CS861: Theoretical Foundations of Machine Learning Lecture 12 - 09/30/2024

University of Wisconsin–Madison, Fall 2024

Lecture 12: Constructing Alternatives for Fano's Method

Lecturer: Kirthevasan Kandasamy *Scribed by: Alejandro Calle-Saldarriaga, Guy Zamir*

In this lecture we will give a proof for Fano's method, using Fano's inequality which was proved last lecture. 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. Unfortunately, proposing multiple alternatives for Fano's method can be more difficult than producing two alternatives for Le Cam's method, so in this class, we also give a way of producing multiple hypotheses via the Varshamov-Gilbert Lemma.

1 Fano's Method

In this section we present two variants of Fano's method, which we will prove and comment on.

Theorem 1 (Fano's method)**.** *Let S be drawn from some $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subseteq \mathcal{P}$. Denote the minimax risk as*

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

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

1. **Global Fano's method**. Let $\bar{P} = \frac{1}{N}\sum_{i=1}^N P_i$, an equally weighted mixture of the alternatives. Then

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

2. **Local Fano's method**.

$$R^* \geq \Phi(\delta/2)\left(1 - \frac{\frac{1}{N^2}\sum_{j,h} \mathrm{KL}(P_j, P_h) + \log(2)}{\log(N)}\right)$$

Between the two variants of Fano's method, the global version is stronger (tighter), but the local version is easier to compute since $\mathrm{KL}(P_j, P_h)$ is easier to compute than $\mathrm{KL}(P_j, \bar{P})$, since it can be hard to compute the density of a mixture $\bar{P}$. As with Le Cam's method, we would like to obtain the strongest possible lower bound, so we want δ to be as big as possible – that is, for our alternatives $P_1, \ldots, P_N$ to be well separated. On the other hand, we also want to minimize the KL divergence between them, which requires the alternatives to be statistically indistinguishable. Now we prove the result.

Proof [Fano's method]

Let us consider the following data generating process. First, define a uniform prior on $\{P_1, \ldots, P_n\}$, with V denoting the selection of alternative, that is, given $V = j$, S is sampled from P_j, i.e.

$$\mathbb{P}(V = j) = \frac{1}{N}$$

and by the law of total probability, we can write the marginal distribution of the data as

$$\mathbb{P}(S \in A) = \sum_{j=1}^{N} \mathbb{P}(S \in A | V = j)\mathbb{P}(V = j) = \frac{1}{N}\sum_{j=1}^{N} P_j(A) = \bar{P}(A) \tag{1}$$

since $\mathbb{P}_{S\sim P_j}(S \in A) = P_j$. Now, by the reduction to testing theorem, we can write

$$\begin{aligned} R^* &\geq \Phi(\delta/2) \inf_{\psi} \max_{j\in[N]} P_j(\psi(S) \neq j) \\ &\geq \Phi(\delta/2) \inf_{\psi} \mathbb{P}_{S,V}(\psi(S) \neq V), \end{aligned}$$

where we used the idea that max is greater than average. We can write this more explicitly as

$$\begin{aligned} \mathbb{P}_{S,V}(\psi(S) \neq V) &= \sum_{i=1}^{N} \mathbb{P}(\psi(S) \neq V | V = j)\mathbb{P}(V = j) \\ &= \frac{1}{N}\sum_{i=1}^{N} P_j(\psi(S) \neq j) \\ &\leq \max_{j\in[N]} P_j(\psi(S) \neq j) \end{aligned}$$

for any test ψ. Now, since we can form the Markov chain $V \to S \to \psi$, we can use Fano's inequality to get

$$\mathbb{P}(\psi(S) \neq V) \geq \frac{H(V|S) - \log(2)}{\log(N)},$$

so by the identity $I(X, Y) = H(X) - H(X|Y)$,

$$\mathbb{P}(\psi(S) \neq V) \geq \frac{H(V) - I(V|S) - \log(2)}{\log(N)} = 1 - \frac{I(V, S) + \log(2)}{\log(N)}$$

because $H(V) = \log(N)$ for the discrete uniform distribution. Then, we have that

$$R^* \geq \Phi(\delta/2)\left(1 - \frac{I(V, S) + \log(2)}{\log(N)}\right). \tag{2}$$

It remains to relate $I(V, S)$ and KL divergences. Let p_j be the density of P_j, $\bar{p}$ the density of $\bar{P}$, and the p the joint density of S, V. Then

$$I(S, V) = \mathbb{E}_{S,V}\left[\log\left(\frac{p(s, v)}{p(s)p(v)}\right)\right] = \sum_{j=1}^{N}\int_S \mathbb{P}(S = s|V = v)\mathbb{P}(V = j)\log\left(\frac{\mathbb{P}(S = s|V = j)\mathbb{P}(V = j)}{\mathbb{P}(S = s)\mathbb{P}(V = j)}\right),$$

where here we use generalized integrals to compute the expectations (since S is continuous we integrate with respect to the Lebesegue measure, since V is discrete we integrate with respect to the counting measure, i.e. a sum). Now, by noticing that $\mathbb{P}(S = s|V = v) = \mathbb{P}_j(S)$, $\mathbb{P}(V = j) = \frac{1}{N}$, and $\mathbb{P}(S = s) = \bar{\mathbb{P}}(S)$ by using the marginal distribution derivation computed in Equation 1, and simplifying, we get

$$I(V, S) = \frac{1}{N}\sum_{j=1}^{N}\int_S p_j(S)\log\left(\frac{p_j(S)}{\bar{p}(S)}\right) = \frac{1}{N}\sum_{j=1}^{N}\mathrm{KL}(P_j, \bar{P}), \tag{3}$$

and we also claim

$$\frac{1}{N}\sum_{j=1}^{N}\mathrm{KL}(P_j, \bar{P}) \leq \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathrm{KL}(P_j, P_k). \tag{4}$$

Notice that by combining Equations 2 and 3 we get the global Fano method, while combining Equations 2, 3 and 4 we get the local Fano method.

What is left to prove is then the claim in Equation 4, so write

$$\mathrm{KL}(P_j, \bar{P}) = \mathbb{E}_{S\sim P_j}\left[\log\left(\frac{p_j(S)}{\frac{1}{N}\sum_{j=1}^{N} p_j(S)}\right)\right].$$

Noting that $\log(1/x)$ is convex, and that $\frac{1}{N}\sum_{j=1}^{N} P_j(S)$ is an average over the discrete uniform distribution over $\{1,\dots,N\}$, we apply Jensen's inequality to obtain

$$\mathrm{KL}(P_j, \bar{P}) \leq \mathbb{E}_{S\sim P_j}\left[\sum_{i=1}^{N}\log\left(\frac{p_i(S)}{p_j(S)}\right)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}(P_j, P_i).$$

Then

$$\frac{1}{N}\sum_{j=1}^{N}\mathrm{KL}(P_j, \bar{P}) \leq \frac{1}{N}\frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}(P_j, P_i) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{KL}(P_j, P_i),$$

which proves the claim and completes the proof.

□

With Fano's method, we have an alternative to Le Cam's method to produce lower bounds for our minimax risk. It is a convenient tool to have if just considering two alternatives is not enough for our estimation problem at hand. We now present a convenient Corollary that will give us a convenient way of applying the local Fano method when the data is i.i.d.

Corollary 1 (Local Fano for i.i.d. data)**.** *Let S be an i.i.d. dataset of size n drawn from some distribution $P \in \mathcal{P}$. Let $\{P_1,\dots,P_N\}$ be our set of alternaties such that $N \geq 16$. Let $\delta \geq \min_{j\neq k}\rho(\theta(P_j),\theta(P_k))$ and let $\mathrm{KL}(P_j, P_k) \leq \frac{\log(N)}{4n}$. Then,*

$$R^* \geq \frac{1}{2}\Phi(\delta/2)$$

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

$$R_n^* \geq \Phi(\delta/2)\left(1 - \frac{\frac{1}{N^2}\sum_{j,h}^{N}\mathrm{KL}(P_j^n, P_h^n) + \log(2)}{\log(N)}\right).$$

Note that $\mathrm{KL}(P_j, P_k) \leq \frac{log(N)}{4n}$ implies $\mathrm{KL}(P_j^n, P_k^n) \leq \frac{log(N)}{4}$ by the decomposition of KL divergence for i.i.d. data, so we get that

$$\begin{aligned}
R_n^* &\geq \Phi(\delta/2)\left(1 - \frac{\frac{1}{N^2}\sum_{j,h}\frac{log(N)}{4} - \log(2)}{\log(N)}\right) \\
&\geq \Phi(\delta/2)\left(1 - \frac{log(N)/4}{log(N)} - \frac{\log(2)}{\log(N)}\right) \\
&\geq \Phi(\delta/2)\left(1 - 1/4 - \frac{\log(2)}{\log(16)}\right) \\
&= \frac{1}{2}\Phi(\delta/2)
\end{aligned}$$

where we used the condition $N \geq 16$.

□

Note that the condition $N \geq 16$ was not really necessary, it is just a technical condition that allows us to get an expression with "nice" looking constants. For smaller N, we can get a similar bound with a different constant multiplying $\Phi(\delta/2)$. This result is what might make local Fano appealing, as we have this simple form of the bound that might be easier to apply. However, picking $\delta \geq \min_{j \neq k} \rho(\theta(P_j), \theta(P_k))$ such that $\max_{j \neq k} \mathrm{KL}(P_j, P_j)$ is small can be harder than in the binary case with Le Cam's method, so we have to be a little more careful while constructing alternatives. Some guidelines on how we might design this alternatives are given in the next section.

2 Constructing Alternatives

Fano's method provides us with a stronger tool to prove lower bounds, but there remains difficulty in selecting a set of "alternatives" $\{P_1, \ldots, P_N\}$. In particular, we need $\delta = \min_{j \neq k} \rho\left(\theta\left(P_j\right), \theta\left(P_k\right)\right)$ to be sufficiently large while simultaneously keeping $\max_{j \neq k} \mathrm{KL}\left(P_j, P_k\right)$ sufficiently small. If we cannot achieve both, then the lower bound of the minimax risk given by Fano's method will be small.

We will discuss two common tools that are used in the construction: the Varshamov-Gilbert Lemma and tight packings. We start with the Varshomov-Gilbert Lemma, and we will cover tight packings in the next lecture.

3 The Varshamov–Gilbert Lemma

To get a lower bound for the minimax risk, one may find it convenient to consider alternatives indexed by a hypercube:

$$\mathcal{P}' = \left\{P_\omega : \omega = (\omega_1, \ldots, \omega_d) \in \{0,1\}^d\right\} \subset \mathcal{P}.$$

However, in a high-dimensional hypercube, the minimum distance will likely be too small relative to the largest KL. To illustrate this problem, we examine the following example.

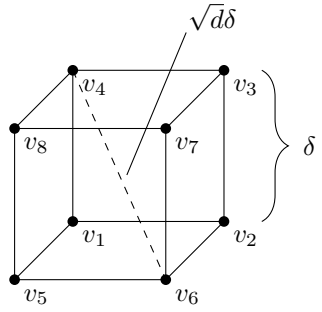

Figure 1: A hypercube that we will use to generate our alternatives by removing a few vertices from the cube.

Example 2. (Normal mean estimation in $\mathbb{R}^d$)

Let P be drawn from the family $\mathcal{P} = \left\{ \left(\mu, \sigma I_d\right), \mu \in R^d \right\}$, where σ^2 is known. Let $S = \{X_1, \ldots, X_n\}$ be n i.i.d samples from $P \in \mathcal{P}$. Additionally, define $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ by $\rho(x, y) = |x - y|$ and $\Phi(x) = x^2$. We want to estimate the mean

$$\theta(P) = \mathbb{E}_{X \sim P}\left[X\right]$$

with respect to the minimax risk

$$R^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\|\theta(P) - \hat{\theta}(S)\|_2^2\right] = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi \circ \rho\left(\theta(P), \hat{\theta}(S)\right)\right].$$

To establish an upper bound to R_n^*, consider the estimator $\hat{\theta}(S) = \frac{1}{n}\sum_{i=1}^n X_i$. As $Var(X_{ij}) \leq \sigma^2$, we have

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

Now, to get a good lower bound for R_n^*, we first consider alternatives from a hypercube (see Figure 1):

$$\mathcal{P}' = \left\{\mathcal{N}\left(\delta\omega, \sigma^2 I_d\right); \omega \in \{0,1\}^d\right\}.$$

For these alternatives, we can calculate the following values:

$$\begin{aligned}
\min_{\omega \neq \omega'} \rho\left(\theta\left(P_\omega\right), \theta\left(P_{\omega'}\right)\right) &= \min_{\omega \neq \omega'} \|\delta\omega - \delta\omega'\|_2 \\
&= \delta, \qquad (\omega \text{ and } \omega' \text{differ on only one coordinate}) \\
\max_{\omega, \omega'} \mathrm{KL}\left(P_\omega, P_{\omega'}\right) &= \frac{\max_{\omega,\omega'} \|\delta\omega - \delta\omega'\|_2^2}{2\sigma^2} \\
&= \frac{d\delta^2}{2\sigma^2} \qquad (\omega \text{ and } \omega' \text{differ on all coordinates}) \\
&\leq \frac{\log 2^d}{4n} \qquad \left(\text{choose } \delta = \sigma\sqrt{\frac{\log 2}{2n}}\right)
\end{aligned}$$

The max KL divergence is too large relative to the minimum distance. Indeed, applying Corollary 1 gives us

$$\begin{aligned}
R_n^* &\geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) \\
&= \frac{\delta^2}{8} \\
&= C\frac{\sigma^2}{n},
\end{aligned}$$

which is not enough to show that our upper bound – which depends on d – is tight.

This example motivates us to introduce Varshamov-Gilbert Lemma. The Varshamov-Gilbert lemma states that we can find a large subset of $\{0,1\}^d$ such that the minimum distance between any two points in the subset is also large. To state the lemma, we must first define the Hamming distance.

Definition 1. *(Hamming distance) The hamming distance between two binary vectors ω, ω' is $H\left(\omega, \omega'\right) = \sum_{j=1}^d \mathbf{1}\left\{\omega_j \neq \omega_j'\right\}$ for $\omega, \omega' \in \mathbb{R}^d$. In other words, $H(\omega, \omega')$ is the number of coordinate where ω_j and ω_j' differ.*

As promised, we now state the lemma.

Lemma 1. *(Varshamov-Gilbert) Let $m \geq 8$. Then there exists $\Omega_m \subseteq \{0,1\}^m$ such that the following are true:*

(i) $|\Omega_m| \geq 2^{m/8}$.

(ii) $\forall \omega, \omega' \in \Omega_m, H\left(\omega, \omega'\right) \geq m/8$.

We will call Ω_m the Varshamov-Gilbert pruned hypercube of $\{0,1\}^m$.

We will not prove this Lemma in class, but a proof can be found in Duchi (2024). Now, we revisit the normal mean estimation example to illustrate the usefulness of the Varshamov-Gilbert lemma.

Example 3. (A lower bound for normal mean estimation in $\mathbb{R}^d$ using the VG lemma)

Consider the same normal mean estimation setting as in Example 2, and let Ω_d be the Varshamov-Gilbert pruned hypercube of $\{0,1\}^d$. We use the alternatives

$$\mathcal{P}' = \left\{ \mathcal{N}\left(\sqrt{\frac{8}{d}}\delta\omega, \sigma^2 I \right); \omega \in \Omega_d \right\}.$$

For these alternatives, we have the bound

$$\begin{aligned}
\min_{\omega,\omega' \in \Omega_d; \omega \neq \omega'} \rho\left(\theta\left(P_\omega\right), \theta\left(P_{\omega'}\right)\right) &= \min_{\omega \neq \omega'} \sqrt{\sum_{j=1}^{d} \left(\sqrt{\frac{8}{d}}\delta\omega_j - \sqrt{\frac{8}{d}}\delta\omega'_j \right)^2} \\
&= \sqrt{\frac{8}{d}}\delta \min_{\omega \neq \omega'} \sqrt{H\left(\omega, \omega'\right)} \\
&\geq \sqrt{\frac{8}{d}}\delta\sqrt{\frac{d}{8}} \\
&= \delta,
\end{aligned}$$

where the inequality follows from the property (ii) of the Varshamov-Gilbert pruned hypercube. Since the maximum ℓ_2-distance over a hypercube is the length of a diagonal, we also have

$$\max_{\omega,\omega' \in \Omega_d} \mathrm{KL}\left(P_\omega, P_{\omega'}\right) = \frac{\left(\sqrt{d}\left(\sqrt{\frac{8}{d}}\delta\right)\right)^2}{2\sigma^2} = \frac{4\delta^2}{\sigma^2}.$$

Choose $\delta = \sigma\sqrt{\frac{d\log(2)}{128n}}$. Then,

$$\begin{aligned}
\max_{\omega,\omega' \in \Omega_d} \mathrm{KL}\left(P_\omega, P_{\omega'}\right) &= \frac{4\delta^2}{\sigma^2} \\
&= \frac{d\log(2)}{32n} \\
&= \frac{\log\left(2^{d/8}\right)}{4n} \\
&\leq \frac{\log\left(|\mathcal{P}'|\right)}{4n},
\end{aligned}$$

where the inequality follows from $|\mathcal{P}'| = |\Omega_d| \geq 2^{d/8}$ (property (i) of the Varshamov-Gilbert pruned hypercube). Therefore, when $d \geq 32$, Corollary 1 gives us

$$\begin{aligned}
R_n^* &\geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) \\
&= \frac{\log(2)}{1024} \cdot \frac{d\sigma^2}{n},
\end{aligned}$$

which proves that $\frac{\sigma^2 d}{n}$ is the minimax optimal rate.

References

Duchi, J. (2024). Lecture notes on information theory.

Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Yuya Shimizu, Keran Chen and Elliot Pickens.