Techniques for bounding minimax risk: the Le Cam and Fano methods

Lekshmi Ramesh

Signal Processing for Communications Lab
IISc, Bangalore

January 30, 2016

Outline

- Estimation problem
  - General set up
  - Minimax, Bayes risk
- Techniques for lower bounding minimax risk
  - Le Cam's method
  - Fano's method
- Connections with Bayes risk

Estimation problem

- Set up: We have a class $\mathcal{P}$ of distributions. For example,

$$\mathcal{P} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}.$$

We observe samples $X_1, \ldots, X_n \in \mathcal{X}$ drawn i.i.d. from some $P \in \mathcal{P}$.
We want to estimate some parameter $\theta \in \Theta$ that depends on the true distribution P.
We design an estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$ of θ based on the observations:

$$\hat{\theta} \equiv \hat{\theta}(X_1, \ldots, X_n)$$

- To assess the quality of a given estimator, we define the risk associated with an estimator.
  - We first define a loss function $\ell : \Theta \times \Theta \to \mathbb{R}_+$
  - The risk of $\hat{\theta}$ is then defined as the expected loss, i.e.,

$$R(\theta, \hat{\theta}) = \mathbb{E}_P \; \ell(\theta, \hat{\theta})$$

- Example: $X \sim \mathcal{N}(\theta, 1)$, estimate θ from a single observation of X.
  (a) Let $\hat{\theta}(X) = X$. Then, the risk function under squared loss is

$$R(\theta, \hat{\theta}) = \mathbb{E}(\theta - X)^2 = var(X) = 1.$$

(b) For $\hat{\theta}(X) = 2$,the risk function under squared loss is

$$R(\theta, \hat{\theta}) = \mathbb{E}(\theta - 2)^2 = (\theta - 2)^2.$$

- Comparing two estimators
  - If $R(\theta, \hat{\theta}_1) > R(\theta, \hat{\theta}_2), \ \forall \theta \in \Theta$, then $\hat{\theta}_2$ is the better estimator
  - In all other cases, we need to quantify the estimators by a number to compare them

- Two ways to do this:
  - Minimax: find the maximum risk $\sup_{\theta \in \Theta} R(\theta, \hat{\theta})$, or
  - Bayes: find the average risk $\mathbb{E}_{\theta \sim \pi} R(\theta, \hat{\theta})$

- Find the estimator that minimizes the Maximum risk/Bayes risk. Such an estimator is called the Minimax estimator/Bayes estimator $\hat{\theta}^*$

- In the minimax case, the risk of $\hat{\theta}^*$ is called the minimax risk, denoted $R_n(\Theta)$:

$$R_n(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_P \ell(\theta, \hat{\theta})$$

Computing the minimax risk

- Computing $R_n(\Theta)$ directly can be difficult, the usual technique is to bound it from above and below
- Upper bound: The maximum risk of an arbitrary estimator will give an upper bound on $R_n(\Theta)$, since

$$R_n(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_P \ell(\theta, \hat{\theta}) \leq \sup_{\theta \in \Theta} \mathbb{E}_P \ell(\theta, \hat{\theta})$$

- Lower bound: We discuss two techniques: Le Cam and Fano

- Example: $X \sim \mathcal{N}(\theta, \sigma^2) = P$, σ^2 known, estimate θ from n i.i.d. observations $X_1, \ldots, X_n$
  Minimax risk under squared error loss is

$$R_n(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_P(\theta - \hat{\theta})^2.$$

  Upper bound: Pick any estimator. Say $\hat{\theta}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then,

$$\begin{aligned} R_n(\Theta) &\leq \sup_{\theta \in \Theta} \mathbb{E}_P(\theta - \frac{1}{n} \sum_{i=1}^{n} X_i)^2 \\ &= \sup_{\theta \in \Theta} var(\frac{1}{n} \sum_{i=1}^{n} X_i) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

- General techniques for finding lower bounds on $R_n(\Theta)$:
    - Le Cam's method
    - Fano's method

Le Cam's method

- Design a test using the estimator.
  Consider the binary hypothesis testing problem with
  $$\mathcal{H}_0 : X \sim P_{\theta_0}$$
  $$\mathcal{H}_1 : X \sim P_{\theta_1},$$
  where $\theta_0, \theta_1 \in \Theta$.
  Given an estimator $\hat{\theta} : \mathcal{X} \to \Theta$, define the test $T : \mathcal{X} \to \{0, 1\}$ with
  $$T(x) = \begin{cases} 0, & \text{if } \|\theta_0 - \hat{\theta}(x)\|_2 \leq \|\theta_1 - \hat{\theta}(x)\|_2, \\ 1, & \text{else.} \end{cases}$$

- To get a lower bound on $R_n(\Theta)$:
  - Find probability of error P_e for test T in terms of max risk of $\hat{\theta}$
  - Lower bound P_e by the probability of error P_e^* of the best test
  - Get a lower bound on max risk of $\hat{\theta}$
- Find P_e for test T

$$P_e = \frac{1}{2} P_{\theta_0}(T(X) \neq 0) + \frac{1}{2} P_{\theta_1}(T(X) \neq 1)$$

Now,

$$\begin{aligned} P_{\theta_0}(T(X) \neq 0) &= P_{\theta_0}(\|\theta_0 - \hat{\theta}(X)\|_2 \geq \|\theta_1 - \hat{\theta}(X)\|_2) \\ &\leq P_{\theta_0}(\|\theta_0 - \hat{\theta}(X)\|_2 \geq \frac{\|\theta_0 - \theta_1\|_2}{2}) \\ &\leq 4\frac{\mathbb{E}_{P_{\theta_0}}\|\theta_0 - \hat{\theta}(X)\|_2^2}{\|\theta_0 - \theta_1\|_2^2} \end{aligned}$$

Similarly,

$$P_{\theta_1}(T(X) \neq 1) \leq 4\frac{\mathbb{E}_{P_{\theta_1}}\|\theta_1 - \hat{\theta}(X)\|_2^2}{\|\theta_0 - \theta_1\|_2^2}$$

Thus,

$$P_e^* \leq P_e \leq \frac{4}{\|\theta_0 - \theta_1\|_2^2}\left(\frac{1}{2}\mathbb{E}_{P_{\theta_0}}\|\theta_0 - \hat{\theta}(X)\|_2^2 + \frac{1}{2}\mathbb{E}_{P_{\theta_1}}\|\theta_1 - \hat{\theta}(X\|_2^2\right)$$
$$\leq \frac{4}{\|\theta_0 - \theta_1\|_2^2}\max_{\{\theta_0,\theta_1\}}\mathbb{E}_{P_\theta}\|\theta - \hat{\theta}(X)\|_2^2$$
$$\leq \frac{4}{\|\theta_0 - \theta_1\|_2^2}\max_{\theta\in\Theta}\mathbb{E}_{P_\theta}\|\theta - \hat{\theta}(X)\|_2^2$$

$$\implies \max_{\theta\in\Theta}\mathbb{E}_{P_\theta}\|\theta - \hat{\theta}(X)\|_2^2 \geq \frac{\|\theta_0 - \theta_1\|_2^2}{4}P_e^*$$

Lower bound on P_e^*

- Consider the binary hypothesis testing problem:
  Random variable X taking values in $\mathcal{X}$
  Null hypothesis: $\mathcal{H}_0 : X \sim P$
  Alternative hypothesis: $\mathcal{H}_1 : X \sim Q$
  Acceptance region $A \subseteq \mathcal{X}$ with

$$A = \{x \in \mathcal{X} : \text{declare } \mathcal{H}_0\}$$

- Probability of error

$$P_e = \frac{1}{2}P(A^c) + \frac{1}{2}Q(A)$$

$$
\begin{aligned}
P_e^* &= \min_{A \subseteq \mathcal{X}} \left(\frac{1}{2}(1 - P(A)) + \frac{1}{2} Q(A) \right) \\
&= \frac{1}{2} - \frac{1}{2} \max_{A \subseteq \mathcal{X}} (P(A) - Q(A)) \\
&= \frac{1}{2}(1 - \|P - Q\|_{TV}) \\
&\geq \frac{1}{2} \left(1 - \sqrt{\frac{D(P||Q)}{8 \log e}} \right) \qquad \text{(Pinsker's inequality)}
\end{aligned}
$$

Thus, we have

$$
\max_{\theta \in \Theta} \mathbb{E}_{P_\theta} \|\theta - \hat{\theta}(X)\|_2^2 \geq \frac{\|\theta_0 - \theta_1\|_2^2}{8} \left(1 - \sqrt{\frac{D(P_{\theta_0}||P_{\theta_1})}{8 \log e}} \right)
$$

Note that there is no dependence on the dimension of the parameter space Θ

Fano's method

- As before, we design a test using the estimator.
  Consider the m-ary hypothesis testing problem with

$$\mathcal{H}_i : X \sim P_{\theta_i}, \quad 1 \le i \le m,$$

where $\theta_1, \ldots, \theta_m$ are chosen such that

$$\min_{i,j} \|\theta_i - \theta_j\|_2 = \alpha.$$

Given an estimator $\hat{\theta} : \mathcal{X} \to \Theta$, consider the following test:

$$T(x) = \arg\min_{1 \le i \le m} \|\hat{\theta}(x) - \theta_i\|_2$$

Now bound the P_e for this test:

$$
\begin{aligned}
P_{\theta_i}(T(X) \neq i) &= P_{\theta_i}(i \neq \arg\min_j \|\hat{\theta}(X) - \theta_j\|_2) \\
&\leq P_{\theta_i}(\|\hat{\theta}(X) - \theta_i\|_2 \geq \frac{\alpha}{2}) \\
&\leq \frac{4}{\alpha^2} \mathbb{E}_{P_{\theta_i}} \|\hat{\theta}(X) - \theta_i\|_2^2, \qquad 1 \leq i \leq m.
\end{aligned}
$$

$$
\begin{aligned}
P_e^* \le P_e &= \sum_{i=1}^{m} \frac{1}{m} P_{\theta_i}(T(X) \neq i) \\
&\le \frac{1}{m} \sum_{i=1}^{m} 4 \frac{\mathbb{E}_{P_{\theta_i}} \|\hat{\theta}(X) - \theta_i\|_2^2}{\alpha^2} \\
&\le \frac{4}{\alpha^2} \max_{\theta \in \Theta} \mathbb{E}_{P_\theta} \|\hat{\theta}(X) - \theta\|_2^2
\end{aligned}
$$

Lower bound on P_e^*:

$$
\begin{aligned}
P_e^* &= \min_T \frac{1}{m} \sum_{i=1}^{m} P(T(X) \neq i) \\
&\ge 1 - \frac{(I(M;X) + 1)}{\log m} \\
&\ge 1 - \frac{(\max_{i,j} D(P_{\theta_i} || P_{\theta_j}) + 1)}{\log m}.
\end{aligned}
$$

Thus,

$$\max_{\theta \in \Theta} \mathbb{E}_{P_\theta} \|\hat{\theta}(X) - \theta\|_2^2 \geq \frac{\alpha^2}{4} \left(1 - \frac{(\max_{i,j} D(P_{\theta_i} || P_{\theta_j}) + 1)}{\log m} \right)$$

- Bound tighter for larger m. Find maximum number of points that can be packed in Θ such that they are separated by at least α.

Connections with Bayes risk

- Recall: Bayes risk

$$R(\pi) = \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi} R(\theta, \hat{\theta})$$

For any prior π

$$\begin{aligned}
\mathbb{E}_{\theta\sim\pi} R(\theta, \hat{\theta}) &= \int R(\theta, \hat{\theta}) \pi d\theta \\
&\leq \int \sup_{\theta} R(\theta, \hat{\theta}) \pi d\theta \\
&\leq \sup_{\theta} R(\theta, \hat{\theta})
\end{aligned}$$

Minimizing over all estimators,

$$\inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi} R(\theta, \hat{\theta}) \le \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}) = R_n(\Theta).$$

- That is, the Bayes risk of any prior gives a lower bound on the minimax risk.
  Maximizing over all priors gives a tighter bound:

$$R_n(\Theta) \ge \sup_{\pi} \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi} R(\theta, \hat{\theta}).$$

References

Chen, X., A. Guntuboyina, and Y. Zhang. "On Bayes risk lower bounds". In: *arXiv:1410.0503v3* (Oct, 2015).

Duchi, J. *Minimax lower bounds: the Fano and Le Cam methods*. URL: https://web.stanford.edu/class/stats311/Lectures/.

Lehmann, E. L. and G. Casella. *Theory of Point Estimation (Springer Texts in Statistics)*. 2nd. Springer, Aug. 1998.

Wasserman, L. *Minimax theory*. URL: http://www.stat.cmu.edu/~larry/=sml/Minimax.pdf.