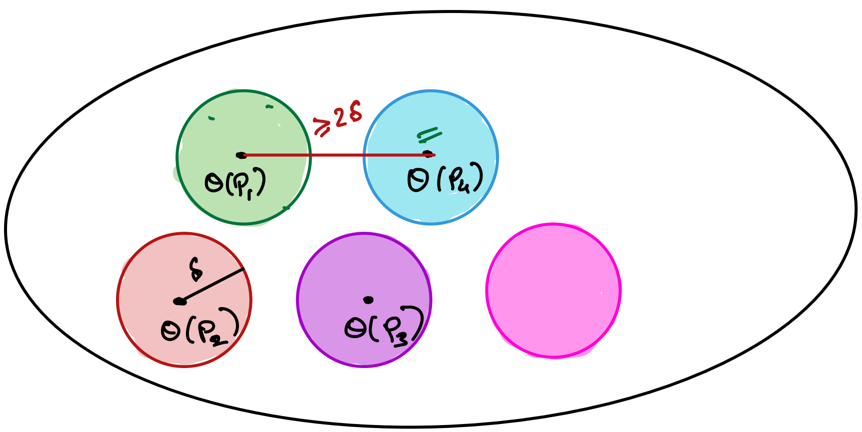
Recap

(Bayesian) Multiple Hypothesis Testing: $\underset{\sim}{V} \to \bar{X} \sim P_V^n \to T(\bar{X})$

↳ uniform over $\mathcal{V}$

$$\underset{\substack{V \in \mathcal{V} \\ \bar{X} \sim P_V^n}}{\mathbb{P}}\left[T(\bar{X}) \neq V\right] \geq 1 - \frac{n \cdot \underset{V_1, V_2 \in \mathcal{V}}{\mathbb{E}}\left[D(P_{V_1} \| P_{V_2})\right] + 1}{\log|\mathcal{V}|}$$

Minimax risk bounds: $\mathcal{M}_n(\Pi, \ell) = \inf_{\hat{\theta}} \sup_{P \in \Pi} \underset{\bar{X} \sim P^n}{\mathbb{E}} \ell(\theta(P), \hat{\theta}(\bar{X}))$



$$\ell(\theta, \hat{\theta}) = \underbrace{\Phi}_{\text{increasing}}(\underbrace{\rho(\theta, \hat{\theta})}_{\text{distance}})$$

$$\mathcal{M}_n(\Pi, \ell) \geq \Phi(\delta) \cdot \inf_T \left\{ \underset{\substack{V \in \mathcal{V} \\ \bar{X} \sim P_V^n}}{\mathbb{P}}\left[T(\bar{X}) \neq V\right] \right.$$

High-dimensional mean estimation

$$\Pi = \{ N(\underbrace{\mu}_{\theta}, I_d) \mid \mu \in \mathbb{R}^d \} \qquad \ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$$

Empirical estimator: $\hat{\theta}(\bar{x}) = \hat{\theta}(x_1 \ldots x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$

$$\text{error} = \mathbb{E}_{\bar{x} \sim (N(\mu, I_d))^n} \left[\| \hat{\theta}(\bar{x}) - \mu \|^2 \right]$$

$$= \mathbb{E}_{\bar{x} \sim (N(\mu, I_d))^n} \left[\left\| \frac{1}{n} \sum (x_i - \mu) \right\|^2 \right]$$

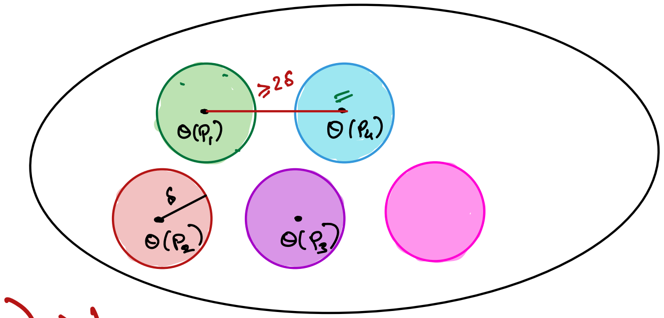
$$= \mathbb{E}_{\bar{x} \sim N(\mu, I_d)} \left[\frac{1}{n^2} \cdot \sum_i \underbrace{\langle x_i - \mu, x_i - \mu \rangle}_{\mathbb{E}[\cdot] = d} + \frac{1}{n^2} \sum_{i \neq j} \underbrace{\langle x_i - \mu, x_j - \mu \rangle}_{\mathbb{E}[\cdot] = 0} \right]$$

$$= \frac{d}{n}$$

Minimax lower bound

$\forall \delta$

$$M_n(\Pi, \mathcal{L}) \geq \Phi(\delta) \cdot \inf_{T} \Big\{ \underbrace{P_{\substack{v \in \mathcal{V} \\ X \sim P_v^n}} \big[T(X) \neq v \big]}_{\geq 1 - \frac{n \cdot \mathbb{E}_{v_1, v_2} D(P_{v_1} \| P_{v_2}) + 1}{\log |\mathcal{V}|}} \Big\}$$



Ex: $D(N(\mu_1, I_d) \,\|\, N(\mu_2, I_d)) = 2 \ln 2 \cdot \|\mu_1 - \mu_2\|^2$

$\forall \delta$

▶ (Packing Lemma): There exists collection $\{\mu_v\}_{v \in \mathcal{V}} \subseteq \mathbb{R}^d$ s.t.

- $|\mathcal{V}| \geq 2^d$
- $\forall v_1 \neq v_2 \quad \underbrace{2\delta}_{c_1 \cdot \delta} \leq \|\mu_{v_1} - \mu_{v_2}\| \leq \underbrace{8\delta}_{c_2 \cdot \delta}$.

Lower bound assuming lemma

$$M_n(\pi, \ell) \geq \Phi(\delta) \cdot \inf_{T} \left\{ \mathbb{P}_{\substack{v \in \mathcal{V} \\ X \sim P_v^n}} [T(X) \neq v] \right\} \geq \Phi(\delta) \cdot \left(1 - \frac{n \cdot \mathbb{E}_{v_1, v_2} D(P_{v_1} \| P_{v_2}) + 1}{\log |\mathcal{V}|} \right)$$

$$M_n(\pi, \ell) \geq \delta^2 \cdot \left(1 - \frac{n \cdot 2 \ln 2 \cdot (8\delta)^2 + 1}{d} \right)$$

$$= \delta^2 \cdot \left(1 - \frac{n}{d} \cdot c\,\delta^2 - \frac{1}{d} \right) \qquad \delta = \sqrt{\frac{c' \cdot d}{n}}$$

$$= c' \cdot \frac{d}{n} \underbrace{\left(1 - c \cdot c' - \frac{1}{d} \right)}_{\geq 1/2}$$

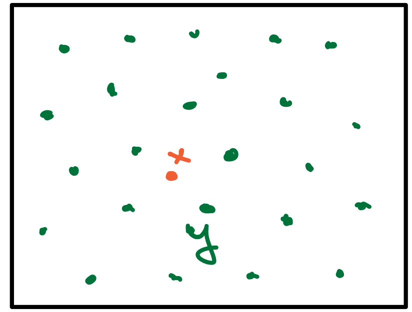
$$\geq \frac{c' d}{2n}$$

[Wainwright]: Metric Entropy and its uses

Covering and Packing

Set S of points. Distance metric ρ

- $C \subseteq S$ is a α-covering of S w.r.t. ρ if

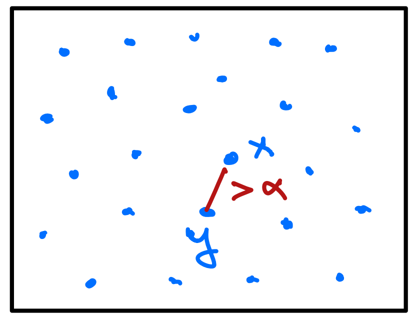


$\forall x \in S \quad \exists y \in C$ s.t. $\rho(x, y) \leq \alpha$

Minimum: $N(\alpha, S, \rho)$

$\log N(\alpha, S, \rho)$: metric entropy

- $P \subseteq S$ is a α-packing w.r.t. ρ if



$\forall x, y \in P \quad x \neq y \quad \rho(x, y) > \alpha$

Maximum: $M(\alpha, S, \rho)$

Ex: $\quad M(2\alpha, S, \rho) \leq N(\alpha, S, \rho) \leq M(\alpha, S, \rho)$

Packing and covering Euclidean balls

$$B_d(x,\alpha) = \{ y \in \mathbb{R}^d \mid \|x-y\| \leq \alpha \}$$

$$\mathrm{Vol}(B_d(0,1)) \leq \sum_{x} \mathrm{Vol}(B_d(s,\alpha))$$

$$= N(\alpha,S,\rho)\cdot \mathrm{Vol}(B_d(\cdot,\alpha))$$

$$S = B_d(0,1) \subseteq \bigcup_{\underbrace{x \in C}_{\text{min } \alpha\text{-covering : } N(\alpha,S,\rho)}} B_d(x,\alpha)$$

$$M(\alpha,S,\rho) \geq N(\alpha,S,\rho) \geq \frac{\mathrm{Vol}(B_d(0,1))}{\mathrm{Vol}(B_d(\cdot,\alpha))}$$

$$= \frac{c_d \cdot 1^d}{c_d \cdot \alpha^d}$$

$$= \left(\frac{1}{\alpha}\right)^d$$

$$\therefore M(\alpha,S,\rho) \geq N(\alpha,S,\rho) \geq 2^d \quad \text{for } \alpha = 1/2 .$$

Proof 2: Random Packings

- Pick i.i.d $u_1, \ldots\ldots, u_N \in \frac{1}{\sqrt{d}}\{-1, 1\}^d$

  $\forall i \ \|u_i\| = 1$

  $\forall i, j \ \|u_i - u_j\| \leq 2$

- $\mathbb{P}[\|u_i - u_j\| \leq \frac{1}{2}] = \mathbb{P}[\|u_i - u_j\|^2 \leq 1/4]$

  $= \mathbb{P}\left[\langle u_i, u_j\rangle \geq \frac{7}{8}\right]$

  $\leq 2^{-d/16}$

- $\mathbb{P}[\exists i \neq j \ \|u_i - u_j\| \leq \frac{1}{2}] \leq N^2 \cdot 2^{-d/16}$

Sparse mean estimation

$$\Pi = \{ N(\mu, I_d) \mid \mu \in \mathbb{R}^d, \underbrace{\|\mu\|_0}_{\text{\# of non-zero coordinates}} = 1 \}$$

$$\bar{x} = (x_1 \ldots\ldots x_n) \sim N(\underbrace{\mu}_{\text{unknown}}, I_d)$$

$$m = \frac{1}{n} \sum x_i$$

$$M_n = \inf_{\hat{\mu}} \sup_{P \in \Pi} \mathbb{E}_{\bar{x} \sim P^n} \left[\|\mu(P) - \hat{\mu}(\bar{x})\|^2 \right]$$

$$j = \arg\max |m_j|$$

$$\hat{\mu}(\bar{x}) = (0 \ldots\ldots m_j \ldots\ldots 0)$$

Analyzing empirical mean

$$\bar{X} = (x_1, \dots, x_n) \sim (N(\mu, I_d))^n, \qquad \eta = \frac{1}{n}\sum_i x_i$$

- $\forall j \in [d]$ $\quad \eta_j - \mu_j \sim N\left(0, \frac{1}{n}\right)$

Proof:

$$\eta_j - \mu_j = \sum_i \underbrace{\frac{1}{n}(x_{ij} - \mu_j)}_{\sim N(0,\, 1/n^2)}$$

$$\sim N(0, 1/n)$$

- $\mathbb{P}[\exists j \ |\eta_j - \mu_j| \geq t] \leq 2d \cdot e^{-nt^2/2}$

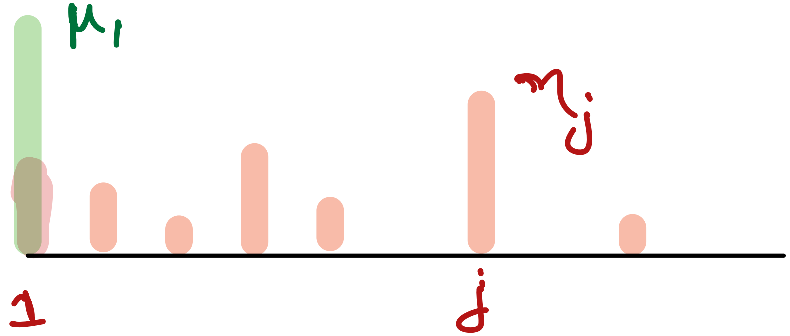
Analyzing estimator

$j = \operatorname{argmax} |\eta_j| \qquad \hat{\mu} = (0 \ldots 0 . \eta_j . 0 \ldots 0)$

- $\mathbb{P}_{\bar{x} \sim (N(\mu, I_d))^n} \left[\underline{\|\mu - \hat{\mu}\|_2} \geq t \right] \leq \mathbb{P}\left[\exists j \ |\mu_j - \eta_j| \geq t/3 \right] \leq 2d \cdot e^{-nt^2/18}$

Proof:

(Say $\mu_1 \neq 0$)



Case 1: $\hat{\mu}_1 \neq 0$

$|\mu_1 - \eta_1| \geq t$

Case 2: $\hat{\mu}_1 = 0$

$|\eta_j| \geq |\eta_1|$

$\geq |\mu_1| - |\eta_1 - \mu_1|$

$|\mu_1| + |\eta_j| \geq t$

$\cancel{|\mu_1|} + 2|\eta_j| \geq t + \cancel{|\mu_1|} - |\eta_1 - \mu_1|$

$|\eta_1 - \mu_1| + 2|\eta_j - \mu_j| \geq t$

Bounding the error

$$\mathbb{P}\left[\|\mu-\hat{\mu}\|^2 \geq s\right] \leq 2d\, e^{-ns/18} \qquad s = t^2$$

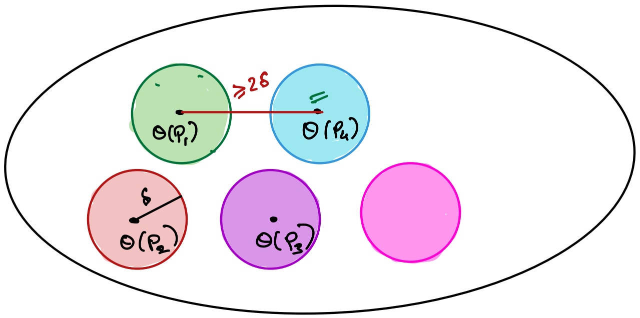
$$\mathbb{E}\left[\|\mu-\hat{\mu}\|^2\right] = \int_0^\infty \mathbb{P}\left[\|\mu-\hat{\mu}\|^2 \geq s\right] ds$$

$$= \int_0^{r} \underbrace{\mathbb{P}\left[\|\mu-\hat{\mu}\|^2 \geq s\right]}_{\leq 1} ds + \int_r^\infty \underbrace{\mathbb{P}[\ldots]}_{2d\cdot e^{-ns/18}} ds$$

$$\leq r - \frac{2d\cdot 18}{n} e^{-ns/18}\Big]_r^\infty \qquad r \simeq c\cdot\frac{\log d}{n}$$

$$= r + \frac{2d\cdot 18}{n} e^{-nr/18}$$

Matching lower bound



Find $\{\mu_v\}_{v \in \mathcal{V}}$. $\|\mu\|_0 = 1$

$$2\delta \geq \|\mu_{v_1} - \mu_{v_2}\| \geq 2\delta$$

$$\sqrt{2}\delta e_1 \quad \cdots \quad \sqrt{2}\delta . e_d$$

$$\mathfrak{M}_n(\pi, \ell) \geq \delta^2 \cdot \left(1 - \frac{n \cdot \mathbb{E}_{v_1, v_2} D(P_{v_1} \| P_{v_2}) + 1}{\log|\mathcal{V}|}\right)$$

$$\geq c \cdot \frac{\log d}{n}$$