The goal of this short note is to record tight bounds on various testing problems for d-dimensional Gaussians. Let $\mathbf{q}$ denote the standard Gaussian, i.e., $\mathbf{q} \coloneqq \mathcal{N}(\mathbf{0}, I_d)$. Any mistake or imprecision is mine.

Problem 1 (TV Testing Under Identity Covariance Assumption). *Given $\varepsilon \in (0,1]$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, I_d)$ with unknown μ, distinguish between $\mathbf{p} = \mathbf{q}$ and $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$.*

Sample complexity: $\Theta\Big(\sqrt{d}/\varepsilon^2\Big)$.

Problem 2 (Mean Norm Estimation Under Identity Covariance Assumption). *Given $\varepsilon \in (0,1]$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, I_d)$ with unknown μ, learn $\|\mu\|_2$ to an additive ε.*

Sample complexity: $\Theta\Big(\sqrt{d}/\varepsilon^2\Big)$.

Problem 3 (Mean Estimation Under Identity Covariance Assumption (a.k.a. Gaussian Location Model)). *Given $\varepsilon \in (0,1]$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, I_d)$ with unknown μ, learn μ to ℓ_2 norm ε.*

Sample complexity: $\Theta\big(d/\varepsilon^2\big)$.

Problem 4 (TV Testing). *Given $\varepsilon \in (0,1]$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$ with unknown μ, Σ, distinguish between $\mathbf{p} = \mathbf{q}$ and $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$.*

Sample complexity: $\Theta\big(d/\varepsilon^2\big)$.

Problem 5 (Mean Testing). *Given $\varepsilon \in (0,1]$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$ with unknown μ, Σ, distinguish between $\mathbf{p} = \mathbf{q}$ and $\|\mu\|_2 > \varepsilon$.*

Sample complexity: $\Theta\Big(\sqrt{d}/\varepsilon^2\Big)$.

Problem 6 (Covariance Norm Estimation, Operator Norm). *Given $\varepsilon \in (0,1], \kappa > 0$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$ with unknown μ, Σ such that $\|\Sigma - I_d\|_{\mathrm{op}} \le \kappa$, learn $\|\Sigma - I_d\|_{\mathrm{op}}$ to an additive ε.*

Sample complexity: $\Theta\big(\kappa^2 d/\varepsilon^2\big)$.

Problem 7 (Covariance Estimation, Operator Norm). *Given $\varepsilon \in (0,1], \kappa > 0$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$ with unknown μ, Σ such that $\|\Sigma - I_d\|_{\mathrm{op}} \le \kappa$, learn Σ to $\|\cdot\|_{\mathrm{op}}$ norm ε.*

Sample complexity: $\Theta\big(\kappa^2 d/\varepsilon^2\big)$.

Problem 8 (Covariance Norm Estimation, Frobenius Norm). *Given $\varepsilon \in (0,1], \kappa > 0$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$ with unknown μ, Σ such that $\|\Sigma - I_d\|_F \le \kappa$, learn $\|\Sigma - I_d\|_F$ to an additive ε.*

Sample complexity: $\Theta\big(\kappa^2 d/\varepsilon^2\big)$.

Problem 9 (Covariance Estimation, Frobenius Norm). *Given $\varepsilon \in (0,1], \kappa > 0$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$ with unknown μ, Σ such that $\|\Sigma - I_d\|_F \le \kappa$, learn Σ to $\|\cdot\|_F$ norm ε.*

Sample complexity: $\Theta\big(\kappa^2 d^2/\varepsilon^2\big)$.

Finally, this one will be rather useful to prove lower bounds, as we will see later, but also makes sense by itself – who has never wanted to test whether the vector of eigenvalues of a covariance matrix had large ℓ_2 norm?

Problem 10 (Covariance Norm Testing, Frobenius Norm). *Given $\varepsilon \in (0,1]$ and i.i.d. samples from some $\mathbf{p} \coloneqq \mathcal{N}(\mathbf{0}, \Sigma)$ with unknown Σ, distinguish between $\mathbf{p} = \mathbf{q}$ and $\|\Sigma - I_d\|_F > \varepsilon$.*

Sample complexity: $\Theta\big(d/\varepsilon^2\big)$.

1 A crucial fact

The starting point of many of those results will be that the total variation distance between two high-dimensional Gaussians – and thus, a fortiori, the distance to the standard one, $\mathbf{q}$ – is characterized by the appropriate distances between their parameters:

Fact 11. *Let* $\mathbf{p} \coloneqq \mathcal{N}(\mu, \Sigma)$. *Then* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq 2\max(\|\mu\|_2, \|\Sigma - I_d\|_F)$. *Conversely, we have* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \geq C\min(1, \max(\|\mu\|_2, \|\Sigma - I_d\|_F))$, *for some absolute constant* $C > 0$.

Proof. There are many references for this, with various degrees of complexity and tightness of the constants; see, e.g., [Li18, Corollary 1.4.6], or [DMR20, Theorem 1.3]. We here give a self-contained proof of the upper bound.[1] We start by the (exact) expression of the Kullback–Leibler divergence between $\mathbf{p}$ and $\mathbf{q}$. Denoting the d eigenvalues of Σ by $0 \leq \lambda_1 \leq \cdots \leq \lambda_d$, we have, using the "folklore" expression for the divergence between two arbitrary multivariate Gaussians,[2]

$$\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = \frac{1}{2}\left(\|\mu\|_2^2 + \operatorname{Tr}\Sigma - d - \log\det\Sigma\right) = \frac{1}{2}\left(\|\mu\|_2^2 + \sum_{i=1}^{d}(\lambda_i - 1 - \log\lambda_i)\right)$$

Assume for now (we will argue later why it is safe to do so) that $|\lambda_i - 1| \leq 1/2$ for all i. In that case, since the very convenient inequality[3] $x - 1 \leq \log x + (x-1)^2$ holds for all $x \geq 1/2$, and we get

$$\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \leq \frac{1}{2}\left(\|\mu\|_2^2 + \sum_{i=1}^{d}(\lambda_i - 1)^2\right) = \frac{1}{2}\left(\|\mu\|_2^2 + \|\Sigma - I_d\|_F^2\right) \leq \max\left(\|\mu\|_2^2, \|\Sigma - I_d\|_F^2\right),$$

where we used that, for positive semi-definite matrices the Frobenius norm is the ℓ_2 norm of the vector of eigenvalues. Now, by Pinsker's inequality, we get

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})} \leq \frac{1}{\sqrt{2}}\max(\|\mu\|_2, \|\Sigma - I_d\|_F)$$

which gives the claim. Almost: it remains to explain why we could assume that $|\lambda_i - 1| \leq 1/2$ for all i. This is just because, otherwise, we have $\|\Sigma - I_d\|_F \geq \max_{1 \leq i \leq d}|\lambda_i - 1| > 1/2$, and since the total variation distance is always at most one we have $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) < 2\|\Sigma - I_d\|_F$, and the claim holds as well. □

Why do we care about this? This will help us find relations between the problems considered, of the type "if I have an algorithm for problem A, then I can use it to solve problem B with the same sample complexity" – in turn allowing us to only prove a few bounds and get the whole picture.

2 Relationship between problems

For instance, suppose that we are under the identity-covariance assumption, i.e., promised that $\Sigma = I_d$ for the unknown Gaussian $\mathbf{p}$. Then $\mathbf{p} = \mathbf{q}$ is equivalent to $\mu = \mathbf{0}$, and by Fact 11 we now know that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$ implies $\|\mu\|_2 > \varepsilon/2$. So we have the following (where $A \preceq B$ means "A requires at most as many samples as B"):

Problem 1 $\preceq$ Problem 2 $\preceq$ Problem 3

We also have the other following relations:

Problem 1 $\preceq$ Problem 4

[1] But would be delighted to be pointed to a simple proof of the general statement, as we do not particularly care about obtaining the tightest constants possible.

[2] Folklore, in that case, is synonymous with Wikipedia [Wik20].

[3] Which is inspired by looking at the Taylor expansion of log around 1: $\log x = (x-1) - \frac{1}{2}(x-1)^2 + o((x-1)^2)$.

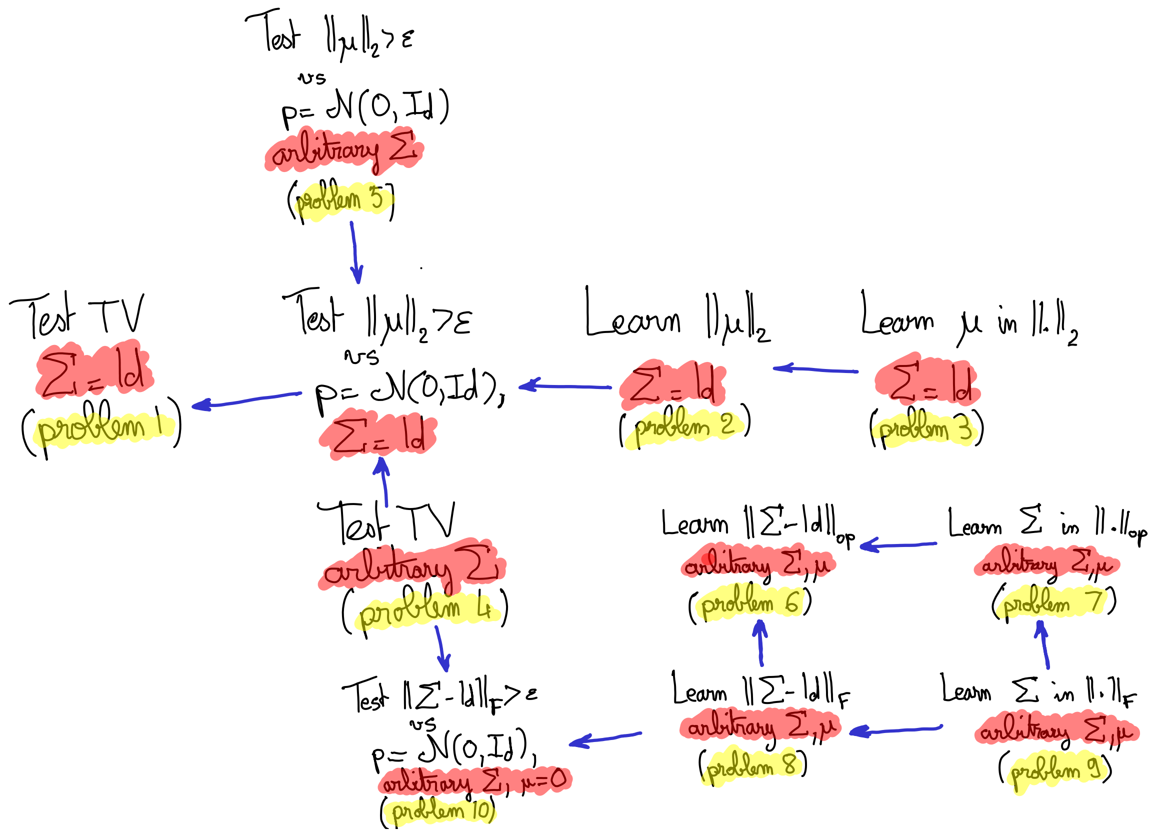

Figure 1: Relationships between the different problems considered here (and one extra, in the middle)

(as an algorithm for the more general problem Problem 4 can be used for the more specific Problem 1); as well as

$$\text{Problem 1} \preceq \text{Problem 5}$$

(because again of Fact 11; can you see why?). Finally, we also have

$$\text{Problem 6} \preceq \text{Problem 7}, \qquad \text{Problem 10} \preceq \text{Problem 8} \preceq \text{Problem 9}$$

(again, check you see why – the idea is that learning the parameter implies learning its norm which implies testing its magnitude), and

$$\text{Problem 6} \preceq \text{Problem 8}, \qquad \text{Problem 7} \preceq \text{Problem 9}$$

this time because $\|\cdot\|_{\text{op}} \leq \|\cdot\|_F$. See Figure 1 for a colorful illustration.

Remark 12. There is also some relation between Problem 4 and the combination (Problem 5+ Problem 10), but it is less obvious and a bit more cumbersome. The idea is that (barring quite a few details) if one can solve both Problem 5 and Problem 10, one can solve Problem 4 as follows:

1. use the algorithm for Problem 5 to detect if $\|\mu\|_2 \gtrsim \varepsilon$ (and "reject" if it is the case)
2. use the algorithm for Problem 10 to detect if $\|\Sigma - I_d\|_F \gtrsim \varepsilon$ (and "reject" if it is the case), *but* on new samples of the form $X' := \frac{X-Y}{\sqrt{2}}$, where $X, Y \sim \mathbf{p}$. This transformation increases the number of samples needed by a factor 2, but as a result the distribution of those new samples is $\mathcal{N}(\mathbf{0}, \Sigma)$ (the mean μ cancels out, the covariance is preserved).

3. if both tests pass, declare $\mathbf{p} = \mathbf{q}$.

Overall, this allows us to distinguish between $\mathbf{p} = \mathbf{q}$ and $\max(\|\mu\|_2, \|\Sigma - I_d\|_F) \ll \varepsilon$, which by Fact 11 is enough to solve Problem 4.

3 The proofs, and where to find them

Given all the relations between problems outlined in the previous section, we don't have as many upper bound lower bounds to prove as one would fear: at least, much fewer that 10 upper and 10 lower bounds.

- The $\Omega\left(\sqrt{d}/\varepsilon^2\right)$ lower bound for Problem 1 is considered folklore, and can be shown, e.g., by Le Cam's two-point method, considering the 2^d distributions

$$\mathbf{p}_z \coloneqq \mathcal{N}\left(\frac{2\varepsilon}{\sqrt{d}} z, I_d\right), z \in \{-1,1\}^d .$$

  Each such $\mathbf{p}_z$ has mean with ℓ_2 norm exactly 2ε, so an algorithm for Problem 1 should allow us to distinguish them from the standard Gaussian $\mathbf{q}$. Even more, if n samples are enough for the task, it should allow us to distinguish between the mixture $\frac{1}{2^d}\sum_{z\in\{-1,1\}^d} \mathbf{p}_z^{\otimes n}$ (pick one $\mathbf{p}_z$ uniformly at random, draw n samples from it) and $\mathcal{N}(\mathbf{0}, I_d)^{\otimes n}$ (n samples from the standard Gaussian). This in turn can be shown to require $n = \Omega\left(\sqrt{d}/\varepsilon^2\right)$ samples (see, for instance, [Wu19, Chapter 23]).
- The $O\left(\sqrt{d}/\varepsilon^2\right)$ upper bound for Problem 2 is also "folklore,"[4] but is achieved by the empirical estimator (considering the squared ℓ_2 norm of the empirical estimator, that is, $\|\frac{1}{n}\sum_{i=1}^n X^{(j)}\|_2^2$, and doing an expectation+variance+Chebyshev analysis). The analysis is not horrendous, although my preference is to divide the n samples in two sets $(X^{(j)})_{j=1}^{n/2}$ and $(Y^{(j)})_{j=1}^{n/2}$ and to use the estimator $\frac{2}{n}\sum_{i=1}^{n/2}\langle X^{(j)}, Y^{(j)}\rangle$. I find the computations cleaner.
- The $O\left(\sqrt{d}/\varepsilon^2\right)$ upper bound for Problem 5 is, oddly, very recent, and can be found in [CCK$^+$19, Section 4]. I find that proof rather suprising and cute.
- The $O\left(d/\varepsilon^2\right)$ upper and lower bound for Problem 3 (a question also known as Gaussian Location Model, or GLM) have many proofs, but I strongly recommend [Wu19, Section 9.1], which provides a more general statement and establishes it in an incredibly elegant way.
- The $\Omega\left(d/\varepsilon^2\right)$ lower bounds for Problem 6 and Problem 10 both follow from the difficult to distinguish between an identity-covariance matrix and one perturbed by a scaled rank-one matrix, i.e., of the form $I_d + \eta vv^\top$. Note that this is a testing (distinguishing) problem for the operator norm, so that implies the same lower bound for Problem 6 (estimating that norm) and Problem 10 (testing in Frobenius, but Frobenius upper bounds operator norm). This lower bound, proven via Le Cam's two-point method combined with Ingster's method,[5] can be found, e.g., in [Wu19, Section 24.2].[6]
- The $O\left(d/\varepsilon^2\right)$ upper bound for Problem 7 can also be found in [Wu19, Section 24.2], and is achieved by the "obvious" estimator: the empirical covariance matrix $\widehat{\Sigma} \coloneqq \frac{1}{n}\sum_{j=1}^n X^{(j)}X^{(j)^\top}$. A more detailed reference (and a very good one!) for that upper bound is [Ver18, Chapter 4.7].
- The $O\left(d/\varepsilon^2\right)$ and $O\left(\kappa^2 d/\varepsilon^2\right)$ upper bounds for Problem 10 and Problem 8 can be found or follow from [CM13]; they are achieved by a unbiased statistic for $\|\Sigma - I_d\|_F^2 = \operatorname{Tr}\left((\Sigma - I_d)^2\right)$.
- All that remains are the $O\left(d^2/\varepsilon^2\right)$ upper and lower bounds for Problem 9. I am not sure which reference is best, but [DKK$^+$19, Section 4.2.2], or [Li18, Corollary 2.1.12] both show that the upper bound is

[4] Meaning it is awfully hard to track down a published reference for it, as nobody appears to know any but will swear there must be dozens.

[5] A fancy way to state we upper bound χ^2 distances in a clever (and very useful to know) fashion.

[6] Really, Yihong Wu's lecture notes are a treasure trove.

achieving (again) by the empirical covariance $\widehat{\Sigma} \coloneqq \frac{1}{n}\sum_{j=1}^{n} X^{(j)} X^{(j)\top}$. I am still tracking down a self-contained reference for the lower bound.

References

[CCK+19] Clément L. Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. Random restrictions of high-dimensional distributions and uniformity testing with subcube conditioning. *CoRR*, abs/1911.07357, 2019. To appear in SODA 2021.

[CM13] T. Tony Cai and Zongming Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019.

[DMR20] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians, 2020.

[Li18] Jerry Zheng Li. *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology, Cambridge, USA, 2018.

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[Wik20] Wikipedia contributors. Kullback–leibler divergence — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Kullback%E2%80%93Leibler_divergence&oldid=983534042, 2020. [Online; accessed 15-October-2020].

[Wu19] Yihong Wu. Lecture notes for ECE598YW: Information-theoretic methods for high-dimensional statistics, 2019.