CS861: Theoretical Foundations of Machine Learning Lecture 12 - 10/02/2023

University of Wi[illegible] Fall 2023

[illegible]2: Fano's method

Lecturer: Kirthe[illegible] *Scribed by: Yuya Shimizu, Keran Chen*

Disclaimer: *These n[illegible]jected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we will prove `Fano's inequality` and apply it to show `Fano's methods`. In the previous lecture, we confirmed that Le Cam's method is only good for the single parameter to derive a lower bound of the minimax risk. While Le Cam's method relies on the Neyman-Pearson test for binary hypotheses, Fano's method considers multiple hypotheses, and it can cover more general situations. We will also introduce a method for `constructing alternatives` for Fano's method using packing numbers.

1 Fano's inequality

Before showing Fano's method, we will prove Fano's inequality using properties in information theory.

Theorem 1. *(Fano's inequality) Let X be a discrete random variable with a finite support $\mathcal{X}$. Let $X, Y, \hat{X}$ form a Markov chain $X \to Y \to \hat{X}$. Denote $P_e = \mathbb{P}(\hat{X} \neq X)$ and*

$$h(P_e) = -P_e \log(P_e) - (1-P_e)\log(1-P_e).$$

Then,

$$H(X \mid Y) \leq H(X \mid \hat{X}) \leq P_e \log(|\mathcal{X}|) + h(P_e). \tag{1}$$

Hence,

$$\mathbb{P}(\hat{X} \neq X) \geq \frac{H(X \mid Y) - \log(2)}{\log(|\mathcal{X}|)}.$$

Remark (Interpretation) We can interpret P_e as a probability of error, and X defines a prior on the alternatives $\{P_1, \cdots, P_{|\mathcal{X}|}\}$. We want to guess X (via $\hat{X}$). If Y uniquely identifies X, we have $H(X \mid Y) = 0$, i.e., no information on X is left after observing Y. Fano's inequality quantifies $\mathbb{P}(X \neq \hat{X})$ in terms of $H(X \mid Y)$. We do not require any restriction on the support of Y. It is instructive to assume that $\hat{X}$ has the same support as X, although this is strictly not necessary.

Proof Let $\mathbb{E} = \mathbf{1}\left\{X \neq \hat{X}\right\}$ (thus, if $E = 1$, there is an error). Using the chain rule in two different ways,

$$\begin{aligned} H(E, X \mid \hat{X}) &= H(X \mid \hat{X}) + H(E \mid X, \hat{X}) \\ &= H(E \mid \hat{X}) + H(X \mid E, \hat{X}). \end{aligned} \tag{2}$$

Here, since E is a function of $X, \hat{X}$,

$$H(E \mid X, \hat{X}) = 0. \tag{3}$$

Also, since conditioning reduces entropy,

$$H(E \mid \hat{X}) \leq H(E) = h(P_e), \tag{4}$$

where the equality follows by the definition of $h(\cdot)$. By the definition of the conditional entropy,

$$H(X \mid \hat{X}) \; [\text{illegible}] \; \underbrace{H(X \mid \hat{X}, E=0)}_{=0} + \underbrace{\mathbb{P}(E=1)}_{=P_e} \underbrace{H(X \mid \hat{X}, E=1)}_{\leq H(X)}$$

$$[\text{illegible}] \; P_e \log(|\mathcal{X}|), \tag{5}$$

where $H(X \mid \hat{X}, E=0)$ [illegible] plies $X = \hat{X}$ and the last inequality follows from the property of discrete X. Thus, ([illegible] hat $H(X \mid \hat{X}) \leq h(P_e) + P_e \log(|\mathcal{X}|)$. We have proved the second inequality in the [illegible]

Next, we will show [illegible] he theorem. By the data processing inequality, $I(X, \hat{X}) \leq I(X, Y)$. Since we also have $I(X, \hat{X}) = H(X) - H(X \mid \hat{X})$ and $I(X, Y) = H(X) - H(X \mid Y)$, the next inequality is implied:

$$H(X \mid Y) \leq H(X \mid \hat{X}).$$

Since $h(P_e)$ is a entropy for a binary random variable, $h(P_e) \leq \log(2)$. This and (1) together imply the third inequality in the theorem:

$$P_e \geq \frac{H(X \mid Y) - \log(2)}{\log(|\mathcal{X}|)}.$$

□

2 Fano's method

Given Fano's inequality, we will derive two different types of lower bounds of the minimax risk. The first bound is called the global Fano's method since it contains a Kullback-Leibler divergence between one distribution and the mixture of all other distributions in the alternatives. The second bound is called the local Fano's method since it contains only the KL divergences between the alternatives.

Theorem 2. *Let S be drawn (not necessary i.i.d.) from some joint distribution $P \in \mathcal{P}$. Let $\{P_1, \cdots, P_N\} \subseteq \mathcal{P}$, and denote $\bar{P} = \frac{1}{N}\sum_{j=1}^{N} P_j$ (an equally weighted mixture distribution). Denote the minimax risk*

$$R^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi \circ \rho\left(\theta(P), \hat{\theta}(S)\right)\right].$$

Let $\delta = \min_{j \neq k} \rho\left(\theta\left(P_j\right), \theta\left(P_k\right)\right)$. Then, the following statements are true:

(I) (Global Fano's method)

$$R^* \geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}\left(P_j, \bar{P}\right) + \log(2)}{\log(N)}\right).$$

(II) (Local Fano's method)

$$R^* \geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{1}{N^2}\sum_{1 \leq j,k \leq N} \mathrm{KL}\left(P_j, P_k\right) + \log(2)}{\log(N)}\right).$$

Remark The global Fano's method is stronger (tighter), but the local Fano's method is easier to apply since $\mathrm{KL}(P_j, P_k)$ is easier to compute than $\mathrm{KL}\left(P_j, \bar{P}\right)$.

Proof Define a uniform prior on $\{P_1, \ldots, P_N\}$ as follows:

$$\mathbb{P}(V = j) = \frac{1}{N}, \text{ for } j = 1, \ldots, N.$$

Given $V = j$, let S be sampled from P_j. Then, the joint distribution of (V, S) is

$$
\begin{aligned}
\ldots j) &= \mathbb{P}(S \in A \mid V = j)\mathbb{P}(V = j) \\
&= \frac{1}{N} P_j(A).
\end{aligned}
$$

By our "basic" theo … estimation to testing,

$$
\begin{aligned}
\ldots &\ldots \Big) \inf_{\psi} \max_{j \in [N]} P_j\left(\psi(S) \neq j\right) \\
&\geq \Phi\left(\frac{\ldots}{2}\right) \inf_{\psi} \mathbb{P}_{V,S}(\psi(S) \neq V), \qquad (6)
\end{aligned}
$$

where the second inequality follows since the maximum value is greater than or equal to the average value.

By Fano's inequality (on) the Markov chain $V \to S \to \psi(S)$,

$$
\begin{aligned}
\mathbb{P}_{V,S}(\psi(S) \neq V) &\geq \frac{H(V \mid S) - \log(2)}{\log(N)} \\
&= \frac{H(V) - I(V,S) - \log(2)}{\log(N)} \\
&= 1 - \frac{I(V,S) + \log(2)}{\log(N)}, \qquad (7)
\end{aligned}
$$

where the second inequality follows from $I(V,S) = H(V) - H(V \mid S)$ and the last equality follows $H(V) = \log(N)$ since V follows from a uniform distribution.

$$
\begin{aligned}
I(V,S) &= I(S,V) \\
&= \mathbb{E}_{S,V}\left[\log\left(\frac{P(S,V)}{P(S)\cdot P(V)}\right)\right] \\
&= \sum_{j=1}^{N} \mathbb{P}(V=j) \int P_j(s) \log\left(\frac{P_j(s)\cdot \mathbb{P}(V=j)}{\bar{P}(s)\cdot \mathbb{P}(V=j)}\right) ds \\
&= \frac{1}{N}\sum_{j=1}^{N} \int P_j(s) \log\left(\frac{P_j(s)}{\bar{P}(s)}\right) ds \\
&= \frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}\left(P_j, \bar{P}\right), \qquad (8)
\end{aligned}
$$

where the first equality follows from the symmetry of mutual information, the second equality follows from the definition of mutual information, the third equality follows from the law of iterated expectations, the fourth equality follows from $\mathbb{P}(V = j) = 1/N$, and the last equality follows from the definition of the Kullback-Leibler divergence. (6)-(8) together imply the global Fano's method.

Moreover, we can further bound (8) as

$$
\frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}\left(P_j, \bar{P}\right) \leq \frac{1}{N^2} \sum_{1 \leq j,k \leq N} \mathrm{KL}\left(P_j, P_k\right), \qquad (9)
$$

since $\mathrm{KL}(Q, P) = \mathbb{E}_Q\left[\log\left(\frac{Q(x)}{P(x)}\right)\right]$ is convex in the second argument, and Jensen's inequality implies that $\mathrm{KL}\left(Q, \frac{1}{N}\sum_{k=1}^{N} P_k\right) \leq \frac{1}{N}\sum_{k=1}^{N} \mathrm{KL}\left(Q, P_k\right)$. (6), (7), and (9) together imply the local Fano's method. $\square$

Remark Similarly to Le Cam's method, Fano's methods also give a lower bound of an average error in the proof. This can be [illegible]bservation:

$$\mathbb{P}_{V,[illegible]}(S) \neq j)\mathbb{P}(V=j) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{P}(\psi(S) \neq j).$$

We will next state t[illegible]f the local Fano method which is convenient to apply.

Corollary 1. *(The local Fano method for iid data) If S contains n i.i.d samples from distribution $P \in \mathcal{P}$, then $S \sim P^n$. Let $\{P_1, \cdots, P_N\} \subseteq \mathcal{P}$. Let $N \geq 16$. Suppose $\underset{j,k}{max} KL\left(P_j, P_k\right) \leq \frac{\log(N)}{4n}$. Define $\delta = \min_{j \neq k} \rho\left(\theta\left(P_j\right), \theta\left(P_k\right)\right)$. Then $R_n^* \geq \frac{1}{2}\Phi(\frac{\delta}{2})$.*

Proof Applying the local Fano method for i.i.d data (product distributions) yields:

$$R^* \geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{1}{N^2}\sum_{1\leq j,k\leq N}\mathrm{KL}\left(P_j^n, P_k^n\right) + \log(2)}{\log(N)}\right)$$

By the fact that $KL\left(P_j^n, P_k^n\right) = nKL\left(P_j, P_k\right)$ and $\underset{j,k}{max} KL\left(P_j, P_k\right) \leq \frac{\log(N)}{4n}$,

$$R^* \geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{1}{N^2}\sum_{1\leq j,k\leq N}\mathrm{KL}\left(P_j^n, P_k^n\right) + \log(2)}{\log(N)}\right) \tag{10}$$

$$= \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{n}{N^2}\sum_{1\leq j,k\leq N}\mathrm{KL}\left(P_j, P_k\right) + \log(2)}{\log(N)}\right) \tag{11}$$

$$\geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{n}{N^2}\left(\frac{log(N)}{4n}N^2\right) + \log(2)}{\log(N)}\right) \tag{12}$$

$$\geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{1}{4} - \frac{\log(2)}{\log(16)}\right) \tag{13}$$

$$= \Phi\left(\frac{\delta}{2}\right)\frac{1}{2} \tag{14}$$

The last inequality is because $N \geq 16$. □

Remark For Global Fano, $\bar{P} = \frac{1}{N}\sum_j P_j^n$ and we do not have a simple form as for the local Fano. This is one reason why the local Fano is easier to apply.

3 Constructing alternatives

When constructing alternatives for Fano's method (and other methods), we need to be careful in our construction of $\{P_1, \ldots, P_N\}$. In particular, we need $\rho\left(\theta\left(P_j\right), \theta\left(P_k\right)\right)$ to be large but $KL\left(P_j, P_k\right)$ to be small. If $\delta = \min_{j\neq k} \rho\left(\theta\left(P_j\right), \theta\left(P_k\right)\right)$ is too small, then the lower bound of the minimax risk will be small. Next, we will discuss two common tools that are used in the construction.

3.1 Method 1: Tight Packings

Definition 1. *Packing*

- An ϵ-packing of a … metric ρ is a set $\{x_1, \dots, x_n\} \in \chi$ s.t $\rho(x_i, x_j) \geq \epsilon, \forall i \neq j$.
- The ϵ-packing nu… ize of the largest ϵ-packing of χ.
- A packing with a … a maximal packing.

Next, we will illustr… kings via a d-dimensional normal mean estimation problem. The following fact will …

Lemma 1. *For an L_2 ball of radius r in R^d, $\chi = \left\{x \in R^d : \|x\|_2 \leq r\right\}$, we have*

$$M(\epsilon, \chi, \rho) \geq \left(\frac{r}{\epsilon}\right)^d.$$

Example 3. Normal mean estimation in R^d.

$\mathcal{P}$ is the family $\left\{N(\mu, \Sigma), \mu \in R^d, \Sigma \preceq \sigma^2 I\right\}$. Let $S = \{X_1, \dots, X_n\}$ be n i.i.d samples from $P \in \mathcal{P}$. The minimax risk:

$$R^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi \circ \rho\left(\theta(P), \hat{\theta}(S)\right)\right].$$

Let $\Phi \circ \rho\left(\theta(P), \hat{\theta}(S)\right) = \|\theta_1 - \theta_2\|_2^2$, $\theta = E_{X \sim P}[X]$.

`Upper bound of minimax risk`: Consider the estimator $\hat{\theta}(S) = \frac{1}{n}\sum_{i=1}^n X_i$. As $Var(x_{ij}) \leq \sigma^2$:

$$R_n^* \leq E\left[\left\|\hat{\theta}(S) - \theta\right\|_2^2\right] = E\left[\sum_{j=1}^d \left(\frac{1}{n}\sum_{i=1}^n X_{ij} - \theta_j\right)^2\right] = \sum_{j=1}^d E\left(\frac{1}{n}\sum_{i=1}^n X_{ij} - \theta_j\right)^2 = \sum_{j=1}^d \frac{Var(x_{ij})}{n} \leq \frac{\sigma^2 d}{n}.$$

`Lower bound of minimax risk`: Let U be a δ-packing of the L_2 ball of radius 2δ. Consider a subset of $\mathcal{P}$: $\mathcal{P}' = \left\{N(\mu, \sigma^2 I), \mu \in U\right\}$. So according to lemma 1, $|\mathcal{P}'| = |U| \geq \left(\frac{2\delta}{\delta}\right)^d = 2^d$. Moreover, by the definition of a δ-packing, we have

$$\min_{P, P' \in \mathcal{P}'} \|\theta(P) - \theta(P')\| = \min \mu, \mu' \in U\|\mu - \mu'\| \geq \delta.$$

Lemma 2. *$N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ are two multidimensional normal random variables, then* $KL(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) = \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$.

Let P_j, P_k be two distributions in $\mathcal{P}'$: $P_j = N\left(\mu_j, \sigma^2 I\right)$, $P_k = N\left(\mu_k, \sigma^2 I\right)$, $\mu_j, \mu_k \in U$. So according to the definition of U, $\|\mu_j - \mu_k\|_2 \leq 4\delta$, and $\|\mu_j - \mu_k\|_2 \geq \delta$. Then

$$KL(P_j, P_k) = \frac{\|\mu_j - \mu_k\|_2^2}{2\sigma^2} \leq \frac{(4\delta)^2}{2\sigma^2} = \frac{8\delta^2}{\sigma^2}.$$

The first equality is by Lemma 2, the inequality is because of $\|\mu_j - \mu_k\|_2 \leq 4\delta$. To apply the Local Fano method (Corollary 1), we want $KL \leq \frac{\log(|\mathcal{P}'|)}{n}$, choose $\delta = \sigma\sqrt{\frac{d\log(2)}{8n}}$ so that $\frac{8\delta^2}{\sigma^2} \leq \frac{\log(2^d)}{n} \leq \frac{d\log(2)}{n}$. Then, by the local Fano method,

$$R_n^* \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = \frac{\log(2)}{64}\frac{\sigma^2 d}{n}.$$