\# 1: A function denoising problem

Let $\theta$ be a discrete function sampled on a regular grid in $[0, 1]$. Namely, for $n \in \mathbb{N}$, we let $\varepsilon = 1/n$, and

$$\theta = (\theta(0), \theta(\varepsilon), \theta(2\varepsilon), \ldots, \theta((n-1)\varepsilon)) \in \mathbb{R}^n.$$  \hfill (1)

We observe noisy measurements of this function $y_k = \theta(k\varepsilon) + z_k$, where $(z_k)_{k \leq n} \sim iid \ N(0, \sigma^2)$, and are interested in estimating $\theta$ with respect to the normalized square loss

$$L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2/n.$$

We define the discrete derivative by letting $\Delta \theta(k\varepsilon) = [\theta((k+1)\varepsilon) - \theta(k\varepsilon)]/\varepsilon$ for $k \in \{0, \ldots, n-2\}$, and $\Delta \theta((n-1)\varepsilon) = [\theta(0) - \theta((n-1)\varepsilon)]/\varepsilon$ (periodic boundary conditions). We consider the following parameter class

$$\Theta(R, n) = \left\{ \theta : \sum_{k=0}^{n-1} \theta(k\varepsilon) = 0, \sum_{k=0}^{n-1} \varepsilon (\Delta \theta(k\varepsilon))^2 \leq R \right\}.$$ \hfill (2)

(a) Give an expression for the linear minimax risk $R_{Lin}(\Theta(R, n))$. 

[Hint: It might be convenient to use the discrete Fourier transform of $\theta$.]

(b) Can you apply Pinsker’s theorem and show that the linear minimax risk is close to the overall minimax risk $R_{st}(\Theta(R, n))$? Justify your answer and state explicitly any eventual condition that you are imposing on $R, n$.

Solution

(a) Starting with this problem, we directly observe that we may write the constraint $\sum_{k=0}^{n-1} \varepsilon (\Delta \theta(k\varepsilon))^2$ in Ellipsoidal form

$$\theta^T A \theta \leq R/n$$

Here:

$$A = \begin{pmatrix}
2 & -1 & 0 & 0 & \cdots & 0 & -1 \\
-1 & 2 & -1 & 0 & \cdots & 0 & 0 \\
0 & -1 & 2 & -1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
-1 & 0 & 0 & 0 & \cdots & -1 & 2
\end{pmatrix}$$
To apply Pinsker’s result, we need to diagonalize $A$. To this end, first consider the DFT matrix $(U_{kl})_{0 \leq k, l \leq n-1}$ with $U_{kl} = \exp\left(-\frac{2\pi i kl}{n}\right)$. Furthermore, recall the following properties: $U^*U = nI_n$, so that $U/\sqrt{n}$ is unitary. We may check that $U/\sqrt{n}$ diagonalizes $A$ with eigenvalues $2(1 - \cos(2\pi j/n))$.

One way to see this is to use Parseval’s identity for the DFT, as well as the Shift identity for the DFT (below we write $\theta \cdot = \theta$ and $\theta \cdot +1 = (\theta_1, \ldots, \theta_{n-1}, \theta_0)$ with $\theta_k = \theta(k \varepsilon)$). More concretely:

\[
n\|\theta_\cdot - \theta_{\cdot +1}\|^2 = \|U \theta_\cdot - U \theta_{\cdot +1}\|^2 \quad \text{Parseval}
= \|U \theta_\cdot - \exp(2i\pi \cdot /n) \cdot U \theta_\cdot\|^2 \quad \text{(coordinatewise product, shift)}
= \sum_{k=0}^{n-1} |1 - \exp(2ik\pi/n)|^2 (U \theta_\cdot)_k^2
= \sum_{k=0}^{n-1} 2(1 - \cos(2\pi k/n))(U \theta_\cdot)_k^2
\]

Since the unitary matrix $U/\sqrt{n}$ diagonalizes $A$, we note that there must exist also an orthogonal (real) matrix $O$ which diagonalizes $A$ and has the same eigenvalues. Furthermore, note that 1st column and row of $U/\sqrt{n}$ just consists of entries $1/\sqrt{n}$, thus also the 1st row of $O$ will consist of these entries. Thus upon mapping $y \mapsto \tilde{y} = Oy$, we observe that if we let $\theta = O\theta$, then $\tilde{y} \sim \mathcal{N}(\theta, \sigma^2)$. Furthermore the constraints turn into:

\[
\tilde{\theta}_0 = \sum_{i=0}^{n-1} \frac{1}{\sqrt{n}} \tilde{\theta}_i = 0
\]

and

\[
\sum_{k=0}^{n-1} 2(1 - \cos(2\pi k/n))\tilde{\theta}_k^2 \leq \frac{R}{n}
\]

Furthermore, since $\|\tilde{\theta} - \theta\|_2^2 = \|O\tilde{\theta} - O\theta\|_2^2$, we see that the transformed and the original estimation problems are equivalent and hence that their (linear) minimax risks must coincide. Also, since we know the first coordinate is 0, by a sufficiency argument we may discard the first observation $Y_0$ without loss of information and find ourselves in a $(n-1)$-dimensional Gaussian problem with the following Ellipsoidal form:

\[
\tilde{Y} \sim \mathcal{N}(\tilde{\theta}, \sigma^2)
\]

\[
\tilde{\theta} \in \tilde{\Theta} = \{\tilde{\theta} \in \mathbb{R}^{n-1} : \tilde{\theta}^\top \tilde{A} \tilde{\theta} \leq 1\}
\]

Here $\tilde{A} = \text{Diag}(\tilde{a}_1^2, \ldots, \tilde{a}_{n-1}^2)$ and $\tilde{a}_j = \sqrt{\frac{2\pi}{R}}(1 - \cos(2\pi j/n))$

We are finally ready to apply Theorem 4.1 from the notes to get (the notes gives us the linear minimax risk for the unnormalized loss so we further divide by $n$):

\[
R_L(\theta) = \frac{1}{n} \inf_{\lambda \geq 0} \left\{ \lambda^2 + \sigma^2 \sum_{i=1}^{n-1} (1 - \lambda \tilde{a}_i)_+^2 \right\}
\]

The minimum is achieved at the unique solution of:
\[ \lambda = \sigma^2 \sum_{j=1}^{n-1} \tilde{a}_j (1 - \lambda \tilde{a}_j)_+ \]

Let us now get a bit more insight into this expression, i.e. what is the minimax rate ignoring constants? We will write \( \asymp \) to denote "rate equality", i.e. we will write \( a_n \asymp b_n \) to mean \( 0 < \liminf a_n/b_n \leq \limsup a_n/b_n < \infty \).

First let us note that (for \( j \) small enough so that the first order Taylor expansion of \( 1 - \cos(x) \approx x^2 / 2 \) is accurate):

\[ \tilde{a}_j \asymp \frac{j}{n^{1/2} R^{1/2}} \]

So with \( \lambda := \lambda(k) \asymp \frac{n^{1/2} R^{1/2}}{k} \) we would get the equality:

\[ \frac{n^{1/2} R^{1/2}}{k} \asymp \sigma^2 \sum_{j=1}^{k} \frac{j}{n^{1/2} R^{1/2}} \asymp \frac{\sigma^2}{n^{1/2} R^{1/2}} k^2 \]

Solve for \( k \) to get:

\[ k^3 \asymp \frac{nR}{\sigma^2} \text{, i.e. } k_* \asymp \frac{n^{1/3} R^{1/3}}{\sigma^{2/3}} \]

So the optimal \( \lambda_* \) satisfies:

\[ \lambda_* \asymp \sigma^{2/3} n^{1/6} R^{1/6} \]

Finally we get the affine minimax risk:

\[ R_L(\Theta) \asymp \sigma^{4/3} R^{1/3} n^{-2/3} \]

In particular, we recover the rate for the nonparametric regression problem over first-order Sobolev ellipsoids (for fixed \( R \)).

(b) Directly applying Pinsker's theorem (Theorem 4.2), recalling that here we are dealing with a normalized loss, we get that for any \( \varepsilon < 1/2 \) we have (for a universal constant \( c_0 \)) that:

\[ R_M \leq R_L \leq (1 + c_0 \varepsilon) R_M + \frac{c_0}{n} \delta(\varepsilon) \]

Here:

\[ \delta(\varepsilon) = \tilde{a}_{\min}^{-2} \exp(-\Lambda_* \varepsilon^2 / 64) \]

\[ \Lambda_* = \frac{\lambda_* / \sigma^2}{\max_{1 \leq i \leq (n-1)} \tilde{a}_i (1 - \lambda_* \tilde{a}_i)_+} \]

Let us again make these terms more concrete, at least in an asymptotic regime. First:

\[ \tilde{a}_{\min} \asymp \frac{1}{n^{1/2} R^{1/2}} \]
\[
\max_{1 \leq i \leq n-1} \tilde{a}_i (1 - \lambda \tilde{a}_i) \approx k_* \approx \frac{n^{1/3} R^{1/3}}{\sigma^{2/3}}
\]

So:
\[
\Lambda_* \approx \sigma^{-2/3} n^{-1/6} R^{-1/6}
\]

Hence the additive term behaves roughly like (for constants \(C_1, C_2\)):
\[
C_1 R \exp(-C_2 \varepsilon^{-2/3} n^{-1/6} R^{-1/6})
\]

Note that if we can make the additive term \(o(R_L)\), we will get \(R_M/R_L \to 1\). One way to achieve this is (taking \(\varepsilon \to 0\)) to require that \(R = o(R_L)\) or in other words \(R = o(n^{-1}\sigma^2)\).

**Remark:** Instead of considering a regime of shrinking radius, the same result also holds in a regime of \(R >> n\), where the radius \(R\) increases at some appropriate rate compared to the sample size \(n\). Both results are not that surprising given that we know that in the 1-dimensional bounded normal mean model in which \(Z \sim N(\mu, 1), \mu \in [-\tau, \tau]\), the minimax risk and affine minimax risk are the same both in the regime where \(\tau \to 0\) and \(\tau \to \infty\).

### #2: A simple application of Le Cam’s method

Let \(f : \mathbb{R}^d \to \mathbb{R}\) be a differentiable probability density function, and assume that there exists another density function \(g : \mathbb{R}^d \to \mathbb{R}\), and a constant \(M\) such that, for all \(x \in \mathbb{R}^d\)
\[
\|\nabla f(x)\|_2 \leq M g(x) .
\]

We will denote by \(P_\theta\) the probability distribution of \(X = \theta + W\) where \(W \sim f(\cdot)\) is noise with density \(f\).

(a) Prove that, for any \(\theta_1, \theta_2 \in \mathbb{R}^d\),
\[
\|P_{\theta_1} - P_{\theta_2}\|_{TV} \leq \frac{M}{2} \|\theta_1 - \theta_2\|_2 .
\]

(b) Consider the problem of estimating \(\theta \in \Theta \equiv \mathbb{R}^d\) from data \(X \sim P_\theta\) under the square loss \(L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2_2\). Use the previous result to derive a lower bound on the minimax risk.

[Hint: It is sufficient to consider two priors \(Q_1, Q_2\) given by Dirac’s deltas.]

(c) Apply this lower bound to the case of Gaussian noise, namely to the case of \(f\) the density of the Gaussian distribution \(N(0, \sigma^2 I_d)\). How does the result compare with the actual minimax risk?

**Solution:**

(a) We first note that \(P_\theta\) has a density w.r.t. Lebesgue measure, namely \(f_\theta(x) = f(x - \theta)\) (i.e. we are dealing with a location family problem). Therefore:
\[ \|P_{\theta_1} - P_{\theta_2}\|_{TV} = \frac{1}{2} \int_{\mathbb{R}^d} |f_{\theta_1}(x) - f_{\theta_2}(x)| dx \]

\[ = \frac{1}{2} \int_{\mathbb{R}^d} |f(x - \theta_1) - f(x - \theta_2)| dx \]

\[ = \frac{1}{2} \int_{\mathbb{R}^d} \left| \int_0^1 \frac{d}{dt} f(x - \theta_1 + t(\theta_1 - \theta_2)) dt \right| dx \]

\[ \leq \frac{1}{2} \int_{\mathbb{R}^d} \int_0^1 \|\nabla f(x - \theta_1 + t(\theta_1 - \theta_2))\|_{\theta_2 - \theta_1} dt dx \]

\[ \leq \frac{\|\theta_2 - \theta_1\|}{2} \int_{\mathbb{R}^d} \int_0^1 M_\theta(x - \theta_1 + t(\theta_1 - \theta_2)) dt dx \]

\[ = \frac{M}{2} \|\theta_2 - \theta_1\| \quad \text{(by Fubini's theorem)} \]

(b) We will directly apply Le Cam's Lemma. To this end, first note that for any \( a \in \mathbb{R}^d \) we have that:

\[ ||a - \theta_1||^2 + ||a - \theta_2||^2 \geq \frac{1}{2} ||\theta_1 - \theta_2||^2 \]

In other words we may take \( d(\theta_1, \theta_2) = \frac{1}{4} ||\theta_1 - \theta_2||^2 \). We want this to be \( \geq 2\delta \).

Hence let us set \( \delta = \frac{1}{4} ||\theta_1 - \theta_2||^2 \), where we will choose these parameters later.

Then:

\[ 1 - \|P_{\theta_1} - P_{\theta_2}\|_{TV} \geq 1 - \frac{M}{2} ||\theta_1 - \theta_2||_2 \]

Le Cam gives the lower bound:

\[ \geq \frac{||\theta_1 - \theta_2||^2}{8} \left(1 - \frac{M}{2} ||\theta_1 - \theta_2||_2 \right) \]

Plugging in \( ||\theta_1 - \theta_2|| = \frac{4}{3M} \) we get the lower bound \( \frac{2}{27M^2} \).

(c)

\[ ||\nabla f(x)|| = \frac{||x||}{\sigma^2(2\pi\sigma^2)^{d/2}} e^{-1/(2\sigma^2)||x||^2} \]

The r.h.s. has finite integral. Letting \( Z \sim N(0, I_d) \), the desired bound holds with

\[ M^{-1} = \int \frac{||x||}{\sigma^2(2\pi\sigma^2)^{d/2}} e^{-1/(2\sigma^2)||x||^2} dx = \frac{1}{\sigma} E[||Z||] \quad (5) \]

We know \( E[||Z||] \approx \sqrt{d} \), so plugging this into the expression from the previous part gives:

\[ R_\text{b}(Q) \geq \frac{2\sigma^2 E[||Z||]^2}{27} \approx \frac{2\sigma^2}{27d} \]

The minimax risk in the problem is \( R_m = \sigma^2 d \), so our argument recovers the correct dependence in \( \sigma^2 \) but not in \( d \).
### 3: Some properties of distances between distributions

(a) Let \( P = P_1 \times P_2 \times \cdots \times P_n \) and \( Q = Q_1 \times Q_2 \times \cdots \times Q_n \) be two product-form distributions (where, for each \( i \leq n \), \( P_i, Q_i \) are probability measures on the same space \( \mathcal{X}_i \)). Show that

\[
\|P - Q\|_{TV} \leq \sum_{i=1}^{n} \|P_i - Q_i\|_{TV}.
\]

[Hint: Start with \( n = 2 \). It is fine to assume that the \( \mathcal{X}_i \)'s are finite sets.]

(b) Prove that there cannot be a reverse Pinsker inequality. Namely, there does not exist any function \( f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \) with \( f(t) > 0 \) for \( t > 0 \) such that, for any two distributions \( P, Q \),

\[
D(P\|Q) \leq f(\|P - Q\|_{TV}).
\]

(c) Assume that \( P \) and \( Q \) are probability distributions over a finite set \( \mathcal{X} \), with probability mass functions \( p, q \), and assume \( q(x) \geq q_{\text{min}} > 0 \) for all \( x \in \mathcal{X} \). Prove that there exists \( g : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \) with \( g(t, s) > 0 \) for \( t, s > 0 \) such that, for any two probability mass functions \( p, q \), we have

\[
D(P\|Q) \leq g(\|P - Q\|_{TV}, q_{\text{min}}).
\]

We would like the function \( g \) to be such that \( \lim_{z \to 0} g(z; q_{\text{min}}) = 0 \) for any \( q_{\text{min}} > 0 \). Give an explicit expression for the function \( g \).

[Hint: Write \( D(P\|Q) = \mathbb{E}_Q(X \log X - X + 1) \), for \( X = \frac{dP}{dQ} \).]

**Solution:**

(a) Consider the case where \( X_1 \in \mathcal{X}_1, X_2 \in \mathcal{X}_2 \) where \( \mathcal{X}_i \) are finite sets. We will show the result in the case where \( n = 2 \), the general case follows by induction.

\[
\|P - Q\|_{TV} = \frac{1}{2} \sum_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} |p_1(x_1)p_2(x_2) - q_1(x_1)q_2(x_2)|
\]

\[
= \frac{1}{2} \sum_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} |(p_1(x_1) - q_1(x_1))p_2(x_2) + (q_2(x_2) - p_2(x_2))q_1(x_1)|
\]

\[
\leq \frac{1}{2} \sum_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} |p_1(x_1) - q_1(x_1)||p_2(x_2)| + |q_2(x_2) - p_2(x_2)||q_1(x_1)|
\]

\[
= \|P_1 - Q_1\|_{TV} + \|P_2 - Q_2\|_{TV}
\]

(b) To show this it suffices to argue that for any \( v > 0 \), there exist \( P, Q \) with \( \|P - Q\|_{TV} = v \) but \( D(P\|Q) = \infty \). Consider \( \mathcal{X} = \{1, 2, 3\} \). Let \( P = v \delta_1 + (1 - v)\delta_2 \) and \( Q = v \delta_1 + (1 - v)\delta_2 \) so that \( \|P - Q\|_{TV} = v \). But \( D(P\|Q) = \infty \) because \( Q(1) = 0 \) and hence \( \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right) = \infty \).

(c) With \( X = \frac{dP}{dQ} \), and using the hint, we write the KL divergence as
\[ D(P||Q) = \mathbb{E}_Q(X \log X - X + 1) \]
\[ \leq \mathbb{E}_Q(X(X - 1) - X + 1) \]
\[ = \mathbb{E}_Q(X^2) - 2\mathbb{E}_Q(X) + 1 \]
\[ = \mathbb{E}_Q(X^2) - 1 \]
\[ = \sum_{x \in \mathcal{X}} \frac{p(x_i)^2}{q(x_i)^2} q(x_i) - 1 \]
\[ = \sum_{x \in \mathcal{X}} \frac{(p(x_i) - q(x_i))^2}{q(x_i)} \]
\[ \leq \frac{1}{q_{\text{min}}} \sum_{x \in \mathcal{X}} (p(x_i) - q(x_i))^2 \]
\[ \leq \frac{1}{q_{\text{min}}} \left( \sum_{x \in \mathcal{X}} |p(x_i) - q(x_i)| \right)^2 \]
\[ = \frac{4||P - Q||_{TV}^2}{q_{\text{min}}} \]

Thus we may choose \( g(t, s) = \frac{4t^2}{s} \) for \( t, s > 0 \).

References