1963. doi: 10.1112/S0025579300004083.

Peter Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31 (1):1–10, 1966. URL https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:1:p:1-10.

Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, Nov 1992. ISSN 1573-1405. doi: 10.1007/BF00129684. URL https://doi.org/10.1007/BF00129684.

S. Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988. doi: 10.1109/34.6778.

Ernesto Araya Valdivia and Hemant Tyagi. Graph matching via convex relaxation to the simplex, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3441–3488. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/wang22a.html.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL https://aclanthology.org/N15-1104.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://aclanthology.org/P17-1179.

A Useful results

We start by proving Lemma 1 which gives the equivalence between PW and geometric graph alignement.

A.1 Proof of Lemma 1

Proof of Lemma 1. We have by Lemma 3 that as soon as we are able to estimate $\pi^\star$ with a small error in PW, we are also capable of doing so $Q^\star$, by perfoming a simple Singular Value Decomposition (SVD). Since one can trivially form an instance $A = X^\top X$ and $B = Y^\top Y$ of geometric graph alignement from an instance (X, Y) of PW from model (1), we can deduce that if we know how to (approximately) solve geometric graph alignement, we know how to (approximately) solve PW.

Conversely, if we are given adjacency matrices $A = X^\top X, B = Y^\top Y$ of two correlated random geometric graphs under the Gaussian model from [Wang et al., 2022, Gong and Li, 2024] where $y_i = x_{\pi^\star(i)} + \sigma z_i$, we can recover $\pi^\star$ via solving PW. Indeed, A is of rank at most d, so we can build $X' = (x'_1|\dots|x'_n) \in \mathbb{R}^{d\times n}$ such that $A = X'^\top X'$. Similarly, we can build $Y' = (y'_1|\dots|y'_n) \in \mathbb{R}^{d\times n}$ such that $B = Y'^\top Y'$. We have $X^\top X = X'^\top X'$, hence $\langle x_i, x_j\rangle = \langle x'_i, x'_j\rangle$, thus there exists $Q_1 \in \mathcal{O}(d)$ such that for all $i \in [n]$, $x'_i = Q_1 x_i$. Similarly, there exists $Q_2 \in \mathcal{O}(d)$ such that for all i, $y'_i = Q_2 y_i$. By multiplying these two orthogonal matrices by independent random uniform orthogonal matrices, we can always assume that they are independent from X and Y. We obtained X', Y' which satisfy $y'_i = Q^\star x'_{\pi^\star(i)} + \sigma z'_i$ for all i, where $Q^\star = Q_2 Q_1^\top$, and $x'_i = Q_1 x_i, z'_i = Q_2 z_i$ are i.i.d. standard Gaussian vectors. This is exactly an instance of the PW problem. If we know how to (approximately) solve the PW problem, we know how to (approximately) recover $\pi^\star$ and thus (approximately) solve the geometric graph alignment problem.

This proves that PW and geometric graph alignement are equivalent. □

A.2 $\varepsilon-$nets of $\mathcal{O}(d)$

Throughout the proofs, we will need to give high probability bounds on quantities for all orthogonal matrices. This is done by covering $\mathcal{O}(d)$ by a finite number of open balls centered at points of $\mathcal{O}(d)$. This is done by considering $\varepsilon-$nets.

Definition 1 ($\varepsilon-$nets of $\mathcal{O}(d)$)**.** *Let $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon \subseteq \mathcal{O}(d)$ is a $\varepsilon-$net of $\mathcal{O}(d)$ for the Frobenius norm if for all $O \in \mathcal{O}(d)$ there exists $O_\varepsilon \in \mathcal{N}_\varepsilon$ such that $\|O - O_\varepsilon\|_F \leqslant \varepsilon$.*

Remark 2. *Note that since $\|\cdot\|_F \leqslant \|\cdot\|_{op}$ by Cauchy-Schwarz any $\varepsilon-$net of $\mathcal{O}(d)$ for the Frobenius norm is also an $\varepsilon-$net of $\mathcal{O}(d)$ for the operator norm.*

We will need $\varepsilon-$nets of $\mathcal{O}(d)$ that are not too large, in order to apply union bounds which will give non-trivial probabilistic controls. Guarantees on such $\varepsilon-$nets are standard in the literature; we give one which will be useful for us in the following Lemma.

Lemma 4 ($\varepsilon-$nets of $\mathcal{O}(d)$ of minimal size, see e.g. Rogers [1963])**.** *There exists a universal constant $C > 0$ such that for all $\varepsilon > 0$, there exists an $\varepsilon-$net $\mathcal{N}_\varepsilon$ of $\mathcal{O}(d)$ such that*

$$|\mathcal{N}_\varepsilon| \leqslant \left(\frac{C\sqrt{d}}{\varepsilon}\right)^{d^2}.$$

B Remaning proofs of Section 2

B.1 Proof of Lemma 2

Proof of Lemma 2. Denote $g(\pi) := \frac{1}{n}\sum_{i=1}^n \left\|x_{\pi(i)} - \hat{Q}^\top y_i\right\|^2$. The proof relies on noticing that for all $\pi \in \mathcal{S}_n$, by definition $g(\hat{\pi}) \leqslant g(\pi)$ and using $\|a+b\|^2 \geqslant \frac{1}{2}\|a\|^2 - \|b\|^2$, one gets

$$g(\hat{\pi}) \geqslant \frac{1}{2}c^2(\hat{\pi}, \pi) - g(\pi),$$

and thus $c^2(\hat{\pi}, \pi) \leqslant 4g(\pi)$. We apply the previous inequality to $\pi = \pi^\star$ and using $\|a+b\|^2 \leqslant 2(\|a\|^2 + \|b\|^2)$, one gets,

$$\begin{aligned} g(\pi^\star) &\leqslant \frac{2}{n}\sum_{i=1}^n \left\|(I_d - \hat{Q}^\top Q^\star)x_i\right\|^2 + \frac{2\sigma^2}{n}\sum_{i=1}^n \left\|\hat{Q}^\top z_i\right\|^2 \\ &= \frac{2}{n}\sum_{i=1}^n \left\|(\hat{Q} - Q^\star)x_i\right\|^2 + \frac{2\sigma^2}{n}\sum_{i=1}^n \|z_i\|^2 , \end{aligned}$$

where we used the fact that the matrices $\hat{Q}, Q^\star$ are orthogonal. Using concentration of Chi squared random variables, we have

$$\mathbb{P}\left(\sum_{i=1}^n \|z_i\|^2 \geqslant nd + 2\sqrt{ndt} + 2t\right) \leqslant e^{-t} ,$$

leading to $\mathbb{P}\left(\sum_{i=1}^n \|z_i\|^2 \geqslant 5nd\right) \leqslant e^{-nd}$ by plugging in $t = nd$. We are now left with $\sum_{i=1}^n \left\|(\hat{Q} - Q^\star)x_i\right\|^2$. We have that for any Q, $\mathbb{E}\left[\sum_{i=1}^n \|(Q-Q^\star)x_i\|^2\right] = n\|Q-Q^\star\|_F^2$; however, $\hat{Q}$ depends on the x_i so we need a uniform upper bound. Using Hanson-Wright inequality, for any $Q \in \mathbb{R}^{d\times d}$,

$$\mathbb{P}\left(\sum_{i=1}^n \|(Q-Q^\star)x_i\|^2 \geqslant n\|Q-Q^\star\|_F^2 + c\max\left(\sqrt{nt\|Q-Q^\star\|_F^2\|Q-Q^\star\|_{\mathrm{op}}^2}, t\|Q-Q^\star\|_{\mathrm{op}}^2\right)\right) \leqslant 2e^{-t} ,$$

which reads as, for Q orthogonal (leading to $\|Q-Q^\star\|_{\mathrm{op}} \leqslant 2$):

$$\mathbb{P}\left(\sum_{i=1}^n \|(Q-Q^\star)x_i\|^2 \geqslant n\|Q-Q^\star\|_F^2 + c'\max\left(\sqrt{nt\|Q-Q^\star\|_F^2}, t\right)\right) \leqslant 2e^{-t} .$$

For $\varepsilon \in (0, 1/2)$, let $\mathcal{N}_\varepsilon$ be an $\varepsilon-$net of $\mathcal{O}(d)$ of minimal cardinality; by Lemma 4 we have $\log(|\mathcal{N}_\varepsilon|) \leqslant Cd^2\ln(d/\varepsilon)$. Using a union bound:

$$\mathbb{P}\left(\sup_{Q\in\mathcal{N}_\varepsilon}\left\{\left|\sum_{i=1}^n \|(Q-Q^\star)x_i\|^2 - n\|Q-Q^\star\|_F^2\right|\right\} \geqslant c'\max\left(\sqrt{nt\|Q-Q^\star\|_F^2}, t\right)\right) \leqslant 2e^{-t+Cd^2\ln(d/\varepsilon)} .$$

Taking $t = \lambda + Cd^2\ln(d/\varepsilon)$, we have with probabiliy $1-2e^{-\lambda}$ that:

$$\sup_{Q\in\mathcal{N}_\varepsilon}\left\{\left|\sum_{i=1}^n \|(Q-Q^\star)x_i\|^2 - n\|Q-Q^\star\|_F^2\right|\right\} \leqslant c'\max\left(\sqrt{n(\lambda + Cd^2\ln(d/\varepsilon))\|Q-Q^\star\|_F^2}, \lambda + Cd^2\ln(d/\varepsilon)\right)$$

Now, if $Q, Q' \in \mathcal{O}(d)$ satisfy $\|Q-Q'\|_F \leqslant \varepsilon$, we have using the orthogonality property of these matrices:

$$\begin{aligned} \sum_{i=1}^n \|(Q-Q^\star)x_i\|^2 - \sum_{i=1}^n \|(Q'-Q^\star)x_i\|^2 &= 2\sum_{i=1}^n \langle (Q'-Q)x_i, Q^\star x_i\rangle \\ &\leqslant 2\sum_{i=1}^n \|(Q'-Q)x_i\|\|Q^\star x_i\| \\ &\leqslant 2\|Q'-Q\|_{\mathrm{op}}\sum_{i=1}^n \|x_i\|^2 \\ &\leqslant 2\varepsilon\sum_{i=1}^n \|x_i\|^2 , \end{aligned}$$

and with probability $1-e^{-nd}$ we have $\sum_{i=1}^n \|x_i\|^2 \leqslant 5nd$. Then,

$$\begin{aligned} \|Q-Q^\star\|_F^2 - \|Q'-Q^\star\|_F^2 &= \langle Q^\star, Q'-Q\rangle \\ &\leqslant \|Q^\star\|_F\|Q'-Q\|_F \\ &\leqslant \sqrt{d}\varepsilon . \end{aligned}$$

Thus,

$$\sup_{Q\in\mathcal{O}(d)}\left\{\left|\sum_{i=1}^n\|(Q-Q^\star)x_i\|^2-n\|Q-Q^\star\|_F^2\right|\right\}\leqslant\sup_{Q\in\mathcal{N}_\varepsilon}\left\{\left|\sum_{i=1}^n\|(Q-Q^\star)x_i\|^2-n\|Q-Q^\star\|_F^2\right|\right\}+10nd\varepsilon+n\sqrt{d}\varepsilon$$
$$\leqslant c'\max\left(\sqrt{n(\lambda+Cd^2\ln(d/\varepsilon))\|Q-Q^\star\|_F^2},\lambda+Cd^2\ln(d/\varepsilon)\right)+10nd\varepsilon+n\sqrt{d}\varepsilon\,,$$

with probability $1-2e^{-nd}-2e^{-\lambda}$. Thus, applying this to $\hat{Q}$:

$$\sum_{i=1}^n\left\|(\hat{Q}-Q^\star)x_i\right\|^2\leqslant n\delta d+c'\max\left(\sqrt{n\delta(\lambda+Cd^2\ln(11/\delta))},\lambda+Cd^2\ln(1\varepsilon)\right)+10nd\varepsilon+n\sqrt{d}\varepsilon\,,$$

and taking $\varepsilon=\delta/11$,

$$\sum_{i=1}^n\left\|(\hat{Q}-Q^\star)x_i\right\|^2\leqslant 2n\delta d+c'\sqrt{n\delta(\lambda+Cd^2\ln(11/\delta))}+c'(\lambda+Cd^2\ln(11/\delta))\,,$$

leading to:

$$c^2(\hat{\pi},\pi^\star)\leqslant 40\sigma^2d+16\delta d+c'\sqrt{\delta\frac{\lambda+Cd^2\ln(11/\delta)}{n}}+c'\frac{\lambda+Cd^2\ln(11/\delta)}{n}\,,$$

hence the result, taking $\lambda=\sqrt{n}+d^2$.

□

B.2 Proof of Lemma 3

Proof of Lemma 3. Denote $g(Q):=\frac{1}{n}\sum_{i=1}^n\left\|x_{\hat{\pi}(i)}-Q^\top y_i\right\|^2$. The proof relies on noticing that for all $Q\in\mathcal{O}(d)$, by definition $g(\hat{Q})\leqslant g(Q^\star)$ and using $\|a+b\|^2\geqslant\frac{1}{2}\|a\|^2-\|b\|^2$, one gets

$$g(\hat{Q})\geqslant\frac{1}{2n}\sum_{i=1}^n\left\|(Q-\hat{Q})y_i\right\|^2-g(Q),$$

and thus $\frac{1}{n}\sum_{i=1}^n\left\|(Q-\hat{Q})y_i\right\|^2\leqslant 4g(Q^\star)$ by applying the previous inequality to $\pi=\pi^\star$. Using $\|a+b\|^2\leqslant 2(\|a\|^2+\|b\|^2)$, one gets:

$$\begin{aligned}g(Q^\star)&=\frac{1}{n}\sum_{i=1}^n\left\|x_{\hat{\pi}(i)}-x_{\pi^\star(i)}-{Q^\star}^\top z_i\right\|^2\\&\leqslant\frac{2}{n}\sum_{i=1}^n\left\|x_{\hat{\pi}(i)}-x_{\pi^\star(i)}\right\|^2+\frac{2\sigma^2}{n}\sum_{i=1}^n\|z_i\|^2\\&=2c^2(\hat{\pi},\pi^\star)+\frac{2\sigma^2}{n}\sum_{i=1}^n\|z_i\|^2\\&\leqslant 2\delta d+\frac{2\sigma^2}{n}\sum_{i=1}^n\|z_i\|^2\,.\end{aligned}$$

With probability $1-e^{-nd}$, $\sum_{i=1}^n\|z_i\|^2\leqslant 5nd$, and we are thus left with lower bounding $\frac{1}{n}\sum_{i=1}^n\left\|(Q-\hat{Q})y_i\right\|^2$. Using results form the previous proof, with probability $1-2e^{-nd}-2e^{-\lambda}$ we have:

$$\frac{1}{1+\sigma^2}\sum_{i=1}^n\left\|(\hat{Q}-Q^\star)y_i\right\|^2\geqslant n\left\|\hat{Q}-Q^\star\right\|_F^2+n\delta d+c'\sqrt{n\delta(\lambda+Cd^2\ln(11/\delta))}+c'(\lambda+Cd^2\ln(11/\delta))\,.$$

Thus, with probability $1-$

$$\ell^2_{\mathrm{ortho}}(\hat{Q},Q^\star)\leqslant\delta d+c'\sqrt{n^{-1}\delta(\lambda+Cd^2\ln(11/\delta))}+c'n^{-1}(\lambda+Cd^2\ln(11/\delta))+8\delta d+40\sigma^2d\,,$$

leading to the desired result for $\lambda=d^2+\sqrt{n}$. □

C Proof of Theorem 1

Proof of Theorem 1. Define $\mathcal{L}(\pi, Q) := \frac{1}{n}\sum_{i=1}^n \left\|x_{\pi(i)} - Q^\top y_i\right\|^2$. Without loss of generality we can assume that $\pi^* = \mathrm{Id}$.

Step 1: using ML estimators. By definition, the ML estimators $(\hat{\pi}, \hat{Q})$ defined in (2) minimize $\mathcal{L}$ and thus $\mathcal{L}(\hat{\pi}, \hat{Q}) \leqslant \mathcal{L}(\pi^\star = \mathrm{Id}, Q^\star)$, which can be expressed as:

$$\frac{1}{n}\sum_{i=1}^n \left\|x_{\hat{\pi}(i)} - \hat{Q}^\top Q^\star x_i - \sigma\hat{Q}^\top z_i\right\|^2 \leqslant \frac{\sigma^2}{n}\sum_{i=1}^n \|z_i\|^2 .$$

Using $\|a\|^2 \leqslant 2(\|a-b\|^2 + \|b\|^2)$, we obtain:

$$\frac{1}{n}\sum_{i=1}^n \left\|x_{\hat{\pi}(i)} - \hat{Q}^\top Q^\star x_i\right\|^2 \leqslant \frac{4\sigma^2}{n}\sum_{i=1}^n \|z_i\|^2 .$$

Then, standard chi-square concentration (see e.g. Laurent and Massart [2000]) entails that for all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n \|z_i\|^2 - nd\right| \geqslant 2\sqrt{ndt} + 2t\right) \leqslant 2e^{-t},$$

so that with probability $1 - 2e^{-n}$, $\frac{4\sigma^2}{n}\sum_{i=1}^n \|z_i\|^2 \leqslant 4\sigma^2(d + 2\sqrt{d} + 2)$, and thus

$$\frac{1}{n}\sum_{i=1}^n \left\|x_{\hat{\pi}(i)} - \hat{Q}^\top Q^\star x_i\right\|^2 \leqslant 4\sigma^2(d + 2\sqrt{d} + 2) \leqslant 5\sigma^2 d\,, \tag{13}$$

for d (or n) large enough.

Step 2: existence of a set $\mathcal{K}$ with prescribed properties. We will now show that the above inequality (13) forces $\mathrm{ov}(\hat{\pi}, \pi^\star = \mathrm{Id})$ to be large. To do so, let us assume that $\mathrm{ov}(\hat{\pi}, \mathrm{Id}) < 1 - \delta$, for some $\delta > 0$ to be specified later: hence, there exist at least $n\delta$ indices $i \in [n]$ such that $\hat{\pi}(i) \neq i$. Let us define

$$\mathcal{I} := \left\{i \in [n] : \left\|x_{\hat{\pi}(i)} - \hat{Q}^\top Q^\star x_i\right\|^2 \leqslant \frac{30}{\delta}\sigma^2 d\right\}.$$

It is clear that under the event $\mathcal{B}_n$ on which (13) holds, we have $(n - |\mathcal{I}|) \times \frac{30}{\delta}\sigma^2 d \leqslant 5n\sigma^2 d$, which gives

$$|\mathcal{I}| \geqslant n(1 - \delta/6)\,.$$

Consequently, denoting

$$\mathcal{J} := \left\{i \in [n] : \pi^\star(i) \neq \hat{\pi}(i),\ \left\|x_{\hat{\pi}(i)} - \hat{Q}^\top Q^\star x_i\right\|^2 \leqslant \frac{30}{\delta}\sigma^2 d\right\},$$

one has $|\mathcal{J}| \geqslant n\delta - (n - |\mathcal{I}|) \geqslant \frac{5}{6}\delta n$. We remark that for all $i \in \mathcal{J}$,

$$|\{j \in \mathcal{J} : \{i, \hat{\pi}(i)\} \cap \{j, \hat{\pi}(j)\} \neq \varnothing\}| \leqslant 4\,.$$

Let us denote $Q := \hat{Q}^\top Q^\star$. Iteratively ruling out at most 3 elements for each $i \in \mathcal{J}$, the above shows that on event $\mathcal{E}_n$ one can build a set $\mathcal{K} := \mathcal{K}(\hat{\pi}, Q) \subseteq [n]$ such that

- (i) $|\mathcal{K}| \geqslant n\delta/6$,
- (ii) for all $i \in \mathcal{K}$, $\hat{\pi}(i) \neq i$, and $(i, \hat{\pi}(i))_{i\in\mathcal{K}}$ are disjoint pairs,
- (iii) for all $i \in \mathcal{K}$, $\left\|x_{\hat{\pi}(i)} - Qx_i\right\|^2 \leqslant \frac{30}{\delta}\sigma^2 d$.

Step 3: upper bounding the probaility of existence of such a set $\mathcal{K}$. We will now bound the probability that such a set $\mathcal{K}$ exists. First, let us fix $i \in [n]$, $Q \in \mathcal{O}(d)$, $\pi \in \mathcal{S}_n$ such that $\pi(i) \neq i$. We have $x_{\pi(i)} - Qx_i \sim \mathcal{N}(0, 2I_d)$. Assume $60\sigma^2 < 1$ and $\delta \geqslant 60\sigma^2$ so that we have $\frac{60}{\delta}\sigma^2 d \leqslant d$. For these fixed Q, π, we have

$$\mathbb{P}\left(\left\|x_{\pi(i)} - Qx_i\right\|^2 \leqslant \frac{60}{\delta}\sigma^2 d\right) \leqslant \mathbb{P}\left(\|\mathcal{N}(0, I_d)\|^2 \leqslant d/2\right) = \mathbb{P}\left(d - \|\mathcal{N}(0, I_d)\|^2 \geqslant d/2\right) \leqslant e^{-d/16},$$

where we applied the one-sided chi-square concentration inequality[4] $\mathbb{P}\left(k - X \geqslant 2\sqrt{kx}\right) \leqslant \exp(-x)$ when $X \sim \chi^2(k)$. This gives that for any given $\mathcal{K} \subset [n]$ satisfying conditions (i) and (ii) above, using independence of the pairs $(x_i, x_{\hat{\pi}(i)})_{i\in\mathcal{K}}$ and recalling that $\delta \geqslant 60\sigma^2$, one has

$$\mathbb{P}\left(\forall i \in \mathcal{K},\ \left\|x_{\pi(i)} - Qx_i\right\|^2 \leqslant \frac{60}{\delta}\sigma^2 d\right) \leqslant e^{-n\delta/6\times d/16} = e^{-\delta n d/96}. \tag{14}$$

Denote by $\mathcal{A}_\delta$ the event

$$\mathcal{A}_\delta := \{\text{there exists } \pi \in \mathcal{S}_n, Q \in \mathcal{O}(d) \text{ and } \mathcal{K} = \mathcal{K}(\pi, Q) \subseteq [n] \text{ which satisfies } (i), (ii) \text{ and } (iii)\}. \tag{15}$$

As previously explained, we want to bound the probability of the event $\mathcal{A}_\delta$, for $\delta \geqslant 60\sigma^2$. For the union bound on $Q \in \mathcal{O}(d)$, we need to use an epsilon-net argument, which is as follows. Let $\varepsilon > 0$ to be specified later. By Lemma 4, there exists $\mathcal{N}_\varepsilon$ an $\varepsilon-$net of $\mathcal{O}(d)$ of cardinality at most $\left(\frac{c_1\sqrt{d}}{\varepsilon}\right)^{d^2}$, which is also an $\varepsilon-$net for the operator norm, see Remark 2. In particular, if we are under event A_δ and take $\pi, Q, \mathcal{K}$ verifying conditions in (15), there exists an element Q_ε of $O_\varepsilon(d)$ such that $\|Q_\varepsilon - Q\|_{op} \leqslant \varepsilon$, which gives

$$\forall i \in \mathcal{K},\ \left\|x_{\pi(i)} - Q_\varepsilon x_i\right\| \leqslant \left\|x_{\pi(i)} - Qx_i\right\| + \|(Q - Q_\varepsilon)x_i\| \leqslant \sqrt{\frac{30}{\delta}\sigma^2 d} + \|Q_\varepsilon - Q\|_{op}\|x_i\|,$$

and applying chi-square concentration again gives that, under an event $\mathcal{C}_n$ with probability $\geqslant 1 - e^{-d+\log n} \geqslant 1 - e^{-d/2}$ since $(d \geqslant 2\log n)$ for all $i \in [n]$, $\|x_i\| \leqslant \sqrt{2d}$ and the above yields

$$\forall i \in \mathcal{K},\ \left\|x_{\pi(i)} - Q_\varepsilon x_i\right\| \leqslant \sqrt{\frac{30}{\delta}\sigma^2 d} + \varepsilon\sqrt{2d} \leqslant \sqrt{\frac{60}{\delta}\sigma^2 d},$$

choosing $\varepsilon = c_2\sqrt{\sigma^2/\delta}$ for some appropriate c_2. Hence, taking a union bound over $\pi \in \mathcal{S}_n, Q_\varepsilon \in O_\varepsilon(d)$ and subsets $\mathcal{K} \subseteq [n]$, and recalling (14), we can bound $\mathbb{P}\left(\mathcal{A}_\delta \,|\, \mathcal{C}_n, \mathcal{B}_n\right)$ by

$$\begin{aligned}
\mathbb{P}\left(\mathcal{A}_\delta \,|\, \mathcal{B}_n, \mathcal{C}_n\right) &\leqslant \frac{1}{\mathbb{P}\left(\mathcal{B}_n, \mathcal{C}_n\right)} \times n! \times \left(\frac{c_1\sqrt{d}}{\varepsilon}\right)^{d^2} \times 2^n \times e^{-\delta n d/96} \\
&\leqslant (1 + o(1))\exp(n\log n + c_3 d^2 \log d + c_4 d^2\sqrt{\delta/(\sigma^2)} + n\log 2 - c_5\delta n d) \\
&\leqslant (1 + o(1))\exp(-c_6\delta n d)
\end{aligned}$$

where we recall that $\delta \geqslant 60\sigma^2$, and the last inequality holds if $c_4 d^2\sqrt{\delta/(\sigma^2)} \leqslant c_7 d^2 \leqslant c_8\delta n d$, and if $c_3 d^2 \log d \leqslant c_8\delta n d$ for which $\delta \geqslant c_9 d\log d/n$ suffices, and if $n\log n \leqslant c_8\delta n d$, for which $\delta \geqslant c_1 0\log n/d$ suffices.

Step 4: conclusion. Now, wrapping things up, we obtain that for $\delta \geqslant \max(60\sigma^2, c_9 d/n, c_1 0\log n/d)$, we have for n large enough

$$\begin{aligned}
\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) < 1 - \delta\right) &\leqslant \mathbb{P}\left(\mathcal{B}_n, \mathcal{C}_n\right)\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) < 1 - \delta \,|\, \mathcal{B}_n, \mathcal{C}_n\right) + \mathbb{P}\left(\bar{\mathcal{B}_n} \cup \bar{\mathcal{C}_n}\right) \\
&\leqslant \mathbb{P}\left(\mathcal{A}_\delta \,|\, \mathcal{B}_n, \mathcal{C}_n\right) + \mathbb{P}\left(\bar{\mathcal{B}_n}\right) + \mathbb{P}\left(\bar{\mathcal{C}_n}\right) \\
&\leqslant (1 + o(1))e^{-c_6\delta n d} + 2e^{-n} + e^{-d/2} = o(1).
\end{aligned}$$

This gives the desired result

$$\mathrm{ov}(\pi^\star, \hat{\pi}) \geqslant 1 - \max\left(60\sigma^2, c_1\frac{d}{n}, c_2\frac{\log n}{d\log d}\right),$$

which remains true when $60\sigma^2 \geqslant 1$.

The desired inequality for $\ell^2(\hat{Q}, Q^\star)$ follows from Lemma 3 (for $\delta = \Theta(d/n)$) and Remark 1. $\square$

[4] again, see e.g. Laurent and Massart [2000].

D Proof of Theorem 2

We here prove that a minimizer $\hat{Q}$ of the conic alignment loss satisfies the following guarantees.

Theorem 3 (Conic alignment minimizer)**.** *Let $\delta_0 \in (0,1)$. Let $q = p^3$. Let $v_1, \dots, v_q$ be i.i.d. uniform directions in $\mathcal{S}^{d-1}$, and assume that $u_1, \dots, u_p$ are independently and uniformly distributed over $\{v_1, \dots, v_q\}$. Let $\mathcal{N}$ be an $\varepsilon-$net of $\mathcal{O}(d)$ for the Frobenius norm of minimal cardinality. Then, there exist constants $C_1, C_2, C_3, C_4, C_5 > 0$ such that, if $\log(n) \geqslant C_1 d \log(1/\delta_0)$, $\varepsilon = C_2 \sigma d^{-1/2}$, $\delta = \delta_0$, $\kappa = \sqrt{\frac{2}{d}}$, $p \geqslant \mathrm{polylog}(1/\sigma, d)$ and $\sigma \leqslant \frac{C_3 \delta_0^2}{\log(1/\delta_0)}$, then, with probability $1 - 6e^{-C_4 d^2}$,*

$$\frac{1}{d}\left\|\hat{Q} - Q^\star\right\|_F^2 \leqslant \delta_0 \,.$$

Then, combinign this result with Lemma 2, setting $\hat{\pi}$ as in Equation (10), we obtain Theorem 2.

D.1 Proof of Theorem 3

Proof of Theorem 3. Recall that $\delta, \kappa, \varepsilon > 0$ are for now any (small) positive number but can be specified later. δ_0 is the target error. We begin by giving a few notations. For the proof, we need to introduce the following probability

$$\beta(\delta, \kappa) := \mathbb{P}\left(X \in \mathcal{C}(u, \delta), \|X\| \geqslant 1/\kappa\right) , \tag{16}$$

where $X \sim \mathcal{N}(0, I_d)$ and u is any unit vector in $\mathbb{R}^d$. Note that $\beta(\delta, \kappa)$ is independent of the choice of u by rotational invariance of Gaussian distribution. It is easy to check that $\mathbb{P}\left(x_i \in \mathcal{C}_{\mathcal{X}}(u, \delta)\right) = \mathbb{P}\left(y_j \in \mathcal{C}_{\mathcal{X}}(u, \delta)\right) = \beta(\delta, \kappa)$ for any i, j and any unit vector u.

Step 1: General strategy. Our goal is to prove that w.h.p. we have

$$F(\hat{Q}) < \inf\left\{F(Q), Q \in \mathcal{N}, \|Q - Q^\star\|_F^2 > \delta_0 d\right\}, \tag{17}$$

for some $\delta_0 > 0$ to be determined. This will entail that $\left\|\hat{Q} - Q^\star\right\|_F^2 \leqslant \delta_0 d$.

Step 2: An upper bound on $\mathbb{E}\left[F(\hat{Q})\right]$. Since $\mathcal{N}$ is an $\varepsilon-$net of $\mathcal{O}(d)$, there exists $Q_\varepsilon^\star \in \mathcal{N}$ such that $\|Q_\varepsilon^\star - Q^\star\|_F \leqslant \varepsilon$. Note that by optimality of $\hat{Q}$, one has $F(\hat{Q}) \leqslant F(Q_\varepsilon^\star)$. We first upper bound the left hand side in (17) by upper bounding the expectation of $F(Q_\varepsilon^\star)$ using the following result.

Lemma 5. *Let $Q \in \mathcal{O}(d)$, $u \in \mathcal{S}^{d-1}$. We have:*

$$\mathbb{E}\left[F(Q)\right] \leqslant 2n\beta(\delta, \kappa)\left(c' d\left[\frac{2B\sigma}{\sqrt{6d}\delta} + \rho(Q^\star - Q)/\delta\right] + 2e^{-B^2/2} + e^{-d}\right) ,$$

for any $B^2 > 0$, where for some matrix M, $\rho(M)$ is defined as its spectral radius.

Lemma 5 (proved in next subsection) gives, for any $B > 0$:

$$\begin{aligned}
F(\hat{Q}) &\leqslant F(Q_\varepsilon^\star) = \mathbb{E}\left[F(Q_\varepsilon^\star)\right] + F(Q_\varepsilon^\star) - \mathbb{E}\left[F(Q_\varepsilon^\star)\right] \\
&\leqslant 2n\beta(\delta, \kappa)\left(c' d\left[\frac{2B\sigma}{\sqrt{6d}\delta} + \rho(Q^\star - Q)/\delta\right] + 2e^{-B^2/2} + e^{-d}\right) + F(Q_\varepsilon^\star) - \mathbb{E}\left[F(Q_\varepsilon^\star)\right] \\
&\leqslant 2n\beta(\delta, \kappa)\left(c' d\left[\frac{2B\sigma}{\sqrt{6d}\delta} + \frac{\varepsilon}{\delta}\right] + 2e^{-B^2/2} + e^{-d}\right) + \sup_{Q \in \mathcal{N}} \left|F(Q) - \mathbb{E}\left[F(Q)\right]\right| ,
\end{aligned} \tag{18}$$

where we used $\rho(Q^\star - Q_\varepsilon^\star) \leqslant \left\|\hat{Q} - Q_\varepsilon^\star\right\|_F \leqslant \varepsilon$ in the above.

Step 3: A lower bound on $\mathbb{E}\left[F(Q)\right]$ for any Q. W lower bound the right hand side in (17) using the following Lemma.

Lemma 6. *Let $Q \in \mathcal{O}(d)$, $u \in \mathcal{S}^{d-1}$. We have, conditionally on the directions $u_1, \dots, u_p$,*

$$\mathbb{E}\left[F(Q)\middle|u_1, \dots, u_p\right] \geqslant 2C_1\beta(\delta,\kappa)\frac{\sum_{k=1}^p 1_{\left\{\|(Q^\star - Q)u_k\|_F^2 > 4\delta + 32\sigma\right\}}}{p}\,.$$

We recall that $\beta(\delta,\kappa)$ is defined in (16) here above. In the sequel, we denote by $\mathbb{P}_U$ (resp. $\mathbb{E}_U$) the probability (resp. expectation) over the directions $u_1, \dots, u_p$. Lemma 6 (proved in next subsection) gives that

$$\begin{aligned}
\inf\left\{F(Q), Q \in \mathcal{N}, \|Q - Q^\star\|_F^2 > \delta_0 d\right\} &\geqslant \inf_{\substack{Q\in\mathcal{N}\\ \|Q-Q^\star\|_F^2>\delta_0 d}} \mathbb{E}\left[F(Q)\right] - \sup_{Q\in\mathcal{N}} |F(Q) - \mathbb{E}\left[F(Q)\right]| \\
&\geqslant \frac{2C_1 n\beta(\delta,\kappa)}{p} \inf_{\substack{Q\in\mathcal{N}\\ \|Q-Q^\star\|_F^2>\delta_0 d}} \mathbb{E}_U\left[\sum_{k=1}^p 1_{\left\{\|(Q-Q^\star)u_k\|^2 > 4\delta + 32\sigma\right\}}\right] \qquad (19)\\
&\quad - \sup_{Q\in\mathcal{N}} |F(Q) - \mathbb{E}\left[F(Q)\right]|\,. \qquad (20)
\end{aligned}$$

Now, if we take $\delta_0 \geqslant 8\delta + 64\sigma$, we have

$$\begin{aligned}
&\inf_{\substack{Q\in\mathcal{N}\\ \|Q-Q^\star\|_F^2>\delta_0 d}} \mathbb{E}_U\left[\sum_{k=1}^p 1_{\left\{\|(Q-Q^\star)u_k\|^2 > 4\delta + 32\sigma\right\}}\right] \qquad (21)\\
&\qquad \geqslant \inf_{\substack{Q\in\mathcal{N}\\ \|Q-Q^\star\|_F^2>\delta_0 d}} \mathbb{E}_U\left[\sum_{k=1}^p 1_{\left\{\|(Q-Q^\star)u_k\|^2 > \delta_0/2\right\}}\right]\\
&\qquad \geqslant \inf_{\substack{Q\in\mathcal{N}\\ \|Q-Q^\star\|_F^2>\delta_0 d}} \mathbb{E}_U\left[\sum_{k=1}^p 1_{\left\{\|(Q-Q^\star)u_k\|^2 > \frac{\|Q-Q^\star\|_F^2}{2d}\right\}}\right]\\
&\qquad = p \inf_{\substack{Q\in\mathcal{N}\\ \|Q-Q^\star\|_F^2>\delta_0 d}} \mathbb{P}_U\left(\|(Q-Q^\star)u_1\|^2 > \frac{\|Q-Q^\star\|_F^2}{2d}\right).
\end{aligned}$$

Note that if $Z \sim \mathcal{N}(0, I_d/d)$, by rotational invariance of the Gaussian, one can always write $Z = Nu_1$ where $N = \|Z\|$ and $u_1 = \frac{Z}{\|Z\|}$ are independent and u_1 is uniform on the sphere. This yields $\frac{\|Q-Q^\star\|_F^2}{d} = \mathbb{E}\left[\|(Q-Q^\star)Z\|^2\right] = \mathbb{E}\left[N^2\right]\mathbb{E}_U[\|(Q-Q^\star)u_1\|^2] = 1 \times \mathbb{E}_U[\|(Q-Q^\star)u_1\|^2]$.

We can lower bound the right hand side of the above using a reverse Markov inequality, namely that $\mathbb{P}\left(X > \mathbb{E}\left[X\right]/2\right) \geqslant \mathbb{E}\left[X\right]/8$ for any X such that $0 \leqslant X \leqslant 4$ a.s. We apply this to $X = \|(Q-Q^\star)u_1\|^2$ and get that for all $Q \in \mathcal{N}$ such that $\|Q - Q^\star\|_F^2 > \delta_0 d$,

$$\mathbb{P}_U\left(\|(Q-Q^\star)u_1\|^2 > \frac{\|Q-Q^\star\|_F^2}{2d}\right) \geqslant \frac{\|Q-Q^\star\|_F^2}{8d} \geqslant \frac{\delta_0}{8},$$

and via Equation (21), Equation (19) becomes

$$\inf\left\{F(Q), Q \in \mathcal{N}, \|Q - Q^\star\|_F^2 > \delta_0 d\right\} \geqslant \frac{C_1 n\beta(\delta,\kappa)\delta_0}{4} - \sup_{Q\in\mathcal{N}} |F(Q) - \mathbb{E}\left[F(Q)\right]|\,. \qquad (22)$$

Step 4: Uniform concentration of $F(Q)$ around its mean. The remaining step is to control the concentration of $F(Q)$ uniformly on $\mathcal{N}$. This is given by the following.

Lemma 7 (Concentration of F)**.** *Let $Q \in \mathcal{O}(d)$ be fixed. Recall that $q = p^3$, that $v_1, \dots, v_q$ are i.i.d. uniformly sampled on the sphere $\mathcal{S}^{d-1}$ and that $u_1, \dots, u_p$ are i.i.d. uniformly sampled in $\{v_1, \dots, v_q\}$. We have, for all $\lambda > 0$:*

$$\mathbb{P}\left(\left|F(Q) - \mathbb{E}\left[F(Q)\right]\right| \geqslant \frac{4\sqrt{2}\lambda\left(3\log(p) + \lambda + \frac{\log(q)^2+\lambda^2}{9n\beta(\delta,\kappa)}\right)n\beta(\delta,\kappa)}{\sqrt{p}}\right) \leqslant 4e^{-\lambda} + 2e^{-\lambda^2}\,.$$

Hence, plugging $\lambda = 2\log(|\mathcal{N}|) > 1$ (assuming $|\mathcal{N}| \geqslant 2$) in Lemma 7, with probability at least $1 - \frac{6}{|\mathcal{N}|}$, we have

$$\sup_{Q\in\mathcal{N}} |F(Q) - \mathbb{E}\left[F(Q)\right]| \tag{23}$$

$$\leqslant 8\sqrt{2}\log(|\mathcal{N}|)\left(3\log(p) + 2\log(|\mathcal{N}|) + \frac{9\log(p)^2 + 4\log(|\mathcal{N}|)^2}{9n\beta(\delta,\kappa)}\right) n\beta(\delta,\kappa)p^{-1/2}$$

$$\leqslant c_4 d^2\log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))n\beta(\delta,\kappa)p^{-1/2} \tag{24}$$

$$+ c_5 d^2\log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))^2 p^{-1/2}\,.$$

where we used $\log(|\mathcal{N}|) \leqslant c_6 d^2\log(1/\varepsilon)$ in the above.

Step 5: Wrapping things up. Putting together the control on deviation in (23), the upper bound (18) and the lower bound (22), one gets that with probability $\geqslant 1 - 6/|\mathcal{N}|$, for any $B > 0$:

$$\inf\left\{F(Q), Q\in\mathcal{N}, \|Q - Q^\star\|_F^2 > \delta_0 d\right\} - F(\hat{Q}) \tag{25}$$

$$\geqslant \frac{C_1 n\beta(\delta,\kappa)\delta_0}{4} - 2n\beta(\delta,\kappa)\left(c' d\left[\frac{2B\sigma}{\sqrt{6d}\delta} + \frac{\varepsilon}{\delta}\right] + 2e^{-B^2/2} + e^{-d}\right)$$

$$- 2\sup_{Q\in\mathcal{N}} |F(Q) - \mathbb{E}\left[F(Q)\right]|$$

$$\geqslant n\beta(\delta,\kappa)\frac{C_1\delta_0}{4}$$

$$- n\beta(\delta,\kappa)\left(c' d\left[\frac{2B\sigma}{\sqrt{6d}\delta} + \frac{\varepsilon}{\delta}\right] + 2e^{-B^2/2} + e^{-d} - c_4 d^2\log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))p^{-1/2}\right)$$

$$- c_5 d^2\log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))^2 p^{-1/2}\,. \tag{26}$$

If this lower bound is positive, we can conclude that $\left\|\hat{Q} - Q^\star\right\|_F^2 \leqslant \delta_0 d$, as desired.

So far, the only constraints on our constants are

$$\text{(A1)} \quad \delta_0 \geqslant 8\delta + 64\sigma\,.$$

While the noise parameter σ is fixed in this proof, we have the freedom to impose some constraints on ε (the granularity of the $\varepsilon-$net), δ (the width of the cones), κ (the truncature parameter), p (the number of directions) and B in order to make the expression in (25) positive (and even $\gg 1$). The remaining step is to show that this is possible ; this is what we shall do now.

Recall that we are in a regime where we need to keepn in our minf that n tends to $+\infty$ and d tends to $+\infty$ with n but with $d \leqslant \log(n)$. We want to show that the positive term in (25) can dominate the others ; first, we would like to have an inequality of the form

$$\frac{C_1\delta_0}{4} - c' d\left[\frac{2B\sigma}{\sqrt{6d}\delta} + \frac{\varepsilon}{\delta}\right] + 2e^{-B^2/2} + e^{-d} - c_4 d^2\log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))p^{-1/2} > c_6\delta_0\,, \tag{27}$$

which is going to be satisfied if:

(A2) $C_1 > 8c_6$,

(A3) $\delta_0\delta \geqslant c_7\sqrt{d}\sigma B$,

(A4) $\delta_0\delta \geqslant c_8 d\varepsilon$,

(A5) $\delta_0 \geqslant c_9(e^{-B^2/2} + e^{-d})$,

(A6) $\delta_0 \geqslant c_{10} d^2\log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))p^{-1/2}$,

where c_7, c_8, c_9, c_{10} are large enough constants. Now,

- (A2) is easily verified by choosing c_6.
- B only appears in (A3) and (A5). (A5) is satisfied if we take $\delta_0 \geqslant 2c_9 e^{-d}$ and $B^2 = c_{11}\max(1, \log(1/\delta_0))$, transforming (A3) into $\frac{\delta_0 \delta}{\sqrt{\max(1,\log(1/\delta_0))}} \geqslant c_{12}\sqrt{d}\sigma$;
- ε appears in (A4) and can be taken as $\varepsilon \leqslant \frac{\delta_0 \delta}{c_8 d}$ for this condition to be satisfied. Combined with (A2), we can simply take $\varepsilon \leqslant c_7\sqrt{d}\sigma B/(c_8 d)$ for this condition to be redundant;
- p only appears in (A6) and can thus be taken as large as desired to have this inequality satisfied (very large p does not degrade any bound).

Consequently, the inequality in Equation (27) is satisfied for parameters that satisfy:

$$\text{(A7)} \quad \varepsilon = c_{16} d^{-1/2}\sigma\,, \qquad \text{(A8)} \quad p^{1/2} \geqslant c_{13}\sigma^{-1} d^{7/2}\log(d/\sigma)^3\,,$$

$$\text{(A9)} \quad \delta = \delta_0\,, \qquad \text{(A10)} \quad \frac{\delta_0^2}{\max(1,\log(1/\delta_0))} \geqslant c_{15}\sqrt{d}\sigma\,,$$

thereby transforming the condition that the RHS in Equation (25) is positive into:

$$n\beta(\delta,\kappa) \geqslant c_{16} d^2 \log(1/\varepsilon)\max(\log(p), d^2\log(1/\varepsilon))^2 p^{-1/2}\left[\sqrt{d}\sigma\log(\sqrt{d}\sigma)\right]^{-1}\,.$$

The RHS of this inequality can be taken smaller than 1 by imposing that p is large enough (recall that p can be taken as large as desired). The final condition thus reads as $n\beta(\delta,\kappa) \geqslant 1$, which is itself satisfied if

$$n \geqslant e^{c' d\log(1/\delta)}(1 - e^{-d/16})^{-1}\,,$$

since $\beta(\delta,\kappa) = \mathbb{P}\left(x_1 \in \mathcal{C}_{\mathcal{X}}(u_0,\delta), \|x_1\| \geqslant 1/\kappa\right)$ and $\mathbb{P}\left(x_1 \in \mathcal{C}_{\mathcal{X}}(u_0,\delta)\right) \geqslant e^{-c' d\log(1/\delta)}$, while $\mathbb{P}\left(\|x_1\| \geqslant 1/\kappa\right) \geqslant 1 - e^{-d/16}$ for

$$\text{(A12)} \quad \kappa^2 = \frac{2}{d}\,.$$

We are now going to use the low-dimensionality assumption $d \ll \log(n)$, since $n \geqslant e^{c' d\log(1/\delta)}(1 - e^{-d/16})$ will be verified for

$$\text{(A13)} \quad \log(n) \geqslant c'' d\log(1/\delta) = c'' d\log(1/\delta_0)\,.$$

Thus, under (A8-A13), Equation (25) is positive, and therefore we have that $\frac{1}{d}\left\|\hat{Q} - Q^\star\right\|_F^2 \leqslant \delta_0$. $\square$

D.2 Misceallenous lemmas on the path to proving Theorem 3

We introduce the (numerical) constants $c, c' > 0$ that verify, for all $\delta' \in (0, 1/4)$ that for x sampled uniformly on $\mathcal{S}^{d-1}$ and any $u \in \mathcal{S}^{d-1}$ we have[5]:

$$\exp\big(-c' d\log(1/\delta')\big) \leqslant \mathbb{P}\left(x \in \mathcal{C}(u,\delta')\right) \leqslant \exp\big(-cd\log(1/\delta')\big)\,.$$

The following lemmas are used to prove Lemma 6 and Lemma 5.

[5]In our model, the probability $\mathbb{P}\left(x \in \mathcal{C}(u,\delta')\right)$ can in fact be computed explicitly. For fixed d and n, the above probability is given by $(1/2)\mathbb{P}\left(X_1^2 \geqslant (1-\delta)^2(X_1^2 + \dots X_d^2)\right)$ where the (X_i) are standard i.i.d. Gaussian variables. It is standard that $\frac{X_1^2}{\|X\|^2}$ is distributed according to the beta distribution $\beta(1/2,(d-1)/2)$, hence

$$\mathbb{P}\left(x \in \mathcal{C}(u,\delta')\right) = \frac{1}{2}\mathbb{P}\left(\beta(1/2,(d-1)/2) \geqslant (1-\delta)^2\right) = \frac{\Gamma(d/2)}{\Gamma(1/2)\Gamma(\frac{d-1}{2})}\int_{(1-\delta)^2}^1 x^{-1/2}(1-x)^{(d-3)/2}dx\,,$$

which is indeed of order $c\delta^d$ when δ is small.

Proof of Lemma 5. We recall that for any i, j, $\mathbb{P}\left(x_i \in \mathcal{C}_{\mathcal{X}}(u,\delta)\right) = \mathbb{P}\left(y_j \in \mathcal{C}_{\mathcal{X}}(u,\delta)\right) = \beta(\delta,\kappa)$ for any unit vector u. Taking the expectation and developing the indicators, we have

$$
\begin{aligned}
&\mathbb{E}\left[F(Q)\right] \\
&= \frac{1}{p}\sum_{k=1}^{p}\sum_{i=1}^{n}\left(\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}(Qu_k,\delta)\right) + \mathbb{P}\left(y_i \in \mathcal{C}_{\mathcal{Y}}(u_k,\delta)\right) - 2\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}(Qu_k,\delta) | y_i \in \mathcal{C}_{\mathcal{Y}}(u_k,\delta)\right)\mathbb{P}\left(y_i \in \mathcal{C}_{\mathcal{Y}}(u_k,\delta)\right)\right) \\
&= \frac{2\beta(\delta,\kappa)}{p}\sum_{k=1}^{p}\sum_{i=1}^{n}\left(1 - \mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Qu_k,\delta), \left\|x_{\pi^\star(i)}\right\| \geqslant 1/\kappa \middle| y_i \in \mathcal{C}(u_k,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right)\right) \\
&= 2n\beta(\delta,\kappa)\left(1 - \mathbb{P}\left(X \in \mathcal{C}(Qu,\delta), \|X\| \geqslant 1/\kappa \middle| Y \in \mathcal{C}(u,\delta), \|Y\| \geqslant \sqrt{1+\sigma^2}/\kappa\right)\right),
\end{aligned}
$$

where $X \sim \mathcal{N}(0, I_d)$, $Y = Q^\star X + \sigma Z$ with $Z \sim \mathcal{N}(0, I_d)$ independent from X, and for any unit vector u. We thus need to bound the last term, which is done by noticing that the two events in the remaining probability become highly positively correlated when Q is close to $Q^\star$. First, we separate the norm component from the direction component in the event $\{X \in \mathcal{C}(Qu,\delta), \|X\| \geqslant 1/\kappa\}$:

$$
\begin{aligned}
&\mathbb{P}\left(X \in \mathcal{C}(Qu,\delta), \|X\| \geqslant 1/\kappa \middle| Y \in \mathcal{C}(u,\delta), \|Y\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&= \mathbb{P}\left(X \in \mathcal{C}(Qu,\delta) \middle| Y \in \mathcal{C}(u,\delta), \|Y\| \geqslant \sqrt{1+\sigma^2}/\kappa\right)\mathbb{P}\left(\|X\| \geqslant 1/\kappa \| \|Y\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&\geqslant \mathbb{P}\left(X \in \mathcal{C}(Qu,\delta) \middle| Y \in \mathcal{C}(u,\delta), \|Y\| \geqslant \sqrt{1+\sigma^2}/\kappa\right)\mathbb{P}\left(\|X\| \geqslant 1/\kappa\right).
\end{aligned}
$$

Now, we have $y_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta) \implies x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}(Q^\top u, \delta + \delta_i(Q,u))$ using Lemma 12, where $\delta_i(Q) = 2\sigma\frac{|\langle z_i, (Q^\star)^\top u\rangle|}{\left\|x_{\pi^\star(i)}\right\|} + \rho(Q^\star - Q)$, so that:

$$
\begin{aligned}
&\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Q,\delta) \middle| y_i \in \mathcal{C}(u,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&= \mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Q,\delta) \middle| y_i \in \mathcal{C}(u,\delta), x_{\pi^\star(i)} \in \mathcal{C}(Q, \delta + \delta_i(Q,u)), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&\geqslant \mathbb{E}\left[\exp\left(-c'd\log\left(1 + 2\sigma\frac{|\langle z_i, (Q^\star)^\top u\rangle|}{\delta\left\|x_{\pi^\star(i)}\right\|} + \rho(Q^\star - Q)/\delta\right)\right) \middle| \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right] \\
&\geqslant \mathbb{E}\left[\exp\left(-c'd\log\left(1 + \frac{2\sigma B'}{\delta\left\|x_{\pi^\star(i)}\right\|} + \rho(Q^\star - Q)/\delta\right)\right) \middle| \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa, |\langle z_i, (Q^\star)^\top u\rangle| \leqslant B'\right] \\
&\qquad \times \mathbb{P}\left(|\langle z_i, (Q^\star)^\top u\rangle| \leqslant B'\right) \\
&\geqslant \exp\left(-c'd\log\left(1 + \frac{2\kappa\sigma B}{\delta} + \rho(Q^\star - Q)/\delta\right)\right)\left(1 - \mathbb{P}\left(|\langle z_i, (Q^\star)^\top u\rangle| > B\right)\right)\left(1 - \mathbb{P}\left(\left\|x_{\pi^\star(i)}\right\| \geqslant 1/\kappa\right)\right).
\end{aligned}
$$

First, $\mathbb{P}\left(\left\|x_{\pi^\star(i)}\right\|^2 > d + 2\sqrt{dt} + t\right) \leqslant e^{-t}$ for any $t > 0$, so that if $1/\kappa^2 \geqslant 3d$, $\mathbb{P}\left(\left\|x_{\pi^\star(i)}\right\| > 1/\kappa\right) \leqslant e^{-\frac{1}{3\kappa^2}+d}$.

Then, $|\langle z_i, (Q^\star)^\top u\rangle| \sim |\mathcal{N}(0,1)|$ since u us unitary, and thus $\mathbb{P}\left(|\langle z_i, (Q^\star)^\top u\rangle| > B\right) \leqslant 2e^{-B^2/2} \leqslant 2e^{-2}$ for $B = 2$, leading to:

$$
\begin{aligned}
&\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Q,\delta) \middle| y_i \in \mathcal{C}(u,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&\geqslant \exp\left(-c'd\log\left(1 + \frac{2B\kappa\sigma}{\delta} + \rho(Q^\star - Q)/\delta\right)\right)(1 - 2e^{-B^2/2})(1 - e^{-\frac{1}{3\kappa^2}+d}) \\
&\geqslant \exp\left(-c'd\log\left(1 + \frac{2N\sigma}{\sqrt{6d}\delta} + \rho(Q^\star - Q)/\delta\right)\right)(1 - 2e^{-B^2/2} - e^{-d}),
\end{aligned}
$$

for $\kappa^2 = 1/(6d)$. Using $\log(1+x) \leqslant x$ and $e^{-x} \geqslant 1-x$ for $x \geqslant 0$,

$$\begin{aligned}
&\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Q,\delta) \middle| y_i \in \mathcal{C}(u,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&\geqslant \exp\left(-c'd\big(\frac{2B\sigma}{\sqrt{6d\delta}} + \rho(Q^\star - Q)/\delta\big)\right)(1 - 2e^{-B^2/2} - e^{-d}) \\
&\geqslant \left(1 - c'd\big(\frac{2B\sigma}{\sqrt{6d\delta}} + \rho(Q^\star - Q)/\delta\big)\right)(1 - 2e^{-B^2/2} - e^{-d}) \\
&\geqslant \left(1 - c'd\left[\frac{2B\sigma}{\sqrt{6d\delta}} + \rho(Q^\star - Q)/\delta\right] - 2e^{-B^2/2} - e^{-d}\right).
\end{aligned}$$

leading to

$$\begin{aligned}
&1 - \mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Q,\delta) \middle| y_i \in \mathcal{C}(u,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa\right) \\
&\leqslant c'd\left[\frac{2B\sigma}{\sqrt{6d\delta}} + \rho(Q^\star - Q)/\delta\right] + 2e^{-B^2/2} + e^{-d},
\end{aligned}$$

and thus to the desired upper bound on $\mathbb{E}\left[F(Q)\right]$. □

Proof of Lemma 7. We first begin by bounding all the terms that appear in the sum of $F(Q)$. Define

$$A(u,Q) = \left|\mathcal{C}_{\mathcal{X}}(Qu,\delta)\right| - \left|\mathcal{C}_{\mathcal{Y}}(u,\delta)\right|,$$

so that $F(Q) = \frac{1}{p}\sum_{k=1}^{p} A(u_k,Q)^2$. Using Bernstein inequality [Vershynin, 2018, Theorem 2.8.4], and writing $\beta(\delta,\kappa) = \mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}(Qu,\delta)\right)$ (so that $\mathbb{E}\left[A(u,Q)\right] \leqslant n\beta(\delta,\kappa)$), we have:

$$\mathbb{P}\left(|A(u,Q)| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2/2}{n\beta(\delta,\kappa) + t/3}\right),$$

so that

$$\begin{aligned}
\mathbb{P}\left(A(u,Q)^2 \geqslant n\beta(\delta,\kappa)t\right) &\leqslant 2\exp\left(-\frac{n\beta(\delta,\kappa)t/2}{n\beta(\delta,\kappa) + \sqrt{n\beta(\delta,\kappa)t}/3}\right) \\
&\leqslant 2\exp\left(-\frac{t}{4}\right) + 2\exp\left(-\frac{3\sqrt{n\beta(\delta,\kappa)t}}{2}\right).
\end{aligned}$$

Now, we have:

$$\begin{aligned}
\mathbb{P}\left(\exists \ell \in [q], \quad A(v_\ell,Q)^2 \geqslant n\beta(\delta,\kappa)t\right) &\leqslant 2\exp\left(-\frac{t}{4} + \log(q)\right) + 2\exp\left(-\frac{3\sqrt{n\beta(\delta,\kappa)t}}{2} + \log(q)\right) \\
&= 4e^{-\lambda},
\end{aligned}$$

for $t = 4(3\log(p) + \lambda) + \frac{4(\log(q)^2+\lambda^2)}{9n\beta(\delta,\kappa)}$. We now use MacDiarmid's inequality [Vershynin, 2018, Theorem 2.9.1], by seeing $F(Q)$ as $F(Q) = f(u_1,\ldots,u_p)$, conditionally on the event $\forall \ell \in [q], A(v_\ell,Q)^2 \leqslant \left(4(3\log(p)+\lambda) + \frac{4(\log(q)^2+\lambda^2)}{9n\beta(\delta,\kappa)}\right)n\beta(\delta,\kappa) = B$, to obtain

$$\mathbb{P}\left(\left|F(Q) - \mathbb{E}\left[F(Q)|V\right]\right| \geqslant t \middle| \forall \ell \in [q], A(v_\ell,Q)^2 \leqslant B\right) \leqslant 2\exp\left(\frac{pt^2}{2B^2}\right),$$

where $V = \{v_1,\ldots,v_q\}$, since the bounded difference inequality is then verified for constant $4B$. Thus,

$$\mathbb{P}\left(\left|F(Q) - \mathbb{E}\left[F(Q)|V\right]\right| \geqslant \frac{\sqrt{2}\left(4(3\log(p)+\lambda) + \frac{4(\log(q)^2+\lambda^2)}{9n\beta(\delta,\kappa)}\right)n\beta(\delta,\kappa)}{\sqrt{p}}\right) \leqslant 4e^{-\lambda} + 2e^{-\lambda^2}.$$

The problem here lies in the fact that $\mathbb{E}[F(Q)|V] = \mathbb{E}[F(Q)]$ may not always hold! Hopefully this is in fact the case:

$$\begin{aligned}\mathbb{E}[F(Q)|V] &= \frac{1}{pq}\sum_{k=1}^{p}\sum_{\ell=1}^{q}\mathbb{E}\left[(|\mathcal{C}_{\mathcal{X}}(Qv_\ell,\delta)| - |\mathcal{C}_{\mathcal{Y}}(v_\ell,\delta)|)^2|u_k = v_\ell\right] \\ &= \mathbb{E}\left[(|\mathcal{C}_{\mathcal{X}}(Qv,\delta)| - |\mathcal{C}_{\mathcal{Y}}(v,\delta)|)^2\right] \qquad \text{for any fixed } v \in \mathcal{S}^{d-1} \\ &= \mathbb{E}[F(Q)]\ ,\end{aligned}$$

concluding the proof. □

Proof of Lemma 6. Let $\varepsilon > 0$ to be determined later and $k \in [p]$ such that $\|(Q-Q^\star)u_k\| > \varepsilon$. We are going to show that $\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Qu_k,\delta)\middle| y_i \in \mathcal{C}(u_k,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa, \|(Q-Q^\star)u_k\| > \varepsilon\right)$ is small.

Using Lemma 11, $y_i \in \mathcal{C}(u_k,\delta)$ implies that $x_{\pi^\star(i)} \in \mathcal{C}(Q^\star u_k, \delta + 2\frac{\sigma\|z_i\|}{\|x_{\pi^\star(i)}\|})$. Then, $\mathcal{C}(Qu_k,\delta) \cap \mathcal{C}(Q^\star u_k, \delta + 2\frac{\sigma\|z_i\|}{\|x_{\pi^\star(i)}\|}) = \varnothing$ provided that $\|Qu_k - Q^\star u_k\|^2 > 4(\delta + \frac{\sigma\|z_i\|}{\|x_{\pi^\star(i)}\|})$ using Lemma 8. Thus, if $\|Qu_k - Q^\star u_k\|^2 > \varepsilon$,

$$\begin{aligned}&\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Qu_k,\delta)\middle| y_i \in \mathcal{C}(u_k,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa, \|(Q-Q^\star)u_k\| > \varepsilon\right) \\ &\leqslant \mathbb{P}\left(4(\delta + \frac{\sigma\|z_i\|}{\|x_{\pi^\star(i)}\|}) > \varepsilon\right) \\ &\leqslant \mathbb{P}\left(\frac{4\sigma\|z_i\|}{\|x_{\pi^\star(i)}\|} > \varepsilon - 4\delta\right)\ .\end{aligned}$$

We have $\mathbb{P}\left(\|z_i\|^2 \geqslant 4d\right) \leqslant e^{-d}$, and $\mathbb{P}\left(\|x_{\pi^\star(i)}\|^2 \leqslant \frac{d}{2}\right) \leqslant e^{-d/16}$, so that if $\varepsilon \geqslant 4\delta + 32\sigma$, we have $\mathbb{P}\left(\frac{4\sigma\|z_i\|}{\|x_{\pi^\star(i)}\|} > \varepsilon - 4\delta\right) \leqslant \mathbb{P}\left(\|z_i\|^2 \geqslant 4d\right) + \mathbb{P}\left(\|x_{\pi^\star(i)}\|^2 \leqslant \frac{d}{2}\right) \leqslant e^{-d} + e^{-16d}$, leading to

$$\mathbb{P}\left(x_{\pi^\star(i)} \in \mathcal{C}(Qu_k,\delta)\middle| y_i \in \mathcal{C}(u_k,\delta), \|y_i\| \geqslant \sqrt{1+\sigma^2}/\kappa, \|(Q-Q^\star)u_k\| > \varepsilon\right) \leqslant e^{-d} + e^{-d/16} \leqslant 1 - C_1\ ,$$

where $C_1 = 1/e + 1/e^{1/16} > 0$ is a numerical constant. This thus gives:

$$\mathbb{E}[F(Q)|U] \geqslant 2C_1\beta(\delta,\kappa)\frac{\sum_{k=1}^{p} 1_{\left\{\|(Q^\star-Q)u_k\|_F^2 > \varepsilon\right\}}}{p}\ .$$

□

Lemma 8 (Cone separation)**.** *For $u, v \in \mathcal{S}^{d-1}$, $\mathcal{C}(u,\delta) \cap \mathcal{C}(v,\delta) \neq \varnothing$ implies that $\|u-v\|^2 \leqslant 8\delta$.*

Proof. Assume $\mathcal{C}(u,\delta) \cap \mathcal{C}(v,\delta) \neq \varnothing$. Take $w \in \mathcal{C}(u,\delta) \cap \mathcal{C}(v,\delta)$: we can always assume that $\|w\| = 1$ by rescaling. Then, by triangle inequality, we have $\|u-v\| \leqslant \|u-w\| + \|v-w\|$. Since $w \in \mathcal{C}(u,\delta)$, $\|u-w\|^2 = 2 - 2\langle v, w\rangle \leqslant 2 - 2(1-\delta) = 2\delta$, and the same is true for $\|v-w\|$. This gives $\|u-v\| \leqslant 2\sqrt{2\delta}$. □

Lemma 9 (Probability that two cones are disjoint)**.** *Let $Q, Q' \in \mathcal{O}(d)$, $\delta \leqslant \frac{1}{12d}\|Q'-Q\|_F^2$ and let u be a random variable uniformly distributed over $\mathcal{S}^{d-1}$. Then,*

$$\mathbb{P}\left(\mathcal{C}(Q'u,\delta) \cap \mathcal{C}(Qu,\delta) = \varnothing\right) \geqslant \delta\ .$$

Proof. Using the previous Lemma, $\mathbb{P}\left(\mathcal{C}(Q'u,\delta) \cap \mathcal{C}(Qu,\delta) \neq \varnothing\right) \leqslant \mathbb{P}\left(\|Qu - Q'u\|^2 \leqslant 8\delta\right)$. Let Z be the random variable $Z = \|Qu - Q'u\|^2$. We have that $\mathbb{E}[Z] = \|Q-Q'\|_F^2/d \geqslant 12\delta$ and $Z \leqslant 4$ almost surely. Thus, using a "reverse Markov" inequality,

$$\begin{aligned}12\delta \leqslant \mathbb{E}[Z] &= \mathbb{E}[Z1_{Z\leqslant 8\delta}]\,\mathbb{P}(Z \leqslant 8\delta) + \mathbb{E}[Z1_{Z>8\delta}]\,\mathbb{P}(Z > 8\delta) \\ &\leqslant 8\delta\mathbb{P}(Z \leqslant 8\delta) + 4\mathbb{P}(Z > 8\delta) \\ &\leqslant 8\delta + 4(1 - \mathbb{P}(X \leqslant 8\delta))\ ,\end{aligned}$$

that is $\mathbb{P}(X \leqslant 8\delta) \leqslant 1 - \delta$, which concludes the proof. □

The following Lemma is easy and does require any proof.

Lemma 10. *For any $u,\ \delta \in (0,1)$, we have $\mathbb{E}\left[|\mathcal{C}_{\mathcal{X}}(u,\delta)|\right] = \mathbb{E}\left[|\mathcal{C}_{\mathcal{Y}}(u,\delta)|\right] = n\mathbb{P}\left(x_1 \in \mathcal{C}(u,\delta), \|x_1\| \geqslant 1/\kappa\right) = n\beta(\delta,\kappa)$ so that $|\mathcal{C}_{\mathcal{X}}(Qu,\delta)| - |\mathcal{C}_{\mathcal{Y}}(u,\delta)|$ in the sum that defines F are all centered.*

$|\mathcal{C}_{\mathcal{X}}(u,\delta)|$ and $|\mathcal{C}_{\mathcal{Y}}(u,\delta)|$ are (correlated) binomial random variables of parameters $(n,\beta(\delta,\kappa))$.

Lemma 11. *For any $u \in \mathcal{S}^{d-1}$, $i \in [n]$, we have $y_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta) \implies x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}((Q^\star)^\top u, \delta + \delta_i)$, where $\delta_i = 2\sigma\frac{\|z_i\|}{\left\|x_{\pi^\star(i)}\right\|}$ and $x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}((Q^\star)^\top u, \delta + \delta_i') \implies y_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta)$, where $\delta_i' = 2\sigma\frac{\|z_i\|}{\|y_i\|}$.*

Proof. Let us prove the first assertion and assume that $y_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta)$. We have $y_i = Q^\star x_{\pi^\star(i)} + \sigma z_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta)$, which writes as:

$$\langle Q^\star x_{\pi^\star(i)} + \sigma z_i, u\rangle \geqslant (1-\delta)\left\|Q^\star x_{\pi^\star(i)} + \sigma z_i\right\|.$$

Thus,

$$\begin{aligned}
\langle x_{\pi^\star(i)}, (Q^\star)^\top u\rangle &\geqslant (1-\delta)\left\|Q^\star x_{\pi^\star(i)} + \sigma z_i\right\| - \sigma\langle z_i, (Q^\star)^\top u\rangle \\
&\geqslant (1-\delta)\left\|Q^\star x_{\pi^\star(i)} + \sigma z_i\right\| - \sigma\|z_i\| \\
&\geqslant (1-\delta)(\left\|Q^\star x_{\pi^\star(i)}\right\| - \sigma\|z_i\|) - \sigma\|z_i\| \\
&\geqslant (1-\delta)\left\|Q^\star x_{\pi^\star(i)}\right\| - 2\sigma\|z_i\| \\
&\geqslant (1-\delta-\delta_i)\left\|Q^\star x_{\pi^\star(i)}\right\|,
\end{aligned}$$

which is the desired result. The second assertion is proved exactly in the same way. □

Lemma 12. *For any $u \in \mathcal{S}^{d-1}$, $i \in [n]$, $Q \in \mathcal{O}(d)$, we have $y_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta) \implies x_{\pi^\star(i)} \in \mathcal{C}_{\mathcal{X}}(Q^\top u, \delta + \delta_i(Q,u))$, where $\delta_i(Q,u) = 2\sigma\frac{|\langle z_i, (Q^\star)^\top u\rangle|}{\left\|x_{\pi^\star(i)}\right\|} + \rho(Q^\star - Q)$.*

Proof. Assume that $y_i \in \mathcal{C}_{\mathcal{Y}}(u,\delta)$. As in the proof of the previous proposition, this reads as:

$$\langle x_{\pi^\star(i)}, (Q^\star)^\top u\rangle \geqslant (1-\delta)\left\|Q^\star x_{\pi^\star(i)} + \sigma z_i\right\| - \sigma\langle z_i, (Q^\star)^\top u\rangle,$$

and thus,

$$\begin{aligned}
\langle x_{\pi^\star(i)}, Q^\top u\rangle &\geqslant \langle x_{\pi^\star(i)}, (Q^\top - (Q^\star)^\top)u\rangle + (1-\delta)\left\|Q^\star x_{\pi^\star(i)} + \sigma z_i\right\| - \sigma\langle z_i, (Q^\star)^\top u\rangle \\
&\geqslant -\left\|x_{\pi^\star(i)}\right\|\rho(Q-Q^\star) + (1-\delta)\left\|Q^\star x_{\pi^\star(i)} + \sigma z_i\right\| - \sigma\langle z_i, (Q^\star)^\top u\rangle \\
&\geqslant -\left\|x_{\pi^\star(i)}\right\|\rho(Q-Q^\star) + (1-\delta-\frac{|\langle z_i, (Q^\star)^\top u\rangle|}{\left\|x_{\pi^\star(i)}\right\|})\left\|x_{\pi^\star(i)}\right\| \\
&= (1-\delta-\delta_i(Q,u))\left\|x_{\pi^\star(i)}\right\|.
\end{aligned}$$

This concludes the proof. □

E Proof of Proposition 1

E.1 Very-fast sorting-based estimator and equivalence with one step of Frank-Wolfe

We have:

$$\nabla f(D) = 2\left(DX^\top XX^\top X - 2Y^\top YDX^\top X + Y^\top YY^\top YD\right),$$

for any bistochastic matrix D, leading to, for $J = \frac{11^\top}{n}$:

$$\nabla f(J) = 2\left(JX^\top XX^\top X - 2Y^\top YJX^\top X + Y^\top YY^\top YJ\right).$$

For any permutation matrix P, we have since $J^\top = J$ and $JP = J$:

$$\begin{aligned}
\langle JX^\top XX^\top X, P\rangle &= \langle X^\top XX^\top X, JP\rangle \\
&= \langle X^\top XX^\top X, J\rangle,
\end{aligned}$$

and similalry:

$$\begin{aligned}\langle Y^\top YY^\top YJ, P\rangle &= \langle JY^\top YY^\top Y, P^\top\rangle \\ &= \langle Y^\top YY^\top Y, JP^\top\rangle \\ &= \langle Y^\top YY^\top Y, J\rangle\,.\end{aligned}$$

Therefore,

$$\operatorname*{arg\,min}_{P\in\mathcal{S}_n}\langle f(J), P\rangle = \operatorname*{arg\,max}_{P\in\mathcal{S}_n}\langle Y^\top YJX^\top X, P\rangle\,.$$

We have $(X^\top X)_{ij} = \langle x_i, x_j\rangle$ and $(JX^\top X)_{ij} = n\langle \bar{x}, x_j\rangle$. Similarly, $(Y^\top YJ)_{ij} = n\langle \bar{y}, y_i\rangle$, and we have $J^2 = J$. Thus,

$$(Y^\top YJX^\top X)_{ij} = n^2\sum_{k=1}^n \langle \bar{y}, y_i\rangle\langle \bar{x}, x_j\rangle\,,$$

leading to:

$$\operatorname*{arg\,min}_{P\in\mathcal{S}_n}\langle f(J), P\rangle = \operatorname*{arg\,max}_{\pi\in\mathcal{S}_n}\sum_{i\in[n]}\sum_{k=1}^n \langle \bar{y}, y_i\rangle\langle \bar{x}, x_{\pi^\star(i)}\rangle\,,$$

and to the following sorting-based estimator, that can be computed very easily in $O(nd\log(n))$ computes.

$$\hat{\pi} \in \operatorname*{arg\,max}_{\pi\in\mathcal{S}_n}\frac{1}{n}\sum_{i=1}^n \langle x_{\pi(i)}, \bar{x}\rangle\langle y_i, \bar{y}\rangle\,, \tag{28}$$

where $\bar{x} = \frac{1}{n}\sum_i x_i$ and $\bar{y} = \frac{1}{n}\sum_i y_i$ are the mean vectors of each point cloud. The idea is that thanks to the scalar product, this estimator gets rid of the orthogonal trasformation. Its strength is that it can be computed in $\mathcal{O}(n\log(n))$ iterations, since it consists in sorting two vectors. We have the following result for this estimator.

Proposition 2. *Let $\delta \in (0,1)$ and $\varepsilon > 0$. If $\sigma \ll n^{\frac{12(1+2\varepsilon)}{\delta}}$, the estimator $\hat{\pi}$ as defined in Equation* (28) *satisfies with high probability:*

$$\mathrm{ov}(\hat{\pi}, \pi^\star) \geqslant 1-\delta\,.$$

Proof. Without loss of generality, we can assume that $\pi^\star = \mathrm{Id}$. Then, for all i,

$$\begin{aligned}\langle y_i, \bar{y}\rangle &= \langle Q^\star x_i + \sigma z_i, Q^\star\bar{x} + \sigma\bar{z}\rangle \\ &= \langle x_i, \bar{x}\rangle + \sigma^2\langle z_i, \bar{z}\rangle + \sigma\langle z_i, Q^\star\bar{x}\rangle + \sigma\langle Q^\star x_i, \bar{z}\rangle\,,\end{aligned}$$

so that:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \langle x_{\pi(i)}, \bar{x}\rangle\langle y_i, \bar{y}\rangle &= \frac{1}{n}\sum_{i=1}^n \langle x_{\pi(i)}, \bar{x}\rangle\langle x_i, \bar{x}\rangle + \frac{1}{n}\sum_{i=1}^n \langle x_{\pi(i)}, \bar{x}\rangle\left[\sigma^2\langle z_i, \bar{z}\rangle + \sigma\langle z_i, Q^\star\bar{x}\rangle + \sigma\langle Q^\star x_i, \bar{z}\rangle\right] \\ &= -\frac{1}{2n}\sum_{i=1}^n \langle x_{\pi(i)} - x_i, \bar{x}\rangle^2 + \frac{1}{n}\sum_{i=1}^n \langle x_i, \bar{x}\rangle^2\langle x_i, \bar{x}\rangle \\ &\quad + \frac{1}{n}\sum_{i=1}^n \langle x_{\pi(i)}, \bar{x}\rangle\left[\sigma^2\langle z_i, \bar{z}\rangle + \sigma\langle z_i, Q^\star\bar{x}\rangle + \sigma\langle Q^\star x_i, \bar{z}\rangle\right]\,.\end{aligned}$$

By definition of $\hat{\pi}$, we have $\frac{1}{n}\sum_{i=1}^n \langle x_{\hat{\pi}(i)}, \bar{x}\rangle\langle y_i, \bar{y}\rangle \geqslant \frac{1}{n}\sum_{i=1}^n \langle x_i, \bar{x}\rangle\langle y_i, \bar{y}\rangle$, that thus writes as:

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^n \langle x_{\pi(i)} - x_i, \bar{x}\rangle^2 &\leqslant \frac{1}{n}\sum_{i=1}^n \langle x_{\hat{\pi}(i)} - x_i, \bar{x}\rangle\left[\sigma^2\langle z_i, \bar{z}\rangle + \sigma\langle z_i, Q^\star\bar{x}\rangle + \sigma\langle Q^\star x_i, \bar{z}\rangle\right] \\ &\leqslant \sup_{i\in[n]}|\langle x_{\hat{\pi}(i)} - x_i, \bar{x}\rangle| \times \frac{1}{n}\sum_{i=1}^n \left|\sigma^2\langle z_i, \bar{z}\rangle + \sigma\langle z_i, Q^\star\bar{x}\rangle + \sigma\langle Q^\star x_i, \bar{z}\rangle\right|\,.\end{aligned}$$

We will first bound this right hand side. First, for all $i, j \in [n]$, $x_i - x_j$ is independent from $\bar{x}$, so that conditionally on $\bar{x}$ we have $\langle x_i - x_j, \bar{x}\rangle \sim \mathcal{N}(0, 2\|\bar{x}\|^2)$, leading to:

$$\mathbb{P}\left(|\langle x_i - x_j, \bar{x}\rangle| > t\|\bar{x}\|\right) \leqslant 2\exp(-t^2/2)\,,$$

and

$$\mathbb{P}\left(\forall i,j\in[n],|\langle x_i-x_j,\bar{x}\rangle|>t\|\bar{x}\|\right)\leqslant 2\exp(-t^2/2+2\log(n))\,,$$

so that with probability $1-2/n^2$, $\sup_{i,j}|\langle x_i-x_j,\bar{x}\rangle|\leqslant 2\sqrt{2\log(n)}\|\bar{x}\|$. Similarly, with probability $1-4/n^2$, $\sup_i|\langle z_i,Q^\star\bar{x}\rangle|\leqslant 2\sqrt{\log(n)}\|\bar{x}\|$ and $\sup_i|\langle Q^\star x_i,\bar{z}\rangle|\leqslant 2\sqrt{\log(n)}\|\bar{z}\|$.

Then, we can write $z_i=z_i'+\bar{z}$ where z_i' is Gaussian (its covariance matrix is the projection on the orthogonal of $\bar{z}$) and independent from $\bar{z}$. Thus, $\sup_i|\langle z_i,\bar{z}\rangle|\leqslant\|\bar{z}\|^2+\sup_i|\langle z_i',\bar{z}\rangle|\leqslant\|\bar{z}\|^2+2\sqrt{\log(n)}\|\bar{z}\|$ with probability $1-2/n^2$.

Thus, with probability $1-8/n^2$ and for $\sigma\leqslant 1$,

$$\begin{aligned}&\sup_{i\in[n]}|\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle|\times\frac{1}{n}\sum_{i=1}^n\left|\sigma^2\langle z_i,\bar{z}\rangle+\sigma\langle z_i,Q^\star\bar{x}\rangle+\sigma\langle Q^\star x_i,\bar{z}\rangle\right|\\&\leqslant 2\sqrt{2\log(n)}\sigma\|\bar{x}\|\left[2\sqrt{\log(n)}\|\bar{x}\|+\|\bar{z}\|^2+4\sqrt{\log(n)}\|\bar{z}\|\right]\end{aligned}$$

Now, $n\|\bar{x}\|^2$ and $n\|\bar{z}\|^2$ are both χ^2_d random variables, so that

$$\mathbb{P}\left(\max(|n\|\bar{x}\|^2-d|,|n\|\bar{z}\|^2-d|)>2t+2\sqrt{dt}\right)\leqslant 4e^{-t}\,.$$

For $t=2(\sqrt{2}-1)d$, this leads to, with probability $4e^{-2(\sqrt{2}-1)d}$:

$$\|\bar{x}\|^2,\|\bar{z}\|^2\in[1/2,3/2]\frac{d}{n}\,.$$

Thus, with probability $1-8/n^2-4e^{-2(\sqrt{2}-1)d}$,

$$\begin{aligned}&\sup_{i\in[n]}|\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle|\times\frac{1}{n}\sum_{i=1}^n\left|\sigma^2\langle z_i,\bar{z}\rangle+\sigma\langle z_i,Q^\star\bar{x}\rangle+\sigma\langle Q^\star x_i,\bar{z}\rangle\right|\\&\leqslant 2\sigma\sqrt{2\log(n)}\|\bar{x}\|^2\left[2\sqrt{\log(n)}+3\sqrt{d/n}+4\sqrt{3\log(n)}\right]\\&=C\log(n)\sigma\|\bar{x}\|^2\,,\end{aligned}$$

for some numerical constant C, if $n\geqslant d$, leading to

$$\frac{1}{2n}\sum_{i=1}^n\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle^2\leqslant C\log(n)\sigma\|\bar{x}\|^2\,.$$

Now, if $i\neq j$ are fixed, for $t\leqslant 1$, $\mathbb{P}\left(\frac{1}{2}\langle x_i-x_j,\bar{x}\rangle^2<t\|\bar{x}\|^2\right)=\mathbb{P}\left(\mathcal{N}(0,1)^2<t\right)\leqslant c\sqrt{t}$, for some constant $c>0$.

We are now going to upper bound the probability of the event $\mathcal{A}=$"there exists $\mathcal{I}\subset[n]$ with $|\mathcal{I}|\geqslant\alpha n$ and π a permutation such that *(i)* for all $i\in\mathcal{I}$, $\pi(i)\neq i$, *(ii)* $\{i,\pi(i)\}_{i\in\mathcal{I}}$ form disjoint pairs and *(iii)* for all $i\in\mathcal{I}$, $\frac{1}{2}\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle^2\leqslant\beta\|\bar{x}\|^2$", for some constants $\alpha,\beta\in(0,1)$ to be fixed later. Let $\mathcal{I}$ and π be fixed. Since $\pi(i)\neq i$, we have $\mathbb{P}\left(\frac{1}{2}\langle x_i-x_{\pi(i)},\bar{x}\rangle^2>\beta\|\bar{x}\|^2\right)\leqslant 2e^{-\beta/2}$, and using *(ii)* all pairs are independent, leading to:

$$\begin{aligned}\mathbb{P}\left(\pi,\mathcal{I}\text{ satisfies }\textit{(i)-(ii)-(iii)}\right)&\leqslant\mathbb{P}\left(\forall i\in\mathcal{I}\,,\quad\frac{1}{2}\langle x_i-x_{\pi(i)},\bar{x}\rangle^2>\beta\|\bar{x}\|^2\right)\\&\leqslant c\sqrt{\beta}\,.\end{aligned}$$

Thus, using a union bound over all possible $\mathcal{I}$ and π, we have that:

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}\right)&\leqslant 2^n n^n e^{-\alpha\beta n/2+\alpha n\log(2)}\\&=e^{\log(c\sqrt{\beta})\alpha n+n\log(n)+(1+\alpha)n\log(2)}\,.\end{aligned}$$

Now, using what we have proved above, denoting $\mathcal{B}$ the event $\frac{1}{2n}\sum_{i=1}^n\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle^2\leqslant C\log(n)\sigma\|\bar{x}\|^2$, we have $\mathbb{P}\left(\mathcal{B}\right)\geqslant 1-8/n^2-4e^{-2(\sqrt{2}-1)d}$. Let $\mathcal{C}$ be the event $\{\mathrm{ov}(\hat{\pi},\pi^\star)\leqslant 1-\delta\}$.

Under $\mathcal{C}\cap\mathcal{B}$, we have the existence of $\mathcal{I}'\subset[n]$ such that for all indices $i\in\mathcal{I}'$, $\hat{\pi}\neq i$ and $|\mathcal{I}'|\geqslant\delta n/6$. Now, since then $\frac{1}{2|\mathcal{I}'|}\sum_{i\in\mathcal{I}'}\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle^2\leqslant\frac{6}{\delta}\times C\log(n)\sigma\|\bar{x}\|^2$, we have that at least half of these indices satisfy $\frac{1}{2}\langle x_{\hat{\pi}(i)}-x_i,\bar{x}\rangle^2\leqslant\frac{12}{\delta}\times C\log(n)\sigma\|\bar{x}\|^2$: we denote by $\hat{\mathcal{I}}$ the set of these indices. Hence, $\hat{\pi},\hat{\mathcal{I}}$ satisfy properties *(i)-(ii)-(iii)* with $\alpha=\frac{\delta}{12}$ and $\beta=\frac{12}{\delta}\times C\log(n)\sigma$, leading to (taking these constants for $\mathcal{A}$):

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}\cap\mathcal{C}\right)&\leqslant\mathbb{P}\left(\mathcal{A}\right)\\&\leqslant e^{\log(c\sqrt{\beta})\alpha n+n\log(n)+(1+\alpha)n\log(2)}\\&=\exp\left(\frac{\delta n\log\left(12C\delta^{-1}\log(n)\sigma\right)}{12}+n\log(n)+2n\log(2)\right).\end{aligned}$$

For this probability to be close to zero, we thus need that $-\frac{\delta n\log\left(12C\delta^{-1}\log(n)\sigma\right)}{12}\geqslant(1+\varepsilon)n\log(n)$, which can be written as:

$$-\log\left(12C\delta^{-1}\log(n)\sigma\right)\geqslant\frac{12(1+\varepsilon)\log(n)}{\delta}=\log\left(n^{\frac{12(1+\varepsilon)}{\delta}}\right),$$

which is satisfied for $\sigma\ll n^{\frac{12(1+2\varepsilon)}{\delta}}$. □

E.2 The "Ace" estimator

Proposition 3 (Ace). *Let $\delta_0>0$. Assume that $\left\|\hat{Q}-Q^\star\right\|_F^2\leqslant 2(1-\delta_0)d$ and $\log n\ll d\ll n$. Then, there exists a constant $C>0$ such that the estimator $\hat{\pi}$ defined in Equation* (4) *satisfies with probability $1-2e^{-d/16}-2n^{-n}$:*

$$\mathrm{ov}(\hat{\pi},\pi^\star)=1-\frac{C}{\delta_0}\max\left(\sqrt{\frac{d\log(d/\delta_0)}{n}+\frac{\log(n)}{d}},\frac{d\log(d/\delta_0)}{n}+\frac{\log(n)}{d}\right).$$

In the $n\gg d\gg\log(n)$ regime: as long as we have non negligible error $\left\|\hat{Q}-Q^\star\right\|_F^2\leqslant 2(1-\varepsilon)d$ (notice that for uniformly random Q, we have $\left\|\hat{Q}-Q^\star\right\|_F^2=2d$), we recover $\pi^\star$ with $1-o(1)$ overlap: doing just a tiny bit better than random for $\hat{Q}$ is enough to recover $\pi^\star$.

Proof of Proposition 3. In this proof we denote $g(\pi):=\frac{1}{n}\sum_{i=1}^n\langle x_{\pi(i)},\hat{Q}^\top y_i\rangle$. By definition, $\hat{\pi}\in\arg\max_{\pi\in\mathcal{S}_n}g(\pi)$. Writing $g(\hat{\pi})\geqslant g(\pi^\star)$ gives

$$\frac{1}{n}\sum_{i=1}^n\langle x_{\hat{\pi}(i)},\hat{Q}^\top Q^\star x_{\pi^\star(i)}\rangle\geqslant\frac{1}{n}\sum_{i=1}^n\langle x_{\pi^\star(i)},\hat{Q}^\top Q^\star x_{\pi^\star(i)}\rangle+\frac{\sigma}{n}\sum_{i=1}^n\langle x_{\pi^\star(i)}-x_{\hat{\pi}(i)},\hat{Q}^\top z_i\rangle. \quad (29)$$

Without loss of generality, we assume $\pi^\star=Id$. The term in the LHS hereabove, for fixed $\hat{\pi},\hat{Q}$, has expectation $\mathrm{ov}(\hat{\pi},\pi^\star)\,\mathrm{Tr}(\hat{Q}^\top Q^\star)$. We are going to compute uniform fluctuations. For some fixed $Q\in\mathcal{O}(d),P\in\mathcal{S}_n$,

$$\sum_{i=1}^n\langle x_{\pi(i)},\hat{Q}^\top Q^\star x_i\rangle=\tilde{X}^\top M\tilde{X},$$

for $\tilde{X}=(x_1^\top,\dots,x_n^\top)^\top\in\mathbb{R}^{nd}$ and $M\in\mathbb{R}^{nd\times nd}$ that writes as $M=\tilde{P}^\top\tilde{Q}$, where $\tilde{Q}\in\mathbb{R}^{nd\times nd}$ is block diagonal with blocks equal to Q, and $\tilde{P}\in\mathbb{R}^{nd\times nd}$ is a block matrix, with blocks of size $n\times n$ that verify $\tilde{P}_{[ij]}=P_{ij}I_n$. Thus, $\|M\|_{\mathrm{op}}=1$ and $\|M\|_F^2=nd$. Using Hanson-Wright inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^n\langle x_{\pi(i)},Qx_i\rangle-n\mathrm{ov}(\pi,Id)\,\mathrm{Tr}(Q)\right|>C(t+\sqrt{ndt})\right)\leqslant 2e^{-t}.$$

Since $d\geqslant\log(n)$, with probability $1-e^{-d/16}$ we have $\sup_{i\in[n]}\|x_i\|\leqslant 2\sqrt{d}$ using Chi concentration. We now work conditionally on this event.

Letting $\mathcal{N}_\delta$ be a $\delta-$net of $\mathcal{O}(d)$,

$$\mathbb{P}\left(\forall \pi \in \mathcal{S}_n\,,\, \forall Q \in \mathcal{N}_\delta\,,\, \left|\sum_{i=1}^n \langle x_{\pi(i)}, Qx_i\rangle - n\mathrm{ov}(\pi, Id)\,\mathrm{Tr}(Q)\right| > C(t+\sqrt{ndt})\right) \leqslant 2e^{-t+n\log(n)+cd^2\log(1/\delta)}\,.$$

Using the fact that $\sum_{i=1}^n \langle x_{\pi(i)}, Qx_i\rangle$ is $n\sup_{i\in[n]}\|x_i\|^2 = 4nd-$Lipschitz in Q, we thus have:

$$\mathbb{P}\left(\forall \pi \in \mathcal{S}_n\,,\, \forall Q \in \mathcal{O}(d)\,,\, \left|\sum_{i=1}^n \langle x_{\pi(i)}, Qx_i\rangle - n\mathrm{ov}(\pi, Id)\,\mathrm{Tr}(Q)\right| > 4nd\delta + C(t+\sqrt{ndt})\right) \leqslant 2e^{-t+n\log(n)+cd^2\log(1/\delta)}\,.$$

Setting $\delta = \frac{\varepsilon}{16}$ and $t = 2n\log(n) + cd^2\log(8/\varepsilon)$, with probability $1-2n^{-n}$, we get that for all π, Q,

$$\left|\sum_{i=1}^n \langle x_{\pi(i)}, Qx_i\rangle - n\mathrm{ov}(\pi, Id)\,\mathrm{Tr}(Q)\right| \leqslant \frac{nd\varepsilon}{4} + C'(n\log(n) + d^2\log(1/\varepsilon) + \sqrt{nd(n\log(n)+d^2\log(1/\varepsilon))})\,.$$

We can thus write, since $\mathrm{Tr}(\hat{Q}^\top Q^\star) \geqslant \varepsilon d$:

$$\frac{1}{n}\sum_{i=1}^n \langle x_{\hat{\pi}(i)}, \hat{Q}^\top Q^\star x_{\pi^\star(i)}\rangle \leqslant \mathrm{Tr}(\hat{Q}^\top Q^\star)d\mathrm{ov}(\pi, Id) + \frac{\varepsilon d}{4} + C'(\log(n) + \frac{d^2\log(1/\varepsilon)}{n} + \sqrt{d(\log(n)+\frac{d^2\log(1/\varepsilon)}{n})})\,.$$

and

$$\frac{1}{n}\sum_{i=1}^n \langle x_{\pi^\star(i)}, \hat{Q}^\top Q^\star x_i\rangle \geqslant \mathrm{Tr}(\hat{Q}^\top Q^\star)d - \frac{\varepsilon d}{4} - C'(\log(n) + \frac{d^2\log(1/\varepsilon)}{n} + \sqrt{d(\log(n)+\frac{d^2\log(1/\varepsilon)}{n})})\,.$$

Similarly than before, we prove that with probability $1-2n^{-n}$, we have for all $\pi \in \mathcal{S}_n, Q \in \mathcal{O}(d)$:

$$\left|\sum_{i=1}^n \langle x_{\pi^\star(i)} - x_{\hat{\pi}(i)}, \hat{Q}^\top z_i\rangle\right| \leqslant \frac{\varepsilon dn}{4} + C'(n\log(n) + d^2\log(1/\varepsilon) + \sqrt{nd(n\log(n)+d^2\log(1/\varepsilon))})\,.$$

Equation (29) thus implies that:

$$\mathrm{Tr}(\hat{Q}^\top Q^\star)d\mathrm{ov}(\pi, Id) \geqslant \mathrm{Tr}(\hat{Q}^\top Q^\star)d - \frac{3\varepsilon d}{4} - 3C'(\log(n) + \frac{d^2\log(1/\varepsilon)}{n} + \sqrt{d(\log(n)+\frac{d^2\log(1/\varepsilon)}{n})})\,,$$

leading to:

$$\mathrm{ov}(\pi, Id) \geqslant 1 - \frac{3\varepsilon}{4\,\mathrm{Tr}(\hat{Q}^\top Q^\star)} - \frac{3C'}{\mathrm{Tr}(\hat{Q}^\top Q^\star)}(\frac{\log(n)}{d} + \frac{d\log(1/\varepsilon)}{n} + \sqrt{\frac{\log(n)}{d} + \frac{d\log(1/\varepsilon)}{n}})\,.$$

Setting $\varepsilon = \frac{\delta_0}{d}$ concludes the proof. □

E.3 Proof of Proposition 1

Proof of Proposition 1. For the first part of Proposition 1, we directly apply Proposition 2 with $\varepsilon = 1/2$ to obtain the result.

For the second part that holds for large dimensions, we apply Proposition 2 for $\varepsilon = 1/2$ and $\delta = 1/8$. Using Lemma 3, the first 'Ping' of **??** 1 leads to $\hat{Q}$ satisfying the assumption of Proposition 3 for some δ_0 bounded away from zero, thus leading to the desired result after the last 'Pong' for $\hat{\pi}$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and scope of the paper are described shortly in the abstract. The introduction section discusses the contributions and the scope more in depth.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   The limitations of the paper are discussed, in particular the fact that our work is mainly theoretical, and we mainly study informational results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides the set of assumptions in every Theorem and Proposition, that are self-contained.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3.3 lists all the needed information to replicate all the experiments in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: see supplementary materials for a notebook.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the details of training procedure and hyperparameters are listed in Section 3.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: the experiments are merely illustrative and due to the n^3 scaling of the LAP, we performed some averagings over 10 runs for each point. But this can be improved easily in a second version.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: In preparing the submission, the authors did not track sufficient information on the computer resources. However, the resources needed to run experiments are minimal, as all can be run on a single CPU. The total compute resources needed are not significant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics and does not pose any potential harm.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is a foundational research paper without any direct societal impact.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses Python with some open-source Python libraries for experiments. There are no other particular existing assets used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no released assets in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.