2 Basic tail and concentration bounds

Exercises

Exercise 2.2 (Mill's ratio) Let $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ be the density function of a standard normal variate $Z \sim \mathcal{N}(0,1)$.

(a) $\phi'(z) = (-z) \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = -z\phi(z)$.

(b) First, the tail can be simplified using part (a) as follows

$$\begin{aligned}\mathbb{P}[Z \geq z] = \int_{u=z}^{\infty} \phi(u)du &\overset{(a)}{=} \int_{u=z}^{\infty} (-u^{-1})\phi'(u)du = -u^{-1}\phi(u)\Big|_{u=z}^{\infty} - \int_{u=z}^{\infty} u^{-2}\phi(u)du \\ &\overset{(a)}{=} z^{-1}\phi(z) + \int_{u=z}^{\infty} u^{-3}\phi'(u)du = (z^{-1} - z^{-3})\phi(z) + 3\int_{u=z}^{\infty} u^{-4}\phi(u)du.\end{aligned}$$

But also,

$$0 \leq 3\int_{u=z}^{\infty} u^{-4}\phi(u)du \overset{(a)}{=} -3\int_{u=z}^{\infty} u^{-5}\phi'(u)du = 3z^{-5}\phi(z) - 15\int_{u=z}^{\infty} u^{-6}\phi(u)du \leq 3z^{-5}\phi(z),$$

and so, $(z^{-1} - z^{-3})\phi(z) \leq \mathbb{P}[Z \geq z] \leq (z^{-1} - z^{-3} + 3z^{-5})\phi(z)$ for all $z > 0$.

Exercise 2.3 (Polynomial Markov vs Chernoff) For $X \geq 0$, assume that the moment generating function exists in an interval around zero. Then, for any $\delta > 0$, and λ small enough to guarantee the existence of the moment generating function, we have

$$\frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\delta}} = \frac{\sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[|X|^k]}{k!}}{\sum_{k=0}^{\infty} \frac{\lambda^k \delta^k}{k!}} = \frac{\sum_{k=0}^{\infty} \frac{\lambda^k \delta^k}{k!} \frac{\mathbb{E}[|X|^k]}{\delta^k}}{\sum_{k=0}^{\infty} \frac{\lambda^k \delta^k}{k!}} \geq \inf_{k \geq 0} \frac{\mathbb{E}[|X|^k]}{\delta^k}.$$

Therefore, we obtain $\inf_{k\geq 0} \frac{\mathbb{E}[|X|^k]}{\delta^k} \leq \inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\delta}}$.

Exercise 2.8 (Bernstein and expectations) Consider a non-negative random variable Z satisfying the tail inequality $\mathbb{P}[Z \geq t] \leq Ce^{-\frac{t^2}{2(v^2+bt)}}$ for all $t > 0$, for positive constants $v, b > 0$ and $C \geq 1$. Let $T > 0$ be such that $v^2 = bT$. Then, for $t \geq T$, $bt \geq bT = v^2$ and for $t < T$, $bt < bT = v^2$. Thus, we have

$$\mathbb{P}[Z \geq t] \leq \begin{cases} Ce^{-\frac{t^2}{4v^2}}, & 0 \leq t < T \\ Ce^{-\frac{t}{4b}}, & t \geq T \end{cases}$$

(a) Since $Z \geq 0$, for T_1, T_2 to be chosen later, we can compute the bound:

$$\begin{aligned}
\mathbb{E}[Z] &= \int_{t=0}^{\infty} \mathbb{P}[Z \geq t]dt = \int_{t=0}^{T} \mathbb{P}[Z \geq t]dt + \int_{t=T}^{\infty} \mathbb{P}[Z \geq t]dt \\
&\leq T_1 + C \int_{t=T_1 \wedge T}^{T} e^{-\frac{t^2}{4\nu^2}} dt + (T_2 - T) + C \int_{t=T_2}^{\infty} e^{-\frac{t}{4b}} dt \\
&\leq T_1 + C \int_{t=T_1}^{\infty} e^{-\frac{t^2}{2(\nu\sqrt{2})^2}} dt + T_2 + C \int_{t=T_2}^{\infty} e^{-\frac{t}{4b}} dt \\
&\leq T_1 + C \cdot \sqrt{2\pi} \cdot \nu \cdot \sqrt{2} \cdot \mathbb{P}[(\nu\sqrt{2})Z \geq T_1] + T_2 + 4b \cdot C \cdot e^{-\frac{T_2}{4b}} \\
&\leq T_1 + 2\nu\sqrt{\pi} \cdot Ce^{-\frac{T_1^2}{4\nu^2}} + T_2 + 4b \cdot Ce^{-\frac{T_2}{4b}}.
\end{aligned}$$

Choose T_1 such that $Ce^{-\frac{T_1^2}{4\nu^2}} = 1$, or $T_1 = 2\nu\sqrt{\ln C}$, and T_2 such that $Ce^{-\frac{T_2}{4b}} = 1$ or $T_2 = 4b \ln C$:

$$\mathbb{E}[Z] \leq 2\nu(\sqrt{\pi} + \sqrt{\ln C}) + 4b(1 + \ln C).$$

Exercise 2.13 (Operations on sub-Gaussian variables) Suppose that X_1 and X_2 are zero mean and sub-Gaussian with parameters σ_1 and σ_2, respectively, i.e., $\mathbb{E}[e^{\lambda X_i}] \leq e^{\lambda^2\sigma_i^2/2}, i = 1, 2$, for all $\lambda \in \mathbb{R}$.

(a) For any $\lambda \in \mathbb{R}$, using independence, $\mathbb{E}[e^{\lambda(X_1+X_2)}] = \mathbb{E}[e^{\lambda X_1}]\mathbb{E}[e^{\lambda X_2}] \leq e^{\lambda^2(\sigma_1^2+\sigma_2^2)/2} \triangleq e^{\lambda^2\sigma^2/2}$, whence $X_1 + X_2$ is sub-Gaussian with parameter $\sigma \triangleq \sqrt{\sigma_1^2 + \sigma_2^2}$.

(b) For any $\lambda \in \mathbb{R}$, by Cauchy-Schwartz inequality applied to $Y_i \triangleq e^{\lambda X_i}, i = 1, 2$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X_1+X_2)}] &= \mathbb{E}[Y_1Y_2] \triangleq \langle Y_1, Y_2\rangle \leq \|Y_1\|_2\|Y\|_2 = (\mathbb{E}[Y_1^2]\mathbb{E}[Y_2^2])^{1/2} \\
&= (\mathbb{E}[e^{2\lambda X_1}]\mathbb{E}[e^{2\lambda X_2}])^{1/2} \leq e^{\lambda^2(2(\sigma_1^2+\sigma_2^2))/2} \triangleq e^{\lambda^2\sigma^2/2},
\end{aligned}$$

whence $X_1 + X_2$ is sub-Gaussian with parameter $\sigma \triangleq \sqrt{2}\sqrt{\sigma_1^2 + \sigma_2^2}$.

Exercise 2.17 (Hanson-Wright inequality) Given $\{X_i\}_{i=1}^n$ and a positive definite matrix $\mathbf{Q} \in \mathcal{S}_+^{n\times n}$, consider the quadratic form $Z = \sum_{i=1}^n \sum_{j=1}^n \mathbf{Q}_{ij}X_iX_j = X^T\mathbf{Q}X$ where $X \triangleq [X_1, \ldots, X_n]^T \in \mathbb{R}^n$. When $X_i \sim \mathcal{N}(0, 1)$ for $i \in [n]$ are i.i.d., by the rotational invariance of X, for the eigen-decomposition $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ with $\mathbf{\Lambda} \triangleq \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, we have that $Z \overset{d}{=} X^T\mathbf{\Lambda}X = \sum_{i=1}^n \lambda_iX_i^2$. Note that, for $i \in [n]$, $Y_i = \lambda_iX_i^2$ is a scaled χ^2-random variable, and is sub-exponential with parameters $(2\lambda_i, 4\lambda_i)$ (by some manipulation of Example 2.4). Hence, $Z = \sum_{i=1}^n Y_i$ is sub-exponential with parameters $\nu = 2\sqrt{\sum_{i=1}^n \lambda_i^2} = 2|||\mathbf{Q}|||_F$ and $\alpha = 4\max_i \lambda_i = 4|||\mathbf{Q}|||_2$ (by the discussion following the Remark after Proposition 2.10). Therefore, noting that $\mathbb{E}[Z] = \sum_{i=1}^n \lambda_i = \mathrm{trace}(\mathbf{Q})$, we have using Proposition 2.9

(Sub-exponential tail bound) that for any $t > 0$:

$$\mathbb{P}[Z \geq \text{trace}(\mathbf{Q}) + t] = \mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \exp\left\{-\min\left(\frac{t}{8|||\mathbf{Q}|||_2}, \frac{t^2}{8|||\mathbf{Q}|||_F^2}\right)\right\}.$$

Exercise 2.18 (Orlicz norms) Given a strictly increasing convex function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$, define the ψ Orlicz norm of a random variable X as $\|X\|_\psi \triangleq \inf\{t > 0 : \mathbb{E}[\psi(t^{-1}|X|)] \leq 1\}$. Consider the function $\psi_q(u) \triangleq e^{u^q} - 1$ for some $q \geq 1$.

(a) Since $\|X\|_{\psi_q} < +\infty$, for any $\lambda^{-1/q} > \|X\|_{\psi_q}$ or $0 < \lambda < \|X\|_{\psi_q}^{-q}$, we have

$$\mathbb{E}[e^{\lambda|X|^q}] = \mathbb{E}[e^{((\lambda^{-1/q})^{-1}|X|)^q}] = 1 + \mathbb{E}[\psi_q((\lambda^{-1/q})^{-1}|X|)] \leq 2.$$

Then, for any $t > 0$, the tail can be bounded using a Chernoff argument as

$$\begin{aligned}\mathbb{P}[|X| > t] = \mathbb{P}[|X|^q > t^q] &\leq \inf_{\lambda>0} e^{-\lambda t^q}\mathbb{E}[e^{\lambda|X|^q}] \\ &\leq \inf_{0<\lambda<\|X\|_{\psi_q}^{-q}} e^{-\lambda t^q}\mathbb{E}[e^{\lambda|X|^q}] \leq 2\inf_{0<\lambda<\|X\|_{\psi_q}^{-q}} e^{-\lambda t^q} = 2e^{-\|X\|_{\psi_q}^{-q} t^q}.\end{aligned}$$

(b) Let the tail bound $\mathbb{P}[|X| > x] = \mathbb{P}[|X|^q > x^q] \leq c_1 e^{-c_2 x^q}$ hold for all $x > 0$ for some $c_2 > 0, c_1 \geq 1$. Then, for any $t > 0$

$$\begin{aligned}1 + \mathbb{E}[\psi_q(t^{-1}|X|)] = \mathbb{E}[e^{t^{-q}|X|^q}] &= \int_{u=0}^{\infty} \mathbb{P}[e^{t^{-q}|X|^q} > u]du = \int_{u=0}^{\infty} \mathbb{P}[|X|^q > t^q(\ln(u))]du \\ &\leq v + c_1\int_{u=v}^{\infty} e^{-c_2 t^q \ln(u)}du = v + c_1\int_{u=v}^{\infty} u^{-c_2 t^q}du,\end{aligned}$$

for any $v > 0$. Using $c_2 t^q = s + 1$ for some $s > 0$, one gets

$$1 + \mathbb{E}[\psi_q(t^{-1}|X|)] \leq v + c_1\int_{u=v}^{\infty} u^{-s-1}du = v + \frac{c_1}{sv^s} \triangleq f(v).$$

Choose v_0 such that $f'(v_0) = 0$, giving $1 - c_1 v_0^{-(s+1)} = 0$ or $v_0 = c_1^{1/(s+1)}$. Therefore, we have

$$1 + \mathbb{E}[\psi_q(t^{-1}|X|)] \leq f(v_0) = c_1^{1/(s+1)} + \frac{c_1}{sc_1^{s/(s+1)}} = c_1^{1/(s+1)}(1 + s^{-1}) \triangleq g(s).$$

If $(1 + s^{-1}) \leq \sqrt{2}$ and $c_1^{1/(s+1)} \leq \sqrt{2}$, then $g(s) \leq 2$, i.e., $s + 1 \geq 1 + 1/(\sqrt{2} - 1) = \sqrt{2}/(\sqrt{2} - 1)$ and $s + 1 \geq 2\ln(c_1)/\ln(2) = 2\log_2 c_1$. Thus, using $s + 1 = c_2 t^q$ back, we get that for any

$$t \geq c_2^{-q^{-1}} \max\{(2\log_2 c_1)^{q^{-1}}, (\sqrt{2}/(\sqrt{2} - 1))^{q^{-1}}\} \triangleq h_q(c_1, c_2),$$

we have $\mathbb{E}[\psi_q(t^{-1}|X|)] \le 1$ whence $\|X\|_{\psi_q} \le h_q(c_1, c_2) < +\infty$.

Exercise 2.19 (Maxima of Orlicz variables) Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of zero mean random variables with finite Orlicz norm $\sigma \triangleq \|X_1\|_\psi$. For any $\lambda > 0$, by Jensen's inequality applied to the convex function ψ (a), and the fact that ψ is strictly increasing (b), one has

$$\psi(\lambda\mathbb{E}[\max_{i\in[n]}|X_i|]) \overset{(a)}{\le} \mathbb{E}[\psi(\lambda\max_{i\in[n]}|X_i|)] \overset{(b)}{\le} \mathbb{E}[\max_{i\in[n]}\psi(\lambda|X_i|)] \le n\mathbb{E}[\psi(\lambda|X_1|)].$$

For $\lambda^{-1} \ge \sigma$ or $0 < \lambda \le \sigma^{-1}$, we have that

$$\psi(\lambda\mathbb{E}[\max_{i\in[n]}|X_i|]) \le n\mathbb{E}[\psi((\lambda^{-1})^{-1}|X_1|)] \le n,$$

and so $\mathbb{E}[\max_{i\in[n]}|X_i|] \le \inf_{0<\lambda\le\sigma^{-1}} \lambda^{-1}\psi^{-1}(n) = \sigma\psi^{-1}(n)$.

Exercise 2.20 (Tail bounds under moment conditions) Suppose that $\{X_i\}_{i=1}^n$ are zero-mean and independent random variables such that for some fixed integer $m \ge 1$, they satisfy the moment bound $\|X_i\|_{2m} \triangleq (\mathbb{E}[X_i^{2m}])^{\frac{1}{2m}} \le C_m$. For any $m \ge 1, \delta > 0$, using Markov's inequality (a) and (the hinted) Rosenthal's inequality (b), one has

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \ge \delta\right] = \mathbb{P}\left[\left|\sum_{i=1}^n X_i\right| \ge n\delta\right] &= \mathbb{P}\left[\Big(\sum_{i=1}^n X_i\Big)^{2m} \ge (n\delta)^{2m}\right] \\
&\overset{(a)}{\le} (n\delta)^{-2m}\mathbb{E}\left[\Big(\sum_{i=1}^n X_i\Big)^{2m}\right] \overset{(b)}{\le} (n\delta)^{-2m}R_m\left\{\sum_{i=1}^n \mathbb{E}[X_i^{2m}] + \Big(\sum_{i=1}^n \mathbb{E}[X_i^2]\Big)^m\right\} \\
&\le (n\delta)^{-2m}R_m(nC_m^{2m} + n^m C_1^{2m}) \le B_m\left(\frac{1}{\sqrt{n}\delta}\right)^{2m},
\end{aligned}$$

where $B_m \triangleq R_m(C_m^{2m} + C_1^{2m})$.

Exercise 2.21 (Concentration and data compression) Let $X = (X_1, \ldots, X_n)$ be the binary vector of i.i.d. Bernoulli random variables with parameter $1/2$.

(a) Let $R < D_2(\delta \,\|\, 1/2) \triangleq \delta \lg \frac{\delta}{1/2} + (1-\delta)\lg\frac{1-\delta}{1/2}$, where lg denotes $\log_2$. Let $N \le 2^{nR}$ be the number of codewords. For any z^j,

$$\mathbb{P}[\rho_H(X, z^j) \le \delta] = \mathbb{P}\left[\sum_{i=1}^n \mathbb{I}[X_i \ne z_i^j] \le n\delta\right] \overset{(a)}{=} \mathbb{P}\left[\sum_{i=1}^n \mathbb{I}[X_i \ne 0] \le n\delta\right] \overset{(b)}{\le} 2^{-nD_2(\delta \,\|\, 1/2)},$$

where (a) follows from the symmetry of X, and (b) follows from Exercise 2.9(a). Since $\{d(X) \le$

$\delta\} = \cup_{j=1}^{N}\{\rho_H(X, z^j) \le \delta\}$, using union bound, and the above upper bound, one has

$$\mathbb{P}[d(X) \le \delta] \le \sum_{j=1}^{N} \mathbb{P}[\rho_H(X, z^j) \le \delta] \le N2^{-nD_2(\delta \,\|1/2)} \le 2^{-n(D_2(\delta \,\|1/2)-R)}.$$

Hence if $R < D_2(\delta \,\|1/2)$, $\mathbb{P}[d(X) \le \delta] \to 0$ as $n \to +\infty$, i.e., the rate cannot achieve the distortion (almost surely).

(b) Let $\Delta R \triangleq R - D_2(\delta \,\|1/2) > 0$. Consider a codebook $\{Z^1, \ldots, Z^N\}$, each Z^j being an i.i.d. Bernoulli random variables with parameter $1/2$ as its entries. Let $V^j \triangleq \mathbb{I}[\rho_H(X, Z^j) \le \delta]$, and $V \triangleq \sum_{j=1}^{N} V^j$. Clearly, $\{d(X) \le \delta\} = \cup_{j=1}^{N}\{\rho_H(X, z^j) \le \delta\} = \{V \ge 1\}$.

(i) From the Cauchy-Schwartz inequality, we have

$$(\mathbb{E}[V])^2 = (\mathbb{E}[V\mathbb{I}[V \ge 1]])^2 \le \mathbb{E}[V^2]\mathbb{E}[(\mathbb{I}[V \ge 1])^2] = \mathbb{E}[V^2]\mathbb{P}[V \ge 1].$$

(ii) By the symmetry of the distributions of X, Z^j, $j \in [N]$, it follows that $V^j, i \in [N]$ are i.i.d. Bernoulli random variables with parameter

$$p_n \triangleq \mathbb{P}[V^1 = 1] = \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{I}[X_i \ne Z_i^1] \le n\delta\right] \ge \frac{2^{-nD_2(\delta \,\|1/2)}}{n+1},$$

where the lower bound is due to the result of Exercise 2.10. Hence, we can compute the mean and second moments of the Binomial random variable V with parameters N, p_n as

$$\mathbb{E}[V] = Np_n \ge \frac{2^{n(R-D_2(\delta \,\|\, 1/2))}}{n+1} = \frac{2^{n\Delta R}}{n+1},$$
$$\mathbb{E}[V^2] = (\mathbb{E}[V])^2 + Np_n(1-p_n) = N^2p_n^2 + Np_n(1-p_n).$$

Thus, using the bound from part (a), one obtains

$$\mathbb{P}[d(X) \le \delta] = \mathbb{P}[V \ge 1] \ge \frac{(\mathbb{E}[V])^2}{\mathbb{E}[V^2]} = \frac{Np_n}{N^2p_n^2 + Np_n(1-p_n)}$$
$$= \frac{1}{1 - \frac{1}{N} + \frac{1}{Np_n}} \ge \frac{1}{1 - \frac{1}{2^{nR}} + \frac{n+1}{2^{n\Delta R}}}.$$

Hence, $\mathbb{P}[d(X) \le \delta] \to 1$ as $n \to +\infty$, i.e., the distortion can be asymptotically controlled by the specified rate almost surely.

Exercise 2.22 (Concentration for spin glasses) For some positive integer $d \ge 2$, consider the collection $\{\theta_{jk}\}_{j \ne k}$ of weights for $j \ne k \in [d]$. We can define a probability distribution over the

Boolean hypercube $\{+1,-1\}^d$ via the mass function

$$\mathbb{P}_\theta(x_1,\ldots,x_d) = \exp\left\{\frac{1}{\sqrt{d}}\sum_{j\neq k}\theta_{jk} - F_d(\theta)\right\},$$

where $F_d : \mathbb{R}^{\binom{d}{2}} \to \mathbb{R}$, known as the *free energy*, is given by

$$F_d(\theta) \triangleq \ln\left(\sum_{x\in\{+1,-1\}^d}\exp\left\{\frac{1}{\sqrt{d}}\sum_{j\neq k}\theta_{jk}x_jx_k\right\}\right),$$

serves to normalize the distribution. Denoting $\mathcal{X} \triangleq \{+1,-1\}^d$, for $x \in \{+1,-1\}^d \triangleq \mathcal{X}$, and $\theta \in \mathbb{R}^{\binom{d}{2}}$, let $h_x(\theta) \triangleq \frac{1}{\sqrt{d}}\sum_{j\neq k}\theta_{jk}x_jx_k$, and observe that $h_x(\cdot)$ is linear in θ for each $x \in \mathcal{X}$.

(a) For any $\lambda \in [0,1]$ and θ_1, θ_2 we have

$$\begin{aligned}
F_d(\lambda\theta_1 + (1-\lambda)\theta_2) &= \ln\left(\sum_{x\in\mathcal{X}}\exp(h_x(\lambda\theta_1 + (1-\lambda)\theta_2))\right)\\
&= \ln\left(\sum_{x\in\mathcal{X}}(\exp(h_x(\theta_1)))^\lambda(\exp(h_x(\theta_2)))^{1-\lambda}\right)\\
&\leq \ln\left(\left(\sum_{x\in\mathcal{X}}\exp(h_x(\theta_1))\right)^\lambda\left(\sum_{x\in\mathcal{X}}\exp(h_x(\theta_2))\right)^{1-\lambda}\right)\\
&= \lambda F_d(\theta_1) + (1-\lambda)F_d(\theta_2),
\end{aligned}$$

where the inequality uses Holder's with $p^{-1} = \lambda = 1 - q^{-1}$, and so F_d is convex.

(b) For any θ, θ', by the log-sum inequality (i.e., $\ln((\sum_i a_i)/(\sum_i b_i)) \leq (\sum_i a_i \ln(a_i/b_i))/(\sum_i a_i)$ for non-negative numbers $\{a_i, b_i\}_{i=1}^n$), we have

$$|F_d(\theta) - F_d(\theta')| = \ln\left(\sum_{x\in\mathcal{X}}\exp(h_x(\theta))\right)\Big/\left(\sum_{x\in\mathcal{X}}\exp(h_x(\theta'))\right) \leq \frac{\sum_{x\in\mathcal{X}}\exp(h_x(\theta))|h_x(\theta) - h_x(\theta')|}{\sum_{x\in\mathcal{X}}\exp(h_x(\theta))}.$$

Denoting $p_x(\theta) \triangleq \frac{\exp(h_x(\theta))}{\sum_{x\in\mathcal{X}}\exp(h_x(\theta))}$, and noting that $\sum_{x\in\mathcal{X}} p_x(\theta) = 1$, $|x_j| = 1$ for any $x \in \mathcal{X}$, $j \in [d]$

$$\begin{aligned}
|F_d(\theta) - F_d(\theta')| &\leq \sum_{x\in\mathcal{X}} p_x(\theta)|h_x(\theta - \theta')| \leq \sum_{x\in\mathcal{X}} p_x(\theta)\frac{1}{\sqrt{d}}\sum_{j\neq k}|\theta_{jk} - \theta'_{jk}|\\
&= \frac{1}{\sqrt{d}}\sum_{j\neq k}|\theta_{jk} - \theta'_{jk}| \overset{(a)}{\leq} \sqrt{d}\|\theta - \theta'\|_2,
\end{aligned}$$

where the last inequality follows by the bound $\|\theta-\theta'\|_1 \leq \sqrt{D}\|\theta-\theta'\|_2$, for $D = \binom{d}{2} = \frac{d(d-1)}{2} \leq d^2$.

(c) Using Jensen's inequality, one has

$$e^{\mathbb{E}[F_d(\theta)]} \leq \mathbb{E}[e^{F_d(\theta)}] = \mathbb{E}\left[\sum_{x\in\mathcal{X}} \exp\left\{\frac{1}{\sqrt{d}}\sum_{j\neq k}\theta_{jk}x_j x_k\right\}\right]$$
$$\leq 2^d \mathbb{E}\left[\exp\left\{\frac{1}{\sqrt{d}}\sum_{j\neq k}\theta_{jk}\right\}\right] = 2^d\left(\mathbb{E}[e^{\theta_{12}/\sqrt{d}}]\right)^D \leq 2^d e^{\frac{\beta^2 D}{2d}} \leq 2^d e^{\frac{\beta^2 d}{4}},$$

and so $\frac{\mathbb{E}[F_d(\theta)]}{d} \leq \ln 2 + \frac{\beta^2}{4}$. From part (b), we get that $f_d(\theta) \triangleq \frac{F_d(\theta)}{d}$ is $L \triangleq \frac{1}{\sqrt{d}}$-Lipschitz with respect to the Euclidean norm. Hence, by Theorem 2.26 concentration of Lipschitz functions of Gaussian random variables, we have for all $t > 0$

$$\mathbb{P}\left[\frac{F_d(\theta)}{d} \geq \ln 2 + \frac{\beta^2}{4} + t\right] \leq \mathbb{P}[f_d(\theta) - \mathbb{E}[f_d(\theta)] \geq t] \leq e^{-t^2/2L^2} = e^{-dt^2/2}.$$

3 Concentration of measure

Exercises

Exercise 3.1 (Shannon entropy and Kullback–Leibler divergence) Given a discrete random variable X with probability mass function p over a set $\mathcal{X}$, its Shannon entropy is given by $H(X) \triangleq -\sum_{x\in\mathcal{X}} p(x)\log(p(x))$. For a random variable Z, the entropy based on $\phi(u) = u\log u$ is $\mathbb{H}(Z) \triangleq \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z]$.

(a) Let U be uniformly distributed over $\mathcal{X}$, and let $Z = p(U)$. Then, noting that $\mathbb{E}[Z] = \mathbb{E}[p(U)] = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} p(x) = \frac{1}{|\mathcal{X}|}$, we have

$$\begin{aligned}\mathbb{H}(Z) &= \mathbb{E}[p(U)\log p(U)] - \mathbb{E}[p(U)]\log\mathbb{E}[p(U)] \\ &= \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} p(x)\log p(x) - \frac{1}{|\mathcal{X}|}\log\frac{1}{|\mathcal{X}|} = \frac{1}{|\mathcal{X}|}\{\log(|\mathcal{X}|) - H(X)\}.\end{aligned}$$

(b) Since by Jensen's inequality, $\mathbb{H}(Z) \geq 0$, we have that $H(X) \leq \log(|\mathcal{X}|) = H(U)$.

(c) The Kullback–Leibler divergence between two mass functions p and q over $\mathcal{X}$ is given by $D(p \parallel q) \triangleq \sum_{x\in\mathcal{X}} p(x)\log(p(x)/q(x))$. Let $(p/q)(X) \triangleq \frac{p(X)}{q(X)}$, where X has the distribution q. Then, $\mathbb{E}[(p/q)(X)] = \sum_{x\in\mathcal{X}} q(x)\cdot\frac{p(x)}{q(x)} = 1$ and so

$$\mathbb{H}((p/q)(X)) = \sum_{x\in\mathcal{X}} q(x)\cdot\frac{p(x)}{q(x)}\cdot\log\frac{p(x)}{q(x)} = D(p \parallel q).$$

Exercise 3.2 (Chain rule and Kullback–Leibler divergence)

Exercise 3.3 (Variational representation for entropy) For $\psi(u) = e^{-u} - 1 + u$, denoting $h(t) \triangleq \mathbb{E}[\psi(\lambda(X-t))e^{\lambda X}] = e^{\lambda t} - \lambda t\mathbb{E}[e^{\lambda X}] + \mathbb{E}[(\lambda X - 1)e^{\lambda X}]$, we set $h(t^*) = \lambda(e^{\lambda t - \mathbb{E}[e^{\lambda X}]}) = 0$ to get the local minimum $t^* = \lambda^{-1}\log\mathbb{E}[e^{\lambda X}]$. Further, since $h''(t) = \lambda^2 e^{\lambda t} \geq 0$, the function h is convex, whence t^* is the (attained) global minimum, and so

$$\inf_{t\in\mathbb{R}} \mathbb{E}[\psi(\lambda(X-t))e^{\lambda X}] = h(t^*) = \mathbb{E}[e^{\lambda X}] - \mathbb{E}[e^{\lambda X}]\log\mathbb{E}[e^{\lambda X}] + \lambda\mathbb{E}[Xe^{\lambda X}] - \mathbb{E}[e^{\lambda X}] = \mathbb{H}[e^{\lambda X}].$$

Exercise 3.4 (Entropy and constant shifts)

(a) For a random variable X and a constant $c \in \mathbb{R}$, for any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{H}(e^{\lambda(X+c)}) &= \lambda\mathbb{E}[(X+c)e^{\lambda(X+c)}] - \mathbb{E}[e^{\lambda(X+c)}]\ln(\mathbb{E}[e^{\lambda(X+c)}]) \\ &= \lambda e^{\lambda c}\mathbb{E}[Xe^{\lambda X}] + \lambda c e^{\lambda c}\mathbb{E}[e^{\lambda X}] - e^{\lambda c}\mathbb{E}[e^{\lambda X}](\lambda c + \ln(\mathbb{E}[e^{\lambda X}])) \\ &= e^{\lambda c}(\lambda\mathbb{E}[Xe^{\lambda X}] - \mathbb{E}[e^{\lambda X}\ln(\mathbb{E}[e^{\lambda X}])]) = e^{\lambda c}\mathbb{H}(e^{\lambda X}).\end{aligned}$$

(b) $\mathbb{H}(e^{\lambda X}) \le \frac{1}{2}\lambda^2\sigma^2\mathbb{E}[e^{\lambda X}] \Leftrightarrow \mathbb{H}(e^{\lambda(X+c)}) \overset{(a)}{=} e^{\lambda c}\mathbb{H}(e^{\lambda X}) \le \frac{1}{2}\lambda^2\sigma^2\mathbb{E}[e^{\lambda(X+c)}]$ for any $c \in \mathbb{R}$.

Exercise 3.5 (Equivalent forms of entropy) For the convex function $\varphi(u) = u\ln u - u$, we can write

$$\mathbb{H}_\varphi(e^{\lambda X}) = \mathbb{E}[\varphi(e^{\lambda X})] - \varphi(\mathbb{E}[e^{\lambda X}]) = \mathbb{E}[\lambda X e^{\lambda X}] - \mathbb{E}[e^{\lambda X}] - \mathbb{E}[e^{\lambda X}]\ln(\mathbb{E}[e^{\lambda X}]) + \mathbb{E}[e^{\lambda X}] = \mathbb{H}(e^{\lambda X}).$$

Exercise 3.6 (Entropy rescaling)

(a) Since $\tilde{X} = X - \mathbb{E}[X]$, we have from Exercise 3.4(a), that $\mathbb{H}(e^{\lambda\tilde{X}}) = e^{-\lambda\mathbb{E}[X]}\mathbb{H}(e^{\lambda X})$, $\varphi_{\tilde{\mathrm{x}}}(\lambda) = e^{-\lambda\mathbb{E}[X]}\varphi_{\mathrm{x}}(\lambda)$ and $\varphi'_{\tilde{\mathrm{x}}}(\lambda) = e^{-\lambda\mathbb{E}[X]}(\varphi'_{\mathrm{x}}(\lambda) - \mathbb{E}[X]\varphi_{\mathrm{x}}(\lambda))$. Therefore,

$$\begin{aligned}&\mathbb{H}(e^{\lambda X}) \le \lambda^2(b\varphi'_{\mathrm{x}}(\lambda) + \varphi_{\mathrm{x}}(\lambda)(\sigma^2 - b\mathbb{E}[X])) \quad \text{for all } \lambda \in [0, 1/b) \\ &\Leftrightarrow e^{-\lambda\mathbb{E}[X]}\mathbb{H}(e^{\lambda X}) \le \lambda^2(b(e^{-\lambda\mathbb{E}[X]}(\varphi'_{\mathrm{x}}(\lambda) - \mathbb{E}[X]\varphi_{\mathrm{x}}(\lambda)) + e^{-\lambda\mathbb{E}[X]}\varphi_{\mathrm{x}}(\lambda)\sigma^2)) \quad \text{for all } \lambda \in [0, 1/b) \\ &\Leftrightarrow \mathbb{H}(e^{\lambda\tilde{X}}) \le \lambda^2(b\varphi'_{\tilde{\mathrm{x}}}(\lambda) + \varphi_{\tilde{\mathrm{x}}}(\lambda)\sigma^2) \quad \text{for all } \lambda \in [0, 1/b).\end{aligned}$$

(b) Let X be a zero-mean random variable and let $\tilde{X} \triangleq X/b$. Then,

$$\begin{aligned}&\mathbb{H}(e^{\lambda X}) \le \lambda^2(b\varphi'_{\mathrm{x}}(\lambda) + \varphi_{\mathrm{x}}(\lambda)\sigma^2) \quad \text{for all } \lambda \in [0, 1/b) \\ &\overset{(\lambda=\tilde{\lambda}/b)}{\Leftrightarrow} \mathbb{H}(e^{\tilde{\lambda}(X/b)}) \le (\tilde{\lambda}^2/b^2)(b\mathbb{E}[Xe^{\tilde{\lambda}(X/b)}] + \mathbb{E}[e^{\tilde{\lambda}(X/b)}]\sigma^2) \quad \text{for all } \tilde{\lambda} \triangleq \lambda b \in [0, 1) \\ &\Leftrightarrow \mathbb{H}(e^{\tilde{\lambda}\tilde{X}}) \le \tilde{\lambda}^2(\varphi'_{\tilde{\mathrm{x}}}(\tilde{\lambda}) + \varphi_{\tilde{\mathrm{x}}}(\tilde{\lambda})\tilde{\sigma}^2) \quad \text{for all } \tilde{\lambda} \in [0, 1),\end{aligned}$$

where $\tilde{\sigma}^2 \triangleq \sigma^2/b^2$.

Exercise 3.7 (Entropy for bounded variables) From Exercise 3.3 (Variational representation for entropy), $\mathbb{H}(e^{\lambda X}) \le \mathbb{E}[\psi(\lambda(X-t))e^{\lambda X}]$ for all $t \in \mathbb{R}$, where $\psi(u) = e^{-u} - 1 + u$. Since $X \in [a, b]$, and ψ is convex, the maximum is attained at the extremal points, i.e., $\psi(\lambda(X-t)) \le \max\{\psi(\lambda(a-t)), \psi(\lambda(b-t))\}$. Now, choose t_0 such that

$$\psi(\lambda(a-t_0)) = \psi(\lambda(b-t_0)) \Leftrightarrow e^{\lambda t_0}(e^{-\lambda a} - e^{-\lambda b}) = \lambda(b-a) \Leftrightarrow e^{-\lambda(b-t_0)} = \frac{\lambda(b-a)}{e^{\lambda(b-a)} - 1} = \frac{2\sigma\lambda}{e^{2\sigma\lambda} - 1}.$$

where $\sigma \triangleq \frac{(b-a)}{2} > 0$. We have that $\psi(\lambda(X - t_0)) \leq \psi(\lambda(b - t_0)) = \frac{2\sigma\lambda}{e^{2\sigma\lambda}-1} - \ln\left(\frac{2\sigma\lambda}{e^{2\sigma\lambda}-1}\right) - 1 \triangleq u_\sigma(\lambda)$. We need to show that $\frac{\lambda^2\sigma^2}{2}$ is an upper bound on $u_\sigma(\lambda)$. To this end, we show that $f_\sigma(\lambda) \triangleq \frac{\lambda^2\sigma^2}{2} - u_\sigma(\lambda) = \frac{\lambda^2\sigma^2}{2} - \frac{2\sigma\lambda}{e^{2\sigma\lambda}-1} + \ln\left(\frac{2\sigma\lambda}{e^{2\sigma\lambda}-1}\right) + 1 \geq 0$ for any $\lambda \in \mathbb{R}$. First, we compute its first derivative as

$$\begin{aligned}
f'_\sigma(\lambda) &= \lambda\sigma^2 - \frac{(e^{2\sigma\lambda}-1)2\sigma - 4\sigma^2\lambda e^{2\sigma\lambda}}{(e^{2\sigma\lambda}-1)^2} + \frac{1}{\lambda} - \frac{2\sigma e^{2\sigma\lambda}}{e^{2\sigma\lambda}-1} \\
&= \frac{(\lambda^2\sigma^2+1)(e^{2\sigma\lambda}-1)^2 - 2\sigma\lambda e^{2\sigma\lambda}(e^{2\sigma\lambda}-1) - (e^{2\sigma\lambda}-1)2\sigma\lambda + 4\lambda^2\sigma^2 e^{2\sigma\lambda}}{\lambda(e^{2\sigma\lambda}-1)^2} \\
&= \frac{e^{4\sigma\lambda}(\lambda^2\sigma^2+1-2\sigma\lambda) + 2e^{2\sigma\lambda}(2\lambda^2\sigma^2-1-\lambda^2\sigma^2) + (\lambda^2\sigma^2+1+2\sigma\lambda)}{\lambda(e^{2\sigma\lambda}-1)^2} \\
&= \frac{(e^{2\sigma\lambda}(\lambda\sigma-1)+(\lambda\sigma+1))^2}{\lambda(e^{2\sigma\lambda}-1)^2}.
\end{aligned}$$

Hence, $f_\sigma(\cdot)$ is increasing for $\lambda > 0$ and decreasing for $\lambda < 0$. Thus, for any $\lambda \in \mathbb{R}$, $f_\sigma(\lambda) \geq f_\sigma(0) \triangleq \lim_{\lambda' \to 0} f_\sigma(\lambda') = 0$, and so $\mathbb{H}(e^{\lambda X}) \leq \mathbb{E}[\psi(\lambda(X - t_0))e^{\lambda X}] \leq u_\sigma(\lambda)\mathbb{E}[e^{\lambda X}] \leq \frac{\lambda^2\sigma^2}{2}\varphi_{\mathrm{x}}(\lambda)$.

Exercise 3.8 (Exponential families and entropy) Consider a random variable $Y \in \mathcal{Y}$ having density $p_\theta(y) = h(y)e^{\langle\theta, T(y)\rangle - \Phi(\theta)}$ with respect to a base measure μ, where $\Phi(\theta) = \log \int_{\mathcal{Y}} \exp(\langle\theta, T(y)\rangle)h(y)\mu(dy)$ for $\theta \in \mathbb{R}^d$ and $T : \mathcal{Y} \to \mathbb{R}^d$ defines a vector of sufficient statistics. Assume that $\|\nabla\Phi(\theta) - \Phi(\theta')\|_2 \leq \|\theta - \theta'\|_2$ for all $\theta, \theta' \in \mathbb{R}^d$.

(a) For a fixed unit-norm vector $v \in \mathbb{R}^d$, consider the random variable $X = \langle v, T(Y)\rangle$. First, since $\exp(\Phi(\theta)) = \int_{\mathcal{Y}} \exp(\langle\theta, T(y)\rangle)h(y)\mu(dy)$, we have $\nabla\Phi(\theta) = e^{-\Phi(\theta)} \int_{y\in\mathcal{Y}} T(y)e^{\langle\theta,T(y)\rangle}\mu(dy) = \mathbb{E}_{Y\sim p_\theta}[T(Y)]$. Using Jensen's inequality to convex function $-\ln(\cdot)$, we have that

$$-\ln \mathbb{E}[e^{\lambda X}] \leq -\lambda\mathbb{E}[X] = -\lambda\langle v, \mathbb{E}_{Y\sim p_\theta}[T(Y)]\rangle = -\lambda\langle v, \nabla\Phi(\theta)\rangle.$$

We can now compute

$$\begin{aligned}
\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = \mathbb{E}_{Y\sim p_\theta}\left[\left\langle v, \frac{T(Y)e^{\lambda\langle v,T(Y)\rangle}}{\mathbb{E}_{Y\sim p_\theta}[e^{\lambda\langle v,T(Y)\rangle}]}\right\rangle\right] &= \left\langle v, \frac{\int_{y\in\mathcal{Y}} T(y)h(y)e^{\langle\lambda v+\theta,T(y)\rangle}\mu(dy)}{\int_{y\in\mathcal{Y}} h(y)e^{\langle\lambda v+\theta,T(y)\rangle}\mu(dy)}\right\rangle \\
&= \langle v, \mathbb{E}_{Y\sim p_{\theta+\lambda v}}[T(Y)]\rangle = \langle v, \nabla\Phi(\theta+\lambda v)\rangle.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\frac{\mathbb{H}(e^{\lambda X})}{\mathbb{E}[e^{\lambda X}]} = \frac{\lambda\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \ln\mathbb{E}[e^{\lambda X}] &\leq \lambda\langle v, \nabla\Phi(\theta+\lambda v) - \Phi(\theta)\rangle \\
&\overset{(a)}{\leq} \lambda\|v\|_2\|\nabla\Phi(\theta+\lambda v) - \nabla\Phi(\theta)\|_2 \overset{(b)}{\leq} \lambda^2 L\|v\|_2^2 = L\lambda^2,
\end{aligned}$$

where (a) uses the Cauchy-Schwartz inequality, (b) uses the Lipschitz condition on the gradient of Φ, and the last equality uses $\|v\|_2 = 1$. Hence, by Proposition 3.2., X is sub-Gaussian with parameter $\sqrt{2L}$.

(b)

Exercise 3.9 (Another variational representation) We need to show that

$$\mathbb{H}(e^{\lambda f(X)}) = \sup_g \{\mathbb{E}[g(X)e^{\lambda f(X)}] \mid \mathbb{E}[e^{g(X)}] \leq 1\}.$$

For a function $h(u)$, denote $h^*(v) = \sup_u\{uv - h(u)\}$ to be its conjugate dual; thus $h(u) + h^*(v) \geq uv$ for any $u \in \mathrm{dom}(h), v \in \mathrm{dom}(h^*)$ by definition. In particular, for $h(u) = e^u$ with $\mathrm{dom}(h) = \mathbb{R}$, we have that $v = h'(u^*) = e^{u^*}$, whence $h^*(v) = v \ln v - v = \varphi(v)$, with $\mathrm{dom}(h^*) = (0, \infty)$. Hence for a random variable X, denoting $u_X = g(X) + \ln \mathbb{E}[e^{\lambda f(X)}], v_X = e^{\lambda f(X)}$, we have using the conjugate inequality that

$$(g(X) + \ln \mathbb{E}[e^{\lambda f(X)}])e^{\lambda f(X)} = u_X v_X \leq h(u_X) + h^*(v_X) = \mathbb{E}[e^{\lambda f(X)}]e^{g(X)} + \varphi(e^{\lambda f(X)}). \tag{1}$$

The conjugate inequality satisfies with equality if $v_X^* = h'(u_X^*) = e^{u_X^*}$. Thus, taking $g^*(X) \triangleq \lambda f(X) - \ln \mathbb{E}[e^{\lambda f(X)}]$, one can verify that $\mathbb{E}[e^{g^*(X)}] = 1$ and $\mathbb{E}[g^*(X)e^{\lambda f(X)}] = \mathbb{H}(e^{\lambda f(X)})$. Therefore, $\mathbb{H}(e^{\lambda f(X)}) \leq \sup_g\{\mathbb{E}[g(X)e^{\lambda f(X)}] \mid \mathbb{E}[e^{g(X)}] \leq 1\}$.

Taking expectations for (1), we get $\mathbb{E}[g(X)e^{\lambda X}] + \mathbb{E}[e^{\lambda X}] \ln \mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{g(X)}]\mathbb{E}[e^{\lambda f(X)}] + \mathbb{E}[\varphi(e^{\lambda f(X)})]$. Finally, using $\mathbb{E}[e^{g(X)}] \leq 1$, one obtains $\mathbb{E}[g(X)e^{\lambda X}] \leq \mathbb{E}[\varphi(e^{\lambda f(X)})] - \varphi(\mathbb{E}[e^{\lambda f(X)}]) = \mathbb{H}_\varphi(e^{\lambda f(X)}) = \mathbb{H}(e^{\lambda f(X)})$, whence $\sup_g\{\mathbb{E}[g(X)e^{\lambda f(X)}] \mid \mathbb{E}[e^{g(X)}] \leq 1\} \leq \mathbb{H}(e^{\lambda f(X)})$.

Exercise 3.10 (Brunn–Minkowski and classical isoperimetric inequality) Consider the Brunn-Minkowski inequality: for any two convex bodies $C, D \subset \mathbb{R}^n$, and $\lambda \in [0, 1]$,

$$[\mathrm{vol}(\lambda C + (1 - \lambda)D)]^{1/n} \geq \lambda[\mathrm{vol}(C)]^{1/n} + (1 - \lambda)[\mathrm{vol}(D)]^{1/n} \tag{2}$$

and also,

$$\mathrm{vol}(\lambda C + (1 - \lambda)D) \geq [\mathrm{vol}(C)]^{\lambda}[\mathrm{vol}(D)]^{1-\lambda} \tag{3}$$

Consider the inequality: for any two convex bodies $A, B \subset \mathbb{R}^n$

$$[\mathrm{vol}(A + B)]^{1/n} \geq [\mathrm{vol}(A)]^{1/n} + [\mathrm{vol}(B)]^{1/n} \tag{4}$$

Note that

$$\text{vol}(\alpha S) = \alpha^n \text{vol}(S), \quad S \subset \mathbb{R}^n, \alpha > 0 \tag{5}$$

(a) Using (2) with $C = A, D = B, \lambda = 1/2$ gives (4), and using (4) with $A = \lambda C, B = (1-\lambda)D$ gives (2), where we used (5).

(b) Using (2), and the concavity of $\log(\cdot)$, one has

$$\begin{aligned}\frac{1}{n}\log(\text{vol}(\lambda C + (1-\lambda)D)) &\geq \log(\lambda[\text{vol}(C)]^{1/n} + (1-\lambda)[\text{vol}(D)]^{1/n}) \\ &\geq \lambda \log([\text{vol}(C)]^{1/n}) + (1-\lambda)\log([\text{vol}(D)]^{1/n}) \\ &= \frac{1}{n}\log([\text{vol}(C)]^{\lambda}[\text{vol}(D)]^{1-\lambda}),\end{aligned}$$

which is the required inequality (3).

(c) For convex bodies, $A, B \subset \mathbb{R}^n$, let $\alpha \triangleq \text{vol}(A), \beta \triangleq \text{vol}(B)$. Now, choosing $C = \frac{1}{\alpha^{1/n}}A, D = \frac{1}{\beta^{1/n}}B$ and $\lambda = \frac{\alpha^{1/n}}{\alpha^{1/n}+\beta^{1/n}}$, we get that $\lambda C = \frac{1}{\alpha^{1/n}+\beta^{1/n}}A$ and $(1-\lambda)D = \frac{1}{\alpha^{1/n}+\beta^{1/n}}B$, and $\text{vol}(C) = \frac{\text{vol}(A)}{\alpha} = 1$, and $\text{vol}(D) = \frac{\text{vol}(B)}{\beta} = 1$. Thus, using (3) with the above parameters, we obtain using (5) that

$$\text{vol}\left(\frac{1}{\alpha^{1/n}+\beta^{1/n}}(A+B)\right) \geq 1 \Leftrightarrow \text{vol}(A+B) \geq ([\text{vol}(A)]^{1/n} + [\text{vol}(B)]^{1/n})^n,$$

which is the inequality (4), which is equivalent to (2) by part (a).

Exercise 3.11 (Concentration on the Euclidean ball) Let $\mathbb{P}$ denote the uniform measure over the Euclidean unit ball $\mathbb{B}_2^n \triangleq \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$.

(a) For any $0 < \epsilon \leq 2$, and for any subset $A \subset \mathbb{B}_2^n$, for $a \in A$ and $b \in \mathbb{B}_2^n \setminus A^\epsilon$, we have that $\|a\|_2 \leq 1, \|b\|_2 \leq 1$, but $\|b-a\|_2 \geq \epsilon$. Thus, we get $2\langle a, b\rangle \leq \|a\|_2^2 + \|b\|_2^2 - \epsilon^2 \leq 2 - \epsilon^2$, giving $\|a+b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 + 2\langle a, b\rangle \leq 4(1-\epsilon^2/4)$, and so $(1/2)\|a+b\|_2 \leq (1-\epsilon^2/4)^{1/2} \leq 1 - \epsilon^2/8$, where the last inequality follows from the fact that $(1-x)^{1/2} \leq 1 - x/2$ for any $0 < x \leq 1$. To see this, $f(x) \triangleq (1-x/2) - (1-x)^{1/2}$. Then, $f(0) = 0$ and $f'(x) = (1/2)(1/(1-x)^{1/2} - 1) > 0$ for $0 < x < 1$, hence, $f(x) \geq f(0) = 0$ for $0 < x < 1$. Thus, $\frac{1}{2}(A + (\mathbb{B}_2^n \setminus A^\epsilon)) \subseteq \left(1 - \frac{\epsilon^2}{8}\right)\mathbb{B}_2^n$.

(b) Using the definition of $\mathbb{P}$, the Brunn-Minkowski inequality (3) with $C = A, D = \mathbb{B}_2^n \setminus A^\epsilon$ and

$\lambda = 1/2$, and part (a), we get

$$\mathbb{P}[A](1 - \mathbb{P}[A^\epsilon]) = \text{vol}(C)\text{vol}(D) \leq \left(\text{vol}\left(\frac{1}{2}(C+D)\right)\right)^2$$
$$\leq \left(\text{vol}\left((1-\epsilon^2/8)\mathbb{B}_2^n\right)\right)^2 = \left(1 - \frac{\epsilon^2}{8}\right)^{2n}.$$

(c) For any A with $\mathbb{P}[A] \geq 1/2$, using part (b), we get $1 - \mathbb{P}[A^\epsilon] \leq 2(1-\epsilon^2/8)^{2n} \leq 2e^{-n\epsilon^2/4}$, where the last inequality uses the fact that $1 + t \leq e^t$ for $t \in \mathbb{R}$ with $t = -\epsilon^2/8$. Hence, for $0 < \epsilon < 2$, $\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) \leq 2e^{-n\epsilon^2/4}$ for $\mathcal{X} = \mathbb{B}_2^n, \rho(\cdot,\star) = \|\cdot - \star\|_2$.

Exercise 3.12 (Rademacher chaos variables) For a symmetric positive definite matrix (PSD) $\mathbf{Q} \in \mathcal{S}_+^{d\times d}$, define the Rademacher chaos variable, $X \triangleq \sum_{i,j=1}^d Q_{ij}\varepsilon_i\varepsilon_j = \varepsilon^T\mathbf{Q}\varepsilon$. Clearly, $\mathbb{E}[X] = \sum_{i=1}^d Q_{ii} + \sum_{i\neq j} Q_{ij}\mathbb{E}[\varepsilon_i]\mathbb{E}[\varepsilon_j] = \text{trace}(\mathbf{Q})$.

(a) Let $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be the eigen decomposition of symmetric PSD $\mathbf{Q}$, denote the matrix $\mathbf{Q}^{1/2} = \mathbf{\Lambda}^{1/2}\mathbf{U}^T$. The function $f(\mathbf{v}) = \sqrt{\mathbf{v}^T\mathbf{Q}\mathbf{v}} = \|\mathbf{Q}^{1/2}\mathbf{v}\|_2$, is clearly convex. Further, for any $\mathbf{v}_1 \neq \mathbf{v}_2$, using the reverse triangle inequality, $|f(\mathbf{v}_1) - f(\mathbf{v}_2)| = |\|\mathbf{Q}^{1/2}\mathbf{v}_1\|_2 - \|\mathbf{Q}^{1/2}\mathbf{v}_2\|_2| \leq \|\mathbf{Q}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 = \sqrt{(\mathbf{v}_1-\mathbf{v}_2)^T\mathbf{Q}(\mathbf{v}_1-\mathbf{v}_2)} \leq |||\mathbf{Q}|||_2{}^{1/2}\|\mathbf{v}_1 - \mathbf{v}_2\|_2$, and so f is Lipschitz with respect to the Euclidean norm, with parameter $L = |||\mathbf{Q}|||_2^{1/2}$. Noting that $f(\varepsilon) = \sqrt{X}$, using Jensen's inequality to the concave function $\sqrt{(\cdot)}$, we obtain that $\mathbb{E}[f(\epsilon)] = \mathbb{E}\left[\sqrt{X}\right] \leq \sqrt{\mathbb{E}[X]} = \sqrt{\text{trace}(\mathbf{Q})}$. Hence, since $\varepsilon_i \in [-1,+1]$ for each $i \in [d]$, we have using Theorem 3.4 that for any $t > 0$:

$$\mathbb{P}[X \geq (\sqrt{\text{trace}(\mathbf{Q})} + t)^2] = \mathbb{P}[f(\varepsilon) \geq \text{trace}(\mathbf{Q}) + t] \leq \mathbb{P}[f(\varepsilon) \geq \mathbb{E}[f(\varepsilon)] + t] \leq \exp\left(-\frac{t^2}{16|||\mathbf{Q}|||_2}\right).$$

(b) For an arbitrary symmetric matrix $\mathbf{M} \in \mathcal{S}^{d\times d}$, consider the decoupled Rademacher chaos $Y \triangleq \sum_{i,j=1}^d M_{ij}\varepsilon_i\varepsilon'_j = \varepsilon^T\mathbf{M}\varepsilon'$, where $\varepsilon, \varepsilon'$ are i.i.d. random vectors with i.i.d. Rademacher random entries. Conditioned on ε, $Y = \sum_{i=1}^d (\mathbf{M}\varepsilon)_i\varepsilon'_i$, which is a linear combination of the zero-mean independent sub-Gaussian random variables ε'_i with parameter 1, is a zero-mean sub-Gaussian variate with parameter $\sigma = \sqrt{\sum_{i=1}^d (\mathbf{M}\varepsilon)_i^2} = \|\mathbf{M}\varepsilon\|_2$. Therefore, by the sub-Gaussian tail bound, for any $\delta > 0$:

$$\mathbb{P}[Y \geq \delta \mid \varepsilon] \leq \exp\left(-\frac{\delta^2}{2\|\mathbf{M}\varepsilon\|_2^2}\right).$$

Now, we need to control $\|\mathbf{M}\varepsilon\|_2^2 = \varepsilon^T\mathbf{M}^T\mathbf{M}\varepsilon$. Using part (a), along with $|||\mathbf{M}^T\mathbf{M}|||_2 = |||\mathbf{M}|||_2^2$ and

trace($\mathbf{M}^T\mathbf{M}$) = $|||\mathbf{M}|||_F^2$, we have

$$\mathbb{P}[|||\mathbf{M}\boldsymbol{\varepsilon}||_2^2 \geq (|||\mathbf{M}|||_F + t)^2] \leq \exp\left(-\frac{t^2}{16|||\mathbf{M}|||_2^2}\right).$$

Since $(|||\mathbf{M}|||_F + t)^2 \leq 2|||\mathbf{M}|||_F^2 + 2t^2$ (uses AM-GM), using $t^2 = \delta|||\mathbf{M}|||_2$, we obtain

$$\mathbb{P}[|||\mathbf{M}\boldsymbol{\varepsilon}||_2^2 \geq 2|||\mathbf{M}|||_F^2 + 2\delta|||\mathbf{M}|||_2] \leq \mathbb{P}[|||\mathbf{M}\boldsymbol{\varepsilon}||_2^2 \geq (|||\mathbf{M}|||_F + \delta|||\mathbf{M}|||_2)^2] \leq \exp\left(-\frac{\delta}{16|||\mathbf{M}|||_2}\right).$$

Finally, since $\{Y \geq \delta\} \subset \{|||\mathbf{M}\boldsymbol{\varepsilon}||_2^2 \leq 2|||\mathbf{M}|||_F^2 + 2\delta|||\mathbf{M}|||_2, Y \geq \delta\} \cap \{|||\mathbf{M}\boldsymbol{\varepsilon}||_2^2 \geq 2|||\mathbf{M}|||_F^2 + 2\delta|||\mathbf{M}|||_2\}$, we have that

$$\begin{aligned}\mathbb{P}[Y \geq \delta] &\leq \exp\left(-\frac{\delta^2}{4|||\mathbf{M}|||_F^2 + 4\delta|||\mathbf{M}|||_2}\right) + \exp\left(-\frac{\delta}{16|||\mathbf{M}|||_2}\right) \\ &\leq 2\exp\left(-\frac{\delta^2}{4|||\mathbf{M}|||_F^2 + 16\delta|||\mathbf{M}|||_2}\right).\end{aligned}$$

Exercise 3.16 (Different forms of functional Bernstein) Let Z be a non-negative random variable satisfying the Bernstein type tail bound as

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq \exp\left(-\frac{n\delta^2}{c_1\gamma^2 + c_2 b\delta}\right),$$

for some positive constants $c_1, c_2 > 0$.

(a) For any $t > 0$, equating $t = \frac{n\delta_t^2}{c_1\gamma^2 + c_2 b\delta_t}$ and solve $\delta_t^2 - \frac{c_2 bt}{n}\delta_t - \frac{c_1 t\gamma^2}{n} = 0$ to obtain the bigger root as $\delta_t = \frac{c_2 bt}{2n} + \sqrt{\left(\frac{c_2 bt}{2n}\right)^2 + \left(\gamma\sqrt{\frac{c_1 t}{n}}\right)^2} \leq \gamma\sqrt{\frac{c_1 t}{n}} + \frac{c_2 bt}{n}$, where we used $(u^2 + v^2) \leq (u + v)^2$ for $u, v \geq 0$. Therefore, we obtain for any $t > 0$:

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + \gamma\sqrt{\frac{c_1 t}{n}} + \frac{c_2 bt}{n}\right] \leq \mathbb{P}[Z \geq \mathbb{E}[Z] + \delta_t] \leq \exp\left(-\frac{n\delta_t^2}{c_1\gamma^2 + c_2 b\delta_t}\right) = e^{-t}.$$

(b) Since $\gamma^2 \leq \sigma^2 + c_3 b\mathbb{E}[Z]$, we get that for any $t > 0, \epsilon > 0$

$$\gamma \leq \sqrt{\sigma^2 + c_3 b\mathbb{E}[Z]} \leq \sigma + \sqrt{c_3 b\mathbb{E}[Z]} = \sigma + \sqrt{\frac{c_1 c_3 bt}{\epsilon n} \cdot \frac{\epsilon\mathbb{E}[Z]}{\frac{c_1 t}{n}}} \overset{\text{(AM–GM)}}{\leq} \sigma + \frac{c_3 b}{2\epsilon}\sqrt{\frac{c_1 t}{n}} + \frac{\epsilon\mathbb{E}[Z]}{\sqrt{\frac{c_1 t}{n}}},$$

and so, $\gamma\sqrt{\frac{c_1 t}{n}} \leq \epsilon\mathbb{E}[Z] + \sigma\sqrt{\frac{c_1 t}{n}} + \frac{c_1 c_3}{2\epsilon}\frac{bt}{n}$. Thus, using the part (a), we have for any $t > 0$:

$$\mathbb{P}\left[Z \geq (1 + \epsilon)\mathbb{E}[Z] + \sigma\sqrt{\frac{c_1 t}{n}} + \left(c_2 + \frac{c_1 c_3}{2\epsilon}\right)\frac{bt}{n}\right] \leq \mathbb{P}\left[Z \geq \mathbb{E}[Z] + \gamma\sqrt{\frac{c_1 t}{n}} + \frac{c_2 bt}{n}\right] \leq e^{-t}.$$

4 Uniform laws of large numbers

Exercises

Exercise 4.1 (Continuity of functionals) Let $\hat{F}_n$ denote the empirical CDF corresponding to n i.i.d. samples from a distribution $\mathbb{P}$ with the CDF, $F(t) = \mathbb{P}[X \le t]$. Let γ be a functional that is continuous with respect to the sup-norm, i.e., for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$, such that $\{\|G - F\|_\infty \le \delta_\epsilon\} \subset \{|\gamma(G) - \gamma(F)| \le \epsilon\}$.

(a) Using the continuity of γ, we have $\mathbb{P}[|\gamma(\hat{F}_n) - \gamma(F)| > \epsilon] \le \mathbb{P}[\|\hat{F}_n - F\|_\infty > \delta_\epsilon]$. By Theorem 4.4 (Glivenko-Cantelli), $\|\hat{F}_n - F\|_\infty \to 0$ almost surely, implying that $\|\hat{F}_n - F\|_\infty \to 0$ in probability, which in turn gives $\mathbb{P}[\|\hat{F}_n - F\|_\infty > \delta_\epsilon] \to 0$ as $n \to \infty$, or $\gamma(\hat{F}_n) \to \gamma(F)$ in probability.

(b) Let $\gamma(F) = \int x \, dF(x)$.

Exercise 4.2 (Failure of Glivenko–Cantelli) Let $\mathcal{S}$ be the collection of all finite subsets of $[0, 1]$ and let $X_i, i \in [n]$ be sample i.i.d. from a distribution over $[0, 1]$. For any finite n, the set $S_{X_1^n,\varepsilon} = \cup_{i=1}^n \{X_i \mid \varepsilon_i = 1\} \in \mathcal{S}$. Note that if $\varepsilon_i = 1$, then $\mathbb{I}_{S_{X_1^n,\varepsilon}}[X_i] = 1 = \mathbb{I}[\varepsilon_i = 1]$. If $\varepsilon_i \neq 1$, then $\mathbb{I}_{S_{X_1^n,\varepsilon}}[X_i] = 0 = \mathbb{I}[\varepsilon_i = 1]$.

$$\mathcal{R}_n(\mathcal{S}) = \mathbb{E}_{X,\varepsilon}\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{I}_S[X_i]\right|\right] \ge \mathbb{E}_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{I}_{S_{X_1^n,\varepsilon}}[X_i]\right|\right] = \mathbb{E}_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{I}[\varepsilon_i = 1]\right|\right]$$
$$= \mathbb{E}_\varepsilon\left[\left|\frac{1}{n}\sum_{i=1}^n \mathbb{I}[\varepsilon_i = 1]\right|\right] \overset{\text{(Jensen's)}}{\ge} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\varepsilon_i}[\mathbb{I}[\varepsilon_i = 1]]\right| = \frac{1}{n}\sum_{i=1}^n \mathbb{P}[\varepsilon_i = 1] = \frac{1}{2}.$$

Since $\mathcal{S}$ is 1-uniformly bounded, from Proposition 4.12, we have that $\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{S} \ge \frac{1}{2}\mathcal{R}_n(\mathcal{S}) - \delta \ge \frac{1}{4} - \delta$ with probability at least $1 - \exp(\frac{-n\delta^2}{2})$. Therefore, for any $0 < \delta < \frac{1}{4}$, $\mathbb{P}[\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{S} \ge \frac{1}{4} - \delta] \to 1$ as $n \to \infty$. Therefore, $\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{S} \nrightarrow 0$ in probability.

Exercise 4.4 (Details of symmetrization argument)

(a) For any $g \in \mathcal{G}$, since $g(X) \le |g(X)| \le \sup_{g\in\mathcal{G}} |g(X)|$. Taking expectations, we get $\mathbb{E}[g(X)] \le \mathbb{E}[\sup_{g\in\mathcal{G}} |g(X)|]$. Therefore, $\sup_{g\in\mathcal{G}} \mathbb{E}[g(X)] \le \mathbb{E}[\sup_{g\in\mathcal{G}} |g(X)|]$, and so (4.17) follows using $g(Y; X) = \frac{1}{n}\sum_{i=1}^n \{f(X_i) - f(Y_i)\}$ with $\mathcal{G} = \mathcal{F}$.

(b) For a convex non-decreasing function Φ, for any $g \in \mathcal{G}$, we have $\Phi(|g(X)|) \le \Phi(\sup_{g\in\mathcal{G}} |g(X)|)$. Taking expectations, using Jensen's inequality, and taking supremum over $g \in \mathcal{G}$ gives the result. Again, using $g(Y; X) = \frac{1}{n}\sum_{i=1}^n \{f(X_i) - f(Y_i)\}$ with $\mathcal{G} = \mathcal{F}$ gives the inequality in the proof of Proposition 4.11.

Exercise 4.5 (Necessity of vanishing Rademacher complexity)

(a) Let $\bar{\mathscr{F}} = \{f - \mathbb{E}[f] \mid f \in \mathscr{F}\}$. Then

$$\begin{aligned}\mathbb{E}_{X,\varepsilon}[\|\mathbb{S}_n\|_{\mathscr{F}}] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right] \\ &\le \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i (f(X_i) - \mathbb{E}[f(X_i)])\right|\right] + \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{E}[f(X_i)]\right|\right] \\ &\le \mathbb{E}_{X,\varepsilon}[\|\mathbb{S}_n\|_{\bar{\mathscr{F}}}] + \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}[f]|}{n}\cdot\mathbb{E}\left[\left|\sum_{i=1}^{n}\varepsilon_i\right|\right].\end{aligned}$$

But, using Jensen's inequality to the concave function $\sqrt{\cdot}$, we have

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}\varepsilon_i\right|\right] = \mathbb{E}\left[\sqrt{\left(\sum_{i=1}^{n}\varepsilon_i\right)^2}\right] \le \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{n}\varepsilon_i\right)^2\right]} = \sqrt{\mathbb{E}\left[\sum_{i=1}^{n}\varepsilon_i^2\right]} = \sqrt{n}. \tag{6}$$

Hence, $\mathbb{E}_{X,\varepsilon}[\|\mathbb{S}_n\|_{\mathscr{F}}] \le \mathbb{E}_{X,\varepsilon}[\|\mathbb{S}_n\|_{\bar{\mathscr{F}}}] + \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}[f]|}{\sqrt{n}}$.

(b) Using equation (4.21), and the concentration for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ around $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}]$ (Theorem 4.10 first part), we have that

$$\mathbb{P}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \ge \frac{1}{2}\mathbb{E}_{X,\varepsilon}[\|S_n\|_{\bar{\mathscr{F}}}] - \delta] \ge \mathbb{P}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \ge \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}] - \delta] \ge 1 - \exp\left(\frac{-n\delta^2}{2b^2}\right).$$

Thus, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \ge \frac{1}{2}\mathcal{R}_n(\mathscr{F}) - \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}[f]|}{2\sqrt{n}} - \delta$ with probability at least $1 - \exp\left(\frac{-n\delta^2}{2b^2}\right)$.

Exercise 4.6 (Too many linear classifiers) Let $\mathscr{F} = \{x \mapsto \text{sign}(\langle\theta, x\rangle) \mid \theta \in \mathbb{R}^d, \|\theta\|_2 = 1\}$ corresponding to the $\{-1, +1\}$-valued classification rules defined by linear functions in $\mathbb{R}^d$ where we define $\text{sign}(0) \triangleq +1$. Let $d \ge n$ and let $\{x_1, \ldots, x_n\}$ be a collection of n linearly independent vectors (and hence, all are non-zero). Denoting $\mathbf{X} \triangleq [x_1, \ldots, x_n]^T \in \mathbb{R}^{n\times d}$, for a given $\varepsilon \in \mathbb{R}^n$, since $n \le d$, there exists a $v_\varepsilon \in \mathbb{R}^d$ such that $\mathbf{X}v_\varepsilon = \varepsilon$. Since $\varepsilon \ne 0$, $v_\varepsilon \ne 0$ and so $\theta_\varepsilon \triangleq v_\varepsilon/\|\varepsilon\|_2$ has unit norm with $\langle\theta_\varepsilon, x_i\rangle = \varepsilon_i$. Hence, noting that $\text{sign}(\langle\theta_\varepsilon, x_i\rangle) = \text{sign}(\varepsilon_i) = \varepsilon_i$, we have

$$\mathcal{R}(\mathscr{F}(x_1^n)/n) = \mathbb{E}_\varepsilon\left[\sup_{\theta\in\mathbb{S}^d}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\text{sign}(\langle\theta, x_i\rangle)\right|\right] \ge \mathbb{E}_\varepsilon\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\text{sign}(\langle\theta_\varepsilon, x_i\rangle)\right|\right] = \mathbb{E}_\varepsilon\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right|\right] = 1.$$

Clearly, $\mathcal{R}(\mathscr{F}(x_1^n)/n) \le 1$, and so we have $\mathcal{R}(\mathscr{F}(x_1^n)/n) = 1$. This states that in the high-dimensional setting, $d \ge n$, i.e., the function class is larger than the sample size, since $\mathcal{R}(\mathscr{F}(x_1^n)/n) \ne o_n(1)$, the empirical risk would not converge to the true risk, and therefore, the model (function class) would be over-fitted to the data.

Exercise 4.7 (Basic properties of Rademacher complexity)

(a) Since $\mathscr{F} \subseteq \text{conv}(\mathscr{F})$, we have that $\mathcal{R}_n(\mathscr{F}) \leq \mathcal{R}_n(\text{conv}(\mathscr{F}))$. Now, for any $\delta > 0$, the definition of the supremum, there must be a $f_\varepsilon \in \text{conv}(\mathscr{F})$, or $f_\varepsilon^1, \ldots, f_\varepsilon^k \in \mathscr{F}$ with $f_\varepsilon = \sum_{j=1}^k \lambda_j f_\varepsilon^j$, with $\lambda_j \geq 0$ and $\sum_{j=1}^k \lambda_j = 1$, such that

$$\mathcal{R}_n(\text{conv}(\mathscr{F})) = \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\text{conv}(\mathscr{F})}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \leq \mathbb{E}_{X,\varepsilon}\left[\left|\sum_{j=1}^k \lambda_j \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_\varepsilon^j(X_i)\right)\right|\right] + \delta$$
$$\leq \mathbb{E}_{X,\varepsilon}\left[\underbrace{\left(\sum_{j=1}^k \lambda_j\right)}_{=1}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \delta = \mathcal{R}_n(\mathscr{F}) + \delta.$$

Since δ was arbitrary, $\mathcal{R}_n(\text{conv}(\mathscr{F})) \leq \mathcal{R}_n(\mathscr{F})$, whence $\mathcal{R}_n(\text{conv}(\mathscr{F})) = \mathcal{R}_n(\mathscr{F})$.

(b) Noting that $\mathscr{F} + \mathscr{G} = \{f + g \mid f \in \mathscr{F}, g \in \mathscr{G}\}$, for any $\delta > 0$, there is an $f_\varepsilon \in \mathscr{F}, g_\varepsilon \in \mathscr{G}$ such that:

$$\mathcal{R}_n(\mathscr{F} + \mathscr{G}) = \mathbb{E}_{X,\varepsilon}\left[\sup_{h\in\mathscr{F}+\mathscr{G}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(X_i)\right|\right] \leq \mathbb{E}_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_\varepsilon(X_i)\right|\right] + \mathbb{E}_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g_\varepsilon(X_i)\right|\right] + \delta$$
$$\leq \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \mathbb{E}_{X,\varepsilon}\left[\sup_{g\in\mathscr{G}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\right|\right] + \delta = \mathcal{R}_n(\mathscr{F}) + \mathcal{R}_n(\mathscr{G}) + \delta.$$

Since δ was arbitrary, $\mathcal{R}_n(\mathscr{F} + \mathscr{G}) \leq \mathcal{R}_n(\mathscr{F}) + \mathcal{R}_n(\mathscr{G})$.

(c) If g was uniformly bounded then, $\mathcal{R}_n(\{g\}) = \frac{1}{n}\mathbb{E}_{X,\varepsilon}\left[\left|\sum_{i=1}^n \varepsilon_i g(X_i)\right|\right] \leq \frac{\|g\|_\infty}{n}\mathbb{E}_\varepsilon\left[\left|\sum_{i=1}^n \varepsilon_i\right|\right] \leq \frac{\|g\|_\infty}{\sqrt{n}}$, where the last step follows from (6). Hence, the result follows from part (b).

Exercise 4.9 (Proof of Lemma 4.14) For each $f \in \mathscr{F}$, $\sum_{i=1}^n \varepsilon_i f(x_i)$ is a zero-mean sub-Gaussian random variable with parameter $\sigma_f = \sqrt{\sum_{i=1}^n f^2(x_i)} \leq \sup_{f\in\mathscr{F}}\sqrt{\sum_{i=1}^n f^2(x_i)} \triangleq \sigma_{\mathscr{F}}$. Thus, using Exercise 2.12(b) (max for sub-Gaussian maxima), and polynomial discrimination of $\mathscr{F}(x_1^n)$, we have that

$$\mathcal{R}(\mathscr{F}(x_1^n)/n) \triangleq \mathbb{E}\left[\sup_{f\in\mathscr{F}(x_1^n)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] \leq \frac{1}{n} 2\sigma_{\mathscr{F}}\sqrt{\log(\text{card}(\mathscr{F}(x_1^n)))}$$
$$\leq 2\frac{\sup_{f\in\mathscr{F}}\sqrt{\sum_{i=1}^n f^2(x_i)}}{\sqrt{n}}\sqrt{\frac{\nu\log(n+1)}{n}} \triangleq 2D(x_1^n)\sqrt{\frac{\nu\log(n+1)}{n}},$$

where $D(x_1^n)$ is the ℓ_2 radius of the set $\mathscr{F}(x_1^n)/\sqrt{n}$.

5 Metric entropy and its uses

Exercises

Exercise 5.2 (Packing and covering)

(a) For any 2δ-packing $\mathbb{U}_{2\delta} \triangleq \{\theta^1, \ldots, \theta^M\} \subset \mathbb{T}$, we have $\rho(\theta^i, \theta^j) > 2\delta$. Denote the $\mathbb{B}(\theta, \delta, \rho) \triangleq \{\theta \in \mathbb{T} \mid \rho(\theta, \theta') \leq \delta\}$. Let $\mathbb{V}_\delta \triangleq \{\tilde{\theta}^1, \ldots, \tilde{\theta}^N\}$ be a δ-cover and so for each $i \in [M]$, $\theta^i \subset \mathbb{B}(\tilde{\theta}^j, \delta, \rho)$ for some j. For $i' \neq i$, $2\delta < \rho(\theta^i, \tilde{\theta}^{i'}) \leq \rho(\theta^i, \tilde{\theta}^j) + \rho(\theta^{i'}, \tilde{\theta}^j) \leq \delta + \rho(\theta^{i'}, \tilde{\theta}^j) \Leftrightarrow \rho(\theta^{i'}, \tilde{\theta}^j) > \delta$. Hence, if $\theta^i \in \mathbb{B}(\tilde{\theta}^j, \delta, \rho)$, then $\theta^{i'} \notin \mathbb{B}(\tilde{\theta}^j, \delta, \rho)$ for all $i' \neq i$. In other words, each ball $\mathbb{B}(\tilde{\theta}^j, \delta, \rho)$ contains at most one element of $\mathbb{U}_{2\delta}$. This implies that the number of elements in $\mathbb{U}_{2\delta}$ must be less than or equal to the number of balls (equivalently, the number of elements in $\mathbb{V}_\delta$), i.e., $|\mathbb{U}_{2\delta}| = M \leq N = |\mathbb{V}_\delta|$ for all packings and coverings respectively, whence $M(2\delta; \mathbb{T}, \rho) = \sup\{|\mathbb{U}_{2\delta}| \mid \mathbb{U}_{2\delta} \text{ is a } 2\delta \text{ packing for } \mathbb{T}\} \leq \inf\{|\mathbb{V}_\delta| \mid \mathbb{V}_\delta \text{ is } \delta \text{ cover for } \mathbb{T}\} = N(\delta; \mathbb{T}, \rho)$.

(b) Let $\mathbb{U}_\delta \triangleq \{\theta^1, \ldots, \theta^{M(\delta;\mathbb{T},\rho)}\} \subset \mathbb{T}$ be a (maximal) δ-packing. For any $\theta \in \mathbb{T} \setminus \mathbb{U}_\delta$, since $\mathbb{U}_\delta \cup \{\theta\}$ is not a δ-packing, we must have $\rho(\theta, \theta^j) \leq \delta$ for some j, and so, $\mathbb{U}_\delta$ is also a δ-cover for $\mathbb{T}$. Thus, $N(\delta; \mathbb{T}, \rho) \leq |\mathbb{U}_\delta| = M(\delta; \mathbb{T}, \rho)$.