Minimax upper and lower bounds

Paul Delatte
delatte@usc.edu
University of Southern California

Last updated: 08 November, 2022

1 Motivation for the minimax paradigm

1. Why minimax? Given a statistical problem with distributions indexed by a class $\mathscr{F}$, we want to find "good" decision rules among all possible decision rules D. For this, we consider a loss function (e.g., a distance between the estimator $\hat{f}$ and the parameter $f_0 \in \mathscr{F}$ to be estimated in the case of estimation) and a risk function defined as the expected loss. (Both are function of the decision rule and the parameter). The general idea is to design or select decision rules that make the risk as small as possible. But since the "true parameter" f_0 is unknown (see also argument against superefficiency as in Pollard minimax notes or Johnstone S.6.5. p.176), risk cannot be used pointwise at f to evaluate the quality of a given decision rule (if the "true parameter" was known, pointwise optimization would be degenerate since taking $\hat{f} = f_0$ a.s. would yield zero risk). Hence, we would like to find a decision rule that minimizes risk uniformly over the parameter space $\mathscr{F}$ (or some neighborhood in it). Unfortunately, this is generally impossible (even when $\mathscr{F}$ is finite-dimensional). Two ways to remove the uniform nature of the problem are often considered (independently): 1. minimizing over D an average of risk taken over an a priori distribution over $\mathscr{F}$ (Bayesian approach); 2. minimizing over D the supremum over $\mathscr{F}$ of the risk (minimax approach). (It is also possible as it happens often in finite-dimensional settings to further simplify the problem by reducing the number $|D|$ of decision rules imposing some constraints on them). The principle of minimax is thus to simplify the optimality problem by reducing it to the selection of the best decision rule in the worst case scenario. (It should be clear that the parameter $f_d \in F$ defining the worst case scenario for a given decision rule d may differ for another decision rule d'. To find the minimax decision rule, the algorithm would be: for each decision rule d, find the parameter f_d that minimizes the maximum risk of d, and then select the decision rule with minimum maximum risk. A minimax rule is minimax independently of the value of the "true parameter" f_0.). One may ask why consider this criterion (and not any other based on optimization, e.g., maximax): first, minimax is attractive from a risk management perspective as selecting the minimax decision rule provides minimal guarantees for any risk-related properties of this decision rule (if we select the minimax decision rule, then we know that it will always achieve lower risk than its worst case risk, and this upper limit in terms of bad performance is the lowest among other decision rules); secondly, it is a simple enough problem to be "solved". (See decision-theory for other possible criteria). A natural limitation of the minimax criterion is that it says nothing about what happens for better scenari: for them, the minimax decision rule could perform much worse than another decision rules.

2. Why only minimax for nonparametric problem (and not Bayesian)? For infinite-dimensional parameter space $\mathscr{F}$, there is in general no natural a priori distributions (see Nemirovsky p.6).

3. Why upper bounds and lower bounds? In general, it is not possible to exactly compute the maximum risk of the minimax decision rule (also known as the minimax risk), hence the need for upper and lower bounds. Upper and lower bounds do not say anything more than the minimax risk could say (in particular, about the exact performance of any decision rule outside of its worst case scenario). If we could easily compute the minimax risk, we would do it, but we generally cannot. For a given sample size n, the objective is then to upper and lower bound the minimax risk by functions of n only differing by some constant. If we can do so, then we approximately know the minimax risk and it is then simple to design or select decision rules based on their maximum risk (as we can now say which one is approximately minimax and which one is not).

2 Minimax risk and general reduction scheme

We consider estimation of a parameter f_0 in some class $\mathscr{F}$ where the loss function is directly expressed through some metric ρ. That is, we have some Borel probability measures $\{\mathbb{P}_f : f \in \mathscr{F}\}$ on a measurable space $(E, \mathcal{E})$ and the risk of an estimator $\hat{f}_n$ of f_0 is given by

$$R(\hat{f}_n, f_0) = \mathbb{E}_{\mathbb{P}_{f_0}}(\rho(\hat{f}_n, f_0))).$$

(It is implicitly assumed that we have a sample of size n i.i.d. according to f_0 so that the expectation is taken with respect to the n-product distribution $\mathbb{P}_{f_0}^{\otimes n}$). The **minimax risk** R_n is defined as the maximum risk of the minimax rule, that is,

$$R_n := \inf_{\hat{f}_n} \sup_{f \in \mathscr{F}} \mathbb{E}_{\mathbb{P}_f}(\rho(\hat{f}_n, f))) = \sup_{f \in \mathscr{F}} \mathbb{E}_{\mathbb{P}_f}(\rho(\hat{\phi}_n, f))$$

where $\hat{\phi}_n$ is the minimax rule. The primary objective is to find upper bounds U_n and lower bounds L_n on R_n, that is,

$$L_n \leq R_n \leq U_n.$$

We do not want any such bounds, but bounds as close as possible to one another. In practice, we want to find a constant-factor approximation of the minimax risk, that is, upper and lower bounds expressible as the same function of n but differing only by a constant. In other words, we want to find a positive sequence (ψ_n) converging to 0 (often, some power functions) such that

$$L_n = c\psi_n \leq R_n \leq C\psi_n = U_n,$$

where $c, C > 0$ are some universal constants. In this case, we write $R_n \asymp \psi_n$ and call ψ_n a **minimax rate of convergence** (which is unique up to multiplicative constants). An estimator $\hat{f}_n$ of f_0 with maximum risk $r_{\hat{f}_n} := \sup_{f \in \mathscr{F}} \mathbb{E}_{\mathbb{P}_f}(\rho(\hat{f}_n, f))$ is said to be

(asymptotically) (minimax) efficient if

$$\lim_{n\to\infty} \frac{r_{\hat{f}_n}}{R_n} = 1.$$

Finding an upper bound is relatively easy since R_n is bounded by the maximum risk of any estimator $\hat{f}_n$ of f_0. That is,

$$R_n = \inf_{\hat{f}_n} \sup_{f\in\mathscr{F}} \mathbb{E}_{\mathbb{P}_f}(\rho(\hat{f}_n, f))) \leq \sup_{f\in\mathscr{F}} \mathbb{E}_{\mathbb{P}_f}(\rho(\hat{f}_n, f)) = U_n$$

for any estimator $\hat{f}_n$ of f_0.

Finding a lower bounds requires more work, but there exists a general "reduction scheme" to find them. The reduction consists in discretizing $\mathcal{F}$ (in a not too patchy way), as optimization over discrete sets is generally more manageable than over uncountable ones. Finding minimax lower bounds then amounts to lower bounding the probability of error in some testing problem. The **reduction from estimation to testing** proceeds as follows:

1. Reduction to a probability bound: using Markov's inequality, we have for all $s > 0$ that

$$\mathbb{E}\left(\rho(\hat{f}_n, f)\right) \geq s\mathbb{P}(\rho(\hat{f}_n, f) \geq s),$$

hence to lower bound R_n, it suffices to lower bound any of the minimax probabilities given for any $s > 0$ by

$$\inf_{\hat{f}_n} \sup_{f\in\mathscr{F}} \mathbb{P}_f(\rho(\hat{f}_n, f) \geq s).$$

2. Discretization of $\mathcal{F}$: for any finite subset $\mathcal{F}_M = \{f_1, \ldots, f_M\}$ of $\mathcal{F}$, we naturally have that

$$\inf_{\hat{f}_n} \sup_{f\in\mathscr{F}} \mathbb{P}_f(\rho(\hat{f}_n, f) \geq s) \geq \inf_{\hat{f}_n} \max_{f\in\{f_1,\ldots,f_M\}} \mathbb{P}_f(\rho(\hat{f}_n, f) \geq s).$$

Intuitively, to make the lower bounds as large (i.e., good) as possible, we need the discretization to be a good approximation of $\mathcal{F}$. (This already indicates that the choice of $\mathcal{F}_M$ is of great importance for the tightness of the bound. As shown in the next step, we should focus on $2s$-separated sets for which the choice of s becomes crucial).

3. Choice of $2s$-separated hypotheses: if $\rho(f_i, f_j) \geq 2s$ for all $i, j = 1, \ldots, M$, $i \neq j$ (i.e., $\mathcal{F}_M$ is a $2s$-separated set), then for any estimator $\hat{f}_n$,

$$\mathbb{P}_{f_i}(\rho(\hat{f}_n, f_i) \geq s) \geq \mathbb{P}_{f_i}(\psi^* \neq i)$$

where $\psi^* \colon E \longrightarrow \{1, \ldots, M\}$ is the minimum distance (multiple deterministic) test defined by

$$\psi^* = \arg \min_{i\in\{1,\ldots,M\}} \rho(\hat{f}_n, f_i).$$

This follows directly from the triangle inequality. Thereafter, the problem reduces to lower bounding the minimax probability of error

$$p^e_{\mathcal{F}_M}(s) := \inf_{\psi} \max_{i\in\{1,\ldots,M\}} \mathbb{P}_{f_i}(\psi \neq i)$$

where the infimum runs over all tests in $\{1, \ldots, M\}$ for appropriately selected $2s$-separated hypotheses $\mathcal{F}_M = \{f_1, \dots, f_M\}$. Indeed, from (1) and (2), we have

$$R_n \geq s p^e_{\mathcal{F}_M}(s).$$

The bound above, which we want as large as possible, is the product of s and the minimax probability of error which depends on s. For s fixed, it is often natural to consider a maximum $2s$-packing for $\mathcal{F}_M$: indeed, we want the minimax probability of error $p^e_{\mathcal{F}_M}(s)$ to be as large as possible; for this we want the testing problem to be more difficult, that is, we want the distributions to be closer to one another, which happens under the separation condition only when there are more of them until no more can be added. The crux of the problem then becomes the choice of s for which we face a clear trade-off: when s increases, the minimax probability of error generally decreases (indeed, if s increases, distributions are more separated, and the testing problem becomes easier); the choice of s (and hence the minimax rate) will then be dictated by the geometry of $\mathcal{F}$.

3 Le Cam's method

The general method developed above applies in the particular case of $M = 2$, in which case, estimation reduces to binary hypothesis testing. The connection between binary testing and total variation allows us to derive lower bounds for the minimax risk. The method generalizes by considering two sets of distributions sufficiently separated.

The method relies on an impossibility result for binary testing due to Le Cam which lower bounds the (sum of) probabilities of error by the total variation distance between the two tested distributions. The intuition of the result is simple: if the distribution are close (i.e., their total variation is small), testing is difficult, and errors are likely large.

Proposition 1. *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be probability distributions on $(E, \mathcal{E})$. For any test $\psi\colon E \longrightarrow \{1, 2\}$, it holds that*

$$\mathbb{P}_1(\psi \neq 1) + \mathbb{P}_2(\psi \neq 2) \geq 1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)$$

with equality if $\psi(x) = \mathbb{1}\{p_2(x) \geq p_1(x)\}$.

Proof. Let $A = \{x \in E : \psi(x) = 1\}$. Then $\mathbb{P}_1(A^c) + \mathbb{P}_2(A) = 1 - (\mathbb{P}_1(A) - \mathbb{P}_2(A)) \geq 1 - \sup_A |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$. Equality follows from the equivalent formulations of the total variation distance. □

This is equivalently formulated by saying that the best binary test (in terms of sum of error probabilities) has sum of error probabilities equal to $1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)$.

Corollary 2. *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be probability distributions on $(E, \mathcal{E})$. Then*

$$\inf_{\psi} [\mathbb{P}_1(\psi \neq 1) + \mathbb{P}_2(\psi \neq 2)] = 1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)$$

where the infimum runs over all test $\psi\colon E \longrightarrow \{1, 2\}$.

Proof. Since any test ψ is uniquely defined by $A = \{x \in E : \psi(x) = 1\}$, the infimum over all binary tests can be taken over all subsets of E. Then taking the infimum over $\mathbb{P}_1(A^c) + \mathbb{P}_2(A) = 1 - (\mathbb{P}_1(A) - \mathbb{P}_2(A))$ yields the result. □

The connection to the reduction scheme of last section is then straighforward. By lower bounding the maximum over the distributions by their average (which always holds), we get

$$\begin{aligned} p^e_{\{\mathbb{P}_{f_1}, \mathbb{P}_{f_2}\}} &= \inf_{\psi} \max_{i \in \{1,2\}} \mathbb{P}_{f_i}(\psi \neq i) \\ &\geq \inf_{\psi} \frac{1}{2}[\mathbb{P}_{f_1}(\psi \neq 1) + \mathbb{P}_{f_2}(\psi \neq 2)] \\ &= \frac{1}{2} - \frac{1}{2}\mathrm{TV}(\mathbb{P}_{f_1}, \mathbb{P}_{f_2}). \end{aligned}$$

As everything we did, this holds in particular for n-product distributions, that is,

$$p^e_{\{\mathbb{P}^{\otimes n}_{f_1}, \mathbb{P}^{\otimes n}_{f_2}\}} = \frac{1}{2} - \frac{1}{2}\mathrm{TV}(\mathbb{P}^{\otimes n}_{f_1}, \mathbb{P}^{\otimes n}_{f_2}).$$

Finally, if f_1 and f_2 are $2s$-separated, then

$$R_n \geq s\left(\frac{1}{2} - \frac{1}{2}\mathrm{TV}\big(\mathbb{P}^{\otimes n}_{f_1}, \mathbb{P}^{\otimes n}_{f_2}\big)\right).$$

The objective is then to find distributions $\mathbb{P}_{f_1}$ and $\mathbb{P}_{f_2}$ such that $\rho(f_1, f_2)$ is large (allowing s to be as large as possible) and $\mathrm{TV}(\mathbb{P}^{\otimes n}_{f_1}, \mathbb{P}^{\otimes n}_{f_2})$ is small. We illustrate this in examples.

Example 3 (Bernoulli Mean Estimation)**.** Consider estimating the mean $\theta \in [-1, 1]$ of a $\{\pm 1\}$-Bernoulli random variable under the squared error loss $\rho(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Denote $\mathrm{Ber}([-1, 1])$ the family of all $\{\pm 1\}$-Bernoulli distributions. Fix $\delta \in (0, 1]$ and for any $a \in [0, 1/\delta]$, define $\mathbb{P}_a \in \mathrm{Ber}([-1, 1])$ by $\mathbb{P}_a(X = 1) = (1 + a\delta)/2$ and $\mathbb{P}_a(X = -1) = (1 - a\delta)/2$. Then the mean of $\mathbb{P}_a$ is $\theta(\mathbb{P}_a) = a\delta$. We have $\rho(\theta(\mathbb{P}_{-1}), \theta(\mathbb{P}_1)) = 4\delta^2$. Then, via Le Cam's method, the minimax risk for the problem is bounded by

$$R_n \geq \delta^2(1 - \mathrm{TV}(\mathbb{P}^{\otimes n}_{-1}, \mathbb{P}^{\otimes n}_{1}))$$

We now bound $\mathrm{TV}(\mathbb{P}^{\otimes n}_{-1}, \mathbb{P}^{\otimes n}_{1})$ in terms of δ and n. By Pinsker's inequality and the tensorization identity for KL-divergence, we have

$$\mathrm{TV}(\mathbb{P}^{\otimes n}_{-1}, \mathbb{P}^{\otimes n}_{1})^2 \leq \frac{1}{2}D_{KL}(\mathbb{P}^{\otimes n}_{-1} \| \mathbb{P}^{\otimes n}_{1}) = \frac{n}{2}D_{KL}(\mathbb{P}_{-1} \| \mathbb{P}_1) = \frac{n}{2}\delta \log \frac{1+\delta}{1-\delta}.$$

Since $\delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, 1/2]$, we have $\mathrm{TV}(\mathbb{P}^{\otimes n}_{-1}, \mathbb{P}^{\otimes n}_{1}) \leq \delta\sqrt{3n/2}$ for $\delta \leq 1/2$. By taking $\delta = 1/\sqrt{6n}$, this guarantees that $\mathrm{TV}(\mathbb{P}^{\otimes n}_{-1}, \mathbb{P}^{\otimes n}_{1}) \leq \frac{1}{2}$, and so

$$R_n \geq \frac{1}{12n}.$$

This is the standard rate of convergence ($1/n$ in squared error) for the parametric estimation. The sample mean $n^{-1}\sum_{i=1}^{n} X_i$ for the above problem achieves mean-squared error $(1 - \theta^2)/n$.

Reference. E.7.7. in Duchi's 311IT notes p.138.

Example 4 (Gaussian Location Model)**.** Consider the problem of estimating the mean $\theta \in \mathbb{R}$ of a normal random variable $N(\theta, \sigma^2)$ with known variance $\sigma^2 > 0$ under either the squared error loss $\rho(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$ or the absolute error loss $\nu(\theta,\hat{\theta}) = |\theta-\hat{\theta}|$. Consider the normal distributions $\mathbb{P}_0 \sim N(0,\sigma^2)$ and $\mathbb{P}_{2\delta} \sim N(2\delta,\sigma^2)$ for some $\delta > 0$. By second-moment bounding Gaussian r.v.s, we have

$$\mathrm{TV}(\mathbb{P}_{2\delta}^n, \mathbb{P}_0^n) \leq \frac{1}{4}\left(e^{4n\delta^2/\sigma^2} - 1\right).$$

Since 2δ and 0 are $4\delta^2$-separated with respect to ρ and 2δ-separated with respect to ν, we have by taking $\delta = \sigma/2\sqrt{n}$ that

$$R_n(\rho) \geq \delta^2\left(1 - \frac{1}{2}\sqrt{e-1}\right) \geq \frac{\delta^2}{3} = \frac{1}{12}\frac{\sigma^2}{n},$$

and

$$R_n(\nu) \geq \frac{\delta}{2}\left(1 - \frac{1}{2}\sqrt{e-1}\right) \geq \frac{\delta^2}{6} = \frac{1}{12}\frac{\sigma}{\sqrt{n}}.$$

This is again an example of standard rates for parametric estimation. The sample mean for the above problem achieves respective risks σ^2/n and $\sqrt{2}\sigma/\sqrt{\pi n}$.

Reference. E.15.4. in Wainwright HDS p.492.

Example 5 (Normal Nonparametric Regression (at a point))**.** Consider the estimation of the function $m\colon [0,1] \to \mathbb{R}$ defined via

$$Y_i = m(X_i) + \varepsilon_i$$

where we observe $(X_1,Y_1),\dots,(X_n,Y_n)$ with $X_i \sim U[0,1]$ (or, equivalently, X_i is deterministic in $[0,1]$) but not $\varepsilon_i \sim N(0,\sigma^2)$. Assume that

$$m \in M = \left\{m\colon [0,1] \to \mathbb{R} : |m(y) - m(x)| \leq L|y-x| \text{ for all } x,y \in [0,1]\right\}.$$

The set $\mathscr{P}$ of distributions for the problem thus comprises all distributions of the form $p(x,y) = p(x)p(y|x) = \phi(y - m(x))$ where $m \in M$. We want to estimate $m(x)$. Without loss of generality, take $x = 0$, so the parameter of interest is $\theta = m(0)$. Consider the absolute error loss $\rho(\theta_0,\theta_1) = |\theta_1 - \theta_0|$. Define $m_0(x) = 0$ for all x. Let $0 \geq \varepsilon \geq 1$ and define

$$m_1(x) = \begin{cases} L(\varepsilon - x) & \text{if } 0 \geq x \geq \varepsilon, \\ 0 & \text{if } x > \varepsilon. \end{cases}$$

Then $m_0, m_1 \in M$ and $\rho(m_0(0), m_1(0)) = L\varepsilon$. Consider the distributions $\mathbb{P}_0, \mathbb{P}_1 \in \mathscr{P}$ respectively associated with $m_0, m_1 \in M$. Their KL-divergence is then given by

$$\begin{aligned} D_{KL}(\mathbb{P}_0 \| \mathbb{P}_1) &= \iint p_0(x,y) \log \frac{p_0(x,y)}{p_1(x,y)}\, dy\, dx \\ &= \iint \phi(y) \log \frac{\phi(y)}{\phi(y - m_1(x))}\, dy\, dx \\ &= \int_0^\varepsilon D_{KL}(N(0,1), N(m_1(x),1))\, dx. \end{aligned}$$

Since $D_{KL}(N(\mu_1, 1), N(\mu_2, 1)) = (\mu_1 - \mu_2)^2/2$, we have

$$D_{KL}(\mathbb{P}_0 \| \mathbb{P}_1) = \frac{L^2}{2} \int_0^{\varepsilon} (\varepsilon - x)^2 \, dx = \frac{L^2 \varepsilon^3}{6}.$$

Then by Pinkser's inequality,

$$\mathrm{TV}(\mathbb{P}_0^n, \mathbb{P}_1^n)^2 \leq \frac{nL^2\varepsilon^3}{12}.$$

By taking $\varepsilon = (\frac{3}{nL^2})^{1/3}$, we have $\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \leq 1/2$, and thus

$$R_n \geq \frac{L}{4}\left(\frac{3}{nL^2}\right)^{1/3} = \left(\frac{c}{n}\right)^{1/3}.$$

The regression histogram $\hat{m}(0)$ for the above problem has risk $(C/n)^{1/3}$. This proves that $R_n \asymp n^{-1/3}$. For the d-dimensional problem, we find similarly that $R_n \asymp n^{-1/(2+d)}$. For the squared error loss, we find $R_n \asymp n^{-2/(2+d)}$.

Reference. E.8. in Larry Wasserman's notes on Minimax or S.5. in Tsybakov INE p.91.

4 Fano's method

Fano's method for minimax lower bounds was developed by R. Z. Has'ininskii. It uses the same reduction scheme introduced initially, but relies directly on multiple hypotheses (i.e., $M \geq 2$), whereas Le Cam's method was inherently based on binary hypotheses (i.e. $M = 2$). For this, it exploits Fano's inequality which provides impossibility result for multiple hypothesis testing in terms of mutual information. Fano's method generally delivers tighter bounds and bounds where Le Cam's method fails (e.g., with L^2 distance $\rho(f, g) = \int (f - g)^2$).

Proposition 6 (Fano's Inequality)**.** *For any Markov chain $V \longrightarrow X \longrightarrow \hat{V}$ taking values in $\mathcal{V}$, it holds that*

$$h_2(\mathbb{P}(\hat{V} \neq V)) + \mathbb{P}(\hat{V} \neq V) \log(|\mathcal{V}| - 1) \geq H(V|\hat{V}),$$

where $h_2(p) = -p \log p - (1 - p) \log(1 - p)$ is the binary entropy function and $H(V|\hat{V})$ is the entropy of V conditioned on $\hat{V}$. In particular, if V is uniform on $\mathcal{V}$, then it holds that

$$\mathbb{P}(\hat{V} \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log(|\mathcal{V}|)},$$

where $I(V; X)$ is the mutual information between V and X.

Proof. P.2.19. and C.2.20. in Duchi IT notes p.27-28. □

This applies in particular when $\hat{V} = \psi(X)$ for any testing function $\psi \colon E \longrightarrow \mathcal{V}$. Moreover, in the uniform setting, the mutual information takes the equivalent representation

$$I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathrm{KL}(\mathbb{P}_v \| \overline{\mathbb{P}}),$$

where $\overline{\mathbb{P}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\mathbb{P}_v$ (the result holds more genrally for any distribution $\mathbb{Q}$ dominating $\mathbb{P}_v$ for all $v \in \mathcal{V}$). We thus get the following corollary which bounds the average probability of error (which can then be used to bound the max probability of error, and hence the minimax probability of error, and thus the minimax risk).

Corollary 7. *Let $\{\mathbb{P}_1, \dots, \mathbb{P}_M\}$ be Borel probability measures on $(E, \mathcal{E})$. For any test function $\psi\colon E \longrightarrow \{1, \dots, M\}$ and any Borel probability measure $\mathbb{Q}$ such that $\mathbb{P} \ll \mathbb{Q}$ for all $i = 1, \dots, M$, it holds that*

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{P}_i(\psi \neq i) \geq 1 - \frac{\frac{1}{M}\sum_{i=1}^{M}\mathrm{KL}(\mathbb{P}_i\|\mathbb{Q}) + \log 2}{\log M}.$$

Proof. This follows directly from Fano's inequality. See L.28. in Larry minimax p.28 or T.3.1. in Giraud HDS p.56. □

As mentioned, this holds in particular for $\mathbb{Q} = \overline{\mathbb{P}}$. Since (we can always lower bound a max by an average – see p.113 in Tsybakov), we finally have that

$$\begin{aligned} p^e_{\mathcal{F}_M} &= \inf_{\psi}\max_{i\in\{1,\dots,M\}}\mathbb{P}_{f_i}(\psi \neq i) \\ &\geq \inf_{\psi}\frac{1}{M}\sum_{i=1}^{M}\mathbb{P}_{f_i}(\psi \neq i) \quad (=: \bar{p}^e_{\mathcal{F}_M}) \\ &\geq 1 - \frac{\frac{1}{M}\sum_{i=1}^{M}\mathrm{KL}(\mathbb{P}_{f_i}\|\overline{\mathbb{P}}) + \log 2}{\log M}. \end{aligned}$$

The next objective is to upper bound $M^{-1}\sum_{i=1}^{M}\mathrm{KL}(\mathbb{P}_{f_i}\|\overline{\mathbb{P}})$. We have that

$$\begin{aligned} \frac{1}{M}\sum_{i=1}^{M}\mathrm{KL}(\mathbb{P}_{f_i}\|\overline{\mathbb{P}}) &\leq \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathrm{KL}(\mathbb{P}_{f_i}\|\mathbb{P}_{f_j}) \\ &\leq \max_{i\neq j}\mathrm{KL}(\mathbb{P}_{f_i}\|\mathbb{P}_{f_j}). \end{aligned}$$

Thereafter, applying the reduction scheme (and using the fact that $\mathrm{KL}(\mathbb{P}_{f_i}^{\otimes n}\|\mathbb{P}_{f_j}^{\otimes n}) = n\mathrm{KL}(\mathbb{P}_{f_i}\|\mathbb{P}_{f_j})$), we can directly lower bound R_n in terms of $\max_{i\neq j}\mathrm{KL}(\mathbb{P}_{f_i}\|\mathbb{P}_{f_j})$ whenever $\mathcal{F}_M$ is $2s$-separated. That is, if $\mathcal{F}_M$ is $2s$-separated, then

$$R_n \geq s\left(1 - \frac{n\max_{i\neq j}\mathrm{KL}(\mathbb{P}_{f_i}\|\mathbb{P}_{f_j}) + \log 2}{\log M}\right).$$

Then, as in Le Cam's method, the objective is to find $\mathcal{F}_M$ such that all the f_i are far apart in terms of ρ, but all the $\mathbb{P}_{f_i}$ are close enough in terms of Kullback–Leibner divergence. To construct such families, we will often use scaled packing of some fixed sets. The two following results will often be used: bounds on the packing number of the unit ball in $\mathbb{R}^d$ with respect to any norm-induced distance; exponentially large packings of the d-dimensional hypercube with respect to the Hamming distance.

Proposition 8. *Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and $B_{\|\cdot\|}$ the corresponding unit ball. Then*

$$\left(\frac{1}{\delta}\right)^d \leq N(B_{\|\cdot\|}, \|\cdot\|, \delta) \leq \mathrm{pack}(B_{\|\cdot\|}, \|\cdot\|, \delta) \leq N(B_{\|\cdot\|}, \|\cdot\|, \delta/2) \leq \left(1 + \frac{4}{\delta}\right)^d.$$

Proof. L.5.4. and L.5.6. in Duchi IT notes. □

Proposition 9 (Gilbert–Varshamov Bound)**.** *Let $d \geq 1$. There is a subset $\mathcal{V}$ of the d-dimensional hypercube $\mathcal{H}^d = \{-1, 1\}^d$ with cardinality $|\mathcal{V}| \geq e^{d/8}$ such that*

$$\|v - v'\|_1 = \sum_{i=1}^{d} \mathbb{1}\{v_i \neq v_i'\} \geq \frac{d}{4}$$

for all $v \neq v'$ with $v, v' \in \mathcal{V}$.

Proof. L.7.5. in Duchi IT notes. □

Example 10 (Gaussian Location Model (bis repetita))**.** Consider the problem of estimating the mean θ in the d-dimensional Gaussian location family $\mathcal{N}_d = \{N_d(\theta, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$ under the squared error loss $\rho(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$. We now construct an appropriate family of probability measures by building a "local packing" of $\Theta = \mathbb{R}^d$. Consider $\mathcal{V}$ a $1/2$-packing of the unit ball in $\mathbb{R}^d$ with respect to the l_2-norm. By standard results, the cardinality of $\mathcal{V}$ is at least 2^d. Fix $\delta > 0$ and for each $v \in \mathcal{V}$, define $\theta_v = \delta v \in \mathbb{R}^d$. Then we have

$$\|\theta_v - \theta_{v'}\|_2 = \delta\|v - v'\|_2 \geq \frac{\delta}{2}$$

for each distinct $v, v' \in \mathcal{V}$. That is, $\{\theta_v : v \in \mathcal{V}\}$ is $\delta^2/4$-separated with respect to ρ. Moreover, $\|\theta_v - \theta_{v'}\|_2 \leq 2\delta$. Define the distribution $\mathbb{P}_v \sim N(\theta_v, \sigma^2 I_d) \in \mathcal{N}_d$. Since the KL-divergence between normal distributions with identical covariance is

$$D_{KL}(N(\theta_1, \Sigma)\|N(\theta_2, \Sigma)) = \frac{1}{2}(\theta_1 - \theta_2)^T \Sigma^{-1} (\theta_1 - \theta_2),$$

we have that

$$D_{KL}(\mathbb{P}_v^n \|\mathbb{P}_{v'}^n) = n D_{KL}(N(\delta v, \sigma^2 I_d)\|N(\delta v', \sigma^2 I_d)) = n\frac{\delta^2}{2\sigma^2}\|v - v'\|_2^2.$$

Since $\|v - v'\|_2 \leq 2$, we have $D_{KL}(\mathbb{P}_v^n\|\mathbb{P}_{v'}^n) \leq 2n\delta^2/\sigma^2$. By taking $\delta^2 = d\sigma^2 \log 2/8n$, we find for $d \geq 2$ that

$$R_n \geq \frac{\delta^2}{16}\left(1 - \frac{2n\delta^2\sigma^{-2} + \log 2}{d \log 2}\right) = \frac{\delta^2}{16}\left(1 - \frac{1}{d} - \frac{1}{4}\right) \geq \frac{d\sigma^2 \log 2}{128n}\frac{1}{4} = c\frac{d\sigma^2}{n}.$$

The sample mean for the above problem achieves risk $Cd\sigma^2/n$, hence $R_n \asymp d\sigma^2/n$.

Reference. E.7.11. in Duchi IT notes. See also Wainwright E.15.13. and Giraud S.3.3.

Example 11 (Fixed Design Normal Linear Regression)**.** Consider estimating $\theta \in \mathbb{R}^d$ in the linear regression model

$$Y = X\theta + \varepsilon$$

where $X \in \mathbb{R}^{n \times d}$ is some fixed matrix and $\varepsilon \sim N(0, \sigma^2 I_n)$. Consider the squared error loss $\rho(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$. Let $\mathscr{P}$ the family of distributions defined by this model, that is,

$$\mathscr{P} = \{N(X\theta, \sigma^2 I_n) : \theta \in \mathbb{R}^d\}.$$

We now construct an appropriate family of probability measures in $\mathscr{P}$ by building a "local packing" of $\Theta = \mathbb{R}^d$. In this case, we use the Gilbert–Varshamov bound: it guarantees the existence of a packing $\mathcal{V}$ of $\{-1,1\}^d$ such that $|\mathcal{V}| \geq e^{d/8}$ and $\|v - v'\|_1 \geq d/4$ for $v \neq v'$. Fix $\delta > 0$ and define $\theta_v := \delta v \in \mathbb{R}^d$. Then we have for $v \neq v'$,

$$\|\theta_v - \theta_{v'}\|_2^2 = \delta^2 \sum_{j=1}^{d} (v_j - v'_j)^2 = \delta^2 \|v - v'\|_1 \geq \frac{d\delta^2}{4}.$$

We have

$$\begin{aligned} D_{KL}(N(X\theta_v, \sigma^2 I_n)\|N(X\theta_{v'}, \sigma^2 I_n)) &= \frac{1}{2\sigma^2}\|X(\theta_v - \theta_{v'})\|_2^2 \\ &\leq \frac{\delta^2}{2\sigma^2}\gamma_{\max}^2(X)\|v - v'\|_2^2 \\ &\leq \frac{\delta^2 d}{2\sigma^2}\gamma_{\max}^2(X) \end{aligned}$$

where $\gamma_{\max}(X)$ denotes the maximal singular value of X. By taking $\delta^2 = \sigma^2/16\gamma_{\max}^2(X)$, then for $d \geq 32$,

$$R_n \geq \frac{d\delta^2}{8}\left(1 - \frac{\delta^2 d\gamma_{\max}^2(X)/2\sigma^2 + \log 2}{d/8}\right) \geq \frac{d\delta^2}{8}\left(1 - \frac{1}{4} - \frac{1}{4}\right) = \frac{1}{256}\frac{d\sigma^2}{\gamma_{\max}^2(X)}.$$

The rate can be rewritten as $d\sigma^2/256n\gamma_{\max}^2(n^{-1/2}X)$. This bound is of the right order in terms of d, n, and σ^2 but our bounding through the maximal singular value of X makes the bound not sharp. An exact calculation shows that the minimax value of the problem is exactly $\sigma^2\mathrm{tr}((X^TX)^{-1})$.

Reference. E.7.12 in Duchi or E.15.14. in Wainwright (for a different metric).

Example 12 (Normal Nonparametric Regression under L_2-distance)**.** Consider again the problem of estimating the function $f\colon [0,1] \to \mathbb{R}$ defined via

$$Y_i = f(X_i) + \varepsilon_i$$

where we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \sim U[0,1]$ (or, equivalently, X is deterministic in $[0,1]$) but do not observe $\varepsilon_i \sim N(0, \sigma^2)$. Assume this time that f lies in the Holder class $\Sigma(\beta, L)$ (which comprises L-Lipschitz functions for $\beta = 1$), that is,

$$f \in \Sigma(\beta, L) = \left\{f\colon [0,1] \to \mathbb{R} : |f^{(l)}(y) - f^{(l)}(x)| \leq L|y - x|^{\beta - l} \text{ for all } x, y \in [0,1]\right\}$$

where l is the greatest integer strictly less than β. The set $\mathscr{P}$ of distributions for the problem is again composed of all distributions of the form $p(x,y) = p(x)p(y|x) = \phi(y - f(x))$ where $f \in \Sigma(\beta, L)$. We want to estimate f under the integrated squared error loss (L_2-distance) $\rho^2(f,g) = \int (f-g)^2$. We now construct an appropriate family F of probability measures in $\mathscr{P}$ by building a "local packing" of $\Sigma(\beta, L)$. Fix $c > 0$, and define $m = \lceil cn^{1/(2\beta+1)}\rceil$. By the Gilbert–Varshamov bound, there exists a packing $\mathcal{V}$ of $\{-1,1\}^m$ such that $|\mathcal{V}| \geq e^{m/8}$ and $\|v - v'\|_1 \geq m/4$ for $v \neq v'$. Define

$$F = \left\{f_v(x) = \sum_{j=1}^{m} v_j \phi_j(x) : v \in \mathcal{V}\right\}$$

where $m = \lceil cn^{1/(2\beta+1)} \rceil$, $h = 1/m$, $\phi_j(x) = Lh^\beta K((x - x_j)/h)$, $x_j = (j - 1/2)/m$, and $K\colon \mathbb{R} \longrightarrow [0, +\infty)$ is any sufficiently smooth function supported on $(-1/2, 1/2)$ such that $F \subseteq \Sigma(\beta, L)$. For any $v \neq v'$ in $\mathcal{V}$, we have

$$\begin{aligned}
\rho(f_v, f_{v'}) &= \left[\int_0^1 \left(\sum_{j=1}^m (v_j - v_j')\phi_j(x) \right)^2 dx \right]^{1/2} \\
&= \left[\sum_{j=1}^m (v_j - v_j')^2 \int_{\Delta_j} \phi_j^2(x)\, dx \right]^{1/2} \\
&= \sqrt{\|v - v'\|_1} L h^{\beta + \frac{1}{2}} \|K\|_2 \\
&\geq c_0 h^\beta
\end{aligned}$$

for some $c_0 > 0$, where $\Delta_j = [(j-1)/m, j/m]$ for $j = 1, 2, \ldots, m$. The second equality follows from the fact that $K((x - x_i)/h)K((x - x_j)/h) = 0$ for $i \neq j$. The last inequality follows by construction, since $\|v - v'\|_1 \geq m/4$. By standard results for the KL-divergence of normals with same variance, we find that

$$\begin{aligned}
D_{KL}(\mathbb{P}_v \| \mathbb{P}_{v'}) &= \int_0^1 D_{KL}(N(f_v(x), 1), N(f_{v'}(x), 1))\, dx \\
&= \frac{1}{2} \int_0^1 (f_v - f_{v'})^2\, dx \\
&\leq c_1 h^{2\beta}
\end{aligned}$$

for some $c_1 > 0$. The last inequality follows from the fact that $\|v - v'\|_1 \leq m$. By taking $h = c^{1/2\beta} c_1^{-1/2\beta} n^{-1/(2\beta+1)}/32$, we find for $m \geq 32$ that

$$R_n(\rho) \geq \frac{c_0 h^\beta}{2} \left(1 - \frac{n c_1 h^{2\beta} + \log 2}{m/8} \right) \geq \frac{c_0 h^\beta}{2} \left(1 - \frac{1}{4} - \frac{1}{4} \right) \geq c_2 n^{-\frac{\beta}{2\beta+1}}$$

for some $c_2 > 0$. Similarly, we directly find that $R_n(\nu) \geq c_2 n^{-\frac{2\beta}{2\beta+1}}$ for some $c_3 > 0$ where $\nu(f, g) = \int (f - g)^2$. It can be shown that there are kernel estimators that achieve these rates, hence $R_n(\rho) \asymp n^{-\beta/(2\beta+1)}$ and $R_n(\nu) \asymp n^{-2\beta/(2\beta+1)}$. A similar calculation in d dimensions shows that $R_n(\nu_d) \asymp n^{-2\beta/(2\beta+d)}$.

Reference. E.16. in Larry minimax or S.2.6. in Tsybakov INE p.95.

Example 13 (Density Estimation)**.** E.15.15. in Wainwright or S.9.2. in Larry minimax.

References

Duchi, J. (2021): "Lecture Notes for Statistics 311/Electrical Engineering 377," *Lecture notes for STAT311/EE377 at Stanford*.

Giraud, C. (2020): *Introduction to high-dimensional statistics*. Manuscript.

Johnstone, I. M. (2019): "Gaussian estimation: Sequence and wavelet models," *Unpublished lecture notes*.

Nemirovski, A. (2000): *Topics in Non-Parametric Statistics. XXVIII Saint-Flour Summer School on Probability Theory*. Springer.

Pollard, D. (2005): "Lecture Notes 18 for Statistics 607b on minimax lower bounds," *Lecture notes for Statistics 607b at Yale*.

Tsybakov, A. B. (2008): *Introduction to nonparametric estimation*. Springer.

Wainwright, M. J. (2019): *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wasserman, L. (2017): "Lecture Notes 10/36-702 on minimax theory," *Lecture notes for Statistics 36-702 at CMU*.