# ECE 901 HW 7: Sparse De-noising Lower Bound

Gautam Dasarathy

## 1 Minimax Lower bound via hypothesis testing

In this homework, you will prove a general minimax lower bound and apply it to sparse signal estimation. There are 8 problems you need to solve, indicated in **red**, below. Recall the setup we had in class:

1. $\mathcal{F}$ is a class of models.
2. To each $f \in \mathcal{F}$, we have an associated probability distribution $P_f$ on $\mathcal{X} \times \mathcal{Y}$ and the data $Z = (X_1, Y_1, \ldots, X_n, Y_n) \in \mathcal{Z}$ is drawn *iid* from one of these distributions, i.e., $(X_i, Y_i) \overset{iid}{\sim} P_f, i = 1, 2, \ldots, n$.
3. $d(\cdot, \cdot)$ is a semi-metric on $\mathcal{F}$.

What follows next is a minimax lower bounding technique based on a multiple-hypothesis test . Fill in the missing steps in the proof of the following theorem.

**Theorem 1.** *Under the above setup, suppose that for $M \geq 2$ there exists a set of models $\{f_1, \ldots, f_M\} \subset \mathcal{F}$ such that $d(f_i, f_j) \geq 2\epsilon_n$, $i \neq j$, and that $\sum_{i,j=1}^{M} \frac{1}{M^2} D(P_{f_i} \| P_{f_j}) \leq \frac{1}{2} \log M - 1$, then the following minimax lower bound holds for the model class $\{(f, P_f) : f \in \mathcal{F}\}$.*

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_{P_f} \left[ d(\hat{f}_n, f) \right] \geq \frac{\epsilon_n}{2} \tag{1}$$

*Proof.* We showed in class that under the hypotheses given we have that

$$\begin{aligned}\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_{P_f} \left[ d(\hat{f}_n, f) \right] &\geq \epsilon_n \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f(d(\hat{f}_n, f) \geq \epsilon_n) \geq \epsilon_n \inf_{\hat{f}_n} \max_{j \in \{1,\ldots,M\}} P_{f_j}(d(\hat{f}_n, f_j) \geq \epsilon_n) \\ &\geq \epsilon_n \inf_{\hat{h}_n} \max_{j \in \{0,\ldots,M\}} P_{f_j}(\hat{h}_n(Z) \neq j) \geq \epsilon_n \inf_{\hat{h}_n} \frac{1}{M} \sum_{j \in \{1,\ldots,M\}} P_{f_j}(\hat{h}_n(Z) \neq j),\end{aligned}$$

where $\hat{h}_n : \mathcal{Z} \to \{1, \ldots, M\}$ is any hypothesis test. (Note that in class, our subset was indexed as $\{f_0, \ldots, f_M\}$; in this homework, we will start the indexing from 1 instead.)

The rest of the proof will focus on obtaining the following lower bound on the average probability of error

$$\bar{P}_{e,M} \triangleq \inf_{\hat{h}_n} \frac{1}{M} \sum_{j \in \{1,\ldots,M\}} P_{f_j}(\hat{h}_n(Z) \neq j) \geq \frac{1}{2}.$$

1/2 is indeed an arbitrary constant but it will suffice for our purposes and it will keep the presentation tidy.

**1. Let $J \sim \mathrm{unf}(\{1, \ldots, M\})$ be a random variable drawn independently of everything else. Show that if the conditional distribution of $Z$ given $J = j$ is $P_{f_j}$, then $\bar{P}_{e,M} = \mathbb{P}(\hat{h}_n(Z) \neq J)$.**

**Background on Information-Theoretic Concepts.** Before we proceed, we need the following definitions and properties:

1. For a random variable $X \sim P_X$ the entropy is defined as $H(X) \triangleq \mathbb{E}_{P_X}[-\log P_X]$. This quantity measures the "amount of randomness" in the random variable $X$. If $X$ takes values in a finite set $\mathcal{X}$, we have that $0 \leq H(X) \leq \log|\mathcal{X}|$; $H(X) = 0$ if $X$ is deterministic and $H(X) = \log|\mathcal{X}|$ if $X \sim \mathrm{uniform}(\mathcal{X})$, i.e., a uniformly distributed random variable has the most "amount of randomness".

   If $X \sim \mathrm{Bernoulli}(p)$, then we write $h_b(p)$ to denote the entropy $H(X)$, i.e.,
   $h_b(p) \triangleq -p\log p - (1-p)\log(1-p)$.

2. Given a pair of random variables $X, Y \sim P_X P_{Y|X} = P_{XY}$, notice that the entropy satisfies the following:

$$H(X,Y) \triangleq \mathbb{E}_{P_{XY}}[-\log P_{XY}] = \mathbb{E}_{P_X}[-\log P_X] + \mathbb{E}_{P_{XY}}\left[-\log P_{Y|X}\right].$$

   The term $\mathbb{E}_{P_{XY}}\left[-\log P_{Y|X}\right]$ is denoted as $H(Y \mid X)$ and is called the *conditional entropy of* $Y$ *given* $X$. It measures the average amount of randomness "left-over" in $Y$ given the value that $X$ takes. This property can be generalized to show that given $n$ random variables $X_1, \ldots, X_n$, we have the following *chain rule* of entropy

$$H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_1, \ldots, X_{i-1}).$$

3. Given two random variables $X, Y \sim P_{XY}$, the KL divergence between their joint distribution $P_{XY}$ and the product of their marginals $P_X \times P_Y$ is called the *mutual information between* $X$ *and* $Y$ and is denoted as $I(X;Y)$. That is

$$I(X;Y) \triangleq D\left(P_{XY} \| P_X P_Y\right).$$

   (Recall that the KL divergence of $P_2$ from $P_1$ is defined as $D(P_1 \| P_2) = \mathbb{E}_{P_1}[\log(P_1/P_2)]$).
   From the above definitions, it can be seen that $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$. Also, since the KL divergence is always positive, we have that $H(X) - H(X \mid Y) = I(X;Y) \geq 0$ which implies that $H(X) \geq H(X \mid Y)$, i.e., conditioning can never increase the amount of randomness.

We will proceed by defining the following "error event" random variable

$$E = \begin{cases} 1 & \text{if } \hat{h}_n(Z) \neq J \\ 0 & \text{o.w.} \end{cases} \tag{2}$$

Let us expand the conditional entropy $H(E, J \mid \hat{h}_n(Z))$ in two different ways as follows:

$$H(E, J \mid \hat{h}_n(Z)) \overset{(a)}{=} H(J \mid \hat{h}_n(Z)) + H(E \mid J, \hat{h}_n(Z)) \overset{(b)}{=} H(J \mid \hat{h}_n(Z)), \tag{3}$$

**2. Explain why equalities $(a)$ and $(b)$ above are true.**

On the other hand, we have

$$H(E, J \mid \hat{h}_n(Z)) \overset{(a)}{=} H(E \mid \hat{h}_n(Z)) + H(J \mid E, \hat{h}_n(Z)) \overset{(b)}{\leq} h_b(\bar{P}_{e,M}) + H(J \mid E, \hat{h}_n(Z)),$$

**3. Prove equality $(a)$ and inequality $(b)$ above.**

$$\begin{aligned} H(J \mid E, \hat{h}_n(Z)) &= (1 - \bar{P}_{e,M}) H(J \mid E = 0, \hat{h}_n(Z)) + \bar{P}_{e,M} H(J \mid E = 1, \hat{h}_n(Z)) \\ &\leq \bar{P}_{e,M} \log M, \end{aligned} \tag{4}$$

where we have used the fact that $H(J \mid E = 0, \hat{h}_n(Z)) = 0$ since given that no error occurs ($E = 0$), the value of $\hat{h}_n(Z)$ completely determines the value of $J$. Also, given that an error does occur ($E = 1$) and the value $\hat{h}_n(Z)$, $H(J \mid E = 1, \hat{h}_n(Z)) = \log(M - 1)$ since $J$ is uniformly distributed on the set $\{1, \ldots, M\} \setminus \{\hat{h}_n(Z)\}$

Combining (3) and (4), we have

$$1 + \bar{P}_{e,M} \log M \geq H(J \mid \hat{h}_n(Z)) \geq H(J \mid Z),$$

where the last inequality follows since $\hat{h}_n(Z)$ is a deterministic function of $Z$, and thus conditioning on $Z$ provides at least as much information as $\hat{h}_n(Z)$). This is the so-called *data processing inequality.* The entropy (amount of randomness) in $J$ can only be reduced if we are given $Z$ instead of $\hat{h}_n(Z)$, since it might be the case that a deterministic function $\hat{h}_n(Z)$ does not have as much "information" about $J$ as $Z$ does.

Now, using the fact that $I(J; Z) = H(J) - H(J \mid Z) = \log M - H(J \mid Z)$, we have the following inequality (which is a version of Fano's inequality)

$$\bar{P}_{e,M} \geq 1 - \frac{I(J; Z) + 1}{\log M}. \tag{5}$$

Finally, we observe that

$$\begin{aligned}
I(J;Z) &\overset{(a)}{=} D(P_{JZ}\|P_J P_Z) \\
&\overset{(b)}{=} \frac{1}{M}\sum_{i=1}^{M} D\left(P_{f_i}\middle\|\sum_{j=1}^{M} P_{f_j}\right) \\
&\overset{(c)}{\leq} \frac{1}{M^2}\sum_{i,j=1}^{M} D\left(P_{f_i}\|P_{f_j}\right)
\end{aligned}$$

where $(a)$ and $(b)$ follow from the definition, and $(c)$ follows from the fact that the KL divergence is convex.

**4. Show that this last inequality, when used in (5), concludes the proof.** □

# 2 Minimax Lower Bound for Sparse Denoising

Consider the problem of estimating a vector $f \in \mathbb{R}^n$ from noisy measurements $z \in \mathbb{R}^n$ of the form

$$z = f + \epsilon, \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $f \in \mathcal{F}_k \triangleq \{x \in \mathbb{R}^n : \|x\|_0 \leq k\}$, i.e., $f$ has at most $k$ non-zero entries.

We saw in class that, using soft-thresholding, it is possible to achieve a mean-squared error upper bound of

$$(2\log n + 1)\left[1 + \sum_{i=1}^{n} \min\left\{|f(i)|^2, 1\right\}\right] = (2\log n + 1)\,[1 + k],$$

where $f(i)$ is the $i-$th coordinate of the vector $f$. Notice that the quantity on the right is no greater than $4k\log n$ if $n, k \geq 3$.

Your mission, if you choose to accept it, is to prove the following minimax lower bound: There exists a constant $C > 0$ such that

$$\mathcal{E}^* \triangleq \inf_{\hat{f}} \sup_{x\in\mathcal{F}_k} \mathbb{E}\left[\left\|\hat{f}(z) - f\right\|_2^2\right] \geq C\ \max\left\{k\log\left(\frac{n}{k}\right), k\right\}, \tag{7}$$

where $\hat{f}$ is any estimator. This shows that the $\log n$ factor is an unavoidable consequence of having to identify the unknown pattern of sparsity. You can use the following outline. First, prove the following simple lower bound.

**5. Show that $k$ is a lower bound on the mean squared error. (Hint: Suppose we are given the support of $f$.)**

The lower bound of $k$ is loose when $k \ll n$, and a tighter lower bound can be obtained using using multiple hypothesis testing (Theorem 1). The idea behind using the above theorem is to identify a set of $M$ examples that demonstrate the hardness of the problem. This will be your first task:

Assume that $n$ and $k$ are fixed, and for simplicity, assume that $k$ is even and that $k < n/2$. We will consider $k$-sparse vectors with non-zero elements equal to $a_n > 0$ in magnitude (will will choose $a_n$ later to maximize the lower bound). Show there exists a set of such vectors $\{f_1, \ldots, f_M\} \subset \mathcal{F}_k$, with $M = \left(\frac{n}{k}\right)^{k/4}$, satisfying

$$\|f_i - f_j\|_2^2 \geq a_n/2 \quad , \quad \forall i \neq j \ . \tag{8}$$

While it is possible show an explicit construction of this set, you can instead use the following outline which employs the probabilistic method.

**6. Consider a set $\mathcal{U} = \left\{ f \in \left\{ -\sqrt{\frac{a_n}{k}}, 0, +\sqrt{\frac{a_n}{k}} \right\}^n : \|f\|_0 = k \right\}$. Show that the cardinality of this set is $|\mathcal{U}| = \binom{n}{k} 2^k$. Also, show that for a fixed $f \in \mathcal{U}$, the following holds**

$$|\{f' \in \mathcal{U} : \|f - f'\|_0 \leq k/2\}| \leq \binom{n}{k/2} 3^{k/2}.$$

**7. Use (i) to argue that if one picks a size $(n/k)^{k/4}$ set of elements at random from $\mathcal{U}$, then with probability greater than $1/2$, this random set has the property (8). For this you might find the following combinatorial inequality (which hold when $n - k > (0.5)k$) useful:**

$$\frac{\binom{n}{k}}{\binom{n}{k/2}} \geq \left( \frac{n}{k} - \frac{1}{2} \right)^{k/2} .$$

**8. Argue that this shows that there exists at least one set of vectors $\{f_1, \ldots, f_M\}$ such that $M + 1 = \left(\frac{n}{k}\right)^{k/4}$ which satisfies property.**

**9. Conclude the proof of the theorem using Theorem 1. (Hint: Compare the results of (a) and (b) with the hypotheses of Theorem 1; this should tell you the right value of $a_n$.)**