The goal of this short note is to record the following lower bounds on ℓ_∞ mean estimation of d-dimensional spherical Gaussians and product binary distributions:

Theorem 1. *For $d \geq 1$ and $\varepsilon \in (0,1]$, estimating the mean of (i) an identity-covariance d-dimensional Gaussian and (ii) a product distribution over $\{-1,1\}^d$ to ℓ_∞ distance ε both has sample complexity $\Theta\Big(\frac{\log d}{\varepsilon^2}\Big)$.*

For convenience, we denote by $\mathrm{Rad}(\lambda)$ the distribution over $\{-1,1\}$ with expectation $\lambda \in [-1,1]$, and let $\mathcal{G}_d \stackrel{\text{def}}{=} \{\, \mathcal{N}(\mu, I_d) \,:\, \mu \in \mathbb{R}^d \,\}$ and $\mathcal{B}_d \stackrel{\text{def}}{=} \{\, \otimes_{i=1}^d \mathrm{Rad}(\mu_i) \,:\, \mu \in [-1,1]^d \,\}$ be the families of identity-covariance Gaussians and product binary distributions considered, respectively.

Upper bound. The upper bound, for both $\mathcal{G}_d$ and $\mathcal{B}_d$, follow from the following simple scheme: given $n = O\big((\log d)/\varepsilon^2\big)$ i.i.d. samples from an unknown $\mathbf{p} \in \mathcal{G}_d$ (resp., $\mathbf{p} \in \mathcal{B}_d$), we can estimate independently the mean μ_i of each marginal to an additive ε, with probability at least $1 - \frac{1}{3d}$, by using the corresponding coordinate of the n samples. By a union bound over all d coordinates, the resulting $\hat{\mu} \in \mathbb{R}^d$ satisfies $\|\hat{\mu} - \mu\|_\infty \leq \varepsilon$ with probability at least $1/3$.

Lower bound. The lower bound is the more interesting part, and shows that this very naive approach (independently deal with each coordinate, then apply a union bound) is essentially the best one can do.

The following argument, which applies to both $\mathcal{G}_d$ and $\mathcal{B}_d$, was communicated to me by Jayadev Acharya. It shows the even stronger statement about *testing* the mean of a Gaussian or product binary distribution, even under the promise that this mean is 1-*sparse*.

Lemma 2 (Gaussian Hide-and-Seek)**.** *For $d \geq 1$ and $\varepsilon \in (0,1]$, distinguishing between $\mathcal{N}(0, I_d)$ and $\mathcal{N}(\mu, I_d)$* where μ is promised to satisfy $\mu = \varepsilon e_i$ for some $i \in [d]$ *requires $\Omega\Big(\frac{\log d}{\varepsilon^2}\Big)$ samples.*

Proof. Fix any number of samples n. Denote by $\mathbf{p}^{(n)}$ the n-fold product of the standard Gaussian $\mathcal{N}(0, I_d)^{\otimes n} = \mathcal{N}(0, I_{nd})$, and by $\mathbf{q}^{(n)}$ the uniform mixture $\mathbf{q}^{(n)} = \frac{1}{d}\sum_{i=1}^d \mathcal{N}(\varepsilon e_i, I_d)^{\otimes n}$. By a standard Le Cam-type argument, any n-sample test for the original problem can be used to distinguish $\mathbf{p}^{(n)}$ and $\mathbf{q}^{(n)}$, which is only possible if $\mathrm{d_{TV}}\big(\mathbf{p}^{(n)}, \mathbf{q}^{(n)}\big) \gtrsim 1$. Since

$$\mathrm{d_{TV}}\Big(\mathbf{p}^{(n)}, \mathbf{q}^{(n)}\Big)^2 \leq \frac{1}{4}\chi^2\Big(\mathbf{q}^{(n)} \;||\; \mathbf{p}^{(n)}\Big)$$

it suffices to bound $\chi^2\big(\mathbf{q}^{(n)} \;||\; \mathbf{p}^{(n)}\big)$. Which is convenient, as we can invoke Lemma 4 to compute this explicitly. Indeed, for any $j \in [n]$ and any $i \in [d]$, the corresponding δ_j^i is independent of j (as all marginals are the same) and equal to

$$\delta_j^i(x) = \frac{\mathcal{N}(\varepsilon e_i, I_d)\,(x) - \mathcal{N}(0, I_d)\,(x)}{\mathcal{N}(0, I_d)\,(x)} = e^{-\frac{\varepsilon^2}{2}} e^{\varepsilon x_i} - 1, \quad x \in \mathbb{R}^d,$$

so that for any two i_1, i_2, the $H_j(i_1, i_2)$ of Lemma 4 (again independent of j) is equal to

$$H_j(i_1, i_2) = \mathbb{E}_{X \sim \mathcal{N}(0, I_d)}[(e^{-\frac{\varepsilon^2}{2}} e^{\varepsilon X_{i_1}} - 1)(e^{-\frac{\varepsilon^2}{2}} e^{\varepsilon X_{i_2}} - 1)] = (e^{\varepsilon^2} - 1)\mathbb{1}_{\{i_1 = i_2\}}\,.$$

This is great as now we get from Lemma 4 that

$$\chi^2\Big(\mathbf{q}^{(n)} \;||\; \mathbf{p}^{(n)}\Big) = \mathbb{E}_{i_1, i_2}[(1 + (e^{\varepsilon^2} - 1)\mathbb{1}_{\{i_1 = i_2\}})^n] - 1 = \frac{d-1}{d} + \frac{e^{n\varepsilon^2}}{d} - 1 = \frac{e^{n\varepsilon^2} - 1}{d}$$

and for this to be $\Omega(1)$ we need $n = \Omega\Big(\frac{\log d}{\varepsilon^2}\Big)$. ∎

This was for Gaussians though. What about product binary distributions? As it turns out, the same argument goes through, nearly unchanged.

Lemma 3 (Bernoullli Hide-and-Seek). *For $d \geq 1$ and $\varepsilon \in (0,1]$, distinguishing between the uniform distribution* $\operatorname{Rad}(0)^{\otimes d}$ *and* $\otimes_{i=1}^d \operatorname{Rad}(\mu_i)$ where μ is promised to satisfy $\mu = \varepsilon e_i$ for some $i \in [d]$ *requires* $\Omega\left(\frac{\log d}{\varepsilon^2}\right)$ *samples.*

Proof. Fix any number of samples n. Denote by $\mathbf{p}^{(n)}$ the n-fold product of the uniform distribution, $(\operatorname{Rad}(0)^{\otimes d})^{\otimes n} = \operatorname{Rad}(0)^{\otimes nd}$, and by $\mathbf{q}^{(n)}$ the uniform mixture $\mathbf{q}^{(n)} = \frac{1}{d}\sum_{i=1}^d \frac{1}{2^{d-1}} \operatorname{Rad}(\varepsilon)$.[1] As in the proof of Lemma 2, it suffices to bound $\chi^2\left(\mathbf{q}^{(n)} \mid\mid \mathbf{p}^{(n)}\right)$, and to do so we will compute it explicitly using Lemma 4. Indeed, for any $j \in [n]$ and any $i \in [d]$, the corresponding δ_j^i is independent of j and can be seen to be equal to

$$\delta_j^i(x) = \frac{\frac{1}{2^{d-1}} \operatorname{Rad}(\varepsilon)(x) - \frac{1}{2^d}}{\frac{1}{2^d}} = \varepsilon x_i, \quad x \in \mathbb{R}^d,$$

so that for any two i_1, i_2, $H_j(i_1, i_2) = \varepsilon^2 \mathbb{E}_X[X_{i_1} X_{i_2}] = \varepsilon^2 \mathbb{1}_{\{i_1 = i_2\}}$ (where the expectation is over X u.a.r. from $\{-1,1\}^d$). From Lemma 4, it follows that

$$\chi^2\left(\mathbf{q}^{(n)} \mid\mid \mathbf{p}^{(n)}\right) = \mathbb{E}_{i_1,i_2}[(1+\varepsilon^2 \mathbb{1}_{\{i_1=i_2\}})^n] - 1 = \frac{(1+\varepsilon^2)^n - 1}{d}$$

and for this to be $\Omega(1)$ we again need $n = \Omega\left(\frac{\log d}{\varepsilon^2}\right)$. $\square$

We finally state the (relatively standard lemma) which allowed us to easily handle the chi square distance between a mixture and a product distribution.

Lemma 4 (See, e.g., [ACT18, Lemma III.5]). *Consider a random variable θ such that for each $\theta = \vartheta$ the distribution Q_ϑ^n is defined as $Q_{1,\vartheta} \times \cdots \times Q_{n,\vartheta}$. Further, let $P^n = P_1 \times \cdots \times P_n$ be a fixed product distribution. Then,*

$$\chi^2(\mathbb{E}_\theta[Q_\theta^n], P^n) = \mathbb{E}_{\theta\theta'}\left[\prod_{j=1}^n (1 + H_j(\theta, \theta'))\right] - 1,$$

where θ' is an independent copy of θ, and with $\delta_j^\vartheta(X_j) = (Q_{j,\vartheta}(X_j) - P_j(X_j))/P_j(X_j)$,

$$H_j(\vartheta, \vartheta') \stackrel{\text{def}}{=} \left\langle \delta_j^\vartheta, \delta_j^{\vartheta'} \right\rangle = \mathbb{E}\left[\delta_j^\vartheta(X_j) \delta_j^{\vartheta'}(X_j)\right],$$

where the expectation is over X_j distributed according to P_j.

References

[ACT18] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: lower bounds from chi-square contraction. *CoRR*, abs/1812.11476, 2018.

[1] Note that for $\mu = \varepsilon e_i$, we have $\otimes_{i=1}^d \operatorname{Rad}(\mu_i) = \frac{1}{2^{d-1}} \operatorname{Rad}(\varepsilon)$ as all coordinates are uniform except one.