Solutions to exercises in Chapter 2

Answer to 2.1: Proving parts (a)–(d) are exercises in algebraic manipulations. We have

\[ \sup_{A \subset \mathcal{X}} |P_1(A) - P_2(A)| = \sup_{A \subset \mathcal{X}} \int_A (p_1 - p_2). \]

It is clear that the final integral is maximized by taking \( A = \{ x \in \mathcal{X} : p_1(x) > p_2(x) \} \), though this is also equal to the negative of the integral over \( A^c = \{ x \in \mathcal{X} : p_1(x) \leq p_2(x) \} \), whence we have

\[ \|P_1 - P_2\|_{TV} = \int_{p_1 > p_2} (p_1 - p_2) = \int_{p_2 > p_1} (p_2 - p_1) = \frac{1}{2} \int_{p_1 > p_2} (p_1 - p_2) + \frac{1}{2} \int_{p_2 > p_1} (p_2 - p_1). \]

Thus we have

\[ \|P_1 - P_2\|_{TV} = \frac{1}{2} \int |p_1 - p_2| = \frac{1}{2} \sup_{\|f\|_{\infty} \leq 1} \int f(p_1 - p_2), \]

since the supremum is attained by setting \( f = 1 \) on the set \( \{ p_1 > p_2 \} \) and \(-1\) otherwise. Let \( a \lor b \) and \( a \land b \) be shorthand for \( \max\{a, b\} \) and \( \min\{a, b\} \), respectively. Noting that for \( a, b \in \mathbb{R} \) we have

\[ |a - b| = a \lor b - a \land b \quad \text{and} \quad a \lor b - a = (b - a)1_{(b > a)}, \]

we have

\[ \int_{p_1 > p_2} (p_1 - p_2) = \int (p_1 \lor p_2 - p_2) = \int p_1 \lor p_2 - 1, \]

and the inequality for \( p_1 \land p_2 \) is similar. This completes the argument for (a)–(d).

To see part (e), note that if at some \( x \in \mathcal{X} \) the functions \( f, g \) satisfy \( f(x) + g(x) > 1 \), decreasing \( f \) and \( g \) until \( f(x) + g(x) = 1 \) (while maintaining positivity of \( f, g \)) can only decrease the infimum. Thus we have \( f + g \equiv 1 \). Moreover, if \( p_1(x) > p_2(x) \), it is clear that for any values of \( f(x), g(x) \) with \( f(x) + g(x) = 1 \) we have

\[ f(x)p_1(x) + g(x)p_2(x) \geq 0 \cdot p_1(x) + 1 \cdot p_2(x). \]

Thus a pair \( f, g \) attaining the infimum can be defined as \( f(x) = 1, g(x) = 0 \) when \( p_1(x) < p_2(x) \) and \( f(x) = 0, g(x) = 1 \) otherwise. In particular, we have

\[ \inf_{f+g \geq 1, f, g \geq 0} \int fp_1 + \int gp_2 = \int_{p_1 < p_2} p_1 + \int_{p_2 < p_1} p_2 \]

\[ = \int p_1 \land p_2 = 1 - \|P_1 - P_2\|_{TV}, \]
where the final equality uses part (d).

\[ \square \]

**Answer to 2.3:** Use the answer to question 2.1, where part (e) implies the result. Part (e) also shows the value is attainable.

\[ \square \]

**Answer to 2.8:** If \( 0 \cdot f(x/0) = +\infty \) for \( t > 0 \), the result is trivial, so we assume that \( a_i > 0 \) for all \( i \). Now, defining \( \lambda_i = 1/n, A = \sum_{i=1}^{n} a_i \) and \( B = \sum_{i=1}^{n} b_i \), we have

\[
\frac{1}{n} Af(B/A) = \left( \sum_{i=1}^{n} \lambda_i a_i \right) f \left( \frac{\sum_{i=1}^{n} \lambda_i b_i}{\sum_{i=1}^{n} \lambda_i a_i} \right) \leq \sum_{i=1}^{n} \lambda_i a_i f \left( \frac{b_i}{a_i} \right) = \frac{1}{n} \sum_{i=1}^{n} a_i f \left( \frac{b_i}{a_i} \right),
\]

where we used the convexity of the perspective transform. Multiplying each side by \( n \) gives the result.

For the second result, note that if we define \( U = \int u(x)dx < \infty \) and \( p(x) = u(x)/U \), then \( p(x) \) is the density of a probability distribution. Then Jensen’s inequality applied to the convex function \( (u,v) \mapsto uf(v/u) \) implies

\[
\int a(x)u(x)dx f \left( \frac{\int b(x)u(x)dx}{\int a(x)u(x)dx} \right) = U\mathbb{E}_P[a(X)] f \left( \frac{\mathbb{E}_P[b(X)]}{\mathbb{E}_P[a(X)]} \right) \leq U\mathbb{E}_P \left[ a(X) f \left( \frac{b(X)}{a(X)} \right) \right]
\]

\[ = \int a(x)u(x)f \left( \frac{b(x)}{a(x)} \right) u(x)dx, \]

which is our desired result. (In the case of a finite measure \( \mu \), we define \( P(A) = \mu(A)/\mu(\mathcal{X}) \), and the result is identical.)

\[ \square \]

**Answer to 2.9:** Let the index sets \( I_1, \ldots, I_m \subset \{1, \ldots, n\} \) be such that \( B_j = \cup_{i \in I_j} A_i \) for each \( j \in \{1, \ldots, m\} \). Then defining \( a_i = P(A_i) \), \( b_i = Q(A_i) \), and \( a_{I_j} = \sum_{i \in I_j} a_i \) and \( b_{I_j} = \sum_{i \in I_j} b_j \), we have by the part (a) of Question 2.8 that

\[
D_f (P\|Q \mid g_1) = \sum_{i=1}^{n} b_i f \left( \frac{a_i}{b_i} \right) \geq \sum_{j=1}^{m} b_{I_j} f \left( \frac{a_{I_j}}{b_{I_j}} \right) = D_f (P\|Q \mid g_2).
\]

For part (b), in the finite case, it is sufficient to note that \( P \) and \( Q \) cannot be further partitioned, so that the discrete \( f \)-divergence is the supremum over all quantizers. If \( \text{card}(\mathcal{X}) = +\infty \), the proof is somewhat more involved. First, we show that

\[
D_f (P\|Q) \leq \sum_{x} q(x) f \left( \frac{p(x)}{q(x)} \right).
\]

Fix any finite partition \( A_1, \ldots, A_n \) of \( \mathcal{X} \). Define the measure \( \mu = P + Q \), and let the function \( a(x) = q(x)/(p(x) + q(x)) \) and \( b(x) = p(x)/(p(x) + q(x)) \). Then

\[
Q(A_i) f \left( \frac{P(A_i)}{Q(A_i)} \right) = \int_{A_i} a d\mu f \left( \frac{\int_{A_i} b d\mu}{\int_{A_i} a d\mu} \right) \leq \int_{A_i} a(x) f \left( \frac{b(x)}{a(x)} \right) d\mu(x) = \sum_{x \in A_i} q(x) f \left( \frac{p(x)}{q(x)} \right),
\]

2
where the inequality follows from Question 2.8(b), as \( \mu(\mathcal{X}) = 2 < \infty \). Thus any finite partition \( \mathcal{A} \) of \( \mathcal{X} \) gives \( D_f(P\|Q \mid \mathcal{A}) \leq \sum_x q(x)f(p(x)/q(x)) \). To see that the sum attains the supremum, we may simply construct a supremum-achieving partition. Let us identify \( \mathcal{X} \) with \( \mathbb{N} \), which is no loss of generality, and let the sequence of partitions \( \mathcal{A}_1, \mathcal{A}_2, \ldots \) be defined by

\[
\mathcal{A}_1 = \{A_0^1, A_1^1\}, \quad \mathcal{A}_2 = \{A_0^2, A_1^2, A_2^2\}, \ldots,
\]

where \( A_i^j = \{i\} \) if \( i < j \) and \( A_j^j = \{j, j+1, \ldots\} \). Then we have that the sequence

\[
D_n := \sum_{i=0}^n Q(A_i^n)f\left(\frac{P(A_i^n)}{Q(A_i^n)}\right)
\]

is nonnegative and increasing, and \( \lim_n D_n = \sum_x q(x)f(p(x)/q(x)) \) as desired. \( \square \)

**Answer to 2.10:** We begin with part (a). By equality (??), we may assume without loss of generality that the space \( Z \) is finite, in which case we have \( K_P(z) = \int k(z, x)p(x)dx \), and similarly for \( K_Q \), for some non-negative function \( k \) satisfying \( \sum_z k(z, x) = 1 \) and for \( z \in \{1, \ldots, m\} \), where \( p, q \) are the densities of \( P \) and \( Q \). Then, using the convexity of the perspective transform (the second part of Question 2.8), we have that

\[
D_f(K_P\|K_Q) = \sum_{z=1}^m K_Q(z)f\left(\frac{K_Q(z)}{K_P(z)}\right) = \sum_{z=1}^m \left(\int_X k(z, x)q(x)dx\right)f\left(\frac{\int_X k(z, x)p(x)dx}{\int_X k(z, x)q(x)dx}\right)
\]

\[
\leq \sum_{z=1}^m \int_X k(z, x)q(x)f\left(\frac{k(z, x)p(x)}{k(z, x)q(x)}\right)dx
\]

\[
= \sum_{z=1}^m \int_X k(z, x)q(x)f\left(\frac{p(x)}{q(x)}\right)dx.
\]

Now note that since \( K \) is a Markov kernel, we have \( \sum_{z=1}^m k(z, x) = 1 \), and thus

\[
D_f(K_P\|K_Q) \leq \int_X q(x)f\left(\frac{p(x)}{q(x)}\right)dx = D_f(P\|Q),
\]

our desired inequality.

For part (b), we simply note that the transition probability from the pair \((X, Y)\) to \((X, Z)\) is given by the product of the identity kernel (on \( X \)) and whatever Markov transition kernel is on pair \( Y \to Z \). Then we apply part (a). \( \square \)

**Answer to 2.14:** We compute the integral

\[
D_{kl}(P_1\|P_2) = \mathbb{E}_{P_1}\left[\frac{1}{2}\log \frac{\text{det}(\Sigma_2)}{\text{det}(\Sigma_1)} - \frac{1}{2}(X - \theta_1)^T\Sigma_1^{-1}(X - \theta_1) + \frac{1}{2}(X - \theta_2)^T\Sigma_2^{-1}(X - \theta_2)\right]
\]

\[
= \frac{1}{2}\log \frac{\text{det}(\Sigma_2)}{\text{det}(\Sigma_1)} - \frac{1}{2}d + \frac{1}{2}\mathbb{E}_{P_1}[(X - \theta_1)^T\Sigma_2^{-1}(X - \theta_1)] + \frac{1}{2}(\theta_2 - \theta_1)^T\Sigma_2^{-1}(\theta_2 - \theta_1)
\]

\[
= \frac{1}{2}\left[\text{tr}(\Sigma_1\Sigma_2^{-1}) - d - \log \text{det}(\Sigma_1\Sigma_2^{-1})\right] + \frac{1}{2}(\theta_2 - \theta_1)^T\Sigma_2^{-1}(\theta_2 - \theta_1).
\]
Solutions to exercises in Chapter 3

Answer to 3.1:

(a) The variance of $X$ is $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. As $\mathbb{E}[X]$ minimizes $t \mapsto \mathbb{E}[(X - t)^2]$, we have

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \inf_t \mathbb{E}[(X - t)^2] \leq \mathbb{E} \left[ \left( X - \frac{a + b}{2} \right)^2 \right] \leq \max \left\{ \left( \frac{b - a + b}{2} \right)^2, \left( \frac{a - a + b}{2} \right)^2 \right\} = \frac{(b - a)^2}{4}. \quad (\ast)$$

(b) These are straightforward calculations. Without loss of generality, we may assume that $\mathbb{E}[X] = 0$ as the calculations are simply shifts otherwise. We have

$$\varphi(0) = \log \mathbb{E}[1] = 0, \quad \varphi'(t) = \frac{\mathbb{E}[X e^{tx}]}{\mathbb{E}[e^{tx}]},$$

and

$$\varphi''(t) = \frac{d}{dt} \left( \frac{\mathbb{E}[X e^{tx}]}{\mathbb{E}[e^{tx}]} \right) + \mathbb{E}[X e^{tx}] \frac{d}{dt} (\mathbb{E}[e^{tx}])^{-1} = \frac{\mathbb{E}[X^2 e^{tx}]}{\mathbb{E}[e^{tx}]} - \frac{\mathbb{E}[X e^{tx}]^2}{\mathbb{E}[e^{tx}]^2}.$$

(c) Let $p(x)$ be the density of $X$ and again assume w.l.o.g. that $\mathbb{E}[X] = 0$. The random variable $Y_t$ with density $q_t$ defined by

$$q_t(y) = \frac{e^{ty}}{\mathbb{E}[e^{tx}]} p(y)$$

satisfies $\int q_t(y) dy = \mathbb{E}[e^{tx}] = 1$, and $Y_t \in [a, b]$ with probability 1. Moreover,

$$\text{Var}(Y_t) = \mathbb{E}[Y_t^2] - \mathbb{E}[Y_t]^2 = \int y^2 q_t(y) dy - \left( \int y q_t(y) dy \right)^2 = \int x^2 \frac{e^{tx}}{\mathbb{E}[e^{tx}]} p(x) dx - \left( \int x \frac{e^{tx}}{\mathbb{E}[e^{tx}]} p(x) dx \right)^2 = \frac{\mathbb{E}[X^2 e^{tx}]}{\mathbb{E}[e^{tx}]} - \frac{\mathbb{E}[X e^{tx}]^2}{\mathbb{E}[e^{tx}]^2}.$$  

(d) Because $\varphi'(0) = \varphi(0) = 0$, we know that

$$\varphi(\lambda) = \int_0^\lambda \varphi'(t) dt = \int_0^\lambda \int_0^t \varphi''(s) ds dt \leq \int_0^\lambda \int_0^t \frac{(b - a)^2}{4} ds dt = \int_0^\lambda t \frac{(b - a)^2}{4} dt = \frac{\lambda^2 (b - a)^2}{8}.$$   

Answer to 3.3:
(a) We bound $\mathbb{E}[L(Z_1^n)]$ by noting that

$$L(z_1, \ldots, z_n) = \sum_{i=1}^{n} L(z_i \mid z_1^{i-1}) \text{ where } L(z_i \mid z_1^{i-1}) = \log \frac{p(z_i \mid z_1^{i-1}, x)}{p(z_i \mid z_1^{i-1}, x')}$$

so that $L(z_i \mid z_1^{i-1}) \in [-\varepsilon, \varepsilon]$. Additionally,

$$\mathbb{E}[L(Z_i \mid z_1^{i-1})] = D_{\text{kl}}(P_{z_i}(\cdot \mid x, z_1^{i-1})\|P_{z_i}(\cdot \mid x', z_1^{i-1}))$$

$$\leq D_{\text{kl}}(P_{z_i}(\cdot \mid x, z_1^{i-1})\|P_{z_i}(\cdot \mid x', z_1^{i-1})) + D_{\text{kl}}(P_{z_i}(\cdot \mid x', z_1^{i-1})\|P_{z_i}(\cdot \mid x, z_1^{i-1}))$$

$$= \sum_{x} (p(z \mid z_1^{i-1}, x) - p(z \mid z_1^{i-1}, x')) \log \frac{p(z \mid z_1^{i-1}, x)}{p(z \mid z_1^{i-1}, x')},$$

and $|p(z \mid z_1^{i-1}, x) - p(z \mid z_1^{i-1}, x'))| \leq p(z \mid z_1^{i-1}, x)(e^\varepsilon - 1)$, while $\log \frac{p(z|x)}{p(z|x')} \leq \varepsilon$ certainly. Thus $\mathbb{E}[L(Z_1, \ldots, Z_n)] \leq n\varepsilon(e^\varepsilon - 1)$.

The increments $L(Z_i \mid Z_1^{i-1}) - \mathbb{E}[L(Z_i \mid Z_1^{i-1}) \mid Z_1^{i-1}]$ form a bounded difference martingale sequence, and as $L(Z_i \mid Z_1^{i-1}) \in [-\varepsilon, \varepsilon]$, we know that

$$\sup_z L(z \mid Z_1^{i-1}) - \mathbb{E}[L(Z_i \mid Z_1^{i-1})] \leq 2\varepsilon + \inf_z L(z \mid Z_1^{i-1}) - \mathbb{E}[L(Z_i \mid Z_1^{i-1})].$$

That is, the increments are $\varepsilon^2$-sub-Gaussian, and we immediately obtain

$$\mathbb{P}(L(Z_1, \ldots, Z_n) \geq n\varepsilon(e^\varepsilon - 1) + t) \leq \exp\left(-\frac{t^2}{2n\varepsilon^2}\right)$$

for all $t \geq 0$ by the Azuma-Hoeffding bound.

(b) The inequality is equivalent to $\|P_x - P_{x'}\|_{TV} \leq \gamma$. Using Pinsker’s inequality, it is sufficient that $D_{\text{kl}}(P_x\|P_{x'}) \leq 2\gamma^2$. Then we compute

$$D_{\text{kl}}(P_x\|P_{x'}) = \mathbb{E}[L(Z_1^n)] \leq n\varepsilon(e^\varepsilon - 1),$$

and so we have (roughly) $\varepsilon = \gamma/\sqrt{n}$ is sufficient for large $n$.

\[\square\]

**Answer to 3.4:** We use the hint, as we know that $\varepsilon^T x$ is $\|x\|^2_2/4$-sub-Gaussian. As $X_i$ are mean zero, we can introduce independent copies $X'_i$ with $\mathbb{E}[X'_i] = 0$, and

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]^p = \mathbb{E}\left[\sum_{i=1}^{n} X_i - \mathbb{E}[X'_i]\right]^p$$

$$\leq (i) \mathbb{E}\left[\sum_{i=1}^{n} X_i - X'_i\right]^p$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} \varepsilon_i (X_i - X'_i)\right]^p$$

$$\leq (ii) 2^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} \varepsilon_i X_i\right]^p + 2^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} \varepsilon_i X'_i\right]^p$$

$$= 2^p \mathbb{E}[\|\varepsilon X\|_p^p]$$.
where inequalities (i) and (ii) are both Jensen’s inequality. Now, we use Theorem 3.10, which yields that
\[ \mathbb{E}[|e^T X|^p] = \mathbb{E}[\mathbb{E}[|e^T X|^p | X]] \leq C p^{p/2} \mathbb{E}[\|X\|_2^p] \]
which is the first result.

If we consider means, we have for \( p \geq 2 \) that
\[ \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)^p \right] \leq \frac{C_p}{n^{p/2}} \mathbb{E} \left[ \left( n^{-1} \sum_{i=1}^{n} X_i^2 \right)^{p/2} \right] \leq \frac{C_p}{n^{p/2}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|X_i|^p] \]
by Jensen’s inequality, as \( t \mapsto t^{p/2} \) is convex.

**Answer to 3.6:** We use Theorem 3.10, showing that there exists a constant \( C \) such that
\[ \mathbb{E}[\exp((f(X) - \mathbb{E}[f(X)])^2/(CL^2\sigma^2))] \leq e. \]
To that end, note that \( t \mapsto e^t \) is a convex function, as its first two derivatives are \( 2te^t \) and \( 2e^t + 4t^2e^t > 0 \). Letting \( Y \) denote an independent copy of \( X \), we thus have by Jensen’s inequality that for any \( \lambda \in \mathbb{R} \),
\[ \mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(X)])^2)] = \mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(Y)])^2)] \leq \mathbb{E}[\exp(\lambda(f(X) - f(Y))^2)] \leq \mathbb{E}[\exp(\lambda L^2(X - Y)^2)]. \]
Now we use that \( (X - Y)^2 \leq 2X^2 + 2Y^2 \) to find that
\[ \mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(X)])^2)] \leq \mathbb{E}[\exp(2\lambda L^2 X^2)] \mathbb{E}[\exp(2\lambda L^2 Y^2)] = \mathbb{E}[\exp(2\lambda L^2 X^2)]^2. \]
As \( X^2 \) is \( \sigma^2 \)-sub-Gaussian, this has bound
\[ \mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(X)])^2)] \leq \frac{1}{\sqrt{1 - 4\lambda L^2 \sigma^2}}, \]
which is less than \( \sqrt{2} \) for \( \lambda = 1/(8L^2 \sigma^2) \).
Solutions to exercises in Chapter 4

Answer to 4.2: If $X$ is a mean-zero random variable satisfying the Bernstein condition (3.1.9), then

$$
\mathbb{E}[\exp(\lambda X/\sigma)] \leq e^{\lambda^2}
$$

for all $|\lambda| \leq \sigma/(2b)$ by Lemma 3.18. Thus if $X_i$ are independent, then

$$
\mathbb{E}[\exp(\lambda \sum_{i=1}^{n} X_i/\sigma)] \leq \exp(n\lambda^2).
$$

By Donsker-Varadhan (Theorem 4.1) we then have for all $\pi, \pi_0$ on $\mathcal{T}$ that

$$
n\lambda \int \frac{P_n \phi_t - P \phi_t}{\max\{C, \sigma(\phi_t)\}} d\pi(t) \leq D_{\text{KL}}(\pi\|\pi_0) + n\lambda^2
$$

for $|\lambda| \leq \inf_t \max\{C, \sigma(\phi_t)\}/(2b)$. Dividing by $n$ and taking expectations by drawing $T$ conditional on observing $X^n_1$, we obtain

$$
\lambda \mathbb{E}\left[ \frac{P_n \phi_T - P \phi_T}{\max\{C, \sigma(\phi_T)\}} \right] \leq \frac{1}{n} I(T; X^n_1) + \lambda^2,
$$

valid for all $|\lambda| \leq C^2/2b$. We may take $\lambda$ to have either positive or negative sign to get the result of the question.

Answer to 4.3: For all $\phi$ we have $P_n(\phi - P_n \phi)^2 \leq P_n(\phi - P \phi)^2$. For $\lambda \in [0,1]$ we have $e^\lambda \leq 1 + 2\lambda$, and thus for any $C > 0$ and $\lambda \in [0, C/4]$ we have $\lambda(\phi(x) - P \phi)^2/\max\{C, \sigma^2\} \leq 4\lambda/C \leq 1$, and so

$$
\mathbb{E}[e^{\lambda n P_n(\phi - P \phi)^2/\max\{C, \sigma^2\}}] = \prod_{i=1}^{n} \mathbb{E}\left[ \exp\left( \frac{\lambda(\phi(X_i) - P \phi)^2}{\max\{C, \sigma^2\}} \right) \right] \leq \left(1 + 2\lambda \frac{P(\phi - P \phi)^2}{\max\{C, \sigma^2\}}\right)^n \leq \exp(2\lambda n).
$$

Consequently, for any prior $\pi_0$ on $t$ we have

$$
\log \int \mathbb{E}[e^{\lambda n (P_n \phi_t - P \phi_t)^2/\max\{C, \sigma^2(\phi_t)\}}] d\pi_0(t) \leq 2\lambda n.
$$

We apply Donsker-Varadhan and that $s_n^2(\phi) = P_n(\phi - P_n \phi)^2 \leq P_n(\phi - P \phi)^2$ to obtain that if $\lambda \in [0, C/4]$, then

$$
n\lambda \mathbb{E}\left[ \frac{P_n \phi_T^2 - (P_n \phi_T)^2}{\max\{C, \sigma^2(\phi_T)\}} \right] \leq n\lambda \mathbb{E}\left[ \frac{P_n(\phi_T - P \phi_T)^2}{\max\{C, \sigma^2(\phi_T)\}} \right] \leq I(T; X^n_1) + 2\lambda n.
$$

Dividing by $n\lambda$ gives the result.

Answer to 4.4: Define the empirical measure $P_{n,-i} = \frac{1}{n-1} \sum_{j \neq i} 1_{x_j}$. Then using this, we have the equalities

$$
\mu - \mu_{-i} = \frac{\phi(X_i) - \mu}{n-1}
$$
and
\[ s^2 - s_{-i}^2 = P_n \phi^2 - P_{n-i} \phi^2 - (\mu^2 - \mu_{-i}^2) = \frac{\phi(X_i)^2 - P_n \phi^2}{n-1} - \frac{\phi(X_i) - \mu}{n-1} (\mu + \mu_{-i}). \]

We always have
\[
\frac{\sigma^2}{\sigma_{-i}^2} \leq \sup_{x \in [0,5]} \frac{\max\{\alpha s^2, \tau^2\}}{\max\{\alpha(s^2 - x/n), \tau^2\}} = \frac{\max\{\alpha s^2, \tau^2\}}{\max\{\alpha(s^2 - 5/n), \tau^2\}} \leq \frac{\tau^2 + 5\alpha/n}{\tau^2},
\]
and similarly
\[
\frac{\sigma^2}{\sigma_{-i}^2} \geq \frac{\tau^2}{\tau^2 + 5\alpha/n}.
\]
Consequently, \( \frac{1}{2} \leq \sigma^2/\sigma_{-i}^2 \leq 2 \) whenever \( \tau^2 \geq \frac{5\alpha}{n} \). We also note that \( \log(1 + x) \geq x - x^2 \) for \( x \geq 0 \), and thus for any \( a \geq b/2 > 0 \) we have
\[
\frac{a}{b} - 1 - \log \frac{a}{b} = \frac{a - b}{b} - \log \left( 1 + \frac{a - b}{b} \right) \leq \frac{a - b}{b} - \frac{a - b}{b} + \frac{(a - b)^2}{b^2}.
\]
Consequently, we obtain that
\[
D_{kl}(N(\mu, \sigma^2) | N(\mu_{-i}, \sigma_{-i}^2)) = \frac{(\mu - \mu_{-i})^2}{2\sigma_{-i}^2} + \frac{1}{2} \left( \frac{\sigma^2}{\sigma_{-i}^2} - 1 - \log \frac{\sigma^2}{\sigma_{-i}^2} \right)
\leq \frac{(\mu - \mu_{-i})^2}{\sigma^2} + \frac{1}{2} \left( \frac{\sigma^2}{\sigma_{-i}^2} \right)^2
\leq \frac{(\mu - \mu_{-i})^2}{\sigma^2} + \left( \frac{\sigma^2 - \sigma_{-i}^2}{\sigma^2} \right)^2
\leq \frac{(\phi(X_i) - \mu)^2}{(n-1)^2\sigma^2} + \left( \frac{\alpha(s^2 - s_{-i}^2)}{\sigma^2} \right)^2
\leq \frac{C}{n^2\sigma^2} \left[ (\phi(X_i) - \mu)^2 + \frac{\alpha^2(\phi(X_i)^2 - P_n \phi^2)^2 + (\phi(X_i) - \mu)^2}{\sigma^2} \right].
\]

Summing over all \( i = 1, \ldots, n \), we obtain
\[
\frac{1}{n} \sum_{i=1}^n D_{kl}(N(\mu, \sigma^2) | N(\mu_{-i}, \sigma_{-i}^2)) \leq \frac{C}{n^2\sigma^2} s^2 + \frac{C\alpha^2}{n^2\sigma^4}(\text{Var}_{P_n}(\phi^2) + s^2).
\]

Now we use that if \( f \) is \( L \)-Lipschitz, then \( \text{Var}(f(X)) \leq L^2 \text{Var}(X) \), because
\[
\text{Var}(f(X)) = \frac{1}{2} \mathbb{E}[(f(X) - f(Y))^2] \leq \frac{L^2}{2} \mathbb{E}[(X - Y)^2] = L^2 \text{Var}(X)
\]
for \( X, Y \) i.i.d. On \([-1, 1]\), the function \( f(t) = t^2 \) is Lipschitz.

\[ \square \]

Answer to 4.5:
(a) Certainly, the \( \phi \) take bounded values, so they satisfy the Bernstein condition (3.1.9), and so Question 4.2 shows that for any \( b > 0 \), then

\[
\mathbb{E} \left[ \frac{P_b \phi_T - P_{\phi_T}}{\max\{b, \sigma(\phi_T)\}} \right] \leq \frac{1}{n\lambda} I(X_1^n; T) + \lambda
\]

(1)

for any random variable \( T \) and \( 0 \leq \lambda \leq \frac{b}{2} \).

Now, define the worst index

\[
k^* = \arg\max_{j \in [k]} |A_j - P_{\phi_T^j}| \quad \text{max}\{b, \sigma(\phi_T^j)\}
\]

and let the triple \((A_*, \phi_*, Z_*)\) be defined by

\[
\phi_* = \text{sign}(A_{k^*} - P_{\phi_{T_{k^*}}}) \quad Z_* = \text{sign}(A_{k^*} - P_{\phi_{T_{k^*}}}) Z_{k^*}, \quad A_* = P_{n\phi_*} + Z_*. 
\]

Note that these can all be computed by a monitor as in the proof of Theorem 4.19, using the observed quantities in Algorithm 4.3. Now, let \( \sigma_* = \sigma_{k^*} \) be the used standard deviation for the Gaussian \( Z \) on this worst query. Then as a consequence, we have

\[
\mathbb{E} \left[ \max_{j \leq k} |A_j - P_{\phi_T^j}| \right] = \mathbb{E} \left[ P_{n\phi_*} - P_{\phi_*} + Z_* \sigma_* \right].
\]

Applying inequality (1), we thus obtain that

\[
\mathbb{E} \left[ \max_{j \leq k} |A_j - P_{\phi_T^j}| \right] \leq \frac{1}{n\lambda} I(X_1^n; T, \ldots, T_k) + \lambda + \mathbb{E} \left[ \max_{j \leq k} |Z_j| \sqrt{\max\{\alpha s_*^2, \tau^2\}} \right].
\]

By Cauchy-Schwarz, we have

\[
\mathbb{E} \left[ \max_{j \leq k} |Z_j| \sqrt{\max\{\alpha s_*^2, \tau^2\}} \right] \leq \sqrt{\mathbb{E}[\max_{j \leq k} Z_j^2]} \sqrt{\mathbb{E} \left[ \max\{\alpha s_*^2, \tau^2\} \right]} \leq \sqrt{2\log(k) + 2} \sqrt{\frac{\alpha}{n\lambda_0} I(X_1^n; T) + 2\alpha + \frac{\tau^2}{b^2}},
\]

where we have used the result of Question 4.3 and Lemma 4.20, and the bound holds for \( \lambda_0 \leq \frac{b}{4} \). Returning to our earlier bounds, we find that for any \( b > 0 \) and appropriate \( \alpha, \tau \) that

\[
\mathbb{E} \left[ \max_{j \leq k} |A_j - P_{\phi_T^j}| \right] \leq \frac{1}{n\lambda} I(X_1^n; T_k^k) + \lambda + \sqrt{2\log(k) e} \sqrt{\frac{4\alpha}{nb} I(X_1^n; T_k^k) + 2\alpha + \frac{\tau^2}{b^2}},
\]

which gives part (a).

(b) We optimize our choices. Let us assume that \( \tau^2 \geq \frac{5\alpha}{n} \) and that \( \alpha^2 \leq \tau^2 \) for simplicity. Then the tensorization (adaptive composition) results of Lemmas 4.16 and 4.17, i.e. Proposition ??, imply that

\[
\frac{1}{n} I(X_1^n; T_k^k) \leq C \frac{k}{n^2 \alpha}.
\]
Choosing \( b^2 = \frac{\sqrt{k \log k}}{n} \), we have from part (a) that

\[
E \left[ \max_{j \leq k} \frac{|A_j - P\phi_j|}{\max\{b, \sigma(\phi_j)\}} \right] \leq C \frac{k}{n^2 \alpha \lambda} + \lambda + C \sqrt{\log k} \sqrt{\frac{k}{n^2 b}} + \alpha + \frac{\tau^2}{b}
\]

\[
= C \frac{k}{n^2 \alpha \lambda} + \lambda + C \sqrt{\frac{\sqrt{k \log k}}{n}} + \alpha + \frac{n \tau^2 \sqrt{\log k}}{\sqrt{k}}.
\]

Now, set \( \alpha = \frac{\sqrt{k \log k}}{n \sqrt{\log k}} \) and \( \tau^2 = \frac{k}{n^2} \), which implies that \( \frac{5n \alpha}{n} \leq \tau^2 \) and \( \alpha^2 \leq \tau^2 \). Thus we obtain

\[
E \left[ \max_{j \leq k} \frac{|A_j - P\phi_j|}{\max\{b, \sigma(\phi_j)\}} \right] \leq C \left[ \frac{\sqrt{k \log k}}{n \lambda} + \lambda + \frac{(k \log k)^{1/4}}{\sqrt{n}} \right].
\]

As \( \lambda \leq b \) and \( b = (k \log k)^{1/4}/\sqrt{n} \), we simply choose \( \lambda = b \) to complete the result.

(c) This is evidently better than Theorem 4.19, because it gives variance-dependent bounds. As \( \sigma(\phi_j) \leq 1 \) always and may be much smaller, roughly what this bound says is that

\[
|A_j - P\phi_j| \lesssim \frac{(k \log k)^{1/4}}{\sqrt{n}} \sigma(\phi_j)
\]

with reasonable probability. This is always better than \( |A_j - P\phi_j| \lesssim \frac{(k \log k)^{1/4}}{\sqrt{n}} \).

\[\square\]
Answers to exercises in Chapter 6

Answer to 6.2:

(a) We know that for any \( P, Q \), we have \( D_\alpha (P | Q) = \log(D_f (P || Q) + 1) \) for \( f(t) = t^\alpha - 1 \), which is convex. We also recall that \( f \)-divergences are jointly convex in their arguments.

Now, let \( s \in \{0, 1\}^n \) or \( s \in \{0, 1\}^{n-1} \) denote bit-wise inclusion vectors. Without loss of generality we may assume that \( X \) and \( X' \) differ only in the last column \( x \) of \( X \) so that \( X = [X' \; x] \). Then we have

\[
Q(\cdot \mid X) = \sum_{s \in \{0, 1\}^n} q^{|s|_1} (1 - q)^{n-|s|_1} P_{Xs}
\]

\[
= \sum_{s \in \{0, 1\}^{n-1}} q^{|s|_1} (1 - q)^{n-1-|s|_1} (q P_{X's+x} + (1 - q) P_{X's})
\]

which is either include the last vector \( x \) in the sum \( \sum_i x_i s_i \) or do not, with probabilities \( q \) and \( 1 - q \), respectively. Similarly,

\[
Q(\cdot \mid X') = \sum_{s \in \{0, 1\}^{n-1}} q^{|s|_1} (1 - q)^{n-1-|s|_1} P_{X's}.
\]

The binomial theorem implies \( \sum_{s \in \{0, 1\}^{n-1}} q^{|s|_1} (1 - q)^{n-1-|s|_1} = 1 \), and so Jensen’s inequality gives for any \( f \)-divergence that

\[
D_f (Q(\cdot \mid X) | Q(\cdot \mid X')) \leq \sum_{s \in \{0, 1\}^{n-1}} q^{|s|_1} (1 - q)^{n-1-|s|_1} D_f (q P_{X's+x} + (1 - q) P_{X's+x'})
\]

\[
\leq \max_{s \in \{0, 1\}^{n-1}} D_f (q P_{X's+x} + (1 - q) P_{X's})
\]

\[
= \max_{s \in \{0, 1\}^{n-1}} D_f (q P_{X's+x} + (1 - q) P_{X's})
\]

and similarly

\[
D_f (Q(\cdot \mid X') \| Q(\cdot \mid X)) \leq \max_{s \in \{0, 1\}^{n-1}} D_f (P_{X's} \| q P_{X's+x} + (1 - q) P_{X's})
\]

Now, as we have an \( f \)-divergence we use the shift-invariant properties of a Gaussian to see that for any \( \mu_1, \mu_0, \Delta \in \mathbb{R}^d \) we have

\[
D_f (q P_{\mu_1 + \Delta} + (1 - q) P_{\Delta} \| q P_{\mu_0 + \Delta} + (1 - q) P_{\Delta}) = D_f (q P_{\mu_1} + (1 - q) P_0 \| q P_{\mu_0} + (1 - q) P_0)
\]

and so we have for any \( \alpha \geq 1 \) that

\[
D_\alpha (Q(\cdot \mid X) \| Q(\cdot \mid X')) \leq D_\alpha (q P_x + (1 - q) P_0 \| P_0)
\]

and

\[
D_\alpha (Q(\cdot \mid X') | Q(\cdot \mid X)) \leq D_\alpha (P_0 \| q P_x + (1 - q) P_0)
\]

11
(b) Let \( \mathbb{E}_0 \) and \( \mathbb{E}_\mu \) denote expectation under \( N(0, \sigma^2 I) \) and \( N(\mu, \sigma^2 I) \), respectively, and let \( p_\mu \) denote the density of \( N(\mu, \sigma^2 I) \).

We first consider \( D_2(qP_\mu + (1 - q)P_0\|P_0) \). We have

\[
\mathbb{E}_0 \left[ \left( \frac{qp_\mu(X) + (1 - q)p_0(X)}{p_0(X)} \right)^2 \right] = \mathbb{E}_0 \left[ \left( 1 + q \frac{p_\mu(X) - p_0(X)}{p_0(X)} \right)^2 \right] = 1 + q^2 \mathbb{E}_0 \left[ \left( \frac{p_\mu(X) - p_0(X)}{p_0(X)} \right)^2 \right] = 1 + q^2 D_f(P_\mu\|P_0)
\]

for the f-divergence with \( f(t) = (t - 1)^2 \), that is, the \( \chi^2 \)-divergence. A calculation (recall Example 6.10) yields that \( D_f(P_\mu\|P_0) = \exp(||\mu||^2_2 / \sigma^2) - 1 \), and so \( D_2(qP_\mu + (1 - q)P_0\|P_0) = \log(1 + q^2(\exp(||x||^2_2 / \sigma^2) - 1)) \).

Now consider \( D_2(P_0\|qP_\mu + (1 - q)P_0) \). We have

\[
\mathbb{E}_{qP_\mu + (1 - q)P_0} \left[ \left( \frac{p_0(X)}{qp_\mu(X) + (1 - q)p_0(X)} \right)^2 \right] = \mathbb{E}_{qP_\mu + (1 - q)P_0} \left[ \left( 1 + q \frac{p_0(X) - p_\mu(X)}{qp_\mu(X) + (1 - q)p_0(X)} \right)^2 \right]
\]

\[
= 1 + q^2 \mathbb{E}_{qP_\mu + (1 - q)P_0} \left[ \left( \frac{p_0(X) - p_\mu(X)}{qp_\mu(x) + (1 - q)p_0(x)} \right)^2 \right]
\]

\[
= 1 + q^2 \int \frac{(p_\mu(x) - p_0(x))^2}{qp_\mu(x) + (1 - q)p_0(x)} \, dx
\]

\[
\leq 1 + \frac{q^2}{1 - q} \int \frac{(p_\mu(x) - p_0(x))^2}{p_0(x)} \, dx
\]

\[
= 1 + \frac{q^2}{1 - q} \left( \exp(||\mu||^2_2 / \sigma^2) - 1 \right),
\]

again using Example 6.10. Taking logarithms gives the result.

(c) We use the composition bounds for Rényi divergences (Theorem 6.23 and Corollary 6.24) to obtain that \( Z_{\text{sub}} \) is \( T \log(1 + \frac{q^2}{1 - q}(\exp(b^2/(2\sigma^2_{\text{sub}})) - 1)) \)-private. As \( \log(1 + t) \leq t \), it is also \( T \frac{q^2}{1 - q}(\exp(b^2/(2\sigma^2_{\text{sub}})) - 1) \)-private.

(d) \( Z_{\text{Gauss}} \) on the other hand is \( \frac{b^2}{n^2\sigma^2_{\text{Gauss}}} \)-private.

(e) We have

\[
\mathbb{E}[\|Z_{\text{sub}} - \mathbf{X}_n\|_2^2] = \frac{1}{T} \mathbb{E}[\|np_{\text{sub}}^{-1}X S - \mathbf{X}_n\|_2^2] + \frac{\sigma^2_{\text{sub}}}{Tn^2q^2} \mathbb{E}[\|W\|_2^2] = 1 - q \frac{1}{T} \sum_{i=1}^n \|x_i\|^2_2 + \frac{d\sigma^2_{\text{sub}}}{Tn^2q^2}.
\]

Using the approximation \( t + O(t^2) = \log(1 + t) \leq t \), we see that it is sufficient that \( T \frac{q^2}{1 - q}(\exp(b^2/(2\sigma^2_{\text{sub}})) - 1) \leq \varepsilon \) to guarantee \( \varepsilon \)-privacy, so we have (at best) that \( 1/(Tq^2) \approx (\exp(b^2/(2\sigma^2_{\text{sub}})) - 1)/\varepsilon \), and

\[
\mathbb{E}[\|Z_{\text{sub}} - \mathbf{X}_n\|_2^2] = \frac{1 - q b^2}{Tnq} + \frac{d}{n^2\varepsilon} \cdot \sigma^2_{\text{sub}} \left[ \exp \left( \frac{b^2}{2\sigma^2_{\text{sub}}} \right) - 1 \right] \approx \frac{b^2}{Tnq} + \frac{db^2}{n^2\varepsilon}
\]
as long as $\sigma^2_{\text{sub}} \gg b^2$.

On the other hand, the Gaussian mechanism (as in the lecture notes) satisfies
\[
E[\|Z_{\text{Gauss}} - \bar{X}_n\|_2^2] = d\sigma_{\text{Gauss}}^2,
\]
and setting $\sigma_{\text{Gauss}}^2 = \frac{b^2}{n\varepsilon}$ yields $\varepsilon$-Rényi privacy and $E[\|Z_{\text{Gauss}} - \bar{X}_n\|_2^2] = \frac{d\sigma^2}{n\varepsilon}$.

Thus, so long as $Tq \gg \frac{n\varepsilon}{d}$, the subsampling mechanism introduces little additional error to the Gaussian mean mechanism. However, it does not provide any better convergence for the mean. 

\[
\text{Answer to 6.3:}
\]

(a) Using the result of Question 6.2, we see that for any $q \in (0,1)$ the scheme (6.7.4) is
\[
\log \left[ 1 + \frac{q^2}{1-q} \left( \exp \left( \frac{M^2}{\sigma_{\text{sub}}^2} \right) - 1 \right) \right] \leq (e-1) \frac{q^2}{1-q} \frac{M^2}{\sigma_{\text{sub}}^2} \leq \frac{2q^2M^2}{\sigma_{\text{sub}}^2},
\]
Rényi private for $q < 1 - 1/e$, where we have used that $\sigma_{\text{sub}} \geq M$ and $e^{1/x} - 1 \leq \frac{(e-1)}{x}$ for $x \geq 1$.

(b) By Question 6.2, we know that after $T$ iterations, the resulting $\bar{\theta}_T$ is evidently $\varepsilon = \frac{2Tq^2}{\sigma_{\text{sub}}^2}$-Rényi private. So any choices of $T, q, \sigma$ such that $\varepsilon \geq \frac{2Tq^2M^2}{\sigma^2}$ are sufficient.

We choose $T = n$, $q = 1/\sqrt{n}$, and $\sigma^2 = 2M^2/\varepsilon$.

(c) With these choices, we have that the variance of the gradients $g$ at $\theta = \theta^*$ satisfy $E[gg^T] = \frac{1}{nq} \text{Cov}(\nabla \ell(\theta^*; X)) + \frac{\sigma_{\text{sub}}^2}{nq} I = \frac{1}{\sqrt{n}} \text{Cov}(\nabla \ell(\theta^*; X)) + \frac{2M^2}{\sqrt{n}} I$. Proposition 6.27 implies that
\[
E[\|\bar{\theta}_T - \hat{\theta}_n\|_2]^{1/2} \leq \sqrt{\frac{\text{tr}(\nabla^2 \log p_{\text{data}}(\hat{\theta}_n) - 1) \Sigma_n \nabla^2 \ell_n(\hat{\theta}_n) - 1}{n^{3/2}}} + \frac{2M^2 \text{tr}(\nabla^2 \log p_{\text{data}}(\hat{\theta}_n) - 2)}{n^{3/2} \varepsilon} + O(n^{1-\beta/2}).
\]

(d) It says it is optimal. 

\[
\text{Answer to 6.3:}
\]

(a) Using the result of Question 6.2, we see that for any $q \in (0,1)$ the scheme (6.7.4) is
\[
\log \left[ 1 + \frac{q^2}{1-q} \left( \exp \left( \frac{M^2}{\sigma_{\text{sub}}^2} \right) - 1 \right) \right] \leq (e-1) \frac{q^2}{1-q} \frac{M^2}{\sigma_{\text{sub}}^2} \leq \frac{2q^2M^2}{\sigma_{\text{sub}}^2},
\]
Rényi private for $q < 1 - 1/e$, where we have used that $\sigma_{\text{sub}} \geq M$ and $e^{1/x} - 1 \leq \frac{(e-1)}{x}$ for $x \geq 1$.

(b) By Question 6.2, we know that after $T$ iterations, the resulting $\bar{\theta}_T$ is evidently $\varepsilon = \frac{2Tq^2}{\sigma_{\text{sub}}^2}$-Rényi private. So any choices of $T, q, \sigma$ such that $\varepsilon \geq \frac{2Tq^2M^2}{\sigma^2}$ are sufficient.

We choose $T = n$, $q = 1/\sqrt{n}$, and $\sigma^2 = 2M^2/\varepsilon$.

(c) With these choices, we have that the variance of the gradients $g$ at $\theta = \theta^*$ satisfy $E[gg^T] = \frac{1}{nq} \text{Cov}(\nabla \ell(\theta^*; X)) + \frac{\sigma_{\text{sub}}^2}{nq} I = \frac{1}{\sqrt{n}} \text{Cov}(\nabla \ell(\theta^*; X)) + \frac{2M^2}{\sqrt{n}} I$. Proposition 6.27 implies that
\[
E[\|\bar{\theta}_T - \hat{\theta}_n\|_2]^{1/2} \leq \sqrt{\frac{\text{tr}(\nabla^2 \log p_{\text{data}}(\hat{\theta}_n) - 1) \Sigma_n \nabla^2 \ell_n(\hat{\theta}_n) - 1}{n^{3/2}}} + \frac{2M^2 \text{tr}(\nabla^2 \log p_{\text{data}}(\hat{\theta}_n) - 2)}{n^{3/2} \varepsilon} + O(n^{1-\beta/2}).
\]

(d) It says it is optimal.
Solutions to exercises in Chapter 7

Answer to 7.1: For part (a), we begin by stating a lemma that will prove useful. We have

**Lemma 1.** Let $P$ and $Q$ be arbitrary distributions. Then

$$D_{\text{kl}}(P || Q) \geq \log \frac{1}{Q(dP > 0)} = \log \frac{1}{Q(\text{supp } P)}.$$  

**Proof** By a generalization of the log-sum inequality, we have

$$D_{\text{kl}}(P || Q) = \int dP \log \frac{dP}{dQ} = \int_{dP > 0} dP \log \frac{dP}{dQ} \geq \left( \int dP \right) \log \frac{\int dP}{\int_{dP > 0} dQ} = \log \frac{1}{Q(dP > 0)},$$

as desired.

With Lemma 1 in place, we can prove the result. We follow a strategy similar to that in the notes, but we must exercise a bit more care as entropy functions no longer exist. In this case, we define $E$ to be the \{0, 1\}-valued indicator for the error that the pair $(V, \hat{V}) \notin \mathcal{N}$. We begin with the mutual information $I(V; \hat{V}, E)$, which by definition is equal to the averaged KL-divergence

$$I(V; \hat{V}, E) = \int D_{\text{kl}} \left( P_{V|\hat{V}=\hat{v}, E=e} || P_{V} \right) dP_{V, E}(\hat{v}, e) \quad (2)$$

$$= \mathbb{P}(E = 1) \int D_{\text{kl}} \left( P_{V|\hat{V}=\hat{v}, E=1} || P_{V} \right) dP_{V|E=1}(\hat{v}) + \mathbb{P}(E = 0) \int D_{\text{kl}} \left( P_{V|\hat{V}=\hat{v}, E=0} || P_{V} \right) dP_{V|E=0}(\hat{v}).$$

But now note that the marginal $P_{V}$ is simply the prior distribution $\pi$, and using Lemma 1 we can give logarithmic bounds similar to our bounds on counts and entropy in the proof of Proposition 7.13. Indeed, we note that we have the support containments

$$\text{supp } P_{V|\hat{V}=\hat{v}, E=1} \subset N_{\hat{v}}^c \quad \text{and} \quad \text{supp } P_{V|\hat{V}=\hat{v}, E=0} \subset N_{\hat{v}}.$$

In particular, we see that

$$\pi(\text{supp } P_{V|\hat{V}=\hat{v}, E=1}) \leq \pi(N_{\hat{v}}^c) = 1 - \pi(N_{\hat{v}}) \quad \text{and} \quad \pi(\text{supp } P_{V|\hat{V}=\hat{v}, E=0}) \leq \pi(N_{\hat{v}}).$$

Revisiting our lower bound on the information $I(V; \hat{V}, E)$, Lemma 1 implies

$$D_{\text{kl}} \left( P_{V|\hat{V}=\hat{v}, E=1} || P_{V} \right) \geq \log \frac{1}{1 - \pi(N_{\hat{v}})} \geq \log \frac{1}{1 - p_{\text{min}}},$$

$$D_{\text{kl}} \left( P_{V|\hat{V}=\hat{v}, E=0} || P_{V} \right) \geq \log \frac{1}{\pi(N_{\hat{v}})} \geq \log \frac{1}{p_{\text{max}}}.$$

Substituting these lower bounds into equation (2), we obtain

$$I(V; \hat{V}, E) \geq \mathbb{P}(E = 1) \log \frac{1}{1 - p_{\text{min}}} + \mathbb{P}(E = 0) \log \frac{p_{\text{max}}}{1 - p_{\text{min}}} = \mathbb{P}(E = 1) \log \frac{p_{\text{max}}}{p_{\text{min}}} \log \frac{1}{p_{\text{max}}} + \log \frac{1}{p_{\text{max}}}.$$  

Noting that $I(V; \hat{V}, E) = I(V; \hat{V}) + I(V; E | \hat{V}) \leq I(V; \hat{V}) + H(E)$ completes the proof of inequality (7.6.2).
For part (b), we use that

\[
\log 2 + P_{err} \log \frac{1}{p_{\text{max}}} \geq h_2(P_{err}) + P_{err} \log \frac{1 - p_{\text{min}}}{p_{\text{max}}} \geq \log \frac{1}{p_{\text{max}}} - I(V; \hat{V}) \geq \log \frac{1}{p_{\text{max}}} - I(V; X),
\]

where we have used the data processing inequality. Rearranging we have

\[
P_{err} \geq \frac{\log \frac{1}{p_{\text{max}}} - I(V; X) - \log 2}{\log \frac{1}{p_{\text{max}}}} = 1 - \frac{I(V; X) + \log 2}{\inf_{\hat{V}} \log \frac{1}{\pi(N_{\hat{V}})}}
\]

as desired.

Part (c) is immediate. □

Answer to 7.2: We answer each part in turn. For part (a), note that

\[
\mathbb{P}(X_{(1)} \geq \theta + t) = \mathbb{P}(X_i \geq \theta + t, \text{ all } i) = [1 - t]^n.
\]

By recognizing that \((X_{(1)} - \theta)^2 \geq t\) if and only if \(X_{(1)} \geq \theta + \sqrt{t}\), we have

\[
E[(X_{(1)} - \theta)^2] = \int_0^\infty \mathbb{P}(X_1 \geq \theta + \sqrt{t}) dt = \int_0^\infty [1 - \sqrt{t}]^n + \int_0^1 (1 - \sqrt{t})^n dt.
\]

With the substitution \(u = \sqrt{t}\), so that \(2\sqrt{t}du = dt = 2udu\), the final integral is equal to

\[
\int_0^1 (1 - \sqrt{t})^n dt = 2 \int_0^1 u(1-u)^n du = 2 \frac{u(1-u)^{n+1}}{n+1} \bigg|_{u=0}^{u=1} + \frac{2}{n+1} \int_0^1 (1-u)^{n+1} du = 0 - \frac{2}{(n+1)(n+2)} (1-u)^{n+2} \bigg|_{u=0}^{u=2} = \left(\frac{n+2}{2}\right)^{-1}.
\]

For part (b), fix \(\delta \geq 0\) and consider the distributions \(P_1 = \text{Uni}(0,1)\) and \(P_2 = \text{Uni}(\delta,1+\delta)\). Comparing their \(n\)-fold products, we have

\[
\|P_1^n - P_2^n\|_{TV} = 1 - \text{Vol}([1-\delta]^n) = 1 - (1-\delta)^n.
\]

The minima of the supports of \(P_1\) and \(P_2\) are distance \(\delta\) from one another, so Le Cam’s method implies that

\[
\mathcal{M}_n(\theta(U), (\cdot)^2) = \sup_{\theta \in \mathbb{R}} \mathbb{E}_{\text{Uni}(\theta,\theta+1)} \left[ (\hat{\theta}(X_1, \ldots, X_n) - \theta)^2 \right] \geq \frac{\delta^2}{8} (1 - \|P_1^n - P_2^n\|_{TV}) \geq \frac{\delta^2}{8} (1 - \delta)^n.
\]

Noting that \(1 - \delta \geq e^{-3\delta^2/2}\) for \(\delta \leq 1/2\), we have

\[
\mathcal{M}_n((\cdot)^2, \theta(U)) \geq \frac{\delta^2}{8} \exp(-3\delta^2/2).
\]

Choosing \(\delta = 1/n\) gives the desired result. □