Lec 2 : KL Divergence

Yaojun Han

Defn. (KL Divergence)

For two probability distributions P, Q over the same space, the Kullback-Leibler divergence (or the relative entropy) of P w.r.t. Q is

$$D_{KL}(P \| Q) = \begin{cases} \mathbb{E}_{X\sim P}\left[\log \frac{dP}{dQ}(X)\right] & \text{if } P \ll Q \\ +\infty & \text{o.w.} \end{cases}$$

Remark: 1. The above defn. covers both discrete and continuous cases, i.e.

$$D_{KL}(P \| Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad \text{if } p, q \text{ are pmfs}$$

and

$$D_{KL}(P \| Q) = \int p(x) \log \frac{p(x)}{q(x)} \, d\mu(x) \quad \text{if } p, q \text{ are pdfs w.r.t. } \mu.$$

2. This is a divergence rather than a distance, i.e. $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$.
For this reason, we write $D_{KL}(P \| Q)$ instead of $D_{KL}(P, Q)$.

3. IT origin: $D_{KL}(P \| Q)$ is the "redundancy" of using Q for source coding while the true distribution is P:

$$D_{KL}(P \| Q) = \underbrace{\sum_x p(x) \log \frac{1}{q(x)}}_{\substack{\text{expected codelength} \\ \text{of using } Q}} - \underbrace{H(P)}_{\substack{\text{optimal expected} \\ \text{codelength} \\ \text{for source } P}}$$

Basic properties.

Property I: $D_{KL}(P \| Q) \geq 0$, with equality iff $P = Q$.

Pf. $D_{KL}(P \| Q) = \mathbb{E}_P\left[\log \frac{dP}{dQ}\right] = \mathbb{E}_P\left[-\log \frac{dQ}{dP}\right] \geq -\log \mathbb{E}_P\left[\frac{dQ}{dP}\right] = 0.$ ☒

Note: this gives the usual proof of

$$I(X;Y) = \mathbb{E}_{P_{XY}}\left[\log \frac{P_{XY}(X,Y)}{P_X(X) P_Y(Y)}\right] = D_{KL}(P_{XY} \| P_X P_Y) \geq 0.$$

Also, equality holds iff $P_{XY} = P_X P_Y$, i.e. X and Y are independent.

Property I: $(P, Q) \mapsto D_{KL}(P \| Q)$ is joint convex.

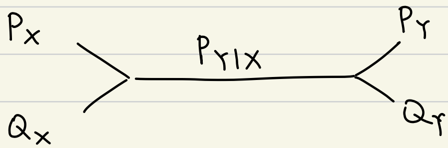
Pf. Follow from the joint convexity of $(x, y) \mapsto x \log \frac{x}{y}$ over $\mathbb{R}_+^2$, whose Hessian is $\begin{bmatrix} 1/x & -1/y \\ -1/y & x/y^2 \end{bmatrix} \succeq 0$. $\square$

Property II (Chain rule): $D_{KL}(P_{X^n} \| Q_{X^n}) = \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}} [D_{KL}(P_{X_i|X^{i-1}} \| Q_{X_i|X^{i-1}})]$.

Pf.
$$D_{KL}(P_{X^n} \| Q_{X^n}) = \mathbb{E}_{P_{X^n}}\left[\log \frac{P_{X^n}}{Q_{X^n}}\right]$$
$$= \mathbb{E}_{P_{X^n}}\left[\sum_{i=1}^n \log \frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}\right]$$
$$= \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}} \underbrace{\mathbb{E}_{P_{X_i|X^{i-1}}}\left[\log \frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}\right]}_{= D_{KL}(P_{X_i|X^{i-1}} \| Q_{X_i|X^{i-1}})} \quad \square$$

Data processing inequality (DPI): an important property of KL divergence

If



then
$$D_{KL}(P_X \| Q_X) \geq D_{KL}(P_Y \| Q_Y)$$
(i.e. distributions become "closer" after processing)

Pf. (Method 1: convexity) Verify that
$$\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\right] = \frac{P_Y}{Q_Y} \quad \text{(exercise)}$$

Then
$$D_{KL}(P_Y \| Q_Y) = \mathbb{E}_{Q_Y}\left[\frac{P_Y}{Q_Y} \log \frac{P_Y}{Q_Y}\right]$$
$$\leq \mathbb{E}_{Q_Y} \mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X} \log \frac{P_X}{Q_X}\right] \quad \text{(Jensen's on } x \log x\text{)}$$
$$= \mathbb{E}_{Q_X}\left[\frac{P_X}{Q_X} \log \frac{P_X}{Q_X}\right] = D_{KL}(P_X \| Q_X)$$

(Method 2: chain rule) Let $P_{XY} = P_X P_{Y|X}$, $Q_{XY} = Q_X P_{Y|X}$.

$$
\begin{aligned}
D_{KL}(P_X \| Q_X) &= D_{KL}(P_X \| Q_X) + \mathbb{E}_{P_X}[\underbrace{D_{KL}(P_{Y|X} \| Q_{Y|X})}_{=0}] \\
&= D_{KL}(P_{XY} \| Q_{XY}) \\
&= D_{KL}(P_Y \| Q_Y) + \mathbb{E}_{P_Y}[\underbrace{D_{KL}(P_{X|Y} \| Q_{X|Y})}_{\geq 0}] \\
&\geq D_{KL}(P_Y \| Q_Y) \qquad \blacksquare
\end{aligned}
$$

Applications of DPI

① DPI of mutual information: if $X - Y - Z$, then

$$I(X;Y) \geq I(X;Z)$$

Pf.

$$P_{XY} \xrightarrow{Id \otimes P_{Z|Y}} P_{XZ} \text{ (by Markov)}$$

$$P_X P_Y \xrightarrow{Id \otimes P_{Z|Y}} P_X P_Z$$

$$\Rightarrow I(X;Y) = D_{KL}(P_{XY} \| P_X P_Y) \geq D_{KL}(P_{XZ} \| P_X P_Z) = I(X;Z) \qquad \blacksquare$$

② Fano's inequality: if $X \sim \mathrm{Unif}([M])$, then

$$\mathbb{P}(X \neq Y) \geq 1 - \frac{I(X;Y) + \log 2}{\log M}.$$

Pf.

$$P_{XY} \xrightarrow{(x,y) \mapsto \mathbb{1}(x=y)} \mathrm{Bern}(\mathbb{P}(X=Y))$$

$$P_X P_Y \xrightarrow{(x,y) \mapsto \mathbb{1}(x=y)} \mathrm{Bern}(\tfrac{1}{M}) \text{ (since } X \sim \mathrm{Unif}([M]))$$

$$
\begin{aligned}
\Rightarrow I(X;Y) &= D_{KL}(P_{XY} \| P_X P_Y) \\
&\geq D_{KL}(\mathrm{Bern}(\mathbb{P}(X=Y)) \| \mathrm{Bern}(\tfrac{1}{M})) \\
&= (1 - \mathbb{P}(X \neq Y)) \log \frac{1 - \mathbb{P}(X \neq Y)}{1/M} + \mathbb{P}(X \neq Y) \log \frac{\mathbb{P}(X \neq Y)}{1 - \frac{1}{M}} \\
&\geq (1 - \mathbb{P}(X \neq Y)) \log M - \log 2 \qquad \blacksquare
\end{aligned}
$$

③ Contiguity : $\forall$ event A : $P(A) \log \frac{P(A)}{e\,Q(A)} \le D_{KL}(P \| Q)$

(so if $D_{KL}(P\|Q) = O(1)$, then $Q(A) \to 0 \implies P(A) \to 0$)

Pf.

$$\begin{matrix} P \\ Q \end{matrix} \xrightarrow{\; x \mapsto \mathbb{1}(x \in A) \;} \begin{matrix} \mathrm{Bern}(P(A)) \\ \mathrm{Bern}(Q(A)) \end{matrix}$$

$$\implies D_{KL}(P\|Q) \ge D_{KL}(\mathrm{Bern}(P(A)) \| \mathrm{Bern}(Q(A))) \ge P(A) \log \frac{P(A)}{e\,Q(A)} \qquad \square$$

Dual representation of KL : move from distributions to functions

Donsker-Varadhan. $D_{KL}(P\|Q) = \sup_f \mathbb{E}_P f - \log \mathbb{E}_Q[e^f]$.

where the sup is taken over all functions f with $\mathbb{E}_Q[e^f] < \infty$.

Pf. ($\le$) Take $f = \log \frac{dP}{dQ}$.

($\ge$) By replacing f by $f - c$, WLOG can assume $\mathbb{E}_Q[e^f] = 1$.

In this case,

$$\tilde{Q}(dx) = e^{f(x)} Q(dx) \text{ is also a distribution.}$$

So

$$D_{KL}(P\|Q) - \mathbb{E}_P f = \mathbb{E}_P\left[\log \frac{dP}{e^f dQ}\right] = \mathbb{E}_P\left[\log \frac{dP}{d\tilde{Q}}\right]$$

$$= D_{KL}(P \| \tilde{Q}) \ge 0 \qquad \square$$

Gibbs variational principle.

$$\log \mathbb{E}_Q[e^f] = \sup_P \mathbb{E}_P f - D_{KL}(P\|Q)$$

Pf. ($\le$) Take $P(dx) = \frac{e^{f(x)} Q(dx)}{\mathbb{E}_Q[e^f]}$.

($\ge$) By Donsker-Varadhan. $\square$

Both results have numerous applications in practice.

Application 1: transportation inequalities.

Example 1.1. Restricting Donsker-Varadhan to $f = \lambda g$ with $\|g\|_\infty \le 1$:

$$D_{KL}(P\|Q) \ge \sup_{\substack{\lambda \in \mathbb{R} \\ \|g\|_\infty \le 1}} \lambda \mathbb{E}_P[g] - \underbrace{\log \mathbb{E}_Q[e^{\lambda g}]}_{\le \lambda \mathbb{E}_Q[g] + \frac{\lambda^2}{2} \text{ by Hoeffding's ineq.}}$$

$$\ge \sup_{\substack{\lambda \in \mathbb{R} \\ \|g\|_\infty \le 1}} \lambda(\mathbb{E}_P[g] - \mathbb{E}_Q[g]) - \frac{\lambda^2}{2}$$

$$= \frac{1}{2}\Big(\sup_{\|g\|_\infty \le 1} \mathbb{E}_P[g] - \mathbb{E}_Q[g]\Big)^2$$

$$= 2 \cdot TV(P, Q)^2,$$

which is Pinsker's inequality (see next lecture, also for an alternative proof)

Example 1.2 (Bobkov & Götze): The following are equivalent:

① $\mathbb{E}_Q[e^{\lambda(f - \mathbb{E}_Q f)}] \le \exp(\frac{\lambda^2}{2} C)$ for all $\underbrace{\text{1-Lip functions}}$ f and $\lambda \in \mathbb{R}$;

② $\underbrace{W_1(P, Q)} \le \sqrt{2C \cdot D_{KL}(P\|Q)}$ holds for all P.

Lipschitz: $|f(x) - f(y)| \le d(x, y)$ for a given metric d.

Wasserstein-1 distance:

$$\inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}[d(X,Y)] = \sup_{f: \text{1-Lip}} \mathbb{E}_P f - \mathbb{E}_Q f.$$

Pf. (① ⇒ ②)

$$D_{KL}(P\|Q) \ge \sup_{\substack{\lambda \in \mathbb{R} \\ f: \text{1-Lip}}} \lambda \cdot \mathbb{E}_P f - \log \mathbb{E}_Q[e^{\lambda f}]$$

$$\ge \sup_{\substack{\lambda \in \mathbb{R} \\ f: \text{1-Lip}}} \lambda(\mathbb{E}_P f - \mathbb{E}_Q f) - \frac{\lambda^2 C}{2}$$

$$= \frac{1}{2C}\Big(\sup_{f: \text{1-Lip}} \mathbb{E}_P f - \mathbb{E}_Q f\Big)^2 = \frac{W_1(P,Q)^2}{2C}.$$

(② ⇒ ①)

$$\log \mathbb{E}_Q[e^{\lambda(f - \mathbb{E}_Q f)}] = \sup_P \mathbb{E}_P[\lambda(f - \mathbb{E}_Q[f])] - D_{KL}(P\|Q)$$

$$\le \sup_P \lambda(\mathbb{E}_P f - \mathbb{E}_Q f) - \frac{(\mathbb{E}_P f - \mathbb{E}_Q f)^2}{2C}$$

$$\le \frac{\lambda^2}{2} C$$

∎

Application 2: variational inference.

Setting: a family of distributions $p_\theta(x^n, y^n)$ where both $p_\theta(x^n)$ and $p_\theta(y^n \mid x^n)$ are tractable

Problem: estimate θ given only y^n (x^n not observable: missing data / latent variables)

Difficulty: $p_\theta(y^n) = \int p_\theta(x^n) p_\theta(y^n \mid x^n)\, dx^n$ often not log-concave or tractable

Evidence Lower Bound (ELBO)

$$\boxed{\log p_\theta(y^n) = \sup_q \mathbb{E}_{x^n \sim q}\left[\log \frac{p_\theta(x^n, y^n)}{q(x^n)}\right]}$$

Pf. Gibbs variational principle

$$\Rightarrow \log p_\theta(y^n) = \log \mathbb{E}_{p_\theta(x^n)} e^{\log p_\theta(y^n \mid x^n)}$$

$$= \sup_q \mathbb{E}_{q(x^n)}[\log p_\theta(y^n \mid x^n)] - D_{KL}(q \,\|\, p_\theta)$$

$$= \text{ELBO} \qquad \blacksquare$$

Example 2.1 (Ising model). $p(y^n) = \frac{1}{Z} \exp\left(\sum_{i<j} A_{ij} y_i y_j + \sum_i b_i y_i\right)$, $y^n \in \{\pm 1\}^n$

Variational inference of $\log Z$:

$$\log Z = \log\left(2^n \mathbb{E}_{y^n \sim \text{Unif}(\{\pm1\}^n)} \exp\left(\sum_{i<j} A_{ij} y_i y_j + \sum_i b_i y_i\right)\right)$$

$$= n \log 2 + \sup_p \left(\mathbb{E}_p\left[\sum_{i<j} A_{ij} y_i y_j + \sum_i b_i y_i\right] - D_{KL}(p \,\|\, \text{Unif}(\{\pm1\}^n))\right)$$

$$= \sup_p \mathbb{E}_p\left[\sum_{i<j} A_{ij} y_i y_j + \sum_i b_i y_i\right] + H(p).$$

Relaxing to $p = \prod_i \text{Bern}(p_i)$ and optimizing over $(p_1, \cdots, p_n)$ yield a tractable lower bound.

Example 2.2 (EM algorithm): aim to find the MLE

$$\arg\max_\theta \log p_\theta(y^n) = \arg\max_\theta \sup_q \mathbb{E}_{x^n\sim q}\left[\log \frac{p_\theta(x^n, y^n)}{q(x^n)}\right].$$

Successive maximization:

- E step: fix $\theta = \theta^{(t)}$, the maximizer is $q^{(t)}(x^n) = \underbrace{p_{\theta^{(t)}}(x^n \mid y^n)}_{\text{factorizable in the missing data case}}$
- M step: fix $q = q^{(t)}$, the maximizer is

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{x^n\sim q^{(t)}}[\underbrace{\log p_\theta(x^n, y^n)}_{\text{no integral; tractable}}]$$

(For example, in exponential families $p_\theta(x,y) \propto \exp(\langle \theta, T(x,y)\rangle - A(\theta))$,
E-step corresponds to the computation of $\mu_i \triangleq \mathbb{E}_{x_i\sim p_{\theta^{(t)}}(\cdot\mid y_i)}[T(x_i, y_i)]$,
and M-step corresponds to the usual MLE computation $\nabla A(\theta^{(t+1)}) = \frac{1}{n}\sum_{i=1}^n \mu_i$)

Example 2.3 (VAE): given images $y_1, \cdots, y_n$, aim to find a generative model:

$$x_i \sim N(0, I), \quad y_i \sim N(\underbrace{\mu_\theta(x_i)}, \underbrace{\sigma_\theta^2(x_i)}I)$$

parametrized by neural nets

Using ELBO:

$$\max_\theta \log p_\theta(y^n) \geq \max_\theta \max_\phi \mathbb{E}_{x^n\sim \underbrace{q_\phi}_{\text{another Gaussian}}}\left[\log \frac{p_\theta(x^n, y^n)}{q_\phi(x^n)}\right]$$

$q_\phi(y_i) = N(\mu_\phi(y_i), \sigma_\phi^2(y_i) I)$

Idea of VAE: ① Replace $\mathbb{E}_{x^n\sim q_\phi}$ by empirical mean of simulated samples

$$x_{ij} \sim N(\mu_\phi(y_i), \sigma_\phi^2(y_i) I), \quad j = 1, 2, \cdots, M;$$

② Compute ∇_θ by the explicit expression of $\log p_\theta(x^n, y^n)$;

③ Compute ∇_ϕ by the reparametrization trick:

$$\nabla_\phi \mathbb{E}_{x\sim N(\mu_\phi, \sigma_\phi^2 I)}[f(x)] = \nabla_\phi \mathbb{E}_{\varepsilon\sim N(0,I)}[f(\mu_\phi + \sigma_\phi \varepsilon)]$$

$$= \mathbb{E}_{\varepsilon\sim N(0,I)}[\nabla_\phi f(\mu_\phi + \sigma_\phi \varepsilon)]$$

$$\approx \frac{1}{M}\sum_{i=1}^M \nabla_\phi f(\mu_\phi + \sigma_\phi \varepsilon_i).$$

Application 3: adaptive data analysis.

Problem: data $X^n \overset{i.i.d.}{\sim} P$, a class of functions $\{\phi_t : \mathcal{X} \to \mathbb{R}\}$

For each given ϕ_t, we have

$$P_n \phi_t := \frac{1}{n}\sum_{i=1}^n \phi_t(X_i) \approx \mathbb{E}_P[\phi_t(X_i)] =: P\phi_t$$

What happens to $P_n\phi_T$ if the index T depends on the data X^n?

Example 3.1 (Russo & Zou'16) If each ϕ_t is σ^2-sub-Gaussian under P, then

$$\left| \mathbb{E}[P_n\phi_T] - \mathbb{E}[P\phi_T] \right| \le \sqrt{\frac{2\sigma^2}{n} I(X^n;T)}.$$

Remark: ① If $I(T;X^n)=0$, i.e. T is independent of X^n, then $P_n\phi_T$ is unbiased for $P\phi_T$.

② If $T\in\{1,\cdots,m\}$, then $I(X^n;T) \le H(T) \le \log m$, and the upper bound $\sqrt{\frac{2\sigma^2\log m}{n}}$ can be shown via union bound

Pf. Define two distributions: $P_{X^n,T}$: the joint distribution in the problem

$Q_{X^n,T} = P_{X^n}P_T$: an auxiliary distribution where X^n and T are independent

Then

$$\mathbb{E}[P_n\phi_T] = \mathbb{E}_{P_{X^n,T}}\left[\frac{1}{n}\sum_i \phi_T(X_i)\right]$$

$$\mathbb{E}[P\phi_T] = \mathbb{E}_{Q_{X^n,T}}\left[\frac{1}{n}\sum_i \phi_T(X_i)\right]$$

Donsker-Varadhan $\Rightarrow$

$$I(X^n;T) = D_{KL}(P_{X^n,T}\| Q_{X^n,T})$$

$$\ge \sup_{\lambda\in\mathbb{R}} \mathbb{E}_{P_{X^n,T}}\left[\frac{\lambda}{n}\sum_i\phi_T(X_i)\right] - \underbrace{\log \mathbb{E}_{Q_{X^n,T}}\left[e^{\frac{\lambda}{n}\sum_i \phi_T(X_i)}\right]}_{\le \lambda\cdot\mathbb{E}[P\phi_T] + \frac{\lambda^2\sigma^2}{2n} \text{ by sub-Gaussian assumption}}$$

$$\ge \sup_{\lambda\in\mathbb{R}} \lambda\left(\mathbb{E}[P_n\phi_T] - \mathbb{E}[P\phi_T]\right) - \frac{\lambda^2\sigma^2}{2n}$$

$$= \left(\mathbb{E}[P_n\phi_T] - \mathbb{E}[P\phi_T]\right)^2 \cdot \frac{n}{2\sigma^2} \qquad \boxed{m}$$

Application 4: PAC-Bayes.

PAC-Bayes inequality. Let $X \sim P$, and consider a class of functions $\{f_\theta : \mathcal{X} \to \mathbb{R}\}$. Fix any prior distribution π of θ. Then w.p. $\geq 1-\delta$ (over the randomness in X), for all distributions ρ over θ,

$$\mathbb{E}_{\theta\sim\rho}[f_\theta(X) - \psi(\theta)] \leq D_{KL}(\rho\|\pi) + \log\frac{1}{\delta},$$

where $\psi(\theta) := \log \mathbb{E}_{X\sim P}\, e^{f_\theta(X)}$.

Remark: ① The exception set depends on π, but not on ρ.

② This inequality holds for all ρ, which generalizes the union bound where ρ is usually taken to be a point mass $\rho = \delta_{\theta_0}$.

③ By taking $\rho = P_{\theta|X}$ to be a data-dependent distribution, we'll have

$$\mathbb{E}_{(\theta,X)\sim P_{\theta X}}[f_\theta(X) - \psi(\theta)] \leq \inf_{\pi} \mathbb{E}_{P_X}[D_{KL}(P_{\theta|X}\|\pi)]$$
$$= I(\theta; X) \quad \text{(exercise!)}$$

Pf. By Markov's inequality, suffices to prove

$$\mathbb{E}_{X\sim P}\left[\sup_\rho \exp\left(\mathbb{E}_{\theta\sim\rho}[f_\theta(X) - \psi(\theta)] - D_{KL}(\rho\|\pi)\right)\right] \leq 1.$$

By Gibbs variational principle, the LHS is

$$\mathbb{E}_{X\sim P}\left[\exp\left(\log \mathbb{E}_{\theta\sim\pi}\, e^{f_\theta(X)-\psi(\theta)}\right)\right]$$
$$= \mathbb{E}_{X\sim P}\,\mathbb{E}_{\theta\sim\pi}\left[e^{f_\theta(X)-\psi(\theta)}\right]$$
$$= \mathbb{E}_{\theta\sim\pi}\Big[\underbrace{\mathbb{E}_{X\sim P}\, e^{f_\theta(X)-\psi(\theta)}}_{=1}\Big] = 1 \qquad \blacksquare$$

Why call it PAC-Bayes? Come from the following application in statistical learning theory:

<u>Example 4.1</u>. Let $f: \mathcal{X} \to [0,1]$, $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P$

$$P_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i), \quad Pf := \mathbb{E}_{X \sim P}[f(X)].$$

For fixed f, sub-Gaussian concentration (Hoeffding's inequality) gives

$$(P_n f - Pf)^2 \le \frac{1}{2n} \log \frac{2}{\delta} \quad \text{w.p.} \ge 1-\delta.$$

By PAC-Bayes, fix a prior π, then w.p. $\ge 1-\delta$, for any ρ:

$$\mathbb{E}_{f \sim \rho}\Big[\lambda(P_n f - Pf)^2 - \underbrace{\log \mathbb{E}_X e^{\lambda(\underbrace{P_n f - Pf}_{\frac{1}{8n}\text{-subGaussian}})^2}}_{\le \frac{1}{2}\log \frac{1}{1-\frac{\lambda}{4n}} \text{ if } \lambda < 4n.}\Big] \le D_{KL}(\rho \| \pi) + \log \frac{1}{\delta}$$

Choosing $\lambda = 2n$ gives $\qquad \mathbb{E}_{f\sim\rho}[(P_n f - Pf)^2] \le \dfrac{D_{KL}(\rho\|\pi) + \log\frac{2}{\delta}}{2n} \quad \forall \rho.$

PAC-Bayes also has surprising applications to concentration inequalities, by choosing ρ and π appropriately.

<u>Example 4.2</u>. If $X \sim N(0, \Sigma)$, then w.p. $\ge 1-\delta$,

$$\|X\|_2 \le \sqrt{\operatorname{Tr} \Sigma} + \sqrt{2\|\Sigma\|_{op} \log \frac{1}{\delta}}.$$

<u>Remark</u>: Try union bound to $\|X\|_2 = \sup_{\|v\|_2 \le 1} \langle X, v\rangle$ yourself!

It's very difficult to make covering / chaining arguments give such sharp bound, because of a general shaped Σ. One need to invoke Talagrand's generic chaining to this example, but it's very difficult to carry out.

Pf. $\|X\|_2 = \sup_{\|v\|_2=1} \langle v, X \rangle$.

To apply PAC-Bayes, we construct a prior ρ_v such that $\mathbb{E}_{\theta\sim\rho_v}[\langle\theta, X\rangle] = \langle v, X\rangle$.
A natural choice is $\rho_v = N(v, \sigma^2 I)$. Then for $\pi = N(0, \sigma^2 I)$: w.p. $\geq 1-\delta$,

$$\sup_{\|v\|_2\leq 1} \mathbb{E}_{\rho_v}\Big[\lambda\langle\theta, X\rangle - \underbrace{\log \mathbb{E}_X e^{\lambda\langle\theta, X\rangle}}_{=\frac{\lambda^2}{2}\theta^T\Sigma\theta}\Big] - \underbrace{D_{KL}(\rho_v\|\pi)}_{=\frac{\|v\|_2^2}{2\sigma^2}=\frac{1}{2\sigma^2}} \leq \log\frac{1}{\delta}.$$

$$\Rightarrow \sup_{\|v\|_2\leq 1} \lambda\langle v, X\rangle - \frac{\lambda^2}{2}\left(v^T\Sigma v + \sigma^2 \mathrm{Tr}(\Sigma)\right) - \frac{1}{2\sigma^2} \leq \log\frac{1}{\delta}.$$

$$\Rightarrow \langle v, X\rangle \leq \frac{\lambda}{2}\Big(\underbrace{v^T\Sigma v}_{\leq\|\Sigma\|_{op}} + \sigma^2\mathrm{Tr}(\Sigma)\Big) + \frac{1}{\lambda}\left(\frac{1}{2\sigma^2} + \log\frac{1}{\delta}\right), \quad \forall v.$$

Optimize over σ^2: $\quad \sigma^2 = \frac{1}{\lambda}\cdot\frac{1}{\sqrt{\mathrm{Tr}(\Sigma)}}$

Optimize over λ: $\quad \lambda = \sqrt{\frac{2\log(1/\delta)}{\|\Sigma\|_{op}}}$

$$\Rightarrow \|X\|_2 \leq \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{2\|\Sigma\|_{op}\log\frac{1}{\delta}}$$
w.p. $\geq 1-\delta$. $\boxed{m}$

<u>Example 4.3</u>. Let $X_1, \cdots, X_n$ be i.i.d with $\mathbb{E}[X_1]=0$, $\mathbb{E}[X_1X_1^T]=\Sigma$, and that v^TX_1 is $v^T\Sigma v$-subGaussian for any $v\in\mathbb{R}^n$. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_iX_i^T$ be the sample covariance. Then w.p. $\geq 1-\delta$,

$$\|\hat{\Sigma} - \Sigma\|_{op} \leq C\|\Sigma\|_{op}\left(\sqrt{\frac{r(\Sigma)+\log\frac{1}{\delta}}{n}} + \frac{r(\Sigma)+\log\frac{1}{\delta}}{n}\right)$$

where $r(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}}$ is called the <u>effective rank</u>.

<u>Remark</u>: This is the result of [Koltchinskii & Lounici '17], where the key challenge is to arrive at the tight factor $r(\Sigma)$. Our proof is taken from [Zhivotovskiy '21] via PAC-Bayes.

Pf. (Throughout the proof, C denotes a large universal constant which may change from line to line).

$$\|\hat{\Sigma}-\Sigma\|_{op} = \sup_{\|u\|_2, \|v\|_2=1} u^T(\hat{\Sigma}-\Sigma)v.$$

Consider $(\theta,\theta') \sim \rho_{u,v} := f_u \otimes f_v$, where f_u is the density of

$$N(u, \sigma^2 I) \text{ conditioned on } (x-u)^T\Sigma(x-u) \le r^2.$$

Clearly $\mathbb{E}_{(\theta,\theta')\sim\rho_{u,v}}[\theta^T(\hat{\Sigma}-\Sigma)\theta'] = u^T(\hat{\Sigma}-\Sigma)v$, and

$$p := \mathbb{P}(z^T\Sigma^{-1}z \le r^2) \ge 1 - \frac{\mathbb{E}[z^T\Sigma z]}{r^2} = 1 - \frac{\sigma^2 Tr(\Sigma)}{r^2} \quad \text{for } z \sim N(0, \sigma^2 I).$$

Let $\pi = N(0,\sigma^2 I) \otimes N(0, \sigma^2 I)$. One can compute

$$D_{KL}(f_u \| N(0,\sigma^2 I)) = \frac{1}{2\sigma^2} + \log\left(\frac{1}{p}\right).$$

so that $D_{KL}(\rho_{u,v}\|\pi) = \frac{1}{\sigma^2} + 2\log\left(\frac{1}{p}\right)$.

Now by PAC-Bayes, w.p. $\ge 1-\delta$,

$$\sup_{\|u\|_2,\|v\|_2=1} \mathbb{E}_{(\theta,\theta')\sim\rho_{u,v}}\Big[\lambda\theta^T(\hat{\Sigma}-\Sigma)\theta' - \underbrace{\log \mathbb{E}e^{\lambda\theta^T(\hat{\Sigma}-\Sigma)\theta'}}_{\lesssim \frac{C\lambda^2}{n}(\theta^T\Sigma\theta + \theta'^T\Sigma\theta')^2}\Big] - D_{KL}(\rho_{u,v}\|\pi) \le \log\frac{1}{\delta}$$

$$\text{for } \lambda \le \frac{n}{C(\theta^T\Sigma\theta + \theta'^T\Sigma\theta')}.$$

Since $\theta^T\Sigma\theta \le \left(\sqrt{u^T\Sigma u} + \sqrt{(\theta-u)^T\Sigma(\theta-u)}\right)^2 \le \left(\sqrt{\|\Sigma\|_{op}} + r\right)^2$, we get

$$\|\hat{\Sigma}-\Sigma\|_{op} \lesssim \frac{C\lambda}{n}\left(\sqrt{\|\Sigma\|_{op}} + r\right)^4 + \frac{1}{\lambda}\left(\frac{1}{\sigma^2} + 2\log\left(\frac{1}{p}\right) + \log\frac{1}{\delta}\right)$$

$$\text{if } \lambda \le \frac{n}{C\left(\sqrt{\|\Sigma\|_{op}} + r\right)^2}.$$

Choose $r^2 = 2\|\Sigma\|_{op}$, $\sigma^2 = \frac{\|\Sigma\|_{op}}{Tr(\Sigma)} = \frac{1}{r(\Sigma)}$, then $p \ge \frac{1}{2}$, and

$$\|\hat{\Sigma}-\Sigma\|_{op} \le C\left(\frac{\lambda}{n}\|\Sigma\|_{op}^2 + \frac{1}{\lambda}\left(r(\Sigma) + \log\frac{1}{\delta}\right)\right), \quad \text{if } \lambda \le \frac{n}{C\|\Sigma\|_{op}}.$$

Finally, choosing

$$\lambda \asymp \begin{cases} \dfrac{n}{\|\Sigma\|_{op}} \sqrt{\dfrac{r(\Sigma)+\log(\frac{1}{\delta})}{n}} & \text{if } \dfrac{r(\Sigma)+\log\frac{1}{\delta}}{n} \leq 1 \\ \dfrac{n}{\|\Sigma\|_{op}} & \text{if } \dfrac{r(\Sigma)+\log\frac{1}{\delta}}{n} > 1 \end{cases}$$

leads to the claimed result. $\blacksquare$