MA6225: Information-Theoretic Methods in Statistical Learning

Lectures 1–11

Vincent Y. F. Tan

AY 2026/27 Semester 1

Conventions. Unless a base is explicitly specified, logarithms are natural and entropy is measured in nats. The prime-number arguments in Lecture 6 use base-2 logarithms and bits, before switching back to natural logarithms for learning theory.

Contents

1 Lecture 1: Entropy, KL Divergence, and Entropy Inequalities **4**
1.1 Shannon entropy . . . 4
1.2 Conditional entropy . . . 4
1.3 Basic properties of entropy . . . 4
1.4 Axiomatic characterization and Rényi entropy . . . 5
1.5 Method of types . . . 6
1.6 Submodularity of entropy . . . 6
1.7 Entropy regions and non-Shannon inequalities . . . 7
1.8 Relative entropy / KL divergence . . . 7
1.9 Differential entropy and Gaussian extremality . . . 8
1.10 Markov kernels, conditional divergence, and the chain rule . . . 8
1.11 A large-deviations estimate from DPI . . . 9

2 Lecture 2: Mutual Information, Fano, and Variational Characterizations **10**
2.1 Mutual information . . . 10
2.2 Gaussian mutual information . . . 10
2.3 AWGN channel . . . 11
2.4 Conditional mutual information and conditional independence . . . 11
2.5 Strong data processing . . . 12
2.6 Sufficient statistics . . . 12
2.7 Probability of error and Fano's inequality . . . 12
2.8 Variational characterizations: the golden formula . . . 13
2.9 A supremum representation and Donsker–Varadhan . . . 13

3 Lecture 3: Gibbs Variational Principle, Softmax, VAE, PAC-Bayes, and Convexity **15**
3.1 Gibbs variational principle . . . 15
3.2 Softmax and temperature . . . 15
3.3 Transformer attention . . . 15
3.4 Variational autoencoders . . . 16
3.5 From Donsker–Varadhan to PAC-Bayes . . . 17
3.6 Four optimization problems in information theory . . . 17
3.7 Convexity of information measures . . . 18
3.8 Minimax and saddle points . . . 19

3.9 Saddle point for mutual information . . . 19

4 Lecture 4: Alternating Optimization, Tensorization, Entropy Rates, and f-Divergences 20
4.1 Recall: the capacity saddle point . . . 20
4.2 Alternating optimization and majorization–minimization . . . 20
4.3 Gibbs variational principle as an alternating update . . . 20
4.4 Blahut–Arimoto / maximizing mutual information . . . 21
4.5 Expectation maximization: convex mixture-weight case . . . 21
4.6 Sinkhorn algorithm . . . 22
4.7 Tensorization of information ratios . . . 22
4.8 Entropy rate of a stationary process . . . 23
4.9 Entropy and symbol (bit) error rate . . . 24
4.10 f-divergences . . . 25
4.11 Basic properties of f-divergences . . . 26
4.12 Monotonicity and data processing for f-divergences . . . 26
4.13 Total variation and hypothesis testing . . . 26

5 Lecture 5: Further Properties of f-Divergences and the Entropy Method 28
5.1 Total variation and maximal coupling . . . 28
5.2 Product experiments and Hellinger distance . . . 29
5.3 Mutual information based on an f-divergence . . . 30
5.4 Local quadratic behavior: f-divergences look like χ^2 . . . 30
5.5 Variational representation of f-divergence . . . 31
5.6 The entropy method in additive combinatorics . . . 32
5.7 Small sumset implies small difference set . . . 33

6 Lecture 6: Combinatorics, Prime Numbers, and Generalization Bounds 34
6.1 Shearer's lemma and counting triangles . . . 34
6.2 Prime numbers: a first entropy lower bound . . . 35
6.3 A sharper bound: the square factor has bounded entropy . . . 35
6.4 A lower bound on the prime-number-theorem scale . . . 36
6.5 Learning-theoretic setup . . . 38
6.6 Random hypotheses, generalization error, and ERM . . . 38
6.7 Expected generalization error as a comparison of two laws . . . 39
6.8 Subgaussian losses and averaging . . . 39
6.9 Decoupling lemma: one application of Donsker–Varadhan . . . 40
6.10 A mutual-information bound on expected generalization . . . 40
6.11 The Gibbs algorithm and symmetrized information . . . 41
6.12 From the exact identity to an explicit Gibbs bound . . . 42

7 Lecture 7: Metric Entropy and Density Estimation 44
7.1 Covering and packing numbers . . . 44
7.2 Finite-dimensional volume bounds . . . 45
7.3 Beyond volume: the ℓ_1-ball under ℓ_2 distance . . . 47
7.4 Sudakov's minoration . . . 50
7.5 Smooth function classes . . . 52
7.6 Density estimation and the approximation–estimation tradeoff . . . 53

8 Lecture 8: Statistical Decision Theory and Minimax Estimation **56**
8.1 Statistical experiments, estimators, and pointwise risk . . . 56
8.2 The Gaussian location model . . . 56
8.3 Bayes risk and Bayes estimators . . . 57
8.4 Quadratic loss: the conditional mean and MMSE . . . 58
8.5 Gaussian prior: posterior by completing the square . . . 58
8.6 Minimax risk and Bayes lower bounds . . . 59
8.7 A summary of the risk notation . . . 60
8.8 Bare convex-optimization essentials and minimax–Bayes duality . . . 60
8.9 The minimax theorem via convex separation . . . 61
8.10 Multiple observations and sample complexity . . . 63

9 Lecture 9: Fisher-Information and Mutual Information Lower Bounds **65**
9.1 A brief recap of the risk criteria . . . 65
9.2 Local and global information . . . 65
9.3 Frequentist HCR and Cramér–Rao bounds . . . 65
9.4 Bayesian HCR and its local Cramér–Rao consequence . . . 67
9.5 A lucid discrete proof of the Bayesian CR bound . . . 69
9.6 Localized priors and the asymptotic minimax lower bound . . . 71
9.7 Maximum likelihood and asymptotic efficiency . . . 73
9.8 The mutual information method . . . 74

10 Lecture 10: Hypothesis-Testing Lower Bounds and Entropic Upper Bounds **76**
10.1 Two complementary geometric viewpoints . . . 76
10.2 Le Cam's two-point method . . . 76
10.3 Assouad's hypercube method . . . 79
10.4 Fano's packing method . . . 81
10.5 Application: sparse high-dimensional linear regression . . . 82
10.6 From lower bounds to entropic upper bounds . . . 83
10.7 A high-level proof of the Yang–Barron theorem . . . 84
10.8 Application to a smooth finite-dimensional model . . . 86

11 Lecture 11: Entropic Bounds for Statistical Estimation **87**
11.1 Recap: Yang–Barron's construction . . . 87
11.2 Hellinger distance: reminders . . . 87
11.3 The global Le Cam–Birgé theorem . . . 88
11.4 Three-step proof of the global theorem . . . 89
11.5 Global entropy and its local refinement . . . 91
11.6 Why local entropy is of order d in regular models . . . 93
11.7 Density estimation over Hölder classes . . . 94

1 Lecture 1: Entropy, KL Divergence, and Entropy Inequalities

1.1 Shannon entropy

Let X be a discrete random variable with PMF $P = \{P(x) : x \in \mathcal{X}\}$.

Definition 1.1 (Entropy)**.** The Shannon entropy of X is

$$H(X) = H(P) = \mathbb{E}\left[\log \frac{1}{P(X)}\right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)}.$$

We use the convention $0 \log(1/0) = 0$.

For random variables $(X_1, \ldots, X_n)$ with joint PMF $P(x_1, \ldots, x_n)$,

$$H(X_1, \ldots, X_n) = \mathbb{E}\left[\log \frac{1}{P(X_1, \ldots, X_n)}\right] = \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) \log \frac{1}{P(x_1, \ldots, x_n)}.$$

Example 1.2 (Bernoulli entropy)**.** If $X \sim \mathrm{Ber}(p)$, then

$$H(X) = h(p) := -p \log p - (1-p) \log(1-p).$$

With base-2 logarithms, $h(p)$ is maximized at $p = 1/2$ with value 1 bit.

Entropy measures the randomness in X.

1.2 Conditional entropy

For $(X, Y) \sim P_{XY}$,

Definition 1.3 (Conditional entropy)**.**

$$H(X \mid Y) = \mathbb{E}\left[\log \frac{1}{P_{X|Y}(X \mid Y)}\right] = \sum_y P_Y(y) H(P_{X|Y=y}).$$

Example 1.4 (Binary symmetric relation)**.** Suppose $X \sim \mathrm{Ber}(1/2)$ and $Y = X \oplus Z$ where $Z \sim \mathrm{Ber}(\delta)$ is independent of X. Then

$$P_{X|Y}(X = Y \mid Y) = 1 - \delta, \qquad P_{X|Y}(X \neq Y \mid Y) = \delta,$$

so

$$H(X \mid Y) = h(\delta).$$

For $\delta = 0$ or 1, $H(X \mid Y) = 0$; for $\delta = 1/2$, $H(X \mid Y) = 1$ bit.

1.3 Basic properties of entropy

Theorem 1.5 (Basic entropy properties)**.** *For discrete random variables:*

(a) $H(X) \geq 0$*, with equality iff* X *is almost surely constant.*

(b) $H(X) = H(f(X))$ *for every bijection* f*.*

(c) $H(X) \leq \log |\mathcal{X}|$ *when* X *takes values in a finite alphabet* $\mathcal{X}$*.*

(d) $H(X \mid Y) \leq H(X)$ *(conditioning reduces entropy).*

(e) $H(X, Y) = H(X) + H(Y \mid X) \leq H(X) + H(Y)$*.*

(f) $H(X) = H(X, f(X)) \geq H(f(X))$.

(g) Chain rule:

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

Proof of $H(X) \leq \log |\mathcal{X}|$. By concavity of log,

$$H(X) = \mathbb{E}\left[\log \frac{1}{P(X)}\right] \leq \log \mathbb{E}\left[\frac{1}{P(X)}\right] = \log \sum_{x:P(x)>0} 1 \leq \log |\mathcal{X}|.$$

□

Proof that conditioning reduces entropy. Since $P_X(x) = \mathbb{E}_Y[P_{X|Y}(x \mid Y)]$ and $z \mapsto z \log(1/z)$ is concave,

$$\begin{aligned} H(X \mid Y) &= \sum_x \mathbb{E}_Y\left[P_{X|Y}(x \mid Y) \log \frac{1}{P_{X|Y}(x \mid Y)}\right] \\ &\leq \sum_x P_X(x) \log \frac{1}{P_X(x)} = H(X). \end{aligned}$$

□

1.4 Axiomatic characterization and Rényi entropy

For a probability vector $(p_1, \ldots, p_m)$, consider a functional $H_m(p_1, \ldots, p_m)$ satisfying the following properties:

(a) permutation invariance;

(b) appending a zero-probability outcome does not change the value;

(c) normalization $H_2(1/2, 1/2) = \log 2$;

(d) subadditivity for a joint law (r_{ij}) with marginals $p_i = \sum_j r_{ij}$ and $q_j = \sum_i r_{ij}$;

(e) additivity for product distributions;

(f) continuity, in particular $H_2(p, 1-p) \to 0$ as $p \downarrow 0$.

Shannon entropy is the unique functional satisfying these axioms.

Remark 1.6 (Rényi entropy). For $\alpha \in (0,1) \cup (1, \infty)$,

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha,$$

with continuous extensions

$$H_1(P) = \sum_i p_i \log \frac{1}{p_i}, \qquad H_\infty(P) = \log \frac{1}{\max_i p_i}.$$

For $\alpha > \beta \geq 0$, $H_\alpha(P) \leq H_\beta(P)$.

1.5 Method of types

Suppose $n_1, \dots, n_k \in \mathbb{N}_0$ with $\sum_i n_i = n$ and type

$$P = \left(\frac{n_1}{n}, \dots, \frac{n_k}{n}\right).$$

Then

Proposition 1.7 (Method-of-types multinomial bound)**.**

$$\frac{1}{(n+1)^k} \exp(nH(P)) \le \binom{n}{n_1, \dots, n_k} \le \exp(nH(P)).$$

For the binary case $n = (\ell, n-\ell)$,

$$\frac{1}{n+1} \exp\left(nh\left(\frac{\ell}{n}\right)\right) \le \binom{n}{\ell} \le \exp\left(nh\left(\frac{\ell}{n}\right)\right).$$

1.6 Submodularity of entropy

A set function $f : 2^S \to \mathbb{R}$ is submodular if for all $T_1, T_2 \subseteq S$,

$$f(T_1 \cup T_2) + f(T_1 \cap T_2) \le f(T_1) + f(T_2).$$

Equivalently, submodularity can be expressed as diminishing returns: if $T' \subseteq T \subseteq S$ and $b \notin T$, then

$$f(T \cup \{b\}) - f(T) \le f(T' \cup \{b\}) - f(T').$$

Theorem 1.8 (Entropy is submodular)**.** *Let $X^n = (X_1, \dots, X_n)$ be any collection of random variables. Then*

$$T \subseteq [n] \longmapsto H(X_T)$$

is a submodular set function.

Proof. Let

$$A = X_{T_1 \setminus T_2}, \qquad B = X_{T_1 \cap T_2}, \qquad C = X_{T_2 \setminus T_1}.$$

We need

$$H(A, B, C) + H(B) \le H(A, B) + H(B, C).$$

Using the chain rule,

$$\begin{aligned} H(A, B, C) + H(B) &= H(A, C \mid B) + 2H(B) \\ &\le H(A \mid B) + H(C \mid B) + 2H(B) \\ &= H(A, B) + H(B, C). \end{aligned}$$

□

Entropy is also monotone: if $T_1 \subseteq T_2$, then

$$H(X_{T_1}) \le H(X_{T_2}).$$

1.7 Entropy regions and non-Shannon inequalities

Let Γ_n denote the cone of nonnegative, monotone, submodular set functions on $[n]$. Let Γ_n^* denote the set of set functions that arise as entropy vectors of n random variables:

$$h(S) = H(X_i : i \in S), \qquad S \subseteq [n].$$

Then

$$\Gamma_n^* \subseteq \Gamma_n.$$

Let $\overline{\Gamma_n^*}$ be the closure of the entropy region. For $n \le 3$, the Shannon inequalities characterize the closure, but for $n \ge 4$ there are additional constraints:

$$\overline{\Gamma_n^*} \subsetneq \Gamma_n.$$

The Zhang–Yeung inequality (1997) is a celebrated non-Shannon inequality showing that entropy set functions are more constrained than arbitrary nonnegative monotone submodular functions. One form is

$$I(X_3; X_4) - I(X_3; X_4 \mid X_1) - I(X_3; X_4 \mid X_2) \le \frac{1}{2} I(X_1; X_2) + \frac{1}{2} I(X_1; X_3, X_4) + \frac{1}{2} I(X_2; X_3, X_4).$$

Non-Shannon inequalities also enter converse arguments in network coding.

1.8 Relative entropy / KL divergence

For discrete P, Q on a finite or countable alphabet $\mathcal{A}$,

Definition 1.9 (Relative entropy)**.**

$$D(P\|Q) = \begin{cases} \displaystyle\sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)}, & \operatorname{supp}(P) \subseteq \operatorname{supp}(Q), \\ +\infty, & \text{otherwise.} \end{cases}$$

For densities p, q on $\mathbb{R}^k$,

$$D(P\|Q) = \int_{\{p>0, q>0\}} p(a) \log \frac{p(a)}{q(a)} \, \mathrm{d}a,$$

provided the part where $p > 0$ and $q = 0$ has Lebesgue measure zero; otherwise $D(P\|Q) = +\infty$.

For general measures, $P \ll Q$ means P is absolutely continuous with respect to Q. By the Radon–Nikodym theorem there exists $\frac{\mathrm{d}P}{\mathrm{d}Q}$ such that

$$P(E) = \int_E \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}Q.$$

Then

$$D(P\|Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]$$

when $P \ll Q$, and $+\infty$ otherwise.

Theorem 1.10 (Entropy as distance from uniformity)**.** *If P is supported on a finite set $\mathcal{A}$ and $U_{\mathcal{A}}$ is uniform on $\mathcal{A}$, then*

$$H(P) = \log |\mathcal{A}| - D(P\|U_{\mathcal{A}}).$$

Theorem 1.11 (Information inequality)**.**

$$D(P\|Q) \ge 0,$$

with equality iff $P = Q$ (under the usual support conditions).

Proof. Let $\varphi(x) = x \log x$, which is strictly convex on $(0, \infty)$. Then

$$D(P\|Q) = \mathbb{E}_Q\left[\varphi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right] \geq \varphi\left(\mathbb{E}_Q \frac{\mathrm{d}P}{\mathrm{d}Q}\right) = \varphi(1) = 0.$$

□

Corollary 1.12 (Cross-entropy minimization)**.** *If $X \sim P$ is discrete and $H(X) < \infty$, then*

$$\min_Q \mathbb{E}_P\left[\log \frac{1}{Q(X)}\right] = H(P),$$

with unique minimizer $Q = P$.

Indeed,

$$\mathbb{E}_P\left[\log \frac{1}{Q(X)}\right] - H(P) = D(P\|Q) \geq 0.$$

1.9 Differential entropy and Gaussian extremality

If X has density p on $\mathbb{R}^k$,

$$h(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right].$$

For $X \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^d$,

$$h(X) = \frac{1}{2} \log\big((2\pi e)^d \det \Sigma\big).$$

In the scalar case $X \sim \mathcal{N}(0, \sigma^2)$,

$$h(X) = \frac{1}{2} \log(2\pi e \sigma^2).$$

Theorem 1.13 (Gaussian maximizes entropy under a covariance constraint)**.** *If $X \in \mathbb{R}^d$ has mean μ and covariance satisfying $\mathrm{Cov}(X) \preceq \Sigma$, then*

$$h(X) \leq \frac{1}{2} \log\big((2\pi e)^d \det \Sigma\big).$$

In particular, if $\mathrm{Var}(X) \leq \sigma^2$,

$$h(X) \leq \frac{1}{2} \log(2\pi e \sigma^2).$$

1.10 Markov kernels, conditional divergence, and the chain rule

A Markov kernel $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ is a random transformation or channel. For kernels $P_{Y|X}, Q_{Y|X}$ and an input law P_X,

$$D(P_{Y|X}\|Q_{Y|X} \mid P_X) := \sum_x P_X(x) D(P_{Y|X=x}\|Q_{Y|X=x}).$$

Theorem 1.14 (Chain rule for KL divergence)**.**

$$D(P_{XY}\|Q_{XY}) = D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X} \mid P_X).$$

More generally,

$$D(P_{X^n}\|Q_{X^n}) = \sum_{i=1}^n D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}} \mid P_{X^{i-1}}).$$

Tensorization for product measures gives

$$D\left(\prod_{i=1}^{n} P_i \middle\| \prod_{i=1}^{n} Q_i\right) = \sum_{i=1}^{n} D(P_i \| Q_i).$$

If the same input P_X is passed through $P_{Y|X}$ and $Q_{Y|X}$ to give P_Y, Q_Y, then

$$D(P_Y \| Q_Y) \le D(P_{Y|X} \| Q_{Y|X} \mid P_X),$$

with the difference equal to a nonnegative reverse-channel divergence.

Theorem 1.15 (Data-processing inequality for KL divergence)**.** *If the same channel $P_{Y|X}$ is applied to P_X and Q_X, producing P_Y, Q_Y, then*

$$D(P_Y \| Q_Y) \le D(P_X \| Q_X).$$

A deterministic mapping $Y = f(X)$ gives

$$D(P_Y \| Q_Y) \le D(P_X \| Q_X),$$

with equality for one-to-one f.

1.11 A large-deviations estimate from DPI

For any measurable event E,

$$d(P(E), Q(E)) \le D(P \| Q),$$

where the binary divergence is

$$d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

This follows by applying DPI to the binary statistic $Y = \mathbb{1}\{X \in E\}$.

Using

$$d(p, q) \ge p \log \frac{1}{q} - h(p),$$

and $h(p) \le \ln 2$ for natural logarithms, one obtains

$$D(P \| Q) \ge P(E) \ln \frac{1}{Q(E)} - \ln 2,$$

and, for $P(E) > 0$,

$$Q(E) \ge \exp\left(-\frac{D(P \| Q) + \ln 2}{P(E)}\right).$$

With base-2 logarithms, the entropy bound is $h(p) \le 1$ and the corresponding lower bound is $2^{-(D(P\|Q)+1)/P(E)}$.

2 Lecture 2: Mutual Information, Fano, and Variational Characterizations

2.1 Mutual information

Recall that for $P \ll Q$,

$$D(P\|Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right].$$

Definition 2.1 (Mutual information)**.** For $(X, Y) \sim P_{XY}$,

$$I(X;Y) := D(P_{XY}\|P_X P_Y).$$

Mutual information measures how dependent X and Y are. Although it is a function of the joint law P_{XY}, it is often useful to view it as a function of the pair $(P_X, P_{Y|X})$.

Theorem 2.2 (Basic properties of mutual information)**.** *(a)*

$$I(X;Y) = D(P_{Y|X}\|P_Y \mid P_X) = \sum_x P_X(x) D(P_{Y|X=x}\|P_Y).$$

(b) $I(X;Y) = I(Y;X)$.

(c) $I(X;Y) \geq 0$.

(d) For every function f,

$$I(f(X);Y) \leq I(X;Y),$$

with equality for bijective f.

(e)

$$I(X_1, X_2; Y) \geq I(X_1; Y).$$

For discrete X,

$$I(X;Y) = H(X) - H(X \mid Y),$$

and similarly

$$I(X;Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y).$$

The same identity holds formally for differential entropies when the quantities are well defined.

As a corollary of $I(X;Y) \geq 0$,

$$H(X \mid Y) \leq H(X).$$

2.2 Gaussian mutual information

Let (X, Y) be jointly Gaussian with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad |\rho| < 1.$$

Write

$$Y = \rho X + Z, \qquad Z \sim \mathcal{N}(0, 1-\rho^2), \quad Z \perp X.$$

Then

$$I(X;Y) = \frac{1}{2} \log \frac{1}{1-\rho^2}.$$

Thus $I(X;Y) \to \infty$ as $|\rho| \uparrow 1$, while $I(X;Y) = 0$ at $\rho = 0$.

2.3 AWGN channel

Consider

$$Y = X + Z, \qquad X \perp Z, \qquad X \sim \mathcal{N}(0, \sigma_X^2), \quad Z \sim \mathcal{N}(0, \sigma_Z^2).$$

Then $Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_Z^2)$ and

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y \mid X) \\ &= \frac{1}{2} \log \frac{\sigma_X^2 + \sigma_Z^2}{\sigma_Z^2} \\ &= \frac{1}{2} \log(1 + \mathrm{SNR}), \end{aligned}$$

where

$$\mathrm{SNR} = \frac{\sigma_X^2}{\sigma_Z^2}.$$

2.4 Conditional mutual information and conditional independence

Definition 2.3 (Conditional mutual information)**.** For $(X, Y, Z) \sim P_{XYZ}$,

$$I(X;Y \mid Z) = \mathbb{E}_Z[I(X;Y \mid Z = z)] = \sum_z P_Z(z) I(X;Y \mid Z = z)$$

in the discrete case.

Conditional independence $X \perp Y \mid Z$ is equivalent to

$$P_{XY|Z} = P_{X|Z} P_{Y|Z}$$

and to

$$I(X;Y \mid Z) = 0.$$

It is also represented by the Markov chain $X - Z - Y$.

Properties include:

(a) $I(X;Y \mid Z) \geq 0$, with equality iff $X \perp Y \mid Z$.

(b) Chain rule:

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + I(X;Z \mid Y).$$

(c) DPI: if $X - Y - Z$, then

$$I(X;Z) \leq I(X;Y),$$

with equality iff $X - Z - Y$ as well.

(d) If $X - Y - Z - W$, then

$$I(X;W) \leq I(Y;Z).$$

(e)

$$I(X;Y_1,\ldots,Y_n) = \sum_{i=1}^{n} I(X;Y_i \mid Y_1,\ldots,Y_{i-1}).$$

(f) One-to-one transformations of either variable preserve mutual information.

Although $H(X \mid Y) \leq H(X)$, the quantities $I(X;Y \mid Z)$ and $I(X;Z)$ are in general incomparable.

2.5 Strong data processing

The ordinary DPI gives

$$D(P_Y\|Q_Y) \le D(P_X\|Q_X).$$

A strong data-processing inequality seeks a contraction coefficient $\eta < 1$ such that

$$D(P_Y\|Q_Y) \le \eta(P_{Y|X})D(P_X\|Q_X).$$

For a binary symmetric channel BSC(δ), the KL contraction coefficient is

$$\eta_{\mathrm{KL}}(\mathrm{BSC}(\delta)) = (1-2\delta)^2.$$

2.6 Sufficient statistics

If $U - X - Y$, then $I(U;Y) \le I(U;X)$. If equality holds,

$$I(U;Y) = I(U;X),$$

then, as far as U is concerned, summarizing X into Y loses no information; Y is a sufficient statistic of X for U.

Example 2.4 (Coin tosses)**.** Let $\Theta \sim \mathrm{Unif}[0,1]$ and, conditionally on $\Theta = \theta$,

$$B_1, \dots, B_n \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta).$$

Then

$$T = \frac{1}{n}\sum_{i=1}^{n} B_i$$

is a sufficient statistic of B^n for Θ.

2.7 Probability of error and Fano's inequality

Let X take values in an alphabet of size $M < \infty$, and let $\widehat{X}$ be a guess of X. Use the error-probability convention for F_M:

$$F_M(x) = x\log(M-1) + h(x), \qquad x \in [0,1].$$

Then

Theorem 2.5. *If $P_e = \mathbb{P}[X \neq \widehat{X}]$, then*

$$H(X \mid \widehat{X}) \le F_M(P_e).$$

This is the elementary form underlying Fano's inequality.

If $X \sim \mathrm{Unif}([M])$, comparing the joint law of $(X, \widehat{X})$ with a law in which X is uniform and independent of $\widehat{X}$, and applying DPI to the success indicator $\mathbb{1}\{X = \widehat{X}\}$ gives the same binary-divergence mechanism used in the proof of Fano's inequality.

Theorem 2.6 (Fano's inequality)**.** *Let $|\mathcal{X}| = M < \infty$ and $X - Y - \widehat{X}$. If*

$$P_e = \mathbb{P}(X \neq \widehat{X}),$$

then

$$H(X \mid Y) \le h(P_e) + P_e\log(M-1).$$

If $W \sim \mathrm{Unif}([M])$ and $W - X - Y - \widehat{W}$, then

$$P_e \ge 1 - \frac{I(W;Y) + h(P_e)}{\log M} \ge 1 - \frac{I(W;Y) + \log 2}{\log M}.$$

Thus if the data carry little information about a large hypothesis index W, the probability of error must be large.

2.8 Variational characterizations: the golden formula

Variational representations provide useful geometric and optimization views of information quantities. Mutual information can be viewed as the average divergence of the family $\{P_{Y|X=x}\}$ from its "center of gravity" P_Y:

$$I(X;Y) = \sum_x P_X(x) D(P_{Y|X=x} \| P_Y).$$

Theorem 2.7 (Golden formula)**.** *For every distribution Q_Y,*

$$D(P_{Y|X} \| Q_Y \mid P_X) = I(X;Y) + D(P_Y \| Q_Y).$$

Discrete proof. Insert and remove P_Y inside the logarithm:

$$\begin{aligned} I(X;Y) &= \mathbb{E}\left[\log \frac{P_{Y|X}(Y \mid X)}{P_Y(Y)}\right] \\ &= \mathbb{E}\left[\log \frac{P_{Y|X}(Y \mid X)}{Q_Y(Y)}\right] + \mathbb{E}\left[\log \frac{Q_Y(Y)}{P_Y(Y)}\right] \\ &= D(P_{Y|X} \| Q_Y \mid P_X) - D(P_Y \| Q_Y). \end{aligned}$$

□

Hence, for every Q_Y,

$$I(X;Y) \le D(P_{Y|X} \| Q_Y \mid P_X),$$

and

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X} \| Q_Y \mid P_X),$$

with unique minimizer $Q_Y = P_Y$ when the relevant divergences are finite.

Mutual information can also be characterized as distance to product distributions:

$$I(X;Y) = \min_{Q_X, Q_Y} D(P_{XY} \| Q_X Q_Y),$$

with minimizer $(Q_X, Q_Y) = (P_X, P_Y)$. A conditional extension is

$$I(X;Y \mid Z) = \min_{Q_{X|Z}, Q_{Y|Z}} D(P_{XYZ} \| Q_{X|Z} Q_{Y|Z} P_Z).$$

2.9 A supremum representation and Donsker–Varadhan

We next turn from infimum to supremum representations.

Theorem 2.8 (Donsker–Varadhan variational formula)**.** *For probability measures P, Q with $P \ll Q$,*

$$D(P\|Q) = \sup_{f:\mathcal{X}\to\mathbb{R}} \left\{ \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[e^{f(X)}] \right\}.$$

Discrete proof. For this proof, assume that the alphabet is finite and P, Q are strictly positive. The general formula uses a supremum; boundary cases need an approximation rather than a finite-valued maximizing function. For arbitrary f, let

$$Z = \mathbb{E}_Q[e^{f(X)}], \qquad Q_f(x) = \frac{Q(x)e^{f(x)}}{Z}.$$

Then

$$D(P\|Q_f) = \sum_x P(x) \log \frac{P(x)}{Q(x)e^{f(x)}/Z}$$
$$= D(P\|Q) - \mathbb{E}_P[f(X)] + \log Z \geq 0.$$

Thus

$$D(P\|Q) \geq \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[e^{f(X)}].$$

Equality is attained by $f(x) = \log(P(x)/Q(x))$. □

Example 2.9 (Maximum of Gaussian variables)**.** Let $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and

$$J \in \arg\max_{1 \leq i \leq n} Z_i.$$

Using DV with a tilted function proportional to Z_J and the bound

$$I(J; Z^n) = H(J) - H(J \mid Z^n) \leq H(J) \leq \log n,$$

one obtains

$$\mathbb{E} \max_{1 \leq i \leq n} Z_i \leq \frac{\log n}{\lambda} + \frac{\lambda}{2}.$$

Optimizing over $\lambda > 0$ gives

$$\mathbb{E} \max_{1 \leq i \leq n} Z_i \leq \sqrt{2 \log n}.$$

This variational formula also leads naturally to the softmax interpretation developed next.

3 Lecture 3: Gibbs Variational Principle, Softmax, VAE, PAC-Bayes, and Convexity

3.1 Gibbs variational principle

Starting from DV,

$$D(P\|Q) = \sup_f \{\mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]\},$$

one obtains the dual Gibbs variational principle.

Theorem 3.1 (Gibbs variational principle)**.** *For a function $f : \mathcal{X} \to \mathbb{R}$ and reference distribution Q,*

$$\log \mathbb{E}_Q[e^{f(X)}] = \sup_P \{\mathbb{E}_P[f(X)] - D(P\|Q)\}.$$

The optimizer is the exponentially tilted distribution

$$P^*(x) = Q_f(x) := \frac{Q(x)e^{f(x)}}{Z}, \qquad Z = \sum_x Q(x)e^{f(x)}.$$

This is a tradeoff between preferring high scores f and paying a penalty for moving away from the reference Q.

3.2 Softmax and temperature

Take Q to be uniform on $[m]$. Since

$$D(P\|U_m) = \log m - H(P),$$

we obtain

$$\log \sum_{i=1}^m e^{s_i} = \max_{p \in \Delta_m} \left\{ \sum_{i=1}^m p_i s_i + H(p) \right\}.$$

Thus log-sum-exp is the optimal value of an entropy-regularized score maximization problem.

Introducing inverse temperature $\beta > 0$,

$$p^* = \arg\max_{p \in \Delta_m} \left\{ \sum_i p_i s_i + \frac{1}{\beta} H(p) \right\},$$

with

$$p_i^* = \frac{e^{\beta s_i}}{\sum_j e^{\beta s_j}} = \mathrm{softmax}(\beta s)_i.$$

As $\beta \to \infty$, the distribution concentrates on maximizers of s_i; as $\beta \downarrow 0$, it approaches the uniform distribution.

3.3 Transformer attention

Suppose an input sequence contains m token representations $x_1, \ldots, x_m \in \mathbb{R}^p$, arranged as a matrix $X \in \mathbb{R}^{m \times p}$. Define

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V.$$

The score matrix is

$$S = \frac{QK^\top}{\sqrt{d_k}},$$

and the attention matrix is obtained by row-wise softmax:

$$A = \mathrm{softmax}(S).$$

The output is

$$\mathrm{Attn}(Q, K, V) = AV.$$

For token i:

- q_i (query): what token i is looking for;
- k_j (key): what token j advertises about itself;
- v_j (value): the information token j can contribute if selected.

The compatibility score is

$$s_{ij} = \frac{q_i^\top k_j}{\sqrt{d_k}}, \qquad \alpha_{ij} = \frac{e^{s_{ij}}}{\sum_\ell e^{s_{i\ell}}}, \qquad z_i = \sum_{j=1}^{m} \alpha_{ij} v_j.$$

The $1/\sqrt{d_k}$ scaling keeps the raw inner products at order one when q_i and k_j have coordinates of order one and $\mathrm{Var}(q_i^\top k_j)$ is of order d_k.

3.4 Variational autoencoders

Consider observations $y_1, \ldots, y_n \overset{\text{iid}}{\sim} P_Y$ in $\mathbb{R}^d$ and a latent parametric model of the form

$$Y = \mu(X; \theta) + D(X; \theta) Z, \qquad X \perp Z, \quad Z \sim \mathcal{N}(0, I_d),$$

where neural networks parameterize the mean vector and diagonal covariance structure. The likelihood objective is

$$\max_\theta \sum_i \log P_Y^{(\theta)}(y_i).$$

Applying the Gibbs variational principle for fixed y gives the ELBO identity

$$\log P_Y(y) = \sup_R \left\{ \mathbb{E}_R[\log P_{Y|X}(y \mid X)] - D(R \| P_X) \right\}.$$

Equivalently,

$$\log P_Y(y) = \sup_R \mathbb{E}_R \left[\log \frac{P_{XY}(X, y)}{R(X)} \right].$$

In a VAE, the variational family is restricted to a parameterized Gaussian recognition model

$$X = \widetilde{\mu}(Y; \phi) + \widetilde{D}(Y; \phi) \widetilde{Z}, \qquad \widetilde{Z} \sim \mathcal{N}(0, I_d),$$

which allows stochastic-gradient optimization over (θ, ϕ).

The three conceptual steps are:

1. **Encode**: map observed data y into latent space using $R_{X|Y}^{(\phi^*)}$.
2. **Decode**: map a latent x back to data space using $P_{Y|X}^{(\theta^*)}$.
3. **Generate**: sample a fresh $X \sim \mathcal{N}(0, I_d)$ and decode it using $P_{Y|X}^{(\theta^*)}$, producing a new sample from the learned marginal

$$P_Y^{(\theta^*)}(y) = \int P_{Y|X}^{(\theta^*)}(y \mid x) P_X(\mathrm{d}x).$$

3.5 From Donsker–Varadhan to PAC-Bayes

Let $S = (Z_1, \dots, Z_n)$ be the training sample with $Z_i \overset{\text{iid}}{\sim} P_Z$, and let W denote a hypothesis. Define the population and empirical risks

$$L(w) = \mathbb{E}[\ell(w, Z)], \qquad \widehat{L}_S(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i).$$

Let π be a fixed prior on hypotheses chosen independently of S, and let ρ_S be a data-dependent posterior.

DV gives, for any function f,

$$\mathbb{E}_{W\sim\rho_S}[f(W)] \le D(\rho_S\|\pi) + \log \mathbb{E}_{W\sim\pi}[e^{f(W)}].$$

Take

$$f(w) = \lambda n\big(L(w) - \widehat{L}_S(w)\big).$$

If $0 \le \ell(w, z) \le 1$, Hoeffding's inequality gives

$$\mathbb{E}_S e^{\lambda n(L(w)-\widehat{L}_S(w))} \le e^{\lambda^2 n/8}.$$

Averaging over the prior and applying Markov's inequality, with probability at least $1 - \delta$,

$$\mathbb{E}_{W\sim\pi} e^{\lambda n(L(W)-\widehat{L}_S(W))} \le \frac{e^{\lambda^2 n/8}}{\delta}.$$

Hence, simultaneously for all posteriors ρ,

$$\mathbb{E}_\rho[L(W)] \le \mathbb{E}_\rho[\widehat{L}_S(W)] + \frac{D(\rho\|\pi) + \log(1/\delta)}{\lambda n} + \frac{\lambda}{8}.$$

The displayed high-probability statement holds for each $\lambda > 0$ chosen independently of the data. Algebraically, its penalty satisfies

$$\inf_{\lambda>0}\left\{\frac{D(\rho\|\pi) + \ln(1/\delta)}{\lambda n} + \frac{\lambda}{8}\right\} = \sqrt{\frac{D(\rho\|\pi) + \ln(1/\delta)}{2n}}.$$

For a data-dependent ρ, this optimization cannot automatically be made on the same high-probability event: selecting λ after seeing the data requires an additional uniform-in-λ argument. The fixed-λ bound above is the guarantee established here. The term $D(\rho\|\pi)$ measures the "price" for moving from the fixed prior to the data-dependent post-training distribution.

3.6 Four optimization problems in information theory

Four recurring optimization problems are as follows.

1. I-projection. Given an empirical law $\widehat{P}_X$ and a constraint set $\mathcal{B}$,

$$\min_{P\in\mathcal{B}} D(P\|Q).$$

This connects to Sanov's theorem, where probabilities of empirical distributions falling in a rare set behave exponentially like

$$\exp\left(-n \inf_{P\in\mathcal{B}} D(P\|Q)\right).$$

2. Reverse I-projection / maximum likelihood. For i.i.d. data $X_1, \dots, X_n$ and a model class $\mathcal{Q}$,

$$\max_{Q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} \log Q(X_i) \quad \Longleftrightarrow \quad \min_{Q \in \mathcal{Q}} D(\widehat{P}_X \| Q).$$

3. Rate distortion. For fixed source law P_X and a set of admissible channels $\mathcal{S}$,

$$\min_{P_{Y|X} \in \mathcal{S}} I(P_X, P_{Y|X}).$$

This is the source-coding optimization underlying lossy compression such as JPEG.

4. Channel capacity. For a channel $P_{Y|X}$ and admissible input laws $\mathcal{P}$,

$$\max_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}).$$

3.7 Convexity of information measures

Theorem 3.2 (Joint convexity of KL divergence)**.** *The map*

$$(P, Q) \mapsto D(P \| Q)$$

is jointly convex.

Introduce a Bernoulli selector $U \sim \mathrm{Ber}(\lambda)$, with

$$P_{X|U=0} = P_0, \quad P_{X|U=1} = P_1,$$

and similarly for Q. Convexity follows by comparing conditional and marginal divergences after averaging over U.

Consequences:

(i) $P_X \mapsto H(P_X)$ is concave, since

$$H(P_X) = \log |\mathcal{X}| - D(P_X \| U_{\mathcal{X}}).$$

(ii) For fixed $P_{Y|X}$, $P_X \mapsto H(X \mid Y)$ is concave.

(iii) On finite alphabets, entropy and conditional entropy are continuous in the input distribution.

Theorem 3.3 (Convexity/concavity of mutual information)**.** *(i) For fixed channel $P_{Y|X}$, the map*

$$P_X \mapsto I(P_X, P_{Y|X})$$

is concave.

(ii) For fixed input distribution P_X, the map

$$P_{Y|X} \mapsto I(P_X, P_{Y|X})$$

is convex.

For (i), use the golden formula

$$I(P_X, P_{Y|X}) = \min_{Q_Y} D(P_{Y|X} \| Q_Y \mid P_X),$$

which is the pointwise minimum of affine functions of P_X. For (ii), write

$$I(P_X, P_{Y|X}) = D(P_{Y|X} \| P_Y \mid P_X)$$

and use joint convexity, noting that P_Y depends linearly on $P_{Y|X}$.

3.8 Minimax and saddle points

For a bivariate function $f(x, y)$,

$$\inf_x \sup_y f(x, y) \geq \sup_y \inf_x f(x, y)$$

always. Equality is guaranteed when there is a saddle point (x^*, y^*) satisfying

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*) \qquad \forall (x, y).$$

This is the saddle orientation for minimizing in x and maximizing in y. A standard minimax theorem applies when the domains are convex and compact, $f(\cdot, y)$ is convex, $f(x, \cdot)$ is concave, and f is continuous.

3.9 Saddle point for mutual information

Let $\mathcal{P}$ be a convex set of input distributions and suppose a capacity-achieving input distribution $P_X^* \in \mathcal{P}$ exists:

$$C = \sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}).$$

Let $P_Y^* = P_{Y|X} \circ P_X^*$ be the corresponding capacity-achieving output distribution (CAOD). Then the saddle inequalities are

$$D(P_{Y|X} \| P_Y^* \mid P_X) \leq D(P_{Y|X} \| P_Y^* \mid P_X^*) \leq D(P_{Y|X} \| Q_Y \mid P_X^*)$$

for all $P_X \in \mathcal{P}$ and all Q_Y. The second inequality follows directly from the golden formula. The first follows by mixing P_X with P_X^* and using concavity of mutual information, then sending the mixing parameter to zero.

If $C < \infty$, the CAOD is unique. The capacity-achieving input distribution need not be unique, but all capacity-achieving input distributions induce the same output distribution.

Consequently,

$$\max_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) = \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y \mid P_X) = \min_{Q_Y} \max_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y \mid P_X).$$

4 Lecture 4: Alternating Optimization, Tensorization, Entropy Rates, and f-Divergences

4.1 Recall: the capacity saddle point

Let $\mathcal{P}$ be a convex set of input distributions and suppose P_X^* is capacity achieving:

$$\sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}) = C.$$

Let

$$P_Y^* = P_{Y|X} \circ P_X^*$$

be the capacity-achieving output distribution. Then (P_X^*, P_Y^*) is a saddle point for

$$(P_X, Q_Y) \mapsto D(P_{Y|X} \| Q_Y \mid P_X),$$

in the sense that

$$D(P_{Y|X} \| P_Y^* \mid P_X) \le D(P_{Y|X} \| P_Y^* \mid P_X^*) \le D(P_{Y|X} \| Q_Y \mid P_X^*)$$

for all $P_X \in \mathcal{P}$ and all Q_Y.

4.2 Alternating optimization and majorization–minimization

Consider the generic optimization problem

$$\min_t f(t)$$

with a variational representation

$$f(t) = \min_s F(t, s).$$

Assume the inner minimizer exists:

$$s^*(t) \in \arg\min_s F(t, s).$$

An alternating scheme is

$$s_k = s^*(t_{k-1}), \qquad t_k = t^*(s_k), \quad t^*(s) \in \arg\min_t F(t, s).$$

This is connected to majorization–minimization (MM): fixing the auxiliary variable creates a surrogate function that is tight at the current iterate and majorizes the objective.

4.3 Gibbs variational principle as an alternating update

For fixed P_Y and a cost $c(y)$,

$$\min_{Q_Y} \{D(Q_Y \| P_Y) + \mathbb{E}_{Q_Y}[c(Y)]\}$$

has Gibbs-form minimizer

$$Q_Y^*(y) = \frac{1}{Z} P_Y(y) e^{-c(y)}.$$

4.4 Blahut–Arimoto / maximizing mutual information

Channel capacity can be written in a two-block variational form:

$$C = \max_{P_X} I(P_X, P_{Y|X}) = \max_{P_X} \max_{Q_{X|Y}} \mathbb{E}_{P_{XY}} \left[\log \frac{Q_{X|Y}(X \mid Y)}{P_X(X)} \right].$$

To justify the inner maximization, write

$$\mathbb{E}_{P_{XY}} \left[\log \frac{Q_{X|Y}(X \mid Y)}{P_X(X)} \right] = I(X;Y) - D(P_{X|Y} \| Q_{X|Y} \mid P_Y),$$

so equality is attained at

$$Q_{X|Y} = P_{X|Y}.$$

Alternating maximization gives the two updates

$$Q_{X|Y}(x \mid y) \leftarrow \frac{P_X(x) P_{Y|X}(y \mid x)}{\sum_{x'} P_X(x') P_{Y|X}(y \mid x')},$$

and, for fixed $Q_{X|Y}$,

$$P_X(x) \leftarrow \frac{1}{Z} \exp \left\{ \sum_y P_{Y|X}(y \mid x) \log Q_{X|Y}(x \mid y) \right\}.$$

The latter follows by writing the objective as $\log Z - D(P_X \| P'_X)$ for an appropriate Gibbs-form P'_X.

4.5 Expectation maximization: convex mixture-weight case

Consider a finite mixture model with known component distributions and unknown mixing weights. For data $x_1, \ldots, x_n$, the empirical distribution is

$$P_X = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Let the known channel $Q_{X|Y=j}$ be, for example,

$$Q_{X|Y=j} = \mathcal{N}(\mu_j, \sigma_j^2),$$

with unknown mixture law Q_Y. For a discrete observation model, the mixture-weight objective can be written as

$$\min_{Q_Y} D(P_X \| Q_{X|Y} \circ Q_Y).$$

For continuous Gaussian components, the empirical P_X above is discrete, so this literal KL divergence is infinite. In that case the intended objective is the negative empirical log likelihood,

$$\min_{Q_Y} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left(\sum_j Q_Y(j) q_{X|Y=j}(x_i) \right) \right\},$$

where $q_{X|Y=j}$ denotes the component density. The responsibility and weight updates below apply to this likelihood objective.

Introduce a conditional distribution $R_{Y|X}$. Using the chain rule, the optimization can be written in an alternating form. The E-step computes responsibilities

$$R_{Y|X}(j \mid x_i) \propto Q_Y(j) Q_{X|Y=j}(x_i),$$

and the M-step updates the mixing weights by

$$Q_Y(j) \leftarrow \frac{1}{n}\sum_{i=1}^{n} R_{Y|X}(j \mid x_i).$$

In the Gaussian notation from the lecture,

$$R_{Y|X}(j \mid x_i) = \frac{w_j\,\mathcal{N}(x_i;\mu_j,\sigma_j^2)}{\sum_\ell w_\ell\,\mathcal{N}(x_i;\mu_\ell,\sigma_\ell^2)}.$$

4.6 Sinkhorn algorithm

Let Q_{XY} be a fixed reference measure and let V_X, V_Y be prescribed marginals. Consider

$$\min_{P_{XY}} D(P_{XY}\|Q_{XY}) \qquad \text{subject to } P_X = V_X,\ P_Y = V_Y.$$

The optimal coupling has multiplicative form

$$P^*_{XY}(x,y) = A(x)Q_{XY}(x,y)B(y)$$

for suitable scaling functions A, B.

For fixed B, enforce the X-marginal by

$$A(x) \leftarrow \frac{V_X(x)}{\sum_y Q_{XY}(x,y)B(y)}.$$

Similarly, for fixed A,

$$B(y) \leftarrow \frac{V_Y(y)}{\sum_x A(x)Q_{XY}(x,y)}.$$

With

$$Q_{XY}(x,y) = c\exp\left(-\frac{\|x-y\|^2}{\varepsilon}\right),$$

the resulting coupling approximates the optimal-transport coupling as $\varepsilon \downarrow 0$ and can be used to approximate Wasserstein distance.

4.7 Tensorization of information ratios

Theorem 4.1 (Tensorization for a memoryless channel)**.** *Suppose*

$$P_{Y^n|X^n}(y^n \mid x^n) = \prod_{i=1}^{n} P_{Y_i|X_i}(y_i \mid x_i).$$

Then

$$I(X^n;Y^n) \le \sum_{i=1}^{n} I(X_i;Y_i),$$

with equality iff $P_{Y^n} = \prod_i P_{Y_i}$*. Consequently,*

$$\max_{P_{X^n}} I(X^n;Y^n) = \sum_{i=1}^{n} \max_{P_{X_i}} I(X_i;Y_i).$$

Theorem 4.2 (Dual tensorization for a memoryless source)**.** *If*

$$P_{X^n}(x^n) = \prod_{i=1}^{n} P_{X_i}(x_i),$$

then

$$I(X^n; Y^n) \geq \sum_{i=1}^{n} I(X_i; Y_i),$$

with equality iff

$$P_{X^n|Y^n} = \prod_i P_{X_i|Y_i} \quad \textit{a.s. in } Y^n.$$

Hence

$$\min_{P_{Y^n|X^n}} I(X^n; Y^n) = \sum_{i=1}^{n} \min_{P_{Y_i|X_i}} I(X_i; Y_i)$$

for appropriately separated constraints.

We have the identity

$$I(X^n; Y^n) - \sum_{i=1}^{n} I(X_i; Y_i) = D(P_{X^n|Y^n} \| \textstyle\prod_i P_{X_i|Y_i} \mid P_{Y^n}) - D(P_{X^n} \| \textstyle\prod_i P_{X_i}),$$

which reduces to a nonnegative quantity when the source is product.

Example 4.3 (Failure for non-product channels)**.** Let $X_1, X_2 \overset{\text{iid}}{\sim} \text{Ber}(1/2)$ and define

$$Y_1 = X_1 \oplus X_2, \qquad Y_2 = X_1.$$

Then

$$I(X_1; Y_1) = 0, \qquad I(X_2; Y_2) = 0,$$

but (Y_1, Y_2) determines (X_1, X_2), so

$$I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) = 2 \ln 2 \quad \text{(two bits)}.$$

Thus the memoryless-channel tensorization statement fails for a general non-product channel.

4.8 Entropy rate of a stationary process

Let $X = (X_1, X_2, \ldots)$ be a stochastic process. Define the entropy rate, when the limit exists, by

$$\overline{H}(X) := \lim_{n \to \infty} \frac{1}{n} H(X^n).$$

For a discrete process with $H(X_1) < \infty$, a sufficient condition is stationarity.

Definition 4.4 (Stationarity)**.** X is stationary if for all $t_1, \ldots, t_n, k \in \mathbb{N}$,

$$(X_{t_1}, \ldots, X_{t_n}) \overset{d}{=} (X_{t_1+k}, \ldots, X_{t_n+k}).$$

Theorem 4.5 (Entropy rate for stationary processes)**.** *For a stationary discrete process with* $H(X_1) < \infty$*:*

(a)
$$H(X_n \mid X^{n-1}) \le H(X_{n-1} \mid X^{n-2}).$$

(b)
$$\frac{1}{n} H(X^n) \ge H(X_n \mid X^{n-1}).$$

(c)
$$\frac{1}{n} H(X^n) \le \frac{1}{n-1} H(X^{n-1}).$$

(d) Therefore $\frac{1}{n} H(X^n)$ *decreases to a limit, and*
$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X^n) = \lim_{n\to\infty} H(X_n \mid X^{n-1}).$$

For (a), stationarity and conditioning give
$$H(X_n \mid X^{n-1}) \le H(X_n \mid X_2^{n-1}) = H(X_{n-1} \mid X^{n-2}).$$

For (b), the chain rule and monotonicity of these conditional entropies yield
$$\frac{1}{n} H(X^n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid X^{i-1}) \ge H(X_n \mid X^{n-1}).$$

For (c),
$$H(X^n) = H(X^{n-1}) + H(X_n \mid X^{n-1}) \le H(X^{n-1}) + \frac{1}{n} H(X^n),$$
which rearranges to the claimed monotonicity. The final identification follows by Cesàro averaging.

4.9 Entropy and symbol (bit) error rate

Let $X = (X_1, X_2, \ldots)$ and $Y = (Y_1, Y_2, \ldots)$ be stationary processes taking values in an alphabet $[M]$, coupled so that
$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{P}(X_j \ne Y_j) \le \varepsilon \le 1 - \frac{1}{M}.$$

Recall
$$F_M(x) = x \log(M-1) + h(x).$$

Then
$$H(X^n \mid Y^n) \le \sum_{j=1}^{n} H(X_j \mid Y_j) \le \sum_{j=1}^{n} F_M(\mathbb{P}[X_j \ne Y_j]) \le n F_M(\varepsilon),$$
where the last step uses concavity of F_M and its monotonicity on $[0, 1 - 1/M]$.

Consequently, if $\mathbb{P}(X_j \ne Y_j) \le \varepsilon$ for all j, then
$$|H(X) - H(Y)| \le F_M(\varepsilon)$$
for single-letter entropies, and for stationary processes
$$|\overline{H}(X) - \overline{H}(Y)| \le F_M(\varepsilon).$$

In particular, the entropy rate is continuous under small symbol error.

4.10 f-divergences

Definition 4.6 (f-divergence)**.** Let $f : (0, \infty) \to \mathbb{R}$ be convex with $f(1) = 0$. If $P \ll Q$, define

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right].$$

For discrete distributions,

$$D_f(P\|Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right).$$

Important examples include:

(a) KL divergence. For

$$f(t) = t \log t,$$

we recover

$$D_f(P\|Q) = D(P\|Q).$$

(b) Total variation. For

$$f(t) = \frac{1}{2}|t - 1|,$$

$$\mathrm{TV}(P, Q) = D_f(P\|Q) = \frac{1}{2}\sum_x |P(x) - Q(x)| = 1 - \sum_x \min\{P(x), Q(x)\}.$$

(c) χ^2 divergence. For

$$f(t) = (t - 1)^2,$$

$$\chi^2(P\|Q) = \sum_x Q(x)\left(\frac{P(x)}{Q(x)} - 1\right)^2 = \sum_x \frac{(P(x) - Q(x))^2}{Q(x)}.$$

(d) Squared Hellinger distance. For

$$f(t) = (1 - \sqrt{t})^2,$$

$$H^2(P, Q) = \sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2.$$

(e) Le Cam distance. For

$$f(t) = \frac{(1 - t)^2}{2(1 + t)}$$

(up to the equivalent normalization used in the lecture),

$$\mathrm{LC}(P, Q) = \frac{1}{2}\sum_x \frac{(P(x) - Q(x))^2}{P(x) + Q(x)}.$$

Moreover, $\sqrt{\mathrm{LC}(P, Q)}$ is a metric.

(f) Jensen–Shannon divergence. With the generator

$$f(t) = t \log \frac{2t}{t + 1} + \log \frac{2}{t + 1},$$

one obtains

$$\mathrm{JS}(P, Q) = D\left(P\middle\|\frac{P + Q}{2}\right) + D\left(Q\middle\|\frac{P + Q}{2}\right),$$

up to the conventional factor $1/2$. Moreover, $\sqrt{\mathrm{JS}(P, Q)}$ is a metric.

4.11 Basic properties of f-divergences

The basic properties are:

(a) Linearity in the generator:
$$D_{f_1+f_2} = D_{f_1} + D_{f_2}.$$

(b)
$$D_f(P\|P) = 0.$$

(c) $D_f(P\|Q) = 0$ for every pair (P, Q) iff $f(x) = c(x-1)$ for some $c \in \mathbb{R}$; affine terms of this form contribute zero divergence.

(d) Conditional decomposition: with P_X fixed,
$$D_f(P_X P_{Y|X}\|P_X Q_{Y|X}) = \mathbb{E}_{X\sim P_X}\big[D_f(P_{Y|X=X}\|Q_{Y|X=X})\big].$$

(e) Appending the same channel preserves divergence:
$$D_f(P_X P_{Y|X}\|Q_X P_{Y|X}) = D_f(P_X\|Q_X).$$

4.12 Monotonicity and data processing for f-divergences

Theorem 4.7 (Monotonicity)**.** *For joint distributions P_{XY} and Q_{XY},*
$$D_f(P_{XY}\|Q_{XY}) \geq D_f(P_X\|Q_X).$$

The proof uses Jensen's inequality. Schematically,
$$D_f(P_{XY}\|Q_{XY}) = \mathbb{E}_{X\sim Q_X}\mathbb{E}_{Y\sim Q_{Y|X}}\left[f\left(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}\frac{\mathrm{d}P_{Y|X}}{\mathrm{d}Q_{Y|X}}\right)\right]$$
and convexity of f allows the conditional expectation over Y to be moved inside f, leaving
$$\mathbb{E}_{Q_X} f\left(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}\right) = D_f(P_X\|Q_X).$$

Theorem 4.8 (Data-processing inequality for f-divergences)**.** *If the same channel $P_{Y|X}$ is applied to P_X and Q_X, producing P_Y, Q_Y, then*
$$D_f(P_Y\|Q_Y) \leq D_f(P_X\|Q_X).$$

Indeed,
$$D_f(P_X P_{Y|X}\|Q_X P_{Y|X}) = D_f(P_X\|Q_X)$$
and monotonicity under marginalization yields the result.

4.13 Total variation and hypothesis testing

Consider binary hypothesis testing
$$H_0 : X \sim P, \qquad H_1 : X \sim Q,$$
with a possibly randomized decision rule $\phi : \mathcal{X} \to \{0, 1\}$. The sum of the two error probabilities is
$$P(\phi(X) = 1) + Q(\phi(X) = 0).$$

Its optimum is characterized by total variation:

$$\min_{\phi}\Big\{P(\phi(X)=1)+Q(\phi(X)=0)\Big\}=1-\mathrm{TV}(P,Q).$$

The total variation distance has the supremum representation

$$\mathrm{TV}(P,Q)=\sup_{E}\{P(E)-Q(E)\}=\frac{1}{2}\sup_{\|f\|_\infty\le 1}\left\{\mathbb{E}_P[f(X)]-\mathbb{E}_Q[f(X)]\right\}.$$

For densities p, q, the maximizing event is

$$E=\{x : p(x)>q(x)\}.$$

Indeed,

$$\int f(x)(p(x)-q(x))\,\mathrm{d}x\le\int|p(x)-q(x)|\,\mathrm{d}x=2\,\mathrm{TV}(P,Q),$$

and choosing $f=2\mathbb{1}_E-1$ attains the bound.

The coupling representation is

$$\mathrm{TV}(P,Q)=\min_{P_{XY}:\,P_X=P,\ P_Y=Q}\mathbb{P}(X\neq Y).$$

The interpretation is that as P moves farther from Q, total variation increases and the smallest possible disagreement probability between coupled versions of P and Q increases.

5 Lecture 5: Further Properties of f-Divergences and the Entropy Method

5.1 Total variation and maximal coupling

Let P, Q be probability distributions on the same alphabet $\mathcal{X}$. In the finite or countable case,

$$\begin{aligned}\mathrm{TV}(P,Q) &= \frac{1}{2}\sum_x |P(x) - Q(x)| \\ &= 1 - \sum_x \min\{P(x), Q(x)\} \\ &= \sup_{E \subseteq \mathcal{X}} \{P(E) - Q(E)\} \\ &= \frac{1}{2}\sup_{f \in \mathcal{F}} \{\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\},\end{aligned}$$

where

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_\infty \leq 1\}.$$

Definition 5.1 (Coupling)**.** A coupling of P and Q is a joint distribution P_{XY} with marginals $P_X = P$ and $P_Y = Q$.

Theorem 5.2 (Coupling representation of total variation)**.**

$$\mathrm{TV}(P,Q) = \min_{P_{XY}:\, P_X = P,\, P_Y = Q} P_{XY}(X \neq Y). \tag{5.1}$$

A coupling attaining the minimum is called a maximal coupling*.*

Proof. First fix any coupling P_{XY} and any $f \in \mathcal{F}$. Since its marginals are P and Q,

$$\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] = \mathbb{E}_{P_{XY}}[f(X) - f(Y)].$$

If $X = Y$, then $f(X) - f(Y) = 0$. Otherwise,

$$f(X) - f(Y) \leq |f(X)| + |f(Y)| \leq 2.$$

Thus, pointwise,

$$f(X) - f(Y) \leq 2\mathbb{1}\{X \neq Y\}.$$

Taking expectations and then the supremum over $f \in \mathcal{F}$ gives

$$\mathrm{TV}(P,Q) \leq P_{XY}(X \neq Y).$$

For the reverse direction, suppose P, Q have densities p, q with respect to a common dominating measure. For PMFs, replace the integrals below by sums. Define the common part and the residual mass by

$$r(x) = \min\{p(x), q(x)\}, \qquad \varepsilon = \mathrm{TV}(P,Q).$$

Then

$$\int r(x)\,\mathrm{d}x = 1 - \varepsilon, \qquad \int (p - r)\,\mathrm{d}x = \int (q - r)\,\mathrm{d}x = \varepsilon.$$

For $0 < \varepsilon < 1$, construct (X, Y) as follows:

(i) With probability $1 - \varepsilon$, draw Z with density $r/(1-\varepsilon)$ and set $X = Y = Z$.

(ii) With probability ε, draw X, Y independently from the residual densities

$$p_+(x) = \frac{p(x) - r(x)}{\varepsilon}, \qquad q_+(y) = \frac{q(y) - r(y)}{\varepsilon}.$$

The marginal density of X is $r + (p - r) = p$, and that of Y is $r + (q - r) = q$. Also, $p - r$ and $q - r$ have disjoint supports: at each point, at least one of them is zero. Consequently, the residual branch always has $X \neq Y$, whereas the common branch always has $X = Y$. Hence

$$P_{XY}(X \neq Y) = \varepsilon = \mathrm{TV}(P, Q).$$

If $\varepsilon = 0$ or 1, omit the branch of probability zero; this avoids normalizing a zero-mass component. □

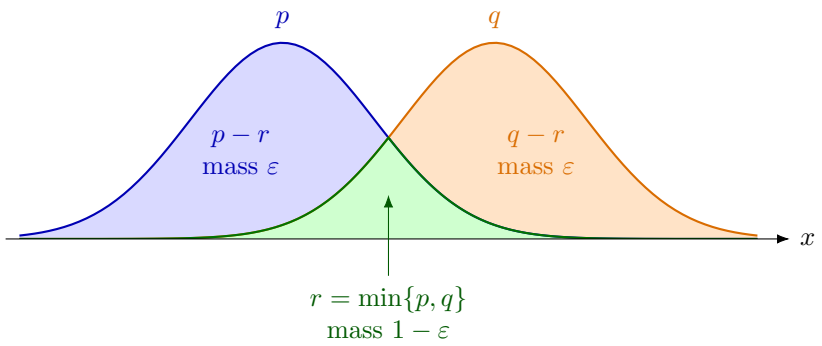


Figure 1: Match the common mass by setting $X = Y$. The two residual parts have disjoint supports and contribute exactly $\mathrm{TV}(P, Q)$ to the probability of disagreement.

5.2 Product experiments and Hellinger distance

For n independent observations, binary hypothesis testing compares

$$H_0 : X^n \sim P^{\otimes n}, \qquad H_1 : X^n \sim Q^{\otimes n}.$$

More generally, the one-observation distributions may depend on n:

$$H_0 : X^n \sim P_n^{\otimes n}, \qquad H_1 : X^n \sim Q_n^{\otimes n}.$$

For instance, the density under the first hypothesis is $\prod_{i=1}^n p_n(x_i)$.

Recall the squared Hellinger distance

$$\begin{aligned} H^2(P, Q) &= \sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 \\ &= 2 - 2\sum_x \sqrt{P(x)Q(x)}. \end{aligned}$$

If $P \ll Q$, this is also

$$H^2(P, Q) = \mathbb{E}_Q\left[\left(1 - \sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}}\right)^2\right] = 2 - 2\mathbb{E}_Q\sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}}.$$

The common-density formula applies even without $P \ll Q$; integrals replace sums on continuous alphabets.

Lemma 5.3 (Comparison with total variation)**.**

$$\frac{1}{2}H^2(P, Q) \leq \mathrm{TV}(P, Q) \leq H(P, Q)\sqrt{1 - \frac{1}{4}H^2(P, Q)} \leq 1.$$

Theorem 5.4 (The one-observation scale for distinguishability)**.** *For any sequence of pairs* (P_n, Q_n)*, as* $n \to \infty$*,*

$$\mathrm{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0 \quad \iff \quad H^2(P_n, Q_n) = o(1/n), \tag{5.2}$$

$$\mathrm{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1 \quad \iff \quad H^2(P_n, Q_n) = \omega(1/n). \tag{5.3}$$

Proof. Write $h_n = H^2(P_n, Q_n)$. The Hellinger affinity tensorizes:

$$\begin{aligned} H^2(P_n^{\otimes n}, Q_n^{\otimes n}) &= 2 - 2 \sum_{x_1,\dots,x_n} \prod_{i=1}^{n} \sqrt{P_n(x_i) Q_n(x_i)} \\ &= 2 - 2 \prod_{i=1}^{n} \left(\sum_{x_i} \sqrt{P_n(x_i) Q_n(x_i)} \right) \\ &= 2 - 2 \left(1 - \frac{h_n}{2}\right)^n. \end{aligned}$$

Equivalently, when $P_n \ll Q_n$, let $X_1, \dots, X_n$ be i.i.d. with law Q_n and factor the expectation

$$\mathbb{E}_{Q_n^{\otimes n}} \prod_{i=1}^{n} \sqrt{\frac{P_n(X_i)}{Q_n(X_i)}} = \prod_{i=1}^{n} \mathbb{E}_{Q_n} \sqrt{\frac{P_n(X_i)}{Q_n(X_i)}}.$$

By Lemma 5.3, total variation tends to zero precisely when the squared Hellinger distance tends to zero, and total variation tends to one precisely when the squared Hellinger distance tends to two. Finally,

$$\left(1 - \frac{h_n}{2}\right)^n \to 1 \iff n h_n \to 0, \qquad \left(1 - \frac{h_n}{2}\right)^n \to 0 \iff n h_n \to \infty.$$

Substitution proves both claims. □

5.3 Mutual information based on an f-divergence

Definition 5.5 (f-mutual information)**.**

$$I_f(X;Y) := D_f(P_{XY} \| P_X P_Y).$$

Theorem 5.6 (Data processing)**.** *If* $U - X - Y$ *is a Markov chain, then*

$$I_f(U;Y) \le I_f(U;X).$$

Proof. Apply data processing for D_f to the stochastic map $(U, X) \mapsto (U, Y)$:

$$I_f(U;X) = D_f(P_{UX} \| P_U P_X) \ge D_f(P_{UY} \| P_U P_Y) = I_f(U;Y).$$

□

5.4 Local quadratic behavior: f-divergences look like χ^2

Recall

$$\chi^2(P\|Q) = \sum_x \frac{(P(x) - Q(x))^2}{Q(x)} = \sum_x Q(x) \left(\frac{P(x)}{Q(x)} - 1\right)^2.$$

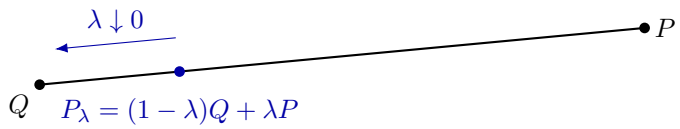
Theorem 5.7 (Local χ^2 expansion)**.** *Let $\mathcal{X}$ be finite, let $Q(x) > 0$ for every x, and let f be convex with $f(1) = 0$ and twice continuously differentiable near 1. For $P_\lambda = (1-\lambda)Q + \lambda P$, as $\lambda \downarrow 0$,*

$$D_f(P_\lambda \| Q) = \frac{f''(1)}{2}\lambda^2 \chi^2(P\|Q) + o(\lambda^2). \tag{5.4}$$

Equivalently,

$$\lim_{\lambda \downarrow 0} \frac{D_f(P_\lambda\|Q)}{\lambda^2} = \frac{f''(1)}{2}\chi^2(P\|Q).$$

The finite-alphabet and positivity assumptions ensure that the Taylor expansion below passes through a finite sum.



Proof. Put $r(x) = P(x)/Q(x)$. Then

$$\frac{P_\lambda(x)}{Q(x)} = 1 + \lambda(r(x) - 1).$$

Expanding f around 1 gives

$$f(1+t) = f(1) + f'(1)t + \frac{1}{2}f''(1)t^2 + o(t^2).$$

Consequently,

$$\begin{aligned} D_f(P_\lambda\|Q) &= \sum_x Q(x) f\big(1 + \lambda(r(x)-1)\big) \\ &= \lambda f'(1) \sum_x Q(x)(r(x)-1) \\ &\quad + \frac{\lambda^2 f''(1)}{2} \sum_x Q(x)(r(x)-1)^2 + o(\lambda^2). \end{aligned}$$

The linear term vanishes, because

$$\sum_x Q(x)(r(x)-1) = \sum_x (P(x) - Q(x)) = 0.$$

The quadratic sum is $\chi^2(P\|Q)$, proving the claim. □

5.5 Variational representation of f-divergence

Let $f : (0,\infty) \to \mathbb{R}$ be convex with $f(1) = 0$. Its *Fenchel–Legendre transform*, or convex conjugate, is

$$f^*(t) = \sup_{u>0}\{ut - f(u)\}, \qquad t \in \mathbb{R}.$$

The function f^* is convex, since it is a supremum of affine functions of t; it is allowed to take the value $+\infty$.

Theorem 5.8 (Finite-alphabet variational representation)**.** *Let P, Q be strictly positive PMFs on a finite alphabet $\mathcal{X}$. Then*

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \left\{\mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))]\right\}. \tag{5.5}$$

A choice with $f^*(g(x)) = +\infty$ contributes value $-\infty$ and can be omitted from the supremum.

Strict positivity keeps every ratio in the domain $(0, \infty)$; it also avoids boundary-value conventions for this classroom version.

Proof. The scalar convex-conjugate identity at $u > 0$ gives

$$f(u) = \sup_{t \in \mathbb{R}} \{ut - f^*(t)\}.$$

Apply this at $u = P(x)/Q(x)$ and multiply by $Q(x)$:

$$Q(x) f\left(\frac{P(x)}{Q(x)}\right) = \sup_{t \in \mathbb{R}} \{P(x)t - Q(x) f^*(t)\}.$$

The value of t can be chosen independently at each x. Since there are only finitely many coordinates,

$$\begin{aligned} D_f(P\|Q) &= \sum_x \sup_t \{P(x)t - Q(x) f^*(t)\} \\ &= \sup_{g: \mathcal{X} \to \mathbb{R}} \sum_x \{P(x) g(x) - Q(x) f^*(g(x))\} \\ &= \sup_g \{\mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))]\}. \end{aligned}$$

□

Remark 5.9 (KL specialization). For natural logarithms, $f(u) = u \ln u$ has conjugate

$$f^*(t) = \sup_{u > 0} \{ut - u \ln u\} = e^{t-1}.$$

Thus the variational formula becomes

$$D(P\|Q) = \sup_g \{\mathbb{E}_P[g(X)] - \mathbb{E}_Q[e^{g(X)-1}]\}.$$

5.6 The entropy method in additive combinatorics

Let $(G, +)$ be an abelian group, for example $\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$. For finite nonempty subsets $A, B \subseteq G$, define

$$A + B = \{a + b : a \in A,\ b \in B\}, \qquad A - B = \{a - b : a \in A,\ b \in B\}.$$

Theorem 5.10 (Ruzsa's sumset inequality / triangle inequality)**.** *For finite nonempty $A, B, C \subseteq G$,*

$$|A - C|\,|B| \le |A - B|\,|B - C|. \tag{5.6}$$

Proof. For each $d \in A - C$, fix one representation

$$d = a_d - c_d, \qquad a_d \in A, \quad c_d \in C.$$

Choose independently

$$D \sim \mathrm{Unif}(A - C), \qquad Y \sim \mathrm{Unif}(B).$$

Then

$$H(D, Y) = H(D) + H(Y) = \log |A - C| + \log |B|.$$

Define

$$U = a_D - Y, \qquad V = Y - c_D.$$

We have $U \in A - B$ and $V \in B - C$. Moreover, (U, V) determines (D, Y):

$$D = U + V, \qquad Y = a_D - U.$$

The fixed representatives (a_d, c_d) are known once $D = d$ is known. Conversely, (D, Y) determines (U, V), so

$$H(U, V) = H(D, Y).$$

Therefore

$$\begin{aligned} \log|A - C| + \log|B| &= H(D, Y) = H(U, V) \\ &\le H(U) + H(V) \\ &\le \log|A - B| + \log|B - C|. \end{aligned}$$

Exponentiation proves (5.6). □

5.7 Small sumset implies small difference set

Set $B = -A$ and $C = A$ in Theorem 5.10. Since $|-A| = |A|$ and $|-A - A| = |A + A|$,

$$|A - A|\,|A| \le |A + A|^2.$$

Thus

$$\boxed{|A + A| \le K|A| \quad \Longrightarrow \quad |A - A| \le K^2|A|.} \tag{5.7}$$

Both $A + A$ and $A - A$ are images of $A \times A$, which contains $|A|^2$ ordered pairs. A small sumset means that many pairs produce the same sum.

Example 5.11 (An arithmetic progression)**.** For $A = \{1, 3, 5, 7\}$,

$$A + A = \{2, 4, 6, 8, 10, 12, 14\}, \qquad A - A = \{-6, -4, -2, 0, 2, 4, 6\}.$$

Both have size 7. More generally, for the first m odd positive integers

$$A_m = \{1, 3, \dots, 2m - 1\},$$

we have

$$|A_m + A_m| = |A_m - A_m| = 2m - 1.$$

Example 5.12 (Powers of two)**.** For

$$A_m = \{1, 2, 4, \dots, 2^{m-1}\},$$

both sets are quadratic in $|A_m| = m$:

$$|A_m + A_m| = \frac{m(m+1)}{2}, \qquad |A_m - A_m| = m(m-1) + 1.$$

6 Lecture 6: Combinatorics, Prime Numbers, and Generalization Bounds

6.1 Shearer's lemma and counting triangles

For three discrete random variables with finite entropies, Shearer's lemma gives

$$H(X_1, X_2, X_3) \le \frac{1}{2}\big(H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)\big). \tag{6.1}$$

Let $G = (V, E)$ be a finite simple undirected graph. Write $N(K_3, G)$ for the number of unordered triangles in G, and define the extremal count

$$N(K_3, m) := \max_{G:\, |E(G)| \le m} N(K_3, G).$$

A complete graph on r vertices has

$$|E(K_r)| = \binom{r}{2} = \frac{r(r-1)}{2}, \qquad N(K_3, K_r) = \binom{r}{3}.$$

Choosing r of order $\sqrt{m}$ shows that $N(K_3, m)$ is at least a constant multiple of $m^{3/2}$ for large m.

Theorem 6.1 (Triangle counting)**.** *Every finite simple graph with m edges satisfies*

$$N(K_3, G) \le \frac{\sqrt{2}}{3}\, m^{3/2}.$$

Consequently,

$$N(K_3, m) = \Theta(m^{3/2}).$$

Proof. If G has no triangles, the claim is immediate. Otherwise choose an ordered triangle uniformly at random and denote its vertices by (X_1, X_2, X_3). There are $3!\, N(K_3, G)$ such ordered triples, so

$$H(X_1, X_2, X_3) = \log\big(3!\, N(K_3, G)\big).$$

Each pair (X_i, X_j) is an oriented edge. There are $2m$ oriented edges, whence

$$H(X_i, X_j) \le \log(2m).$$

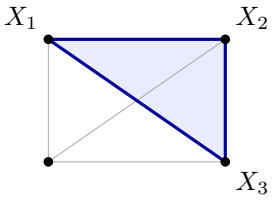
Using (6.1),

$$\log\big(3!\, N(K_3, G)\big) \le \frac{1}{2}\big(H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)\big) \le \frac{3}{2}\log(2m).$$

Therefore

$$N(K_3, G) \le \frac{(2m)^{3/2}}{6} = \frac{\sqrt{2}}{3}\, m^{3/2}.$$

The complete-graph construction supplies the matching order lower bound for the extremal count. □



An ordered triangle determines three oriented edges.

Each vertex appears in two pairs, producing the factor $1/2$.

6.2 Prime numbers: a first entropy lower bound

Logarithm convention for the prime-number arguments. In this subsection and the next two, entropy is in bits and every entropy logarithm is explicitly written as $\log_2$. The notation ln denotes the natural logarithm.

Let $\pi(n)$ be the number of primes at most n, and write them as

$$p_1 < p_2 < \cdots < p_{\pi(n)}.$$

A positive integer is *square-free* if no square of a prime divides it. Every positive integer has a unique decomposition

$$N = L^2 S,$$

where S is square-free and $L \geq 1$ is an integer. For $N \leq n$,

$$S = \prod_{i=1}^{\pi(n)} p_i^{X_i}, \qquad X_i \in \{0,1\}, \qquad L \leq \sqrt{n}.$$

Theorem 6.2 (A logarithmic lower bound)**.** *For every integer $n \geq 2$,*

$$\pi(n) \geq \frac{1}{2} \log_2 n.$$

In particular, there are infinitely many primes.

Proof. Choose $N \sim \mathrm{Unif}\{1, \ldots, n\}$. Uniqueness of the representation gives

$$\begin{aligned} \log_2 n = H(N) = H(L,S) &= H(L, X_1, \ldots, X_{\pi(n)}) \\ &\leq H(L) + \sum_{i=1}^{\pi(n)} H(X_i) \\ &\leq \log_2 \lfloor \sqrt{n} \rfloor + \pi(n) \\ &\leq \frac{1}{2} \log_2 n + \pi(n). \end{aligned}$$

Rearranging proves the bound. □

6.3 A sharper bound: the square factor has bounded entropy

The preceding proof bounds $H(L)$ by the logarithm of its support size. Large values of L are rare, so a distribution-sensitive bound is much smaller.

Theorem 6.3 (Improving the leading coefficient)**.** *For every integer $n \geq 2$,*

$$\pi(n) \geq \log_2 n - 4.$$

Proof. **Step 1: bound the probability of a large square factor.** If $L = \ell$, then $\ell^2 \mid N$. There are exactly $\lfloor n/\ell^2 \rfloor$ multiples of ℓ^2 in $\{1, \ldots, n\}$. Thus

$$P_L(\ell) = \mathbb{P}(L = \ell) \leq \mathbb{P}(\ell^2 \mid N) = \frac{\lfloor n/\ell^2 \rfloor}{n} \leq \frac{1}{\ell^2}.$$

Step 2: compare with an inverse-square reference distribution. Let $M = \lfloor \sqrt{n} \rfloor$ and set

$$Q_M(\ell) = \frac{\ell^{-2}}{Z_M}, \qquad \ell \in \{1, \ldots, M\}, \qquad Z_M = \sum_{\ell=1}^{M} \frac{1}{\ell^2} \leq 2.$$

Nonnegativity of KL divergence, in bits, implies

$$\begin{aligned} 0 \le D(P_L \| Q_M) &= -H(L) - \sum_{\ell=1}^{M} P_L(\ell) \log_2 Q_M(\ell), \\ H(L) &\le \log_2 Z_M + 2\mathbb{E}[\log_2 L] \\ &\le 1 + 2\sum_{\ell=2}^{\infty} \frac{\log_2 \ell}{\ell^2}. \end{aligned}$$

The function $t \mapsto \log_2(t)/t^2$ decreases for $t \ge 2$. Therefore

$$\begin{aligned} \sum_{\ell=2}^{\infty} \frac{\log_2 \ell}{\ell^2} &\le \frac{1}{4} + \int_2^{\infty} \frac{\log_2 t}{t^2}\,\mathrm{d}t \\ &= \frac{1}{4} + \frac{\ln 2 + 1}{2 \ln 2}. \end{aligned}$$

This comparison includes the first summand $\ell = 2$. It follows that

$$H(L) \le 1 + 2\left(\frac{1}{4} + \frac{\ln 2 + 1}{2 \ln 2}\right) = \frac{5}{2} + \frac{1}{\ln 2} < 4.$$

Returning to the entropy decomposition,

$$\log_2 n \le H(L) + \pi(n) < 4 + \pi(n),$$

which proves the stated lower bound. □

6.4 A lower bound on the prime-number-theorem scale

The prime number theorem gives the benchmark

$$\pi(n) \sim \frac{n}{\ln n}.$$

The next argument proves a lower bound of that order, but does not prove the asymptotic equivalence.

Lemma 6.4 (Arithmetic input)**.** *For every integer $n \ge 2$ and $0 \le k \le n$,*

$$\binom{n}{k} \le n^{\pi(n)}.$$

We use this arithmetic lemma in the below proof.

Proof of Lemma 6.4. Fix $0 \le k \le n$. For a prime p, let $\nu_p(m)$ denote the exponent of p in the positive integer m. For example, $72 = 2^3 \cdot 3^2$ gives $\nu_2(72) = 3$ and $\nu_3(72) = 2$. Set $a_p = \nu_p\big(\binom{n}{k}\big)$.

1. Count prime factors in a factorial. Among $1, \dots, n$, the $\lfloor n/p \rfloor$ multiples of p each supply a first factor of p. The $\lfloor n/p^2 \rfloor$ multiples of p^2 each supply an additional factor, and so on. A number divisible by p^a but not p^{a+1} is therefore counted exactly a times. This proves *Legendre's formula*

$$\nu_p(n!) = \sum_{r \ge 1} \left\lfloor \frac{n}{p^r} \right\rfloor. \tag{6.2}$$

Only finitely many terms are nonzero. For instance,

$$\nu_2(10!) = \left\lfloor \frac{10}{2} \right\rfloor + \left\lfloor \frac{10}{4} \right\rfloor + \left\lfloor \frac{10}{8} \right\rfloor = 5 + 2 + 1 = 8.$$

2. Subtract the factors in the denominator. In $\binom{n}{k} = n!/[k!(n-k)!]$, subtract the prime exponents of the two denominator factorials from those of $n!$. Using (6.2),

$$\boxed{a_p = \sum_{r\geq 1} \left[\left\lfloor \frac{n}{p^r} \right\rfloor - \left\lfloor \frac{k}{p^r} \right\rfloor - \left\lfloor \frac{n-k}{p^r} \right\rfloor \right].} \tag{6.3}$$

These are the standard factorial and binomial-exponent formulas. Each bracket is the net contribution at one prime-power level after cancellation.

3. Each bracket is either zero or one. For nonnegative real numbers u, v, write $u = A+\alpha$ and $v = B+\beta$, where A, B are integers and $0 \leq \alpha, \beta < 1$. Then

$$\lfloor u+v \rfloor - \lfloor u \rfloor - \lfloor v \rfloor = \lfloor \alpha + \beta \rfloor \in \{0, 1\}.$$

Apply this with $u = k/p^r$ and $v = (n-k)/p^r$. Since $u+v = n/p^r$, every bracket in (6.3) belongs to $\{0, 1\}$. Although the three factorial counts may be large, their difference is at most one.

4. There are at most $\lfloor \log_p n \rfloor$ contributing levels. If $p^r > n$, all three floors in (6.3) are zero. Therefore

$$a_p \leq \lfloor \log_p n \rfloor, \qquad \boxed{p^{a_p} \leq p^{\lfloor \log_p n \rfloor} \leq n.}$$

Every prime factor of $\binom{n}{k}$ is at most n, because $n! = \binom{n}{k} k!(n-k)!$. Multiplying the preceding bound over the primes up to n proves the promised lemma:

$$\binom{n}{k} = \prod_{p \leq n} p^{a_p} \leq \prod_{p \leq n} n = n^{\pi(n)}. \tag{6.4}$$

□

Theorem 6.5 (A lower bound of order $n/\ln n$)**.** *For every integer $n \geq 2$,*

$$\pi(n) \geq \frac{n - \log_2(n+1)}{\log_2 n} = \frac{n \ln 2 - \ln(n+1)}{\ln n}. \tag{6.5}$$

Proof. Let $B_1, \ldots, B_n$ be independent Ber(1/2) random variables, and put

$$B = (B_1, \ldots, B_n), \qquad K = \sum_{i=1}^{n} B_i.$$

Then $H(B) = n$. Since K is a function of B,

$$n = H(B) = H(B, K) = H(K) + H(B \mid K).$$

The random variable K has at most $n+1$ values, so $H(K) \leq \log_2(n+1)$. Given $K = k$, the vector B is uniformly distributed among the $\binom{n}{k}$ binary vectors of weight k. Consequently,

$$H(B \mid K = k) = \log_2 \binom{n}{k}.$$

Using Lemma 6.4,

$$\begin{aligned} n &\leq \log_2(n+1) + \mathbb{E}_K \left[\log_2 \binom{n}{K} \right] \\ &\leq \log_2(n+1) + \mathbb{E}_K[\log_2(n^{\pi(n)})] \\ &= \log_2(n+1) + \pi(n) \log_2 n. \end{aligned}$$

Rearranging gives (6.5). □

6.5 Learning-theoretic setup

From here onward, logarithms are natural, and information is in nats.

Let $\mathcal{Z}$ be the data space, $\mathcal{W}$ the model or hypothesis class, and

$$\ell : \mathcal{Z} \times \mathcal{W} \to [0, \infty)$$

a loss function. The data distribution p is fixed but unknown. Observe

$$Z^n = (Z_1, \dots, Z_n), \qquad Z_i \overset{\text{iid}}{\sim} p.$$

Definition 6.6 (Population and empirical risk)**.** For a fixed $w \in \mathcal{W}$, the population risk and empirical risk are

$$L(w) = \mathbb{E}_{Z \sim p}[\ell(Z, w)], \qquad L_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, w).$$

The best risk within $\mathcal{W}$ is

$$L^*(\mathcal{W}) = \inf_{w \in \mathcal{W}} L(w).$$

Assume a minimizer $w^* \in \mathcal{W}$ exists, so $L^*(\mathcal{W}) = L(w^*)$.

Example 6.7 (Binary classification)**.** Let $\mathcal{Z} = \mathbb{R}^d \times \{0, 1\}$, with $Z = (X, Y)$. A model is a classifier $w : \mathbb{R}^d \to \{0, 1\}$. With the zero–one loss,

$$\ell((x, y), w) = \mathbb{1}\{y \neq w(x)\}, \qquad L(w) = \mathbb{P}(Y \neq w(X)).$$

Example 6.8 (Regression)**.** Let $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$ and $w : \mathbb{R}^d \to \mathbb{R}$. With squared-error loss,

$$\ell((x, y), w) = (y - w(x))^2, \qquad L(w) = \mathbb{E}[(Y - w(X))^2].$$

Population risk is the performance criterion of interest, but it depends on the unknown p. For each fixed w,

$$\mathbb{E}_{Z^n}[L_n(w)] = L(w).$$

The difficulty is that learning selects w using the same data on which $L_n(w)$ is evaluated.

6.6 Random hypotheses, generalization error, and ERM

A learning algorithm is a channel $P_{W|Z^n}$ from the training dataset to a random hypothesis $W \in \mathcal{W}$. A fresh observation $Z \sim p$ is independent of (Z^n, W).

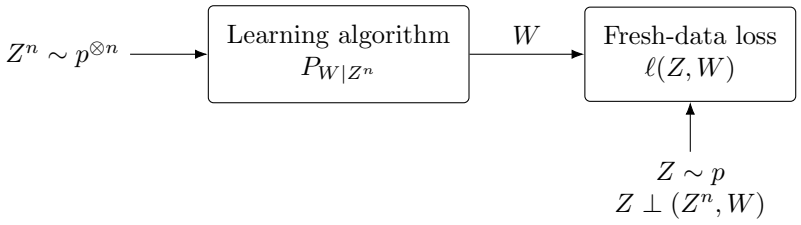


Figure 2: The training data select the hypothesis. Population risk evaluates that hypothesis on an independent observation.

Definition 6.9 (Generalization error and excess risk)**.**

$$\text{gen}(W, Z^n) = L(W) - L_n(W), \qquad \text{excess}(W) = L(W) - L(w^*) \geq 0.$$

Adding and subtracting the two empirical risks gives the exact three-term decomposition

$$\begin{aligned} L(W) - L(w^*) = & \underbrace{L(W) - L_n(W)}_{\text{generalization error}} \\ & + \underbrace{L_n(W) - L_n(w^*)}_{\text{empirical comparison}} \\ & + \underbrace{L_n(w^*) - L(w^*)}_{\text{fluctuation at a fixed model}} . \end{aligned}$$

Definition 6.10 (Empirical risk minimization)**.** The empirical risk minimizer is a data-dependent choice

$$W_{\mathrm{ERM}} = w_{\mathrm{ERM}}(Z^n) \in \arg\min_{w \in \mathcal{W}} L_n(w).$$

6.7 Expected generalization error as a comparison of two laws

To make the dataset argument explicit, write

$$\mathcal{L}(w, z^n) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, w), \qquad L_n(W) = \mathcal{L}(W, Z^n).$$

Introduce a *ghost dataset* $\overline{Z}^n \sim p^{\otimes n}$, independent of (W, Z^n). Since W was not trained on $\overline{Z}^n$,

$$\begin{aligned} \mathbb{E}[\mathcal{L}(W, \overline{Z}^n) \mid W] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(\overline{Z}_i, W) \mid W] \\ &= \frac{1}{n} \sum_{i=1}^{n} L(W) = L(W). \end{aligned}$$

Hence

$$\mathbb{E}[\mathrm{gen}(W, Z^n)] = \mathbb{E}[\mathcal{L}(W, \overline{Z}^n)] - \mathbb{E}[\mathcal{L}(W, Z^n)]. \tag{6.6}$$

This evaluates the *same function* under *two different laws*:

Training evaluation	Independent evaluation
$(W, Z^n) \sim P_{W,Z^n}$	$(W, \overline{Z}^n) \sim P_W \otimes p^{\otimes n}$
The hypothesis depends on the dataset.	The same hypothesis marginal is paired with independent data.

The information-theoretic approach measures the dependence through $I(W; Z^n)$ (Russo–Zou; Xu–Raginsky).

6.8 Subgaussian losses and averaging

Definition 6.11 (Subgaussian random variable)**.** A random variable T is σ^2-subgaussian if

$$\ln \mathbb{E} e^{\lambda(T - \mathbb{E}T)} \le \frac{\lambda^2 \sigma^2}{2} \qquad \text{for every } \lambda \in \mathbb{R}.$$

The parameter σ^2 is the variance proxy.

Standing assumption. For every fixed $w \in \mathcal{W}$, the random loss $\ell(Z, w)$ is σ^2-subgaussian under $Z \sim p$, with the same σ^2 for all w.

For fixed w, independence of the observations gives

$$\begin{aligned}\ln \mathbb{E}_{Z^n} e^{\lambda(\mathcal{L}(w,Z^n)-L(w))} &= \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i} e^{(\lambda/n)(\ell(Z_i,w)-L(w))} \\ &\le \sum_{i=1}^{n} \frac{(\lambda/n)^2\sigma^2}{2} = \frac{\lambda^2\sigma^2}{2n}.\end{aligned}$$

Thus $\mathcal{L}(w, Z^n)$ is (σ^2/n)-subgaussian. The hypothesis is fixed in this calculation; dependence of the learned W on Z^n is handled next.

6.9 Decoupling lemma: one application of Donsker–Varadhan

Lemma 6.12 (Decoupling)**.** *Let $(U, V) \sim P_{UV}$ and $(\overline{U}, \overline{V}) \sim P_U \otimes P_V$. Suppose that for every u, the random variable $\phi(u, V)$ is σ^2-subgaussian under the marginal distribution P_V. Then*

$$\left|\mathbb{E}[\phi(U, V)] - \mathbb{E}[\phi(\overline{U}, \overline{V})]\right| \le \sqrt{2\sigma^2 I(U; V)}. \tag{6.7}$$

Proof. Fix u and put

$$m(u) := \mathbb{E}_{V\sim P_V}[\phi(u, V)], \qquad \Delta(u) := \mathbb{E}[\phi(u, V) \mid U = u] - m(u).$$

Apply the Donsker–Varadhan inequality to $P_{V|U=u}$ and P_V with test function $v \mapsto \lambda\phi(u, v)$. For every $\lambda \in \mathbb{R}$,

$$\begin{aligned}D(P_{V|U=u}\|P_V) &\ge \lambda\mathbb{E}[\phi(u, V) \mid U = u] - \ln \mathbb{E}_{P_V} e^{\lambda\phi(u,V)} \\ &\ge \lambda\Delta(u) - \frac{\lambda^2\sigma^2}{2}.\end{aligned}$$

Optimizing the quadratic over λ yields

$$|\Delta(u)| \le \sqrt{2\sigma^2 D(P_{V|U=u}\|P_V)}.$$

Average over U and use the triangle inequality followed by concavity of the square root:

$$\begin{aligned}\left|\mathbb{E}\phi(U, V) - \mathbb{E}\phi(\overline{U}, \overline{V})\right| &= |\mathbb{E}_U \Delta(U)| \\ &\le \mathbb{E}_U|\Delta(U)| \\ &\le \mathbb{E}_U \sqrt{2\sigma^2 D(P_{V|U}\|P_V)} \\ &\le \sqrt{2\sigma^2 \mathbb{E}_U D(P_{V|U}\|P_V)} \\ &= \sqrt{2\sigma^2 I(U; V)}.\end{aligned}$$

□

6.10 A mutual-information bound on expected generalization

Theorem 6.13 (Expected generalization bound)**.** *Under the standing subgaussian assumption,*

$$|\mathbb{E}[L(W) - L_n(W)]| \le \sqrt{\frac{2\sigma^2}{n} I(W; Z^n)}. \tag{6.8}$$

Proof. For each fixed w, $\mathcal{L}(w, Z^n)$ is (σ^2/n)-subgaussian. Apply Lemma 6.12 with

$$U = W, \qquad V = Z^n, \qquad \phi(w, z^n) = \mathcal{L}(w, z^n),$$

and use (6.6). □

Remark 6.14 (Interpreting the information term). A smaller value of $I(W; Z^n)$ gives a smaller upper bound in (6.8). The factor $1/n$ also improves that bound when the information term is held fixed. These statements concern the bound, not a monotonicity theorem for the actual error.

For a deterministic algorithm with a discrete output, $I(W; Z^n) = H(W)$, which can be large. For continuous deterministic outputs, one cannot use $h(W \mid Z^n) = 0$; the mutual information may be infinite.

Corollary 6.15 (Finite hypothesis class)**.** *If $\mathcal{W}$ is finite, then every learning algorithm satisfies*

$$|\mathbb{E}[L(W) - L_n(W)]| \le \sqrt{\frac{2\sigma^2 \ln |\mathcal{W}|}{n}}.$$

Proof. Use $I(W; Z^n) \le H(W) \le \ln |\mathcal{W}|$ in Theorem 6.13. □

6.11 The Gibbs algorithm and symmetrized information

Write $S = Z^n$ and

$$L_n(w, s) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, w), \qquad L(w) = \mathbb{E}_{Z \sim p}[\ell(Z, w)].$$

Fix a data-independent prior π and an inverse temperature $\beta > 0$. The Gibbs algorithm draws its output from

$$\rho_s(w) := P_{W|S=s}(w) = \frac{\pi(w) e^{-\beta L_n(w,s)}}{Z_\beta(s)}, \qquad Z_\beta(s) = \int e^{-\beta L_n(w,s)} \pi(w)\, \mathrm{d}w. \tag{6.9}$$

For discrete hypotheses, use a sum in the definition of $Z_\beta(s)$. The exponent uses the *average* empirical loss, not the total loss.

As $\beta \downarrow 0$, the Gibbs distribution approaches the prior π; it approaches the uniform distribution when the prior is uniform. As $\beta \to \infty$, it concentrates on empirical minimizers, provided those minimizers are accessible under the prior. For a finite hypothesis class with a positive prior and finite losses, these conclusions follow directly from (6.9).

Define the expected *signed* generalization error

$$g_\beta := \mathbb{E}_{W,S}[L(W) - L_n(W, S)].$$

Recall mutual information and introduce its reverse-KL counterpart:

$$I(W; S) = D(P_{WS} \| P_W \otimes P_S),$$
$$I_{\mathrm{laut}}(W; S) = D(P_W \otimes P_S \| P_{WS}).$$

The second quantity is *lautum information* (Palomar–Verdú). Their sum is the symmetrized KL information:

$$I_{\mathrm{SKL}}(W; S) := I(W; S) + I_{\mathrm{laut}}(W; S).$$

There is no factor of $1/2$ in this convention.

Theorem 6.16 (Exact Gibbs generalization identity)**.** *For the Gibbs algorithm,*

$$\boxed{g_\beta = \frac{I_{\mathrm{SKL}}(W;S)}{\beta}.} \tag{6.10}$$

This is the result of Aminian, Bu, Toni, Rodrigues, and Wornell (NeurIPS 2021). Assume the expectations in the following calculation are finite; finite data and hypothesis spaces with a strictly positive prior and finite losses give a sufficient elementary setting.

Proof. For brevity, write

$$P = P_{WS}, \qquad Q = P_W \otimes P_S.$$

Under P, the model is evaluated on the dataset that trained it; under Q, the same model marginal is evaluated on independent data. Thus

$$\mathbb{E}_P[L_n(W,S)] = \text{expected training loss}, \qquad \mathbb{E}_Q[L_n(W,S)] = \mathbb{E}[L(W)].$$

In particular,

$$g_\beta = \mathbb{E}_Q[L_n(W,S)] - \mathbb{E}_P[L_n(W,S)]. \tag{6.11}$$

Expand the Gibbs log-likelihood ratio:

$$\begin{aligned}\ln \frac{P(w,s)}{Q(w,s)} &= \ln \frac{P_{W|S=s}(w)}{P_W(w)} \\ &= -\beta L_n(w,s) + \underbrace{\ln \frac{\pi(w)}{P_W(w)}}_{a(w)} - \underbrace{\ln Z_\beta(s)}_{b(s)}.\end{aligned}$$

Therefore

$$\begin{aligned}I_{\mathrm{SKL}}(W;S) &= D(P\|Q) + D(Q\|P) \\ &= \mathbb{E}_P\left[\ln \frac{P}{Q}\right] - \mathbb{E}_Q\left[\ln \frac{P}{Q}\right] \\ &= \mathbb{E}_P[-\beta L_n(W,S) + a(W) - b(S)] \\ &\quad - \mathbb{E}_Q[-\beta L_n(W,S) + a(W) - b(S)].\end{aligned}$$

The two laws have the same marginals, so

$$\mathbb{E}_P[a(W)] = \mathbb{E}_Q[a(W)], \qquad \mathbb{E}_P[b(S)] = \mathbb{E}_Q[b(S)].$$

Both marginal terms cancel, leaving

$$I_{\mathrm{SKL}}(W;S) = -\beta\big(\mathbb{E}_P[L_n(W,S)] - \mathbb{E}_Q[L_n(W,S)]\big) = \beta g_\beta.$$

This proves (6.10). □

6.12 From the exact identity to an explicit Gibbs bound

Theorem 6.17 (Gibbs expected generalization bound)**.** *Under the standing σ^2-subgaussian assumption,*

$$0 \le g_\beta \le \frac{2\beta\sigma^2}{n}. \tag{6.12}$$

In particular, if $0 \le \ell(z,w) \le 1$, then

$$\boxed{0 \le \mathbb{E}[L(W) - L_n(W,S)] \le \frac{\beta}{2n}.} \tag{6.13}$$

Proof. The mutual-information bound gives

$$|g_\beta| \le \sqrt{\frac{2\sigma^2}{n} I(W;S)}.$$

By the exact Gibbs identity and nonnegativity of both KL divergences,

$$g_\beta \ge 0, \qquad I(W;S) \le I_{\mathrm{SKL}}(W;S) = \beta g_\beta.$$

Consequently,

$$g_\beta \le \sqrt{\frac{2\sigma^2}{n} I(W;S)} \le \sqrt{\frac{2\beta\sigma^2}{n} g_\beta}.$$

When $g_\beta > 0$, squaring and dividing by g_β gives $g_\beta \le 2\beta\sigma^2/n$. The case $g_\beta = 0$ is immediate.

If $0 \le \ell \le 1$, Hoeffding's lemma yields, for each fixed w,

$$\ln \mathbb{E}_Z e^{\lambda(\ell(Z,w) - L(w))} \le \frac{\lambda^2}{8} = \frac{\lambda^2 (1/4)}{2}.$$

Thus one can take $\sigma^2 = 1/4$ in (6.12), giving

$$\frac{2\beta\sigma^2}{n} = \frac{2\beta(1/4)}{n} = \frac{\beta}{2n}.$$

□

Remark 6.18 (Temperature and fitting). Large β makes Gibbs resemble ERM and can increase the tendency to fit the training data. Small β makes the output resemble the independent prior; for bounded losses, (6.13) then forces the expected gap to zero. The theorem controls the expected signed gap, not the gap on every dataset and not the population risk by itself.

7 Lecture 7: Metric Entropy and Density Estimation

7.1 Covering and packing numbers

Let (V, d) be a metric space and let $\Theta \subseteq V$.

Definition 7.1 (Covering number)**.** An *ε-covering*, or *ε-net*, of Θ is a finite set $\{v_1, \dots, v_N\} \subseteq V$ such that

$$\Theta \subseteq \bigcup_{i=1}^{N} B(v_i, \varepsilon), \qquad B(v, \varepsilon) := \{u \in V : d(u, v) \le \varepsilon\}.$$

The covering number is

$$N(\Theta, d, \varepsilon) := \min\bigl\{N : \text{there exists an } \varepsilon\text{-covering of } \Theta \text{ of size } N\bigr\}.$$

Its logarithm $\log N(\Theta, d, \varepsilon)$ is called the *metric entropy* at resolution ε.

Definition 7.2 (Packing number)**.** An *ε-packing* of Θ is a finite set $\{\theta_1, \dots, \theta_M\} \subseteq \Theta$ such that

$$\min_{i \neq j} d(\theta_i, \theta_j) > \varepsilon.$$

Equivalently, the balls $B(\theta_i, \varepsilon/2)$ are pairwise disjoint. The packing number is

$$M(\Theta, d, \varepsilon) := \max\bigl\{M : \text{there exists an } \varepsilon\text{-packing of } \Theta \text{ of size } M\bigr\}.$$

The quantity $\log M(\Theta, d, \varepsilon)$ is also called the ε-capacity.

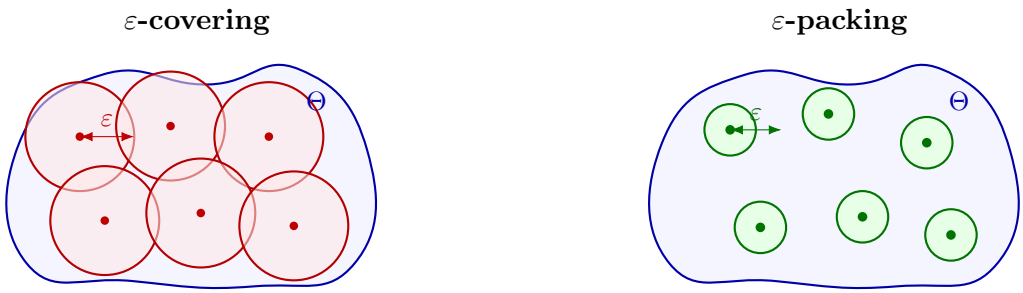


Figure 3: Covering asks how many radius-ε balls suffice; packing asks how many points can be kept more than ε apart.

Theorem 7.3 (Kolmogorov–Tikhomirov comparison)**.** *For every $\varepsilon > 0$,*

$$M(\Theta, d, 2\varepsilon) \le N(\Theta, d, \varepsilon) \le M(\Theta, d, \varepsilon). \tag{7.1}$$

Proof. For the right inequality, take a maximal ε-packing $E = \{\theta_1, \dots, \theta_M\}$. If some $\theta \in \Theta$ were more than ε from every point of E, then θ could be added to the packing, contradicting maximality. Thus the balls $B(\theta_i, \varepsilon)$ cover Θ, so $N(\Theta, d, \varepsilon) \le M(\Theta, d, \varepsilon)$.

For the left inequality, let $\{x_1, \dots, x_N\}$ be an ε-cover and let $\{\theta_1, \dots, \theta_M\}$ be a 2ε-packing. If $M > N$, then two packing points lie in the same covering ball, say $\theta_i, \theta_j \in B(x_k, \varepsilon)$. The triangle inequality gives

$$d(\theta_i, \theta_j) \le d(\theta_i, x_k) + d(x_k, \theta_j) \le 2\varepsilon,$$

contradicting the packing property. Hence $M \le N$. □

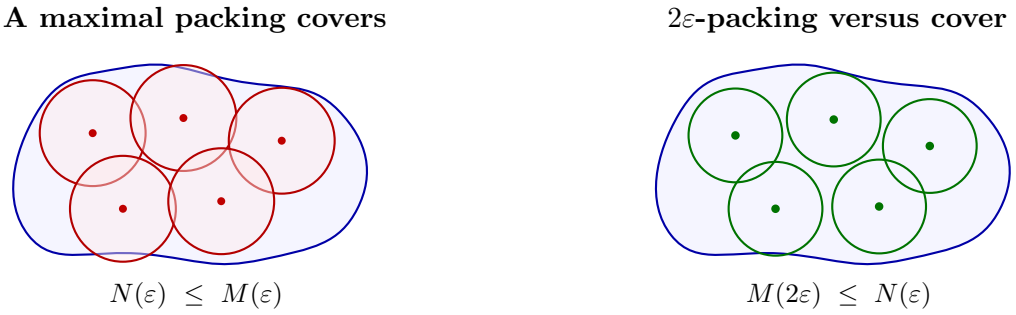


Figure 4: The two geometric ideas behind (7.1).

The comparison has two distinct uses. In fixed dimension one often studies $\varepsilon \downarrow 0$. In high-dimensional problems, d and ε may vary together, and the behavior can change near a critical scale such as $\varepsilon \asymp d^{-1/2}$.

7.2 Finite-dimensional volume bounds

Let $V = \mathbb{R}^d$ with an arbitrary norm $\|\cdot\|$, and write

$$B := \{x \in \mathbb{R}^d : \|x\| \le 1\}.$$

For sets $A, C \subseteq \mathbb{R}^d$, their Minkowski sum is

$$A + C := \{a + c : a \in A,\ c \in C\}.$$

Theorem 7.4 (Volume bounds)**.** *Let $\Theta \subseteq \mathbb{R}^d$ be measurable and let $\varepsilon > 0$. Then*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B)} \le N(\Theta, \|\cdot\|, \varepsilon) \le M(\Theta, \|\cdot\|, \varepsilon) \le \frac{\mathrm{vol}(\Theta + \frac{\varepsilon}{2}B)}{\mathrm{vol}(\frac{\varepsilon}{2}B)}. \tag{7.2}$$

If, in addition, Θ is convex and $\varepsilon B \subseteq \Theta$, then

$$M(\Theta, \|\cdot\|, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B)}. \tag{7.3}$$

Proof. If Θ is covered by N radius-ε balls, then

$$\mathrm{vol}(\Theta) \le \sum_{i=1}^{N} \mathrm{vol}(B(v_i, \varepsilon)) = N\varepsilon^d \,\mathrm{vol}(B),$$

which proves the first inequality. The middle inequality is (7.1).

For an ε-packing $\{\theta_1, \dots, \theta_M\}$, the balls $B(\theta_i, \varepsilon/2)$ are pairwise disjoint and lie in $\Theta + (\varepsilon/2)B$. Hence

$$M\ \mathrm{vol}\Big(\frac{\varepsilon}{2}B\Big) \le \mathrm{vol}\Big(\Theta + \frac{\varepsilon}{2}B\Big),$$

which proves the last inequality in (7.2).

Now suppose that Θ is convex and $\varepsilon B \subseteq \Theta$. Then $(\varepsilon/2)B \subseteq (1/2)\Theta$ and

$$\Theta + \frac{1}{2}\Theta = \frac{3}{2}\Theta.$$

Indeed, convexity gives $\frac{2}{3}\theta_1 + \frac{1}{3}\theta_2 \in \Theta$, so $\theta_1 + \frac{1}{2}\theta_2 \in \frac{3}{2}\Theta$; the reverse inclusion follows from $\frac{3}{2}\theta = \theta + \frac{1}{2}\theta$. Therefore

$$\Theta + \frac{\varepsilon}{2}B \subseteq \frac{3}{2}\Theta,$$

and substituting this into the packing bound gives (7.3). □

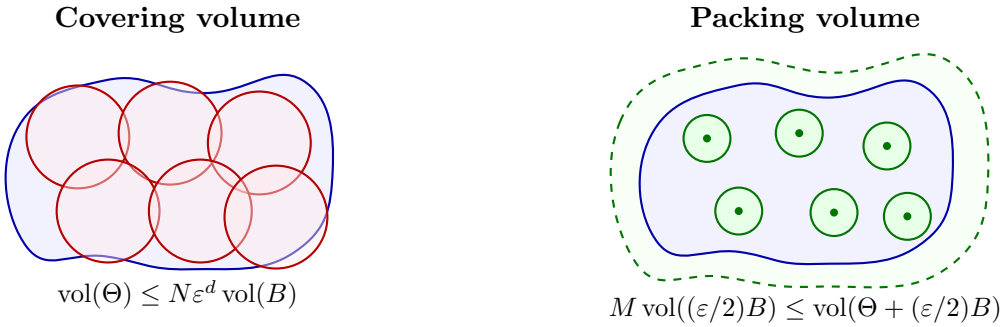


Figure 5: The volume argument compares either a cover with Θ, or disjoint packing balls with a thickening of Θ.

Corollary 7.5 (Balls and spheres)**.** *Let*

$$B = \{x : \|x\| \le 1\}, \qquad S = \{x : \|x\| = 1\}.$$

For $0 < \varepsilon < 1$,

$$\left(\frac{1}{\varepsilon}\right)^d \le N(B, \|\cdot\|, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^d, \tag{7.4}$$

and

$$N(S, \|\cdot\|, \varepsilon) \le 2d \left(1 + \frac{2}{\varepsilon}\right)^{d-1}. \tag{7.5}$$

If $\|\cdot\|$ *is an absolute norm, then also*

$$\left(\frac{1}{2\varepsilon}\right)^{d-1} \le N(S, \|\cdot\|, \varepsilon). \tag{7.6}$$

Here a norm is *absolute* if

$$\|(x_1, \ldots, x_d)\| = \|(|x_1|, \ldots, |x_d|)\|.$$

Such norms are monotone under coordinatewise increase of absolute values. Every ℓ_p norm is absolute.

Proof. The ball lower bound follows directly from (7.2). For the upper bound,

$$B + \frac{\varepsilon}{2}B = \left(1 + \frac{\varepsilon}{2}\right) B,$$

so the packing upper bound gives (7.4).

For the sphere, every ball $B(s, \varepsilon/2)$ centered at $s \in S$ lies in the shell

$$\left(1 + \frac{\varepsilon}{2}\right) B \setminus \operatorname{int}\left(1 - \frac{\varepsilon}{2}\right) B.$$

Consequently,

$$\begin{aligned} N(S, \|\cdot\|, \varepsilon) &\le M(S, \|\cdot\|, \varepsilon) \\ &\le \frac{(1 + \varepsilon/2)^d - (1 - \varepsilon/2)^d}{(\varepsilon/2)^d} \\ &= \frac{d}{(\varepsilon/2)^d} \int_{-\varepsilon/2}^{\varepsilon/2} (1 + t)^{d-1} \, \mathrm{d}t \\ &\le 2d \left(1 + \frac{2}{\varepsilon}\right)^{d-1}. \end{aligned}$$

For completeness, assume the norm is absolute and define on $\mathbb{R}^{d-1}$ the restricted norm $\|x\|' = \|(x,0)\|$, with unit ball B'. Every $x \in B'$ can be lifted vertically to a point $x^{\uparrow} = (x, t_x) \in S$ with $t_x \geq 0$. Coordinatewise monotonicity gives

$$\|x^{\uparrow} - y^{\uparrow}\| \geq \|(x-y,0)\| = \|x-y\|'.$$

Thus a 2ε-packing of B' lifts to a 2ε-packing of S. By the covering–packing comparison and the $(d-1)$-dimensional volume lower bound,

$$N(S, \|\cdot\|, \varepsilon) \geq M(S, \|\cdot\|, 2\varepsilon) \geq N(B', \|\cdot\|', 2\varepsilon) \geq \left(\frac{1}{2\varepsilon}\right)^{d-1}.$$

□



Figure 6: The shell gives the sphere upper bound. Under an absolute norm, vertical lifting gives the lower bound.

For a compact set Θ with nonempty interior and fixed d,

$$N(\Theta, \|\cdot\|, \varepsilon) \asymp M(\Theta, \|\cdot\|, \varepsilon) \asymp \varepsilon^{-d} \qquad (\varepsilon \downarrow 0),$$

where the constants may depend on Θ, the norm, and d.

Definition 7.6 (Minkowski dimension)**.** When the limit exists, the Minkowski or box-counting dimension is

$$\dim_{\mathrm{M}}(\Theta) := \lim_{\varepsilon \downarrow 0} \frac{\log N(\Theta, \|\cdot\|, \varepsilon)}{\log(1/\varepsilon)}.$$

For a smooth manifold, the exponent is its intrinsic dimension rather than the dimension of the ambient Euclidean space. For example,

$$\mathrm{O}(d) := \{Q \in \mathbb{R}^{d \times d} : Q^{\top} Q = I_d\}$$

has dimension $d(d-1)/2$: its tangent space at the identity consists of skew-symmetric matrices, which have one free entry for each unordered pair of coordinates.

7.3 Beyond volume: the ℓ_1-ball under ℓ_2 distance

Let

$$B_1^d := \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}, \qquad B_2^d := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}.$$

The volume bounds give

$$\frac{\mathrm{vol}(B_1^d)}{\mathrm{vol}(\varepsilon B_2^d)} \leq N(B_1^d, \|\cdot\|_2, \varepsilon) \leq M(B_1^d, \|\cdot\|_2, \varepsilon) \leq \frac{\mathrm{vol}(B_1^d + (\varepsilon/2)B_2^d)}{\mathrm{vol}((\varepsilon/2)B_2^d)}.$$

The relevant volumes are

$$\mathrm{vol}(B_1^d) = \frac{2^d}{d!}, \qquad \mathrm{vol}(B_2^d) = \frac{\pi^{d/2}}{\Gamma(1+d/2)}.$$

The Gamma function extends the factorial: $\Gamma(m+1) = m!$ for integers $m \geq 0$. Stirling's formula gives

$$\mathrm{vol}(B_1^d)^{1/d} \asymp \frac{1}{d}, \qquad \mathrm{vol}(B_2^d)^{1/d} \asymp \frac{1}{\sqrt{d}},$$

and hence

$$\left(\frac{\mathrm{vol}(B_1^d)}{\mathrm{vol}(B_2^d)}\right)^{1/d} \asymp \frac{1}{\sqrt{d}}.$$

Therefore the volume lower bound has the form

$$M(B_1^d, \|\cdot\|_2, \varepsilon) \gtrsim \left(\frac{c}{\varepsilon\sqrt{d}}\right)^d.$$

For the upper bound, Cauchy–Schwarz gives $B_2^d \subseteq \sqrt{d}\, B_1^d$, so

$$B_1^d + \frac{\varepsilon}{2} B_2^d \subseteq \left(1 + \frac{\varepsilon\sqrt{d}}{2}\right) B_1^d.$$

Consequently,

$$M(B_1^d, \|\cdot\|_2, \varepsilon) \lesssim \left[C\left(1 + \frac{1}{\varepsilon\sqrt{d}}\right)\right]^d.$$

These bounds are informative when $\varepsilon \lesssim d^{-1/2}$. When $\varepsilon \gg d^{-1/2}$, the lower bound becomes trivial and the upper bound can be exponentially loose.

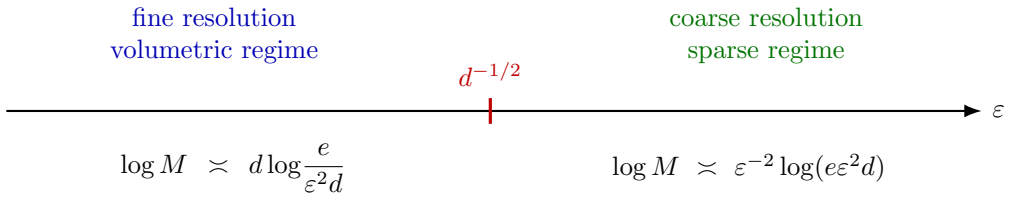


Figure 7: The complexity changes character at $\varepsilon \asymp d^{-1/2}$.

Theorem 7.7 (Two regimes for B_1^d)**.** *For* $0 < \varepsilon < 1$,

$$\log M(B_1^d, \|\cdot\|_2, \varepsilon) \asymp \begin{cases} d \log \dfrac{e}{\varepsilon^2 d}, & 0 < \varepsilon \leq d^{-1/2}, \\ \dfrac{1}{\varepsilon^2} \log(e\varepsilon^2 d), & d^{-1/2} \leq \varepsilon < 1. \end{cases} \tag{7.7}$$

The fine-resolution regime follows, up to universal constants, from the volume calculation. We now prove the upper and lower bounds in the coarse regime.

Upper bound: coarse approximation creates sparsity

Fix $x \in B_1^d$ and choose the mesh size

$$\delta := \varepsilon^2.$$

Quantize each coordinate toward zero:

$$\widehat{x}_i := \delta \, \mathrm{sgn}(x_i) \left\lfloor \frac{|x_i|}{\delta} \right\rfloor. \tag{7.8}$$

Thus positive coordinates are rounded down and negative coordinates are rounded up to the adjacent multiple of δ. In particular,

$$|x_i - \widehat{x}_i| < \delta, \qquad |\widehat{x}_i| \le |x_i|.$$

Because the quantization is toward zero,

$$\|x - \widehat{x}\|_1 = \|x\|_1 - \|\widehat{x}\|_1 \le 1, \qquad \|x - \widehat{x}\|_\infty \le \delta.$$

Using $\|v\|_2^2 \le \|v\|_1 \|v\|_\infty$,

$$\|x - \widehat{x}\|_2^2 \le \|x - \widehat{x}\|_1 \|x - \widehat{x}\|_\infty \le \delta = \varepsilon^2.$$

Hence $\|x - \widehat{x}\|_2 \le \varepsilon$.

Write $\widehat{x}_i = \delta m_i$ with $m_i \in \mathbb{Z}$. Since $\|\widehat{x}\|_1 \le 1$,

$$\sum_{i=1}^{d} |m_i| = \frac{\|\widehat{x}\|_1}{\delta} \le \frac{1}{\delta} =: k.$$

Every nonzero coordinate consumes at least one unit of this integer budget, so

$$|\operatorname{supp}(\widehat{x})| \le k \asymp \varepsilon^{-2}. \tag{7.9}$$

There is no second approximation: after forming $\widehat{x}$, we merely encode its support, signs, and integer magnitudes.

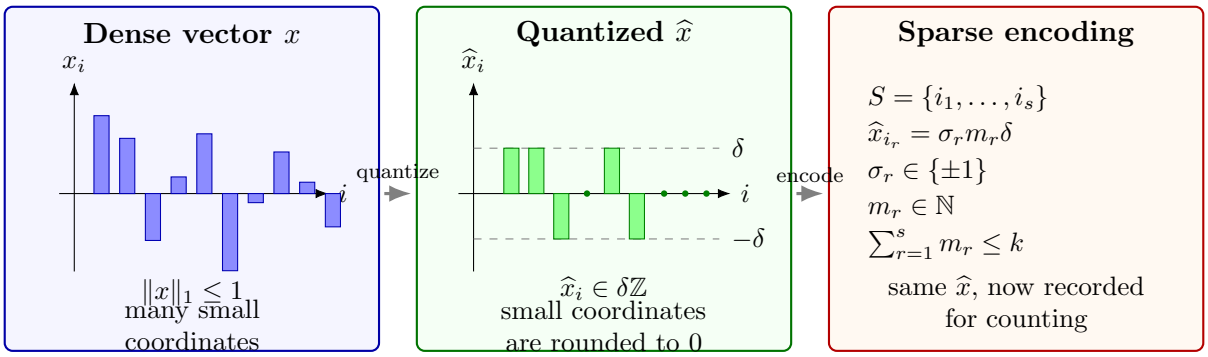


Figure 8: At coarse ℓ_2 resolution, quantization removes small coordinates. The third panel rewrites the same quantized vector $\widehat{x}$ as support, signs, and integer magnitudes, which is the form needed for counting.

A quantized vector can be written as a sum of at most k signed coordinate atoms:

$$\widehat{x} = \delta \sum_{r=1}^{k} a_r, \qquad a_r \in \{0, \pm e_1, \dots, \pm e_d\}.$$

The order of the atoms is irrelevant. The number of possible multisets is at most

$$\binom{2d+k}{k}.$$

Thus these vectors form an ε-net. In the coarse regime $k \lesssim d$, and therefore

$$\begin{aligned} \log N(B_1^d, \|\cdot\|_2, \varepsilon) &\le \log \binom{2d+k}{k} \\ &\le k \log \frac{e(2d+k)}{k} \\ &\lesssim k \log \frac{ed}{k} \\ &\asymp \frac{1}{\varepsilon^2} \log(e\varepsilon^2 d). \end{aligned}$$

The covering–packing comparison converts this into the same-order packing upper bound after changing ε by a universal factor.

Lower bound: sparse vectors from a Hamming packing

Lemma 7.8 (Constant-weight Gilbert–Varshamov bound)**.** *For $1 \le k \le d/8$, there exists*

$$\mathcal{U} \subseteq \{u \in \{0,1\}^d : \|u\|_0 = k\}$$

such that

$$d_H(u,v) \ge \frac{k}{2} \quad (u \ne v), \qquad \log|\mathcal{U}| \ge ck \log \frac{d}{k},$$

where $c > 0$ is universal.

For $u \in \mathcal{U}$, define

$$\theta_u := \frac{u}{k}.$$

Then $\|\theta_u\|_1 = 1$, so $\theta_u \in B_1^d$, and

$$\|\theta_u - \theta_v\|_2^2 = \frac{d_H(u,v)}{k^2} \ge \frac{1}{2k}.$$

Choosing $k \asymp \varepsilon^{-2}$, with the constant chosen so that $k \le d/8$, produces an ε-packing. Hence

$$\log M(B_1^d, \|\cdot\|_2, \varepsilon) \gtrsim k \log \frac{d}{k} \asymp \frac{1}{\varepsilon^2} \log(\varepsilon^2 d),$$

which matches the coarse-regime upper bound up to universal constants.

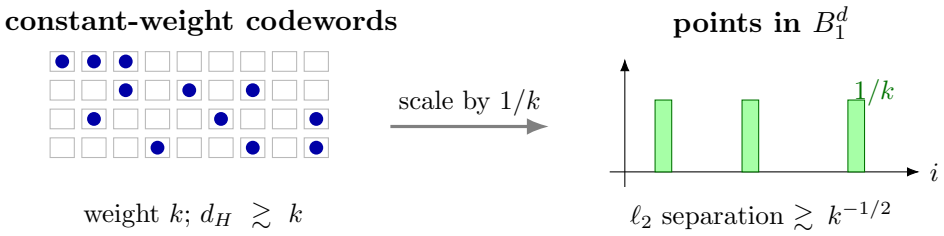


Figure 9: Hamming separation becomes Euclidean separation after scaling.

7.4 Sudakov's minoration

Definition 7.9 (Gaussian width)**.** For $\Theta \subseteq \mathbb{R}^d$, its Gaussian width is

$$w(\Theta) := \mathbb{E}\left[\sup_{\theta \in \Theta} \langle \theta, Z \rangle\right], \qquad Z \sim \mathcal{N}(0, I_d).$$

Theorem 7.10 (Sudakov's minoration)**.** *For every $\Theta \subseteq \mathbb{R}^d$,*

$$w(\Theta) \gtrsim \sup_{\varepsilon > 0} \varepsilon \sqrt{\log M(\Theta, \|\cdot\|_2, \varepsilon)}. \tag{7.10}$$

We use two standard Gaussian facts.

Lemma 7.11 (Slepian comparison)**.** *Let $X_1, \dots, X_M$ and $Y_1, \dots, Y_M$ be centered Gaussian random variables. If*

$$\mathbb{E}(Y_i - Y_j)^2 \le \mathbb{E}(X_i - X_j)^2 \qquad \text{for all } i, j,$$

then

$$\mathbb{E} \max_i Y_i \le \mathbb{E} \max_i X_i.$$

Lemma 7.12 (Maximum of independent Gaussians)**.** *If $G_1, \dots, G_M$ are independent $\mathcal{N}(0,1)$ random variables, then, for $M \geq 2$,*

$$\mathbb{E} \max_{1 \leq i \leq M} G_i \asymp \sqrt{\log M}.$$

Proof of Sudakov's minoration. Fix $\varepsilon > 0$ and let $\{\theta_1, \dots, \theta_M\}$ be an ε-packing of Θ. Set

$$X_i := \langle \theta_i, Z \rangle, \qquad Z \sim \mathcal{N}(0, I_d),$$

and let $Y_1, \dots, Y_M$ be independent $\mathcal{N}(0, \varepsilon^2/2)$ random variables. For $i \neq j$,

$$\mathbb{E}(X_i - X_j)^2 = \|\theta_i - \theta_j\|_2^2 > \varepsilon^2 = \mathbb{E}(Y_i - Y_j)^2.$$

Slepian's lemma therefore gives

$$\mathbb{E} \max_i X_i \geq \mathbb{E} \max_i Y_i.$$

Since $Y_i \overset{d}{=} (\varepsilon/\sqrt{2}) G_i$,

$$\begin{aligned} w(\Theta) &= \mathbb{E} \sup_{\theta \in \Theta} \langle \theta, Z \rangle \\ &\geq \mathbb{E} \max_{1 \leq i \leq M} X_i \\ &\geq \mathbb{E} \max_{1 \leq i \leq M} Y_i \\ &\gtrsim \varepsilon \sqrt{\log M}. \end{aligned}$$

Taking the supremum over ε proves (7.10). □

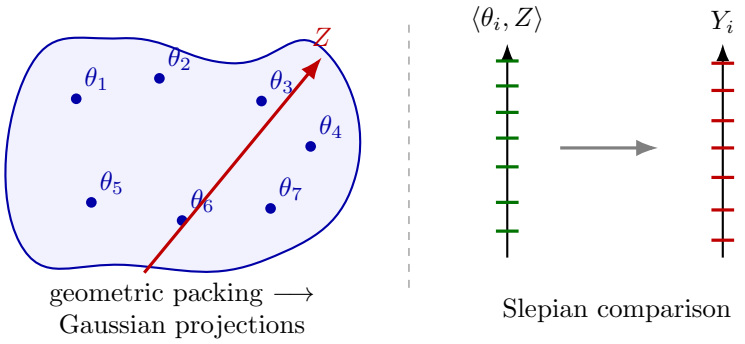


Figure 10: Sudakov's proof turns metric separation into a lower bound on a Gaussian maximum.

Corollary 7.13 (Urysohn-type volume lower bound)**.** *For measurable $\Theta \subseteq \mathbb{R}^d$,*

$$w(\Theta) \gtrsim \sqrt{d} \left(\frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B_2^d)} \right)^{1/d} \asymp d \, \mathrm{vol}(\Theta)^{1/d}. \tag{7.11}$$

Proof. Let

$$r_\Theta := \left(\frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B_2^d)} \right)^{1/d}.$$

The volume lower bound gives

$$M(\Theta, \|\cdot\|_2, \varepsilon) \geq \left(\frac{r_\Theta}{\varepsilon} \right)^d.$$

Sudakov's minoration therefore yields

$$w(\Theta) \gtrsim \varepsilon \sqrt{d \log \frac{r_\Theta}{\varepsilon}} \qquad (0 < \varepsilon < r_\Theta).$$

Taking $\varepsilon = e^{-1/2} r_\Theta$ proves the first comparison in (7.11); the second follows from $\mathrm{vol}(B_2^d)^{1/d} \asymp d^{-1/2}$. □

7.5 Smooth function classes

Metric entropy becomes particularly important in infinite-dimensional spaces.

Definition 7.14 (Hölder class)**.** Let $\beta = \ell + \alpha$, where $\ell \in \mathbb{Z}_{\geq 0}$ and $\alpha \in (0, 1]$. Let $\mathcal{F}_\beta(A, L)$ be the collection of probability densities on $[0, A]^d$ that are ℓ times continuously differentiable and satisfy

$$\|D^{(\ell)} f(x) - D^{(\ell)} f(y)\|_\infty \leq L\|x - y\|_\infty^\alpha \qquad (x, y \in [0, A]^d).$$

Theorem 7.15 (Metric entropy of Hölder classes)**.** *For fixed A, L, d, β and every $1 \leq p \leq \infty$,*

$$\log N(\mathcal{F}_\beta(A, L), \|\cdot\|_p, \varepsilon) = \Theta\Big(\varepsilon^{-d/\beta}\Big). \tag{7.12}$$

The ratio d/β is the effective complexity exponent. Increasing the dimension increases complexity, while increasing smoothness decreases it. A useful heuristic partitions $[0, A]^d$ into cells of side length

$$h \asymp \varepsilon^{1/\beta}.$$

There are

$$h^{-d} \asymp \varepsilon^{-d/\beta}$$

cells, corresponding to the number of local descriptions visible at resolution ε.



Figure 11: The exponent d/β reflects the number of local pieces needed at accuracy ε.

Definition 7.16 (Sobolev class)**.** For an integer $s \geq 1$ and $1 \leq r \leq \infty$, the Sobolev ball $W^{s,r}(A, L)$ consists of functions $f : [0, A]^d \to \mathbb{R}$ whose weak partial derivatives $D^\alpha f$ exist for all multi-indices $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $|\alpha| := \alpha_1 + \cdots + \alpha_d \leq s$ and satisfy

$$\|f\|_{W^{s,r}} := \begin{cases} \left(\displaystyle\sum_{|\alpha| \leq s} \|D^\alpha f\|_r^r \right)^{1/r}, & 1 \leq r < \infty, \\ \max_{|\alpha| \leq s} \|D^\alpha f\|_\infty, & r = \infty, \end{cases} \leq L.$$

If we use Sobolev balls as density classes, we also impose $f \geq 0$ and $\int_{[0,A]^d} f(x)\,\mathrm{d}x = 1$.

Proposition 7.17 (Metric entropy of Sobolev balls)**.** *Sobolev balls have essentially the same entropy exponent as Hölder classes, with the smoothness parameter β replaced by s. In the basic Hilbert case,*

$$\log N\big(W^{s,2}(A, L), \|\cdot\|_2, \varepsilon\big) \asymp \left(\frac{L}{\varepsilon}\right)^{d/s}. \tag{7.13}$$

More generally, under the usual Sobolev-embedding conditions, the same exponent d/s governs the L_p covering numbers.

The difference from the Hölder class is conceptual. A Hölder class controls derivatives *pointwise* and uniformly, so it encodes a uniform modulus of continuity. A Sobolev class controls derivatives only in an *integrated* sense through L_r norms of weak derivatives. Thus Hölder smoothness is stronger and more local, whereas Sobolev smoothness is weaker but often more natural in analysis and PDE. For metric entropy, however, both classes lead to the same form

$$\log N(\cdot, \varepsilon) \asymp \varepsilon^{-(\text{dimension})/(\text{smoothness})}.$$

In other words, the effective complexity exponent is d/β for Hölder classes and d/s for Sobolev classes.

7.6 Density estimation and the approximation–estimation tradeoff

We now state the statistical problem precisely. Let $X_1, \dots, X_n$ be i.i.d. observations from an unknown density $f \in \mathcal{F}_\beta(A, L)$ on $[0, A]^d$, with respect to Lebesgue measure. An estimator is any measurable density-valued function

$$\widehat{f}_n = \widehat{f}_n(X_1, \dots, X_n).$$

Under squared L_2 loss,

$$R(\widehat{f}_n, f) := \mathbb{E}_f \|\widehat{f}_n - f\|_2^2, \qquad \|g\|_2^2 := \int_{[0,A]^d} g(x)^2 \, \mathrm{d}x.$$

The minimax risk is

$$R_n^*(\mathcal{F}_\beta) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}_\beta(A,L)} \mathbb{E}_f \|\widehat{f}_n - f\|_2^2. \tag{7.14}$$

For a transparent upper-bound argument, first restrict to a subclass on which every density is bounded above and away from zero:

$$0 < c \le f(x) \le C < \infty.$$

On this subclass, KL divergence and squared L_2 distance are equivalent up to constants:

$$D(f\|q) \asymp \|f - q\|_2^2.$$

The boundedness assumption is a technical bridge between the KL aggregation argument and squared L_2 risk.

Proposition 7.18 (Finite-dictionary aggregation)**.** *Let $q_1, \dots, q_N$ be candidate densities, and let $X_1, \dots, X_n$ be i.i.d. from a density f. There exists an estimator $\overline{q}_n$ such that*

$$\mathbb{E}_f D(f\|\overline{q}_n) \le \min_{1 \le j \le N} D(f\|q_j) + \frac{\log N}{n}. \tag{7.15}$$

Proof. Put the uniform prior on the candidates and form the mixture density on the entire sample,

$$Q_n(x^n) := \frac{1}{N} \sum_{j=1}^N q_j^{\otimes n}(x^n).$$

For every j, $Q_n(x^n) \ge N^{-1} q_j^{\otimes n}(x^n)$, so

$$D(f^{\otimes n}\|Q_n) \le n D(f\|q_j) + \log N.$$

Factor Q_n sequentially into predictive densities $\widehat{q}_t(\cdot \mid X^t)$ and define

$$\overline{q}_n := \frac{1}{n}\sum_{t=0}^{n-1} \widehat{q}_t(\cdot \mid X^t).$$

The chain rule for relative entropy and convexity of $D(f\|\cdot)$ give

$$\mathbb{E}_f D(f\|\overline{q}_n) \le \frac{1}{n} D(f^{\otimes n}\|Q_n) \le D(f\|q_j) + \frac{\log N}{n}.$$

Minimizing over j proves (7.15). □

Take an ε-net $\{q_1, \dots, q_{N(\varepsilon)}\}$ of the density class under L_2. For every f, an oracle can select q_{j^*} satisfying

$$\|f - q_{j^*}\|_2 \le \varepsilon, \qquad D(f\|q_{j^*}) \lesssim \varepsilon^2.$$

The first term is the approximation error. Because the estimator does not know j^*, learning among the $N(\varepsilon)$ candidates costs $\log N(\varepsilon)/n$.

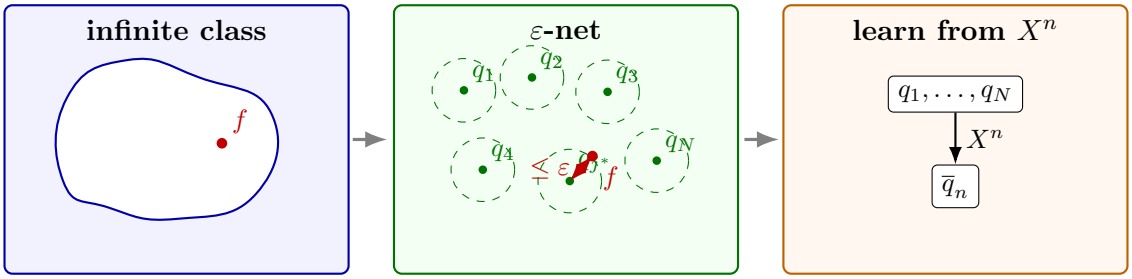


Figure 12: Discretization replaces an infinite class by $N(\varepsilon)$ candidates. Approximation costs ε^2; selection among candidates costs $\log N(\varepsilon)/n$.

Combining the two costs gives the entropic upper bound

$$R_n^*(\mathcal{F}_\beta) \lesssim \inf_{\varepsilon>0}\left\{\varepsilon^2 + \frac{1}{n}\log N(\mathcal{F}_\beta, \|\cdot\|_2, \varepsilon)\right\}. \tag{7.16}$$

Using (7.12),

$$R_n^*(\mathcal{F}_\beta) \lesssim \inf_{\varepsilon>0}\left\{\varepsilon^2 + \frac{1}{n}\varepsilon^{-d/\beta}\right\}.$$

Let $a = d/\beta$. For $g(\varepsilon) = \varepsilon^2 + n^{-1}\varepsilon^{-a}$,

$$g'(\varepsilon) = 2\varepsilon - \frac{a}{n}\varepsilon^{-a-1}.$$

The minimizing resolution satisfies

$$\varepsilon_n^{a+2} \asymp \frac{1}{n}, \qquad \varepsilon_n \asymp n^{-\beta/(2\beta+d)}.$$

At this resolution the two terms have the same order, giving

$$R_n^*(\mathcal{F}_\beta) \lesssim n^{-2\beta/(2\beta+d)}. \tag{7.17}$$

Why the rate is best possible. A standard local bump construction produces an L_2 packing of the Hölder class with separation ε and

$$\log M(\varepsilon) \gtrsim \varepsilon^{-d/\beta},$$

while the KL divergence between the corresponding n-sample laws is of order at most $n\varepsilon^2$. Fano's inequality prevents reliable identification whenever

$$n\varepsilon^2 \lesssim \log M(\varepsilon).$$

Balancing the two sides again yields $\varepsilon^2 \asymp n^{-2\beta/(2\beta+d)}$, and therefore

$$R_n^*(\mathcal{F}_\beta) \gtrsim n^{-2\beta/(2\beta+d)}. \tag{7.18}$$

Together, (7.17) and (7.18) show the minimax-optimal rate

$$\boxed{R_n^*(\mathcal{F}_\beta) \asymp n^{-2\beta/(2\beta+d)}.}$$

Larger d makes estimation harder (the curse of dimensionality), while larger β makes it easier because the density is smoother.

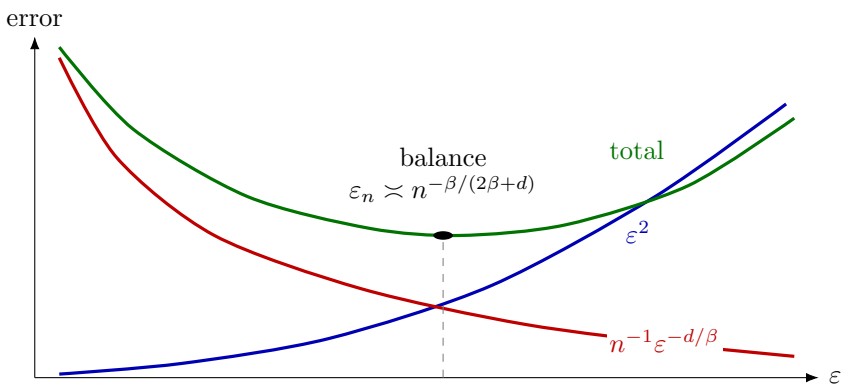


Figure 13: A coarse cover has large approximation error; a fine cover has large statistical complexity. The optimal resolution balances them.

Lecture 7 summary

1. Covering and packing numbers agree up to a constant change of radius.
2. In fixed finite dimension, volume gives metric entropy of order $d\log(1/\varepsilon)$.
3. For B_1^d under ℓ_2 distance, coarse resolution reveals sparsity and produces the sharper entropy $\varepsilon^{-2}\log(e\varepsilon^2 d)$.
4. Sudakov's minoration converts metric separation into Gaussian width.
5. A β-Hölder density class in dimension d has entropy exponent d/β, leading to the minimax density-estimation rate $n^{-2\beta/(2\beta+d)}$.

8 Lecture 8: Statistical Decision Theory and Minimax Estimation

8.1 Statistical experiments, estimators, and pointwise risk

Definition 8.1 (Statistical experiment)**.** A statistical experiment, or statistical model, is a family of probability distributions

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

on a common observation space $\mathcal{X}$. The unknown parameter is $\theta \in \Theta$, where Θ is the parameter space.

A statistical decision problem has four ingredients.

(i) Nature chooses a parameter $\theta \in \Theta$ and generates an observation $X \sim P_\theta$.

(ii) The statistician observes X and wishes to estimate either θ itself or an estimand $T(\theta)$, where $T : \Theta \to \mathcal{Y}$.

(iii) An estimator $\widehat{T}$ maps the data to an action. It may be deterministic or randomized. Formally, a randomized estimator is described by a Markov kernel $P_{\widehat{T}|X}(d\widehat{t} \mid x)$.

(iv) A loss function $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}$ assigns the cost $\ell(T(\theta), \widehat{T})$.

Definition 8.2 (Pointwise risk)**.** The risk of a fixed estimator $\widehat{T}$ at the fixed truth θ is

$$\begin{aligned} R_\theta(\widehat{T}) &:= \mathbb{E}_\theta\Big[\ell\big(T(\theta), \widehat{T}\big)\Big] \\ &= \int_{\mathcal{X}} \int_{\widehat{\mathcal{Y}}} \ell\big(T(\theta), \widehat{t}\big) P_{\widehat{T}|X}(d\widehat{t} \mid x) P_\theta(dx). \end{aligned} \tag{8.1}$$

Here the expectation is over the observation and any internal randomization of the estimator; the parameter θ is held fixed.

For the remainder of this lecture, we usually estimate the parameter itself, so $T(\theta) = \theta$ and the estimator is denoted by $\widehat{\theta}$.

8.2 The Gaussian location model

The Gaussian location model (GLM) is

$$\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2 I_d) : \theta \in \Theta\}, \qquad X = \theta + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I_d), \tag{8.2}$$

where the noise variance $\sigma^2 > 0$ is known. Typical parameter spaces include

$$\Theta = \mathbb{R}^d, \qquad \Theta = \{\theta : \|\theta\|_p \le r\}, \qquad \Theta = \{\theta : \|\theta\|_0 \le k\},$$

and, for matrix-valued parameters,

$$\Theta = \{\theta \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(\theta) \le k\}.$$

Recall that

$$\|\theta\|_p = \left(\sum_{i=1}^{d} |\theta_i|^p\right)^{1/p}, \quad 1 \le p < \infty, \qquad \|\theta\|_\infty = \max_i |\theta_i|,$$

whereas

$$\|\theta\|_0 := \sum_{i=1}^{d} \mathbb{1}\{\theta_i \neq 0\}$$

counts the nonzero coordinates and is not a norm.

A common loss is

$$\ell(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_p^{\alpha}, \qquad 1 \le p \le \infty, \quad \alpha > 0.$$

For quadratic loss, $p = \alpha = 2$. With no structural restriction on θ, the maximum-likelihood estimator is

$$\widehat{\theta}_{\mathrm{ML}} = X.$$

For $d \ge 3$, the James–Stein estimator is

$$\widehat{\theta}_{\mathrm{JS}}(X) = \left(1 - \frac{(d-2)\sigma^2}{\|X\|_2^2}\right) X.$$

It has strictly smaller quadratic risk than X at every finite $\theta \in \mathbb{R}^d$. Show this.

8.3 Bayes risk and Bayes estimators

Risk profiles can cross: one estimator may be better for some parameter values and worse for others. The Bayes criterion averages the profile under a prior, whereas the minimax criterion takes its largest value.

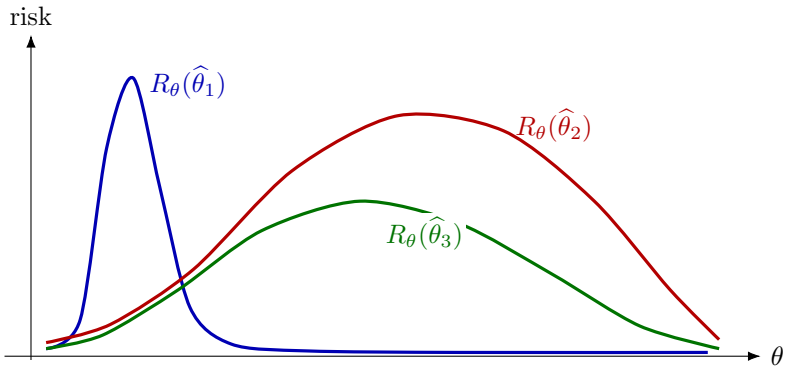


Figure 14: Risk profiles can cross. Bayes risk compares a prior-weighted average, while minimax risk compares the peak heights.

Let π be a probability distribution on Θ.

Definition 8.3 (Average risk and Bayes risk)**.** The average risk of a fixed estimator under the prior π is

$$R_\pi(\widehat{\theta}) := \mathbb{E}_{\theta\sim\pi}\Big[R_\theta(\widehat{\theta})\Big] = \mathbb{E}_{\theta,X}\Big[\ell(\theta, \widehat{\theta}(X))\Big]. \tag{8.3}$$

The Bayes risk is

$$R_\pi^\star := \inf_{\widehat{\theta}} R_\pi(\widehat{\theta}). \tag{8.4}$$

An estimator attaining this infimum is called a Bayes estimator.

Proposition 8.4 (A Bayes estimator may be taken deterministic)**.** *Randomization cannot improve the Bayes risk.*

Proof. Write a randomized estimator as $\widehat{\theta} = \varphi(X, U)$, where U is an independent random seed. Then

$$R_\pi(\widehat{\theta}) = \mathbb{E}_U\big[R_\pi\big(\varphi(\cdot, U)\big)\big] \ge \inf_u R_\pi\big(\varphi(\cdot, u)\big).$$

For every fixed u, the map $x \mapsto \varphi(x, u)$ is deterministic. Hence some deterministic estimator has average risk no larger than that of the randomized estimator. □

8.4 Quadratic loss: the conditional mean and MMSE

Proposition 8.5 (Conditional mean as the Bayes estimator)**.** *Assume* $\mathbb{E}\big[\|\theta\|_2^2\big] < \infty$*. Under quadratic loss*

$$\ell(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_2^2,$$

the Bayes estimator is

$$\widehat{\theta}_\pi(X) = \mathbb{E}[\theta \mid X]\,.$$

Moreover,

$$R_\pi^\star = \mathbb{E}\big[\|\theta - \mathbb{E}[\theta \mid X]\|_2^2\big] = \mathbb{E}\big[\mathrm{Tr}\big(\mathrm{Cov}(\theta \mid X)\big)\big]\,. \tag{8.5}$$

Proof. Let $m(X) := \mathbb{E}[\theta \mid X]$. For any estimator $\widehat{\theta} = \widehat{\theta}(X)$,

$$\begin{aligned}\mathbb{E}\Big[\|\theta - \widehat{\theta}\|_2^2\Big] &= \mathbb{E}\big[\|\theta - m(X)\|_2^2\big] + \mathbb{E}\Big[\|m(X) - \widehat{\theta}\|_2^2\Big] \\ &\quad + 2\mathbb{E}\Big[(\theta - m(X))^\top (m(X) - \widehat{\theta})\Big]\,.\end{aligned}$$

The cross term is zero because

$$\begin{aligned}&\mathbb{E}\Big[(\theta - m(X))^\top (m(X) - \widehat{\theta})\Big] \\ &\quad = \mathbb{E}\Big[(m(X) - \widehat{\theta})^\top \mathbb{E}[\theta - m(X) \mid X]\Big] = 0.\end{aligned}$$

Therefore

$$\mathbb{E}\Big[\|\theta - \widehat{\theta}\|_2^2\Big] = \mathbb{E}\big[\|\theta - m(X)\|_2^2\big] + \mathbb{E}\Big[\|m(X) - \widehat{\theta}\|_2^2\Big]\,, \tag{8.6}$$

which is minimized by $\widehat{\theta} = m(X)$. Finally,

$$\begin{aligned}\mathbb{E}\big[\|\theta - m(X)\|_2^2 \mid X\big] &= \mathbb{E}\Big[\mathrm{Tr}\big((\theta - m(X))(\theta - m(X))^\top\big) \mid X\Big] \\ &= \mathrm{Tr}\big(\mathrm{Cov}(\theta \mid X)\big).\end{aligned}$$

Taking expectations proves (8.5). □

Equation (8.6) is a conditional Pythagorean identity: conditional expectation is the best mean-square predictor.

8.5 Gaussian prior: posterior by completing the square

Consider the scalar GLM

$$X = \theta + Z, \qquad Z \sim \mathcal{N}(0, \sigma^2), \qquad \theta \sim \mathcal{N}(0, s),$$

where Z and θ are independent and $s > 0$. Bayes' rule gives

$$\begin{aligned}p_{\theta \mid X = x}(\theta) &\propto p_{X \mid \theta}(x \mid \theta) p_\theta(\theta) \\ &\propto \exp\!\left(-\frac{(x - \theta)^2}{2\sigma^2} - \frac{\theta^2}{2s}\right).\end{aligned}$$

Expanding the exponent,

$$-\frac{(x - \theta)^2}{2\sigma^2} - \frac{\theta^2}{2s} = -\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{s}\right)\theta^2 + \frac{x}{\sigma^2}\theta - \frac{x^2}{2\sigma^2}.$$

Define

$$v := \left(\frac{1}{\sigma^2} + \frac{1}{s}\right)^{-1} = \frac{s\sigma^2}{s + \sigma^2}, \qquad m := v\frac{x}{\sigma^2} = \frac{s}{s + \sigma^2}x.$$

Since $1/v = 1/\sigma^2 + 1/s$ and $m/v = x/\sigma^2$,

$$-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{s}\right)\theta^2 + \frac{x}{\sigma^2}\theta = -\frac{1}{2v}(\theta^2 - 2m\theta)$$
$$= -\frac{(\theta-m)^2}{2v} + \frac{m^2}{2v}.$$

The last term, together with $-x^2/(2\sigma^2)$, does not depend on θ and is absorbed into the normalizing constant. Hence

$$\boxed{\theta \mid X = x \sim \mathcal{N}\left(\frac{s}{s+\sigma^2}x, \frac{s\sigma^2}{s+\sigma^2}\right)}. \tag{8.7}$$

Consequently,

$$\widehat{\theta}_\pi(X) = \frac{s}{s+\sigma^2}X, \qquad R^\star_\pi = \frac{s\sigma^2}{s+\sigma^2}.$$

In d dimensions, with prior $\mathcal{N}(0, sI_d)$,

$$R^\star_\pi = d\frac{s\sigma^2}{s+\sigma^2}. \tag{8.8}$$

The limits are intuitive: as $s \to \infty$, the prior becomes diffuse and $\mathbb{E}[\theta \mid X] \to X$; as $s \downarrow 0$, the prior concentrates at zero and $\mathbb{E}[\theta \mid X] \to 0$.

8.6 Minimax risk and Bayes lower bounds

Definition 8.6 (Minimax risk)**.** The minimax risk is

$$R^\star := \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R_\theta(\widehat{\theta}). \tag{8.9}$$

An estimator attaining the infimum is minimax optimal.

To prove an upper bound on $R^\star$, exhibit one estimator whose risk is uniformly small. To prove a lower bound, show that every estimator has large risk at at least one parameter value. Often one seeks only a constant-factor characterization

$$c\Psi \le R^\star \le C\Psi, \qquad 0 < c \le C < \infty,$$

written $R^\star \asymp \Psi$.

Theorem 8.7 (Bayes lower bound)**.** *Let $\mathcal{P}(\Theta)$ be the set of priors on Θ. Then*

$$R^\star \ge R^\star_{\text{Bayes}} := \sup_{\pi\in\mathcal{P}(\Theta)} R^\star_\pi. \tag{8.10}$$

A prior attaining the supremum is called a least favourable prior.

Proof. For every estimator and every prior,

$$R_\pi(\widehat{\theta}) = \mathbb{E}_{\theta\sim\pi}\Big[R_\theta(\widehat{\theta})\Big] \le \sup_{\theta\in\Theta} R_\theta(\widehat{\theta}).$$

Taking the infimum over $\widehat{\theta}$ and then the supremum over π gives (8.10). □

Proposition 8.8 (Exact quadratic minimax risk in the GLM)**.** *For one observation from*

$$X \sim \mathcal{N}(\theta, \sigma^2 I_d), \qquad \theta \in \mathbb{R}^d,$$

under squared Euclidean loss,

$$R^\star = d\sigma^2. \tag{8.11}$$

Proof. For the upper bound, use $\widehat{\theta} = X$:

$$R_\theta(X) = \mathbb{E}\big[\|X-\theta\|_2^2\big] = \mathbb{E}\big[\|Z\|_2^2\big] = d\sigma^2 \qquad \text{for every } \theta.$$

For the lower bound, take the prior $\theta \sim \mathcal{N}(0, sI_d)$. By (8.8),

$$R^\star \geq R^\star_\pi = d\frac{s\sigma^2}{s+\sigma^2}.$$

Letting $s \to \infty$ yields $R^\star \geq d\sigma^2$, matching the upper bound. □

8.7 A summary of the risk notation

Notation	Definition	Meaning
$\ell(\theta, a)$	loss for action a when the truth is θ	one realized cost
$R_\theta(\widehat{\theta})$	$\mathbb{E}_\theta\Big[\ell(\theta,\widehat{\theta})\Big]$	risk of a fixed estimator at fixed θ
$R_\pi(\widehat{\theta})$	$\mathbb{E}_{\theta\sim\pi}\Big[R_\theta(\widehat{\theta})\Big]$	average risk under prior π
$R^\star_\pi$	$\inf_{\widehat{\theta}} R_\pi(\widehat{\theta})$	Bayes risk for prior π
$\sup_{\theta\in\Theta} R_\theta(\widehat{\theta})$	worst coordinate of the risk profile	worst-case risk of a fixed estimator
$R^\star$	$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta} R_\theta(\widehat{\theta})$	minimax risk
$R^\star_{\text{Bayes}}$	$\sup_\pi R^\star_\pi$	worst-case Bayes risk over priors
$R^\star_n(\Theta)$	n-sample version of $R^\star$	minimax risk with n observations
$n^\star(\varepsilon)$	$\min\{n : R^\star_n(\Theta) \leq \varepsilon\}$	sample complexity

A useful reading rule is

θ : fixed truth, π : prior average, $\star$: optimized over estimators, n : sample size.

8.8 Bare convex-optimization essentials and minimax–Bayes duality

For finite Θ, write the minimax problem in epigraph form:

$$\begin{aligned} R^\star = \inf_{P_{\widehat{\theta}|X},\, t} \quad & t \\ \text{subject to} \quad & R_\theta(\widehat{\theta}) \leq t, \qquad \theta \in \Theta. \end{aligned} \tag{8.12}$$

Associate a multiplier $\pi_\theta \geq 0$ with the constraint indexed by θ. The Lagrangian is

$$\begin{aligned} L(P_{\widehat{\theta}|X}, t; \pi) &= t + \sum_{\theta\in\Theta} \pi_\theta\big(R_\theta(\widehat{\theta}) - t\big) \\ &= \left(1 - \sum_{\theta\in\Theta}\pi_\theta\right) t + \sum_{\theta\in\Theta}\pi_\theta R_\theta(\widehat{\theta}). \end{aligned} \tag{8.13}$$

For a general constrained problem, the Lagrange dual function is the infimum of the Lagrangian over the primal variables. It always gives a lower bound on the primal optimum; this is *weak duality*. Equality of the primal and dual optimal values is called *strong duality*. Finite-dimensional linear programs satisfy strong duality whenever the optimal values are finite.

In (8.13), the infimum over $t \in \mathbb{R}$ is $-\infty$ unless

$$\sum_{\theta \in \Theta} \pi_\theta = 1.$$

Together with $\pi_\theta \geq 0$, this says that a dual-feasible multiplier is exactly a prior on Θ. For such a prior,

$$\inf_{P_{\widehat{\theta}|X}, t} L(P_{\widehat{\theta}|X}, t; \pi) = \inf_{\widehat{\theta}} \sum_{\theta \in \Theta} \pi_\theta R_\theta(\widehat{\theta}) = R^\star_\pi.$$

Thus the dual problem is

$$\sup_{\pi \in \mathcal{P}(\Theta)} R^\star_\pi = R^\star_{\text{Bayes}},$$

and weak duality is exactly

$$R^\star \geq R^\star_{\text{Bayes}}.$$

The primal variable is the randomized estimator; the dual variable is a prior.

If both Θ and $\mathcal{X}$ are finite, the conditional probabilities $P_{\widehat{\theta}|X}(\widehat{\theta} \mid x)$ are finitely many variables, every risk constraint is linear in these variables, and (8.12) is a finite-dimensional linear program. Linear programming strong duality then gives equality of the minimax and worst-case Bayes risks.

8.9 The minimax theorem via convex separation

Theorem 8.9 (Separating hyperplane theorem: form used here)**.** *Let $C, D \subseteq \mathbb{R}^m$ be nonempty, convex, and disjoint, with D open. Then there exist $a \neq 0$ and $c \in \mathbb{R}$ such that*

$$\inf_{x \in C} \langle a, x \rangle \geq c \geq \sup_{y \in D} \langle a, y \rangle.$$

Theorem 8.10 (Minimax theorem)**.** *The equality*

$$R^\star = R^\star_{\text{Bayes}}$$

holds in either of the following cases:

(i) Θ and $\mathcal{X}$ are finite;

(ii) Θ is finite and the loss is bounded below.

The first case follows from finite-dimensional linear-programming strong duality. We prove the second case.

Proof. Let $m = |\Theta|$. First suppose $R^\star = +\infty$. For every estimator, at least one coordinate risk must be $+\infty$; otherwise its maximum risk would be finite and the definition of $R^\star$ would give a finite upper bound. Since the loss is bounded below and Θ is finite, the uniform prior then gives infinite average risk for every estimator. Thus $R^\star_{\text{Bayes}} = R^\star = +\infty$.

Now assume $R^\star < \infty$. Estimators with a nonfinite risk coordinate cannot improve the finite minimax value, so define the set of finite achievable risk vectors

$$S := \left\{ R(\widehat{\theta}) = (R_\theta(\widehat{\theta}))_{\theta \in \Theta} \in \mathbb{R}^m : \widehat{\theta} \text{ is randomized} \right\}$$

and the open lower orthant

$$T := \{ t \in \mathbb{R}^m : t_\theta < R^\star \text{ for every } \theta \in \Theta \}.$$

The set S is convex. Indeed, if $\widehat{\theta}_0$ and $\widehat{\theta}_1$ are randomized estimators and $B \sim \mathrm{Ber}(\lambda)$ is independent of everything else, use $\widehat{\theta}_1$ when $B = 1$ and $\widehat{\theta}_0$ when $B = 0$. The resulting estimator $\widehat{\theta}_\lambda$ satisfies

$$R(\widehat{\theta}_\lambda) = (1 - \lambda)R(\widehat{\theta}_0) + \lambda R(\widehat{\theta}_1).$$

The set T is convex because it is an open lower orthant.

We now prove carefully that S and T are disjoint. Suppose, for a contradiction, that some estimator has $R(\widehat{\theta}) \in S \cap T$. Then

$$R_\theta(\widehat{\theta}) < R^\star \qquad \text{for every } \theta \in \Theta.$$

Because Θ is finite,

$$M := \max_{\theta \in \Theta} R_\theta(\widehat{\theta}) < R^\star.$$

On the other hand, by the definition

$$R^\star = \inf_{\widetilde{\theta}} \max_{\theta \in \Theta} R_\theta(\widetilde{\theta}),$$

the infimum over all estimators is no larger than the value at this particular estimator. Hence

$$R^\star \le \max_{\theta \in \Theta} R_\theta(\widehat{\theta}) = M < R^\star,$$

a contradiction. Therefore

$$S \cap T = \varnothing.$$

By the separating hyperplane theorem, there exist a nonzero $\pi \in \mathbb{R}^m$ and $c \in \mathbb{R}$ such that

$$\inf_{s \in S} \langle \pi, s \rangle \ge c \ge \sup_{t \in T} \langle \pi, t \rangle. \tag{8.14}$$

Every coordinate of π must be nonnegative. If $\pi_{\theta_0} < 0$, we could send $t_{\theta_0} \to -\infty$ while remaining in T, making $\langle \pi, t \rangle \to +\infty$, contrary to (8.14). Since $\pi \neq 0$, normalize it so that $\sum_\theta \pi_\theta = 1$; it is now a prior.

The left side of (8.14) is

$$\inf_{s \in S} \langle \pi, s \rangle = \inf_{\widehat{\theta}} \sum_{\theta \in \Theta} \pi_\theta R_\theta(\widehat{\theta}) = R^\star_\pi.$$

The right side is $R^\star$: every $t \in T$ satisfies $\langle \pi, t \rangle < R^\star$, while the constant vector $t_\theta = R^\star - \varepsilon$ belongs to T and has inner product $R^\star - \varepsilon$. Thus

$$\sup_{t \in T} \langle \pi, t \rangle = R^\star.$$

Consequently,

$$R^\star_{\mathrm{Bayes}} \ge R^\star_\pi \ge R^\star.$$

Together with the Bayes lower bound $R^\star_{\mathrm{Bayes}} \le R^\star$, this proves equality. □

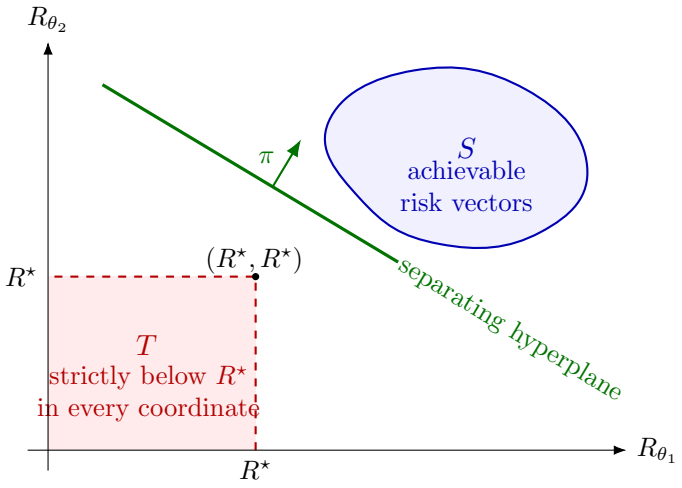


Figure 15: For two parameter values, T consists of risk vectors that would beat $R^\star$ in both coordinates. The achievable set S cannot intersect T. The normal of a separating hyperplane becomes a prior.

The finiteness of Θ is used when pointwise strict inequalities $R_\theta(\widehat{\theta}) < R^\star$ are converted into $\max_\theta R_\theta(\widehat{\theta}) < R^\star$. For an infinite parameter space, all values may be strictly below $R^\star$ while their supremum is still $R^\star$.

8.10 Multiple observations and sample complexity

For $n \geq 1$, the independent-sampling experiment is

$$\mathcal{P}_n = \{P_\theta^{\otimes n} : \theta \in \Theta\},$$

and $X^n = (X_1, \ldots, X_n)$ consists of independent observations from P_θ.

Definition 8.11 (n-sample minimax risk and sample complexity)**.** The minimax risk with n observations is

$$R_n^\star(\Theta) := \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\ell\big(\theta, \widehat{\theta}_n(X^n)\big) \right]. \tag{8.15}$$

The sample complexity at target risk ε is

$$n^\star(\varepsilon) := \min\{n \in \mathbb{N} : R_n^\star(\Theta) \leq \varepsilon\}. \tag{8.16}$$

The sequence $n \mapsto R_n^\star(\Theta)$ is nonincreasing because an estimator can discard extra observations. In a regular parametric model, one typically sees an order-$1/n$ risk.

Example 8.12 (n-sample quadratic GLM)**.** Suppose

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2 I_d), \qquad \theta \in \mathbb{R}^d.$$

The sample mean is sufficient and

$$\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n} I_d\right).$$

Applying (8.11) with noise variance σ^2/n gives

$$R_n^\star(\mathbb{R}^d) = \frac{d\sigma^2}{n}, \qquad n^\star(\varepsilon) = \left\lceil \frac{d\sigma^2}{\varepsilon} \right\rceil. \tag{8.17}$$

Lecture 8 summary

1. Pointwise risk fixes the truth; Bayes risk averages over a prior; minimax risk takes the worst parameter value.

2. Under quadratic loss, the posterior mean is Bayes and the Bayes risk is the MMSE.

3. For a Gaussian prior and Gaussian likelihood, completing the square gives an explicit Gaussian posterior.

4. Every Bayes risk is a minimax lower bound. In the unconstrained Gaussian location model, diffuse Gaussian priors prove the exact minimax risk $d\sigma^2$.

5. In minimax–Bayes duality, randomized estimators are primal variables and priors are dual variables.

6. For finite parameter spaces, convex separation identifies a prior whose Bayes risk equals the minimax risk.

7. With n Gaussian observations, the quadratic minimax risk is $d\sigma^2/n$ and the sample complexity is $\lceil d\sigma^2/\varepsilon \rceil$.

9 Lecture 9: Fisher-Information and Mutual Information Lower Bounds

9.1 A brief recap of the risk criteria

For a fixed estimator $\widehat{\theta}$ and fixed truth θ, the pointwise risk is

$$R_\theta(\widehat{\theta}) := \mathbb{E}_\theta\Big[\ell\big(\theta, \widehat{\theta}(X)\big)\Big].$$

For a prior π on the parameter space, the average and Bayes risks are

$$R_\pi(\widehat{\theta}) := \mathbb{E}_{\Theta\sim\pi}\Big[R_\Theta(\widehat{\theta})\Big], \qquad R_\pi^* := \inf_{\widehat{\theta}} R_\pi(\widehat{\theta}).$$

The minimax risk is

$$R^*(\Theta) := \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R_\theta(\widehat{\theta}).$$

Every prior gives a minimax lower bound:

$$R^*(\Theta) \geq R_\pi^*. \tag{9.1}$$

Thus both the Fisher-information and mutual-information methods below first lower-bound a Bayes risk and then invoke (9.1).

9.2 Local and global information

For observations $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$ and quadratic loss, a central fixed-dimensional benchmark is

$$R_n^*(\Theta) := \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\|\widehat{\theta} - \theta\|_2^2\Big] \gtrsim \frac{1}{n} \sup_{\theta\in\Theta} \operatorname{Tr}\big(J_F(\theta)^{-1}\big). \tag{9.2}$$

The Fisher-information method studies infinitesimal perturbations of a smooth parametric model. The mutual-information method is global: it compares the information required to reconstruct a random parameter at a prescribed distortion with the information supplied by the experiment.

9.3 Frequentist HCR and Cramér–Rao bounds

9.3.1 Score and Fisher information

For a scalar parameter θ, let p_θ be a density or pmf of P_θ. Its score is

$$S_\theta(x) := \frac{\partial}{\partial\theta} \log p_\theta(x), \tag{9.3}$$

and the Fisher information is

$$J_F(\theta) := \mathbb{E}_\theta\big[S_\theta(X)^2\big]. \tag{9.4}$$

For a vector parameter $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, write

$$\nabla_\theta := \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_d}\right)^\top.$$

Then

$$J_F(\theta) := \mathbb{E}_\theta\Big[\nabla_\theta \log p_\theta(X) \nabla_\theta \log p_\theta(X)^\top\Big]. \tag{9.5}$$

Under the usual differentiation-under-the-integral condition, the score has mean zero. In the discrete case,

$$\mathbb{E}_\theta[S_\theta(X)] = \sum_x p_\theta(x) \frac{\partial}{\partial\theta} \log p_\theta(x) = \frac{\partial}{\partial\theta} \sum_x p_\theta(x) = 0.$$

We use the convention

$$\chi^2(P\|Q) := \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2\right].$$

Its variational representation is

$$\chi^2(P\|Q) = \sup_{f:\, 0<\mathrm{Var}_Q(f)<\infty} \frac{(\mathbb{E}_P[f] - \mathbb{E}_Q[f])^2}{\mathrm{Var}_Q(f)}. \tag{9.6}$$

9.3.2 The Hammersley–Chapman–Robbins bound

Theorem 9.1 (HCR lower bound)**.** *Let $\Theta \subseteq \mathbb{R}$. For every possibly randomized estimator $\widehat{\theta}$ and every $\theta \in \Theta$,*

$$R_\theta(\widehat{\theta}) := \mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] \geq \mathrm{Var}_\theta(\widehat{\theta}) \geq \sup_{\substack{\theta' \in \Theta \\ \theta' \neq \theta}} \frac{\left(\mathbb{E}_{\theta'}\left[\widehat{\theta}\right] - \mathbb{E}_\theta\left[\widehat{\theta}\right]\right)^2}{\chi^2(P_{\theta'}\|P_\theta)}. \tag{9.7}$$

Proof. The bias–variance decomposition gives

$$\mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] = \mathrm{Var}_\theta(\widehat{\theta}) + \left(\mathbb{E}_\theta\left[\widehat{\theta}\right] - \theta\right)^2.$$

Fix $\theta' \neq \theta$. Let $P_{\theta'}^{\widehat{\theta}}$ and $P_\theta^{\widehat{\theta}}$ denote the laws of the estimator under $P_{\theta'}$ and P_θ. The estimator is obtained from X through the same Markov kernel under both models, so data processing gives

$$\chi^2(P_{\theta'}\|P_\theta) \geq \chi^2\left(P_{\theta'}^{\widehat{\theta}} \middle\| P_\theta^{\widehat{\theta}}\right).$$

Applying (9.6) to the identity function gives

$$\chi^2\left(P_{\theta'}^{\widehat{\theta}} \middle\| P_\theta^{\widehat{\theta}}\right) \geq \frac{\left(\mathbb{E}_{\theta'}\left[\widehat{\theta}\right] - \mathbb{E}_\theta\left[\widehat{\theta}\right]\right)^2}{\mathrm{Var}_\theta(\widehat{\theta})}.$$

Rearrange and take the supremum over θ'. □

9.3.3 The frequentist Cramér–Rao lower bound

Suppose now that $\widehat{\theta}$ is unbiased:

$$\mathbb{E}_\theta\left[\widehat{\theta}\right] = \theta \qquad \text{for every } \theta \in \Theta.$$

Then the HCR bound becomes

$$\mathrm{Var}_\theta(\widehat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\theta' - \theta)^2}{\chi^2(P_{\theta'}\|P_\theta)}.$$

Under regularity conditions,

$$\chi^2(P_{\theta+h}\|P_\theta) = h^2 J_F(\theta) + o(h^2), \qquad h \to 0. \tag{9.8}$$

Taking $h \to 0$ yields the classical bound

$$\boxed{\mathrm{Var}_\theta(\widehat{\theta}) \geq \frac{1}{J_F(\theta)}.} \tag{9.9}$$

For n i.i.d. observations, Fisher information adds: $J_{F,n}(\theta) = nJ_F(\theta)$. In the multivariate case, an unbiased estimator satisfies

$$\mathrm{Cov}_\theta(\widehat{\theta}) \succeq \frac{1}{n} J_F(\theta)^{-1}, \qquad \mathbb{E}_\theta\Big[\|\widehat{\theta} - \theta\|_2^2\Big] \geq \frac{1}{n} \operatorname{Tr}\big(J_F(\theta)^{-1}\big)\,. \tag{9.10}$$

The frequentist Cramér–Rao inequality is a pointwise lower bound for unbiased estimators. Minimax estimation permits biased procedures, and bias can be beneficial; the James–Stein phenomenon is a particularly striking example. The Bayesian bounds below apply to all estimators.

9.4 Bayesian HCR and its local Cramér–Rao consequence

Let $\Theta \sim \pi$, and conditionally on $\Theta = \theta$, let $X \sim P_\theta$. Under quadratic loss, the Bayes risk is

$$R_\pi^* := \inf_{\widehat{\theta}} \mathbb{E}\Big[(\widehat{\theta} - \Theta)^2\Big]\,. \tag{9.11}$$

9.4.1 A finite-shift Bayesian HCR bound

For a shift δ, construct two joint laws of (Θ, X). It is clearest to start from a common $U \sim \pi$:

	Θ	X
$\mathbb{Q}$	U	$X \sim P_U$
$\mathbb{P}_\delta$	$U + \delta$	$X \sim P_U$.

Equivalently,

$$\mathbb{Q}(\mathrm{d}\theta, \mathrm{d}x) = \pi(\mathrm{d}\theta) P_\theta(\mathrm{d}x), \qquad \mathbb{P}_\delta(\mathrm{d}\theta, \mathrm{d}x) = (T_\delta \pi)(\mathrm{d}\theta) P_{\theta-\delta}(\mathrm{d}x).$$

The marginal law of X is identical under the two constructions, while

$$\mathbb{E}_{\mathbb{P}_\delta}[\Theta] = \mathbb{E}_{\mathbb{Q}}[\Theta] + \delta.$$

Theorem 9.2 (Bayesian HCR lower bound)**.** *For every admissible nonzero shift δ for which $\mathbb{P}_\delta \ll \mathbb{Q}$,*

$$R_\pi^* \geq \frac{\delta^2}{\chi^2(\mathbb{P}_\delta \| \mathbb{Q})}. \tag{9.12}$$

Consequently,

$$\boxed{R_\pi^* \geq B_{\mathrm{HCR}}(\pi) := \sup_{\delta \neq 0} \frac{\delta^2}{\chi^2(\mathbb{P}_\delta \| \mathbb{Q})}.} \tag{9.13}$$

Proof. The Markov chain

$$(\Theta, X) \longrightarrow (\Theta, \widehat{\theta}) \longrightarrow \Theta - \widehat{\theta}$$

and (9.6) imply

$$\chi^2(\mathbb{P}_\delta \| \mathbb{Q}) \geq \frac{\Big(\mathbb{E}_{\mathbb{P}_\delta}\Big[\Theta - \widehat{\theta}\Big] - \mathbb{E}_{\mathbb{Q}}\Big[\Theta - \widehat{\theta}\Big]\Big)^2}{\mathrm{Var}_{\mathbb{Q}}(\Theta - \widehat{\theta})}.$$

Because X, and hence $\widehat{\theta}$, has the same marginal law under $\mathbb{P}_\delta$ and $\mathbb{Q}$, the numerator equals δ^2. Moreover,

$$\mathbb{E}_{\mathbb{Q}}\Big[(\widehat{\theta} - \Theta)^2\Big] \geq \mathrm{Var}_{\mathbb{Q}}(\widehat{\theta} - \Theta).$$

The conclusion holds for every estimator, so take the infimum over $\widehat{\theta}$ and then the supremum over δ. □

Under suitable smoothness conditions, the joint divergence has the local expansion

$$\chi^2(\mathbb{P}_\delta\|\mathbb{Q}) = \delta^2\left(J(\pi) + \mathbb{E}_{\Theta\sim\pi}[J_F(\Theta)] + o(1)\right), \qquad \delta \to 0. \tag{9.14}$$

Here $J(\pi)$ is the Fisher information of the prior. Hence

$$\boxed{R_\pi^* \geq B_{\mathrm{HCR}}(\pi) \geq B_{\mathrm{BCR}}(\pi) := \frac{1}{J(\pi) + \mathbb{E}_{\Theta\sim\pi}[J_F(\Theta)]}.} \tag{9.15}$$

The Bayesian HCR bound optimizes over all finite shifts; the Bayesian Cramér–Rao bound keeps only the infinitesimal shift.

9.4.2 Gaussian example: both bounds are exact

We first record the elementary Gaussian calculation that will be used repeatedly. If

$$P = \mathcal{N}(m+\Delta, v), \qquad Q = \mathcal{N}(m, v),$$

then, for $X \sim Q$,

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(X) = \exp\left(\frac{\Delta(X-m)}{v} - \frac{\Delta^2}{2v}\right).$$

Consequently,

$$1 + \chi^2(P\|Q) = \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(X)\right)^2\right] \tag{9.16}$$

$$= e^{-\Delta^2/v}\mathbb{E}\left[\exp\left(\frac{2\Delta Z}{v}\right)\right] = e^{\Delta^2/v}, \qquad Z \sim \mathcal{N}(0, v). \tag{9.17}$$

Thus

$$\chi^2(\mathcal{N}(m+\Delta, v)\|\mathcal{N}(m, v)) = e^{\Delta^2/v} - 1.$$

Example 9.3 (Gaussian location with a Gaussian prior)**.** Let

$$X_1, \ldots, X_n \mid \Theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\Theta, 1), \qquad \Theta \sim \mathcal{N}(0, s).$$

Under the original joint law $\mathbb{Q}$,

$$Q_\Theta = \mathcal{N}(0, s), \qquad Q_{X^n|\Theta=\theta} = \mathcal{N}(\theta\mathbf{1}_n, I_n).$$

Under the shifted law $\mathbb{P}_\delta$,

$$P_{\delta,\Theta} = \mathcal{N}(\delta, s), \qquad P_{\delta,X^n|\Theta=\theta} = \mathcal{N}((\theta-\delta)\mathbf{1}_n, I_n).$$

By (9.17),

$$\chi^2(P_{\delta,\Theta}\|Q_\Theta) = e^{\delta^2/s} - 1, \qquad \chi^2(P_{\delta,X^n|\theta}\|Q_{X^n|\theta}) = e^{n\delta^2} - 1. \tag{9.18}$$

The chain rule for χ^2 divergence can be written as

$$1 + \chi^2(\mathbb{P}_\delta\|\mathbb{Q}) = \mathbb{E}_{Q_\Theta}\left[\left(\frac{\mathrm{d}P_{\delta,\Theta}}{\mathrm{d}Q_\Theta}(\Theta)\right)^2\left(1 + \chi^2(P_{\delta,X^n|\Theta}\|Q_{X^n|\Theta})\right)\right].$$

The conditional divergence in (9.18) does not depend on θ, and

$$\mathbb{E}_{Q_\Theta}\left[\left(\frac{\mathrm{d}P_{\delta,\Theta}}{\mathrm{d}Q_\Theta}\right)^2\right] = 1 + \chi^2(P_{\delta,\Theta}\|Q_\Theta) = e^{\delta^2/s}.$$

Therefore

$$1 + \chi^2(\mathbb{P}_\delta \| \mathbb{Q}) = e^{\delta^2/s} e^{n\delta^2}, \tag{9.19}$$

$$\chi^2(\mathbb{P}_\delta \| \mathbb{Q}) = e^{(n+1/s)\delta^2} - 1. \tag{9.20}$$

Equivalently, one may first reduce X^n to the sufficient statistic $\overline{X}$, whose conditional variance is $1/n$.

The Bayesian HCR bound now gives

$$\begin{aligned} R_\pi^* &\geq \sup_{\delta \neq 0} \frac{\delta^2}{e^{(n+1/s)\delta^2} - 1} \\ &= \frac{1}{n + 1/s} \sup_{u>0} \frac{u}{e^u - 1} = \frac{s}{sn+1}. \end{aligned}$$

The last supremum equals one because $e^u \geq 1 + u$ and equality is approached as $u \downarrow 0$. Since the prior information is $J(\pi) = 1/s$ and the n observations supply Fisher information n,

$$B_{\mathrm{BCR}}(\pi) = \frac{1}{n + 1/s}.$$

The posterior variance from Lecture 8 is also $s/(sn+1)$, so

$$R_\pi^* = B_{\mathrm{HCR}}(\pi) = B_{\mathrm{BCR}}(\pi) = \frac{s}{sn+1}.$$

9.5 A lucid discrete proof of the Bayesian CR bound

We now give the rigorous proof needed in this lecture, assuming that the observation alphabet is finite or countable. This avoids measure notation and exposes the central identity directly.

9.5.1 Fisher information of the prior

Let the random parameter Θ have a differentiable density π on $[a, b]$. Since π is a function of the single variable t, we use an ordinary derivative for its score:

$$S_\pi(t) := \frac{\mathrm{d}}{\mathrm{d}t} \log \pi(t) = \frac{1}{\pi(t)} \frac{\mathrm{d}\pi(t)}{\mathrm{d}t}. \tag{9.21}$$

By definition, Fisher information is the expectation of the squared score:

$$J(\pi) := \mathbb{E}\big[S_\pi(\Theta)^2\big] \tag{9.22}$$

$$= \int_a^b \pi(t) \left(\frac{1}{\pi(t)} \frac{\mathrm{d}\pi(t)}{\mathrm{d}t} \right)^2 \mathrm{d}t = \boxed{\int_a^b \frac{\left(\frac{\mathrm{d}\pi(t)}{\mathrm{d}t} \right)^2}{\pi(t)} \, \mathrm{d}t}. \tag{9.23}$$

Thus the formula

$$\int_a^b \frac{\left(\frac{\mathrm{d}\pi(t)}{\mathrm{d}t} \right)^2}{\pi(t)} \, \mathrm{d}t$$

is not an additional definition; it is simply the expected square of the prior score.

9.5.2 The theorem and proof

Theorem 9.4 (Bayesian Cramér–Rao bound, discrete observations)**.** *Let Θ have a differentiable density π on $[a,b]$, with $\pi(a)=\pi(b)=0$ and $J(\pi)<\infty$. Given $\Theta=t$, suppose that X has pmf $p_t(x)$ on a finite or countable alphabet $\mathcal{X}$. Assume $p_t(x)$ is differentiable in t and that differentiation may be interchanged with summation, so*

$$\sum_{x\in\mathcal{X}} \frac{\partial}{\partial t} p_t(x) = 0. \tag{9.24}$$

Then every estimator satisfies

$$\mathbb{E}\Big[(\widehat{\theta}-\Theta)^2\Big] \geq \frac{1}{\mathbb{E}_{\Theta\sim\pi}[J_F(\Theta)] + J(\pi)}. \tag{9.25}$$

Consequently, the same inequality holds for the Bayes risk R^*_π.

Proof. It is enough to consider a deterministic estimator $\widehat{\theta}=\widehat{\theta}(X)$. Define the joint density/pmf of (Θ, X) by

$$q(t,x) := \pi(t)p_t(x)$$

and its joint score by

$$S(t,x) := \frac{\partial}{\partial t}\log q(t,x) = \frac{\partial}{\partial t}\log p_t(x) + \frac{\mathrm{d}}{\mathrm{d}t}\log \pi(t). \tag{9.26}$$

Step 1: an integration-by-parts identity. Using $S(t,x)q(t,x) = \frac{\partial}{\partial t}q(t,x)$,

$$\begin{aligned}
&\mathbb{E}\Big[(\widehat{\theta}-\Theta)S(\Theta,X)\Big] \\
&\quad= \sum_{x\in\mathcal{X}} \int_a^b (\widehat{\theta}(x)-t)\frac{\partial}{\partial t}q(t,x)\,\mathrm{d}t \\
&\quad= \sum_{x\in\mathcal{X}} \Big[(\widehat{\theta}(x)-t)q(t,x)\Big]_a^b + \sum_{x\in\mathcal{X}}\int_a^b q(t,x)\,\mathrm{d}t \\
&\quad= 1.
\end{aligned}$$

The boundary term vanishes because $\pi(a)=\pi(b)=0$.

Step 2: Cauchy–Schwarz. Therefore,

$$1 \leq \mathbb{E}\Big[(\widehat{\theta}-\Theta)^2\Big]\,\mathbb{E}\big[S(\Theta,X)^2\big],$$

so

$$\mathbb{E}\Big[(\widehat{\theta}-\Theta)^2\Big] \geq \frac{1}{\mathbb{E}[S(\Theta,X)^2]}. \tag{9.27}$$

Step 3: likelihood information and prior information add. Write

$$S(\Theta,X) = S_X(\Theta,X) + S_\pi(\Theta), \qquad S_X(t,x) := \frac{\partial}{\partial t}\log p_t(x).$$

The cross term vanishes because

$$\mathbb{E}[S_X(\Theta,X)\mid \Theta=t] = \sum_x p_t(x)\frac{\partial}{\partial t}\log p_t(x) = \frac{\partial}{\partial t}\sum_x p_t(x) = 0.$$

Hence

$$\mathbb{E}\big[S(\Theta,X)^2\big] = \mathbb{E}_{\Theta\sim\pi}[J_F(\Theta)] + J(\pi),$$

which, together with (9.27), proves the claim. □

Takeaway

The proof has one line of intuition:

$$\underbrace{\text{Bayes MSE}}_{\mathbb{E}[(\widehat{\theta}-\Theta)^2]} \times \underbrace{(\text{data information} + \text{prior information})}_{\mathbb{E}_\pi[J_F(\Theta)]+J(\pi)} \geq 1.$$

The likelihood score and prior score are orthogonal, which is why their Fisher informations add.

Corollary 9.5 (Multivariate product-prior form)**.** *Let $\Theta \in \mathbb{R}^d$ have product density $\pi(\theta) = \prod_{i=1}^d \pi_i(\theta_i)$, where each factor satisfies the preceding boundary assumptions. Under the corresponding score regularity condition,*

$$R_\pi^* \geq \operatorname{Tr}(\mathbb{E}_{\Theta\sim\pi}[J_F(\Theta)] + J(\pi))^{-1}, \qquad J(\pi) := \operatorname{diag}(J(\pi_1), \ldots, J(\pi_d)). \tag{9.28}$$

9.6 Localized priors and the asymptotic minimax lower bound

The prior average in (9.28) becomes a pointwise quantity by concentrating a smooth prior near an arbitrary parameter θ_0.

9.6.1 The squared-cosine density

Define

$$g(u) := \cos^2\Big(\frac{\pi u}{2}\Big)\, \mathbb{1}\{|u| \leq 1\}. \tag{9.29}$$

It is a probability density because

$$\int_{-1}^{1} \cos^2\Big(\frac{\pi u}{2}\Big)\, \mathrm{d}u = 1,$$

and it vanishes at the boundary. Moreover,

$$J(g) = \int_{-1}^{1} \frac{\left(\frac{\mathrm{d}g(u)}{\mathrm{d}u}\right)^2}{g(u)}\, \mathrm{d}u \tag{9.30}$$

$$= \pi^2 \int_{-1}^{1} \sin^2\Big(\frac{\pi u}{2}\Big)\, \mathrm{d}u = \pi^2. \tag{9.31}$$

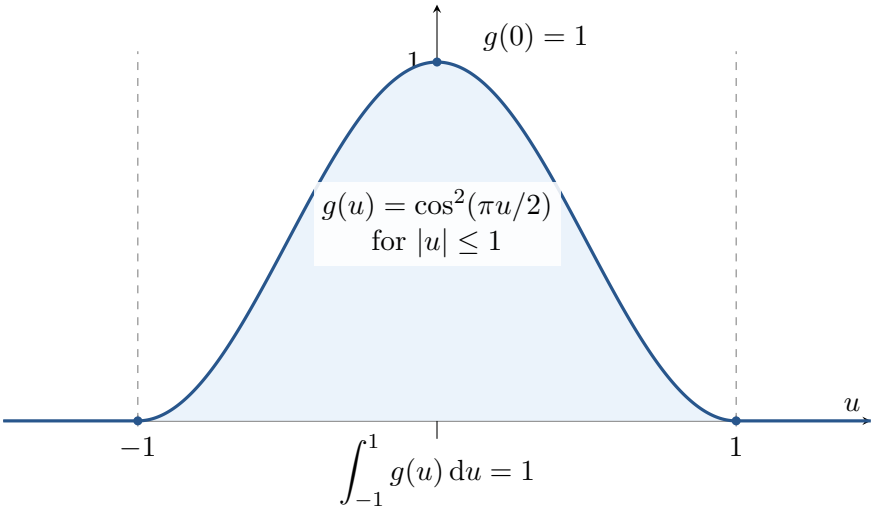


Figure 16: The squared-cosine density is smooth at the endpoints, supported on $[-1, 1]$, and has the minimum Fisher information among sufficiently regular densities on this interval.

For a center t_0 and width δ, define

$$g_{\delta,t_0}(t) := \frac{1}{\delta} g\left(\frac{t - t_0}{\delta}\right).$$

Fisher information obeys the scaling rule

$$J(g_{\delta,t_0}) = \frac{\pi^2}{\delta^2}. \tag{9.32}$$

9.6.2 From a local Bayes bound to a worst-case minimax bound

Theorem 9.6 (Asymptotic minimax lower bound)**.** *Let $\Theta \subseteq \mathbb{R}^d$ be open. Suppose that $J_F(\theta)$ is continuous and positive definite. For n i.i.d. observations,*

$$\liminf_{n\to\infty} n R_n^*(\Theta) \ge \sup_{\theta\in\Theta} \operatorname{Tr}\big(J_F(\theta)^{-1}\big). \tag{9.33}$$

Proof. Fix an arbitrary $\theta_0 \in \Theta$. For all sufficiently small δ, the cube $B_\infty(\theta_0, \delta) \subseteq \Theta$. Put an independent squared-cosine prior on each coordinate:

$$\varrho_{\delta,\theta_0}(\theta) = \prod_{i=1}^{d} \frac{1}{\delta} g\left(\frac{\theta_i - \theta_{0,i}}{\delta}\right).$$

Its prior Fisher information matrix is

$$J(\varrho_{\delta,\theta_0}) = \frac{\pi^2}{\delta^2} I_d.$$

The n-sample BCR bound and the inequality "minimax risk $\ge$ Bayes risk" give

$$R_n^*(\Theta) \ge \frac{1}{n} \operatorname{Tr}\left(\mathbb{E}_{\varrho_{\delta,\theta_0}}[J_F(\Theta)] + \frac{\pi^2}{n\delta^2} I_d\right)^{-1}. \tag{9.34}$$

Choose $\delta_n = n^{-1/4}$. Then $\delta_n \to 0$ while $n\delta_n^2 = \sqrt{n} \to \infty$. Continuity of J_F yields

$$\mathbb{E}_{\varrho_{\delta_n,\theta_0}}[J_F(\Theta)] \longrightarrow J_F(\theta_0), \qquad \frac{\pi^2}{n\delta_n^2} I_d \longrightarrow 0.$$

Hence

$$\liminf_{n\to\infty} n R_n^*(\Theta) \ge \operatorname{Tr}\big(J_F(\theta_0)^{-1}\big).$$

Since θ_0 was arbitrary, take the supremum over θ_0. □

9.7 Maximum likelihood and asymptotic efficiency

Let

$$L_n(\theta) := \sum_{i=1}^{n} \log p_\theta(X_i), \qquad \widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} L_n(\theta).$$

Write

$$V(\theta, x) := \nabla_\theta \log p_\theta(x), \qquad H(\theta, x) := \nabla_\theta^2 \log p_\theta(x).$$

Remark 9.7 (A standard sufficient set of regularity conditions). The following conditions are convenient and deliberately stronger than necessary.

(i) The true parameter θ_0 is an interior point of an open parameter set, and the model is identifiable.

(ii) All P_θ are dominated by a common measure and have common support in a neighborhood of θ_0.

(iii) For almost every x, $\theta \mapsto \log p_\theta(x)$ is three times continuously differentiable near θ_0.

(iv) Differentiation may be interchanged with integration, and the score, Hessian, and local third derivatives admit integrable envelopes. In particular,

$$\mathbb{E}_{\theta_0}[V(\theta_0, X)] = 0, \qquad -\mathbb{E}_{\theta_0}[H(\theta_0, X)] = J_F(\theta_0).$$

(v) $J_F(\theta_0)$ is finite and positive definite.

(vi) A consistent sequence of local or global MLEs exists: $\widehat{\theta}_{\mathrm{MLE}} \to \theta_0$ in probability. A local uniform law of large numbers holds for the Hessian.

For convergence of the quadratic risk, rather than merely convergence in distribution, one also assumes uniform integrability of $n\|\widehat{\theta}_{\mathrm{MLE}} - \theta_0\|_2^2$; a suitable $(2+\eta)$-moment bound is one sufficient condition. To convert the pointwise conclusion into a uniform minimax upper bound, uniform versions of these assumptions are needed over the relevant parameter set, with boundary effects excluded.

Under these conditions, a Taylor expansion of the score equation around θ_0 yields the local asymptotic normality expansion

$$L_n\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - L_n(\theta_0) = h^\top \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V(\theta_0, X_i) - \frac{1}{2} h^\top J_F(\theta_0) h + o_{\mathbb{P}_{\theta_0}}(1). \tag{9.35}$$

The central limit theorem and the local uniform law of large numbers give

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V(\theta_0, X_i) \Rightarrow \mathcal{N}(0, J_F(\theta_0)), \qquad -\frac{1}{n} \sum_{i=1}^{n} H(\theta_0, X_i) \to J_F(\theta_0).$$

Consequently,

$$\sqrt{n}\big(\widehat{\theta}_{\mathrm{MLE}} - \theta_0\big) \Rightarrow \mathcal{N}\big(0, J_F(\theta_0)^{-1}\big)\,. \tag{9.36}$$

If the squared errors are uniformly integrable, then

$$n\,\mathbb{E}_{\theta_0}\Big[\|\widehat{\theta}_{\mathrm{MLE}} - \theta_0\|_2^2\Big] \longrightarrow \mathrm{Tr}\big(J_F(\theta_0)^{-1}\big)\,. \tag{9.37}$$

Thus the MLE attains the pointwise Cramér–Rao constant. Under uniform regularity, this matches the asymptotic minimax lower bound derived from localized priors.

9.8 The mutual information method

The Fisher-information method is local. The mutual information method instead asks a global question: how much information must an estimator retain about a random parameter in order to attain a prescribed loss?

9.8.1 Rate–distortion as information required

Let $\Theta \sim \pi$, and let $\ell(\theta, \widehat{\theta})$ be a loss.

Definition 9.8 (Rate–distortion function)**.** The rate–distortion function of the prior π is

$$\mathcal{R}_\pi(D) := \inf_{P_{\widetilde{\Theta}|\Theta}:\, \mathbb{E}\left[\ell(\Theta,\widetilde{\Theta})\right] \le D} I(\Theta; \widetilde{\Theta}). \tag{9.38}$$

It is the minimum mutual information that any representation $\widetilde{\Theta}$ must retain about Θ in order to reproduce it with expected distortion at most D.

The operational coding theorem is not needed here; only the variational quantity in (9.38) is used.

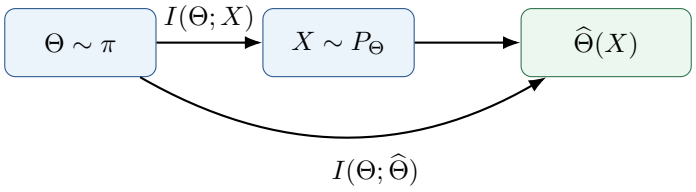


Figure 17: The mutual information method compares the information required for distortion D with the information supplied by the experiment.

Theorem 9.9 (Mutual information method)**.** *Suppose an estimator satisfies*

$$\mathbb{E}\Big[\ell(\Theta, \widehat{\Theta})\Big] \le D.$$

Then

$$\mathcal{R}_\pi(D) \le I(\Theta; \widehat{\Theta}) \le I(\Theta; X). \tag{9.39}$$

Consequently, with the generalized inverse

$$\mathcal{R}_\pi^{-1}(C) := \inf\{D : \mathcal{R}_\pi(D) \le C\},$$

we have

$$R_\pi^* \ge \mathcal{R}_\pi^{-1}(I(\Theta; X))\,. \tag{9.40}$$

If $I(\Theta; X) \le C$, then $R^* \ge R_\pi^* \ge \mathcal{R}_\pi^{-1}(C)$.

Proof. The first inequality follows from the definition of rate–distortion; the conditional law of $\widehat{\Theta}$ given Θ is an admissible reproduction channel. The second is data processing for $\Theta \to X \to \widehat{\Theta}$. □

9.8.2 Two Gaussian formulas

Proposition 9.10 (Gaussian rate–distortion)**.** *Let $\Theta \sim \mathcal{N}(0, sI_d)$ and use total squared-error distortion $\mathbb{E}\Big[\|\Theta - \widetilde{\Theta}\|_2^2\Big] \le D$. Then*

$$\boxed{\mathcal{R}_\pi(D) = \begin{cases} \dfrac{d}{2} \log \dfrac{ds}{D}, & 0 < D < ds, \\ 0, & D \ge ds. \end{cases}} \tag{9.41}$$

For a scalar Gaussian of variance s, the formula is $\frac{1}{2}\log(s/D)$. Independent coordinates and total distortion D allocate distortion D/d to each coordinate, yielding $d \cdot \frac{1}{2}\log(s/(D/d))$.

Proposition 9.11 (Gaussian channel capacity)**.** *For the additive Gaussian channel*

$$Y = U + Z, \qquad Z \sim \mathcal{N}(0, NI_d), \qquad \mathbb{E}\big[\|U\|_2^2\big] \le P,$$

the capacity in nats is

$$\boxed{C_{\mathrm{AWGN}}(P, N) = \frac{d}{2}\log\left(1 + \frac{P}{dN}\right).} \tag{9.42}$$

Indeed,

$$I(U; Y) = h(Y) - h(Z).$$

Among random vectors with a prescribed covariance trace, a Gaussian has maximum differential entropy. Hence

$$h(Y) \le \frac{d}{2}\log\left(2\pi e\left(N + \frac{P}{d}\right)\right), \qquad h(Z) = \frac{d}{2}\log(2\pi e N),$$

and equality is attained by $U \sim \mathcal{N}(0, (P/d)I_d)$.

9.8.3 Gaussian location revisited

Let

$$X_i = \Theta + Z_i, \qquad Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d), \qquad \Theta \sim \mathcal{N}(0, sI_d).$$

The sufficient statistic is

$$\overline{X} = \Theta + \overline{Z}, \qquad \overline{Z} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} I_d\right).$$

Using (9.42) with $P = ds$ and $N = \sigma^2/n$,

$$I(\Theta; X^n) = I(\Theta; \overline{X}) = \frac{d}{2}\log\left(1 + \frac{ns}{\sigma^2}\right). \tag{9.43}$$

Combining (9.41) and (9.43) gives

$$R^*_{\pi,n} \ge ds \exp\left(-\frac{2}{d} I(\Theta; X^n)\right) \tag{9.44}$$

$$= \frac{ds}{1 + ns/\sigma^2} = \frac{ds\sigma^2}{\sigma^2 + ns}. \tag{9.45}$$

This is exactly the posterior trace covariance, so MIM is sharp here. Letting $s \to \infty$ yields the minimax lower bound

$$R^*_n(\mathbb{R}^d) \ge \frac{d\sigma^2}{n},$$

which is attained by the sample mean.

Lecture 9 summary

The principal formulas from this lecture are

Frequentist CR:	$\mathrm{Var}_\theta(\widehat{\theta}) \ge \dfrac{1}{J_F(\theta)}$	for unbiased estimators,
Bayesian HCR:	$R^*_\pi \ge \sup\limits_{\delta \ne 0} \dfrac{\delta^2}{\chi^2(\mathbb{P}_\delta \Vert \mathbb{Q})}$,	
Bayesian CR:	$R^*_\pi \ge \dfrac{1}{J(\pi) + \mathbb{E}_\pi[J_F(\Theta)]}$,	
Mutual information method:	$\mathcal{R}_\pi(D) \le I(\Theta; \widehat{\Theta}) \le I(\Theta; X)$.	

10 Lecture 10: Hypothesis-Testing Lower Bounds and Entropic Upper Bounds

This lecture develops three generic minimax lower-bound methods by reducing estimation to hypothesis testing: Le Cam's two-point method, Assouad's hypercube method, and Fano's packing method. It then turns in the opposite direction and proves the Yang–Barron entropic upper bound, which constructs a predictive distribution using the capacity and KL covering number of the model class.

10.1 Two complementary geometric viewpoints

Let

$$R^*(\Theta) := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[\ell(\theta, \widehat{\theta}) \Big]$$

be the minimax risk. If $\Theta' \subseteq \Theta$, then

$$R^*(\Theta) \geq R^*(\Theta'). \tag{10.1}$$

For lower bounds, we restrict the parameter space to a finite hard subset. The Yang–Barron result goes in the opposite direction: it constructs an estimator and proves an upper bound by approximating the entire model class with a finite KL cover.

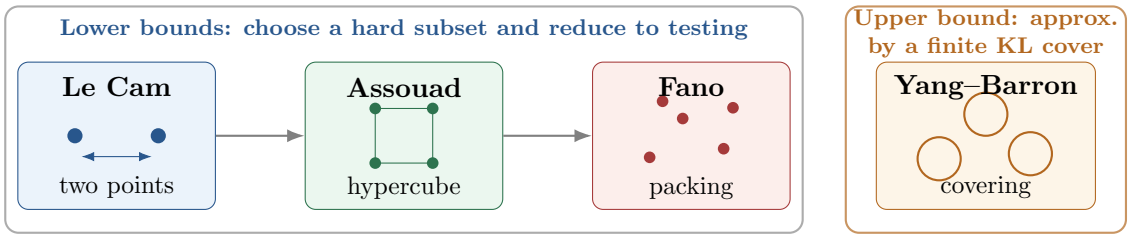


Figure 18: The three left-hand panels are minimax lower-bound methods based on hard subsets and hypothesis testing. Yang–Barron, shown separately on the right, is an entropic upper-bound construction. Packing and covering are related geometric notions, but the proof directions are opposite.

Takeaway

The arrows connect only the three lower-bound methods:

$$\text{two points} \longrightarrow \text{hypercube} \longrightarrow \text{general packing}.$$

Yang–Barron is not a fourth lower-bound step. It uses a KL cover to build a predictive estimator and establish achievability.

10.2 Le Cam's two-point method

10.2.1 Statement

Assume that $\ell : \Theta \times \Theta \to \mathbb{R}_+$ satisfies $\ell(\theta, \theta) = 0$ and the α-triangle inequality

$$\ell(\theta_0, \theta_1) \leq \alpha\big(\ell(\theta_0, \theta) + \ell(\theta_1, \theta)\big) \qquad \forall \theta_0, \theta_1, \theta \in \Theta. \tag{10.2}$$

For a metric loss one may take $\alpha = 1$, while for squared metric loss one may take $\alpha = 2$.

Theorem 10.1 (Le Cam's two-point method)**.** *Under* (10.2)*,*

$$R^*(\Theta) \geq \sup_{\theta_0,\theta_1\in\Theta} \frac{\ell(\theta_0,\theta_1)}{2\alpha}\left(1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1})\right). \tag{10.3}$$

10.2.2 Proof by a randomized reduction to testing

Fix θ_0, θ_1, and let $\widehat{\theta}$ be any estimator. Conditionally on $\widehat{\theta}$, write

$$a := \ell(\theta_0, \widehat{\theta}), \qquad b := \ell(\theta_1, \widehat{\theta}), \qquad L := \ell(\theta_0, \theta_1).$$

Construct a randomized binary test $\widetilde{\theta} \in \{\theta_0, \theta_1\}$ by

$$\mathbb{P}(\widetilde{\theta} = \theta_0 \mid \widehat{\theta}) = \frac{b}{a+b}, \qquad \mathbb{P}(\widetilde{\theta} = \theta_1 \mid \widehat{\theta}) = \frac{a}{a+b}. \tag{10.4}$$

If $a + b = 0$, then (10.2) forces $L = 0$, and the claim is trivial; the test may be defined arbitrarily in that case.

Under θ_0, the conditional probability of error is

$$\mathbb{P}(\widetilde{\theta} \neq \theta_0 \mid \widehat{\theta}) = \frac{a}{a+b}.$$

Since $L \leq \alpha(a+b)$,

$$a \geq \frac{L}{\alpha}\frac{a}{a+b} = \frac{L}{\alpha}\mathbb{P}(\widetilde{\theta} \neq \theta_0 \mid \widehat{\theta}).$$

Taking expectation under P_{θ_0},

$$R_{\theta_0}(\widehat{\theta}) \geq \frac{L}{\alpha} P_{\theta_0}(\widetilde{\theta} = \theta_1). \tag{10.5}$$

Similarly,

$$R_{\theta_1}(\widehat{\theta}) \geq \frac{L}{\alpha} P_{\theta_1}(\widetilde{\theta} = \theta_0). \tag{10.6}$$

Under the uniform prior on $\{\theta_0, \theta_1\}$, averaging (10.5)–(10.6) gives

$$R_\pi(\widehat{\theta}) \geq \frac{L}{\alpha}\mathbb{P}(\widetilde{\theta} \neq \Theta).$$

The minimum average error probability for testing P_{θ_0} against P_{θ_1} with equal priors is

$$\frac{1}{2}\left(1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1})\right).$$

Hence

$$R_\pi(\widehat{\theta}) \geq \frac{L}{2\alpha}\left(1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1})\right).$$

Taking the infimum over estimators and then the supremum over the two parameters proves (10.3).

10.2.3 Gaussian location examples

Example 10.2 (One-dimensional Gaussian location)**.** Suppose

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), \qquad \theta \in \mathbb{R},$$

and use squared loss. The sample mean is sufficient, with $\overline{X} \sim \mathcal{N}(\theta, 1/n)$. Since squared distance satisfies (10.2) with $\alpha = 2$, Le Cam gives

$$R_n^* \geq \sup_{\theta_0,\theta_1} \frac{(\theta_0 - \theta_1)^2}{4}\left[1 - \mathrm{TV}(\mathcal{N}(\theta_0, 1/n), \mathcal{N}(\theta_1, 1/n))\right].$$

Set $t = \sqrt{n}\,|\theta_1 - \theta_0|$. Using

$$\mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(t,1)) = 2\Phi(t/2) - 1,$$

we obtain

$$R_n^* \geq \frac{1}{n} \sup_{t>0} \frac{t^2}{2}\big(1 - \Phi(t/2)\big) = \frac{c_{\mathrm{LC}}}{n} \tag{10.7}$$

for a universal constant $c_{\mathrm{LC}} > 0$. Thus Le Cam recovers the correct $1/n$ rate.

Example 10.3 (Why two points miss dimension)**.** Consider the d-dimensional reduced Gaussian experiment

$$X \sim \mathcal{N}(\theta, n^{-1} I_d), \qquad \theta \in \mathbb{R}^d,$$

under squared Euclidean loss. Since squared loss has $\alpha = 2$, Le Cam's bound gives, for any pair $\theta_0, \theta_1 \in \mathbb{R}^d$,

$$R_n^*(\mathbb{R}^d) \geq \frac{\|\theta_0 - \theta_1\|_2^2}{4} \left[1 - \mathrm{TV}\big(\mathcal{N}(\theta_0, n^{-1} I_d), \mathcal{N}(\theta_1, n^{-1} I_d)\big)\right]. \tag{10.8}$$

Let $h := \theta_1 - \theta_0$. Translation invariance of TV reduces the two laws to

$$\mathcal{N}(0, n^{-1} I_d) \quad \text{and} \quad \mathcal{N}(h, n^{-1} I_d).$$

Choose an orthogonal matrix U satisfying $Uh = \|h\|_2 e_1$. Since TV is unchanged by a common bijective transformation and $U I_d U^\top = I_d$,

$$\mathrm{TV}\big(\mathcal{N}(0, n^{-1} I_d), \mathcal{N}(h, n^{-1} I_d)\big) = \mathrm{TV}\big(\mathcal{N}(0, n^{-1} I_d), \mathcal{N}(\|h\|_2 e_1, n^{-1} I_d)\big).$$

The latter two distributions have identical last $d-1$ coordinates. More explicitly, they factor as

$$\mathcal{N}(0, 1/n) \otimes R \quad \text{and} \quad \mathcal{N}(\|h\|_2, 1/n) \otimes R, \qquad R := \mathcal{N}(0, n^{-1} I_{d-1}).$$

Tensoring both distributions with the same factor does not change TV, so

$$\begin{aligned} &\mathrm{TV}\big(\mathcal{N}(\theta_0, n^{-1} I_d), \mathcal{N}(\theta_1, n^{-1} I_d)\big) \\ &\quad = \mathrm{TV}(\mathcal{N}(0, 1/n), \mathcal{N}(\|h\|_2, 1/n)) \\ &\quad = 2\Phi\left(\frac{\sqrt{n}\|h\|_2}{2}\right) - 1. \end{aligned} \tag{10.9}$$

Set

$$t := \sqrt{n}\|h\|_2.$$

Substituting (10.9) into (10.8) gives

$$\begin{aligned} R_n^*(\mathbb{R}^d) &\geq \frac{1}{n} \sup_{t>0} \frac{t^2}{4} \left[1 - \{2\Phi(t/2) - 1\}\right] \\ &= \frac{1}{n} \sup_{t>0} \frac{t^2}{2}\big[1 - \Phi(t/2)\big] = \frac{c_{\mathrm{LC}}}{n}, \end{aligned} \tag{10.10}$$

where $c_{\mathrm{LC}} > 0$ is the same universal constant as in the one-dimensional example. Thus even after optimizing over all pairs of parameters, the two-point method gives only order $1/n$, with no dependence on d.

By contrast, the exact quadratic minimax risk is d/n. Geometrically, any chosen pair lies along a single direction in $\mathbb{R}^d$; the two-point argument therefore probes only one scalar degree of freedom. Assouad's hypercube method remedies this by perturbing all d coordinates simultaneously.

10.3 Assouad's hypercube method

Theorem 10.4 (Assouad's lemma)**.** *Assume that ℓ satisfies (10.2). Suppose that Θ contains a hypercube*

$$\Theta' = \{\theta_b : b \in \{0,1\}^d\}$$

such that, for some $\beta > 0$,

$$\ell(\theta_b, \theta_{b'}) \geq \beta\, d_H(b, b') \qquad \forall b, b' \in \{0,1\}^d. \tag{10.11}$$

Then

$$R^*(\Theta) \geq \frac{\beta d}{4\alpha}\left[1 - \max_{d_H(b,b')=1} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}})\right]. \tag{10.12}$$

Proof. Put the uniform prior on $B \in \{0,1\}^d$. Given any estimator $\widehat{\theta}$, round it to the nearest hypercube vertex according to the loss:

$$\widehat{B} \in \operatorname*{arg\,min}_{b' \in \{0,1\}^d} \ell(\theta_{b'}, \widehat{\theta}).$$

By (10.11), the α-triangle inequality, and the defining property of $\widehat{B}$,

$$\beta d_H(B, \widehat{B}) \leq \ell(\theta_B, \theta_{\widehat{B}}) \leq \alpha\big[\ell(\theta_B, \widehat{\theta}) + \ell(\theta_{\widehat{B}}, \widehat{\theta})\big] \leq 2\alpha \ell(\theta_B, \widehat{\theta}).$$

Since

$$d_H(B, \widehat{B}) = \sum_{i=1}^{d} \mathbf{1}\{\widehat{B}_i \neq B_i\},$$

taking expectations gives

$$\mathbb{E}\Big[\ell(\theta_B, \widehat{\theta})\Big] \geq \frac{\beta}{2\alpha} \sum_{i=1}^{d} \mathbb{P}(\widehat{B}_i \neq B_i). \tag{10.13}$$

We now lower-bound the error probability for each coordinate separately. Fix $i \in [d]$, and let

$$\mathcal{B}_{i,0} := \{b \in \{0,1\}^d : b_i = 0\}, \qquad \mathcal{B}_{i,1} := \{b \in \{0,1\}^d : b_i = 1\}.$$

Because B is uniform, the conditional laws of the observation are the uniform mixtures

$$P_{i,0} := P_{X|B_i=0} = \frac{1}{2^{d-1}} \sum_{b \in \mathcal{B}_{i,0}} P_{\theta_b}, \qquad P_{i,1} := P_{X|B_i=1} = \frac{1}{2^{d-1}} \sum_{b \in \mathcal{B}_{i,1}} P_{\theta_b}. \tag{10.14}$$

Thus $\widehat{B}_i$ is a binary test between $P_{i,0}$ and $P_{i,1}$ with equal priors. Consequently,

$$\begin{aligned} \mathbb{P}(\widehat{B}_i \neq B_i) &= \frac{1}{2} P_{i,0}(\widehat{B}_i = 1) + \frac{1}{2} P_{i,1}(\widehat{B}_i = 0) \\ &\geq \frac{1}{2}\big(1 - \mathrm{TV}(P_{i,0}, P_{i,1})\big). \end{aligned} \tag{10.15}$$

It remains to compare the TV distance between the two mixtures with the TV distances along hypercube edges. Pair every $b \in \mathcal{B}_{i,0}$ with $b \oplus e_i \in \mathcal{B}_{i,1}$. Using (10.14) and the joint convexity inequality

$$\mathrm{TV}\left(\sum_j \lambda_j P_j, \sum_j \lambda_j Q_j\right) \leq \sum_j \lambda_j\, \mathrm{TV}(P_j, Q_j), \qquad \lambda_j \geq 0, \quad \sum_j \lambda_j = 1,$$

we obtain

$$\begin{aligned}\mathrm{TV}(P_{i,0}, P_{i,1}) &= \mathrm{TV}\left(\frac{1}{2^{d-1}}\sum_{b\in\mathcal{B}_{i,0}} P_{\theta_b}, \frac{1}{2^{d-1}}\sum_{b\in\mathcal{B}_{i,0}} P_{\theta_{b\oplus e_i}}\right)\\ &\le \frac{1}{2^{d-1}}\sum_{b\in\mathcal{B}_{i,0}} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b\oplus e_i}})\\ &\le \max_{d_H(b,b')=1} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}). \end{aligned} \tag{10.16}$$

For completeness, joint convexity follows immediately from the variational definition of TV: for every measurable set A,

$$\left|\sum_j \lambda_j P_j(A) - \sum_j \lambda_j Q_j(A)\right| \le \sum_j \lambda_j \,\mathrm{TV}(P_j, Q_j),$$

and then one takes the supremum over A.

Substituting (10.15) and (10.16) into (10.13) proves the Bayes lower bound in (10.12). Finally, the minimax risk dominates the Bayes risk under the uniform prior on the hypercube. □

Example 10.5 (The d-dimensional Gaussian location model)**.** Consider the reduced Gaussian experiment

$$X \sim \mathcal{N}(\theta, n^{-1}I_d), \qquad \theta \in \mathbb{R}^d,$$

under squared Euclidean loss

$$\ell(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_2^2.$$

We apply Assouad's lemma to the hypercube

$$\theta_b = \varepsilon b, \qquad b \in \{0,1\}^d,$$

where $\varepsilon > 0$ will be chosen below.

First, for any two vertices,

$$\ell(\theta_b, \theta_{b'}) = \|\theta_b - \theta_{b'}\|_2^2 = \varepsilon^2 d_H(b, b'). \tag{10.17}$$

Thus the separation constant in Assouad's lemma is $\beta = \varepsilon^2$. Moreover, squared Euclidean loss satisfies the α-triangle inequality with $\alpha = 2$, because

$$\|u - v\|_2^2 \le 2\|u - w\|_2^2 + 2\|v - w\|_2^2.$$

Next consider adjacent vertices b, b', so that $d_H(b, b') = 1$. If they differ in coordinate i, then $\theta_b - \theta_{b'} = \pm\varepsilon e_i$. By translation invariance of TV,

$$\mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}) = \mathrm{TV}\big(\mathcal{N}(0, n^{-1}I_d), \mathcal{N}(\varepsilon e_i, n^{-1}I_d)\big)\,.$$

The two distributions have identical coordinates except for coordinate i. Tensoring both one-dimensional laws with the same $(d-1)$-dimensional Gaussian factor does not change TV, and hence

$$\begin{aligned}\mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}) &= \mathrm{TV}(\mathcal{N}(0, 1/n), \mathcal{N}(\varepsilon, 1/n))\\ &= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{2}\right) - 1. \end{aligned} \tag{10.18}$$

Here we used the exact equal-variance Gaussian formula

$$\mathrm{TV}\big(\mathcal{N}(0,\sigma^2),\mathcal{N}(\mu,\sigma^2)\big) = 2\Phi\left(\frac{|\mu|}{2\sigma}\right) - 1.$$

Substituting $\beta = \varepsilon^2$, $\alpha = 2$, and (10.18) into Assouad's bound gives

$$\begin{aligned} R_n^*(\mathbb{R}^d) &\geq \frac{\varepsilon^2 d}{8}\left[1 - \left\{2\Phi(\varepsilon\sqrt{n}/2) - 1\right\}\right] \\ &= \frac{\varepsilon^2 d}{4}\left[1 - \Phi(\varepsilon\sqrt{n}/2)\right]. \end{aligned} \tag{10.19}$$

Set

$$\varepsilon = \frac{t}{\sqrt{n}}$$

for any fixed $t > 0$. Then

$$R_n^*(\mathbb{R}^d) \geq \frac{d}{n}\frac{t^2}{4}\left[1 - \Phi(t/2)\right]. \tag{10.20}$$

For example, taking $t = 1$ yields the explicit universal bound

$$R_n^*(\mathbb{R}^d) \geq \frac{1 - \Phi(1/2)}{4}\frac{d}{n} \approx 0.077\frac{d}{n}.$$

Consequently,

$$R_n^*(\mathbb{R}^d) \gtrsim \frac{d}{n}. \tag{10.21}$$

The sample mean has risk exactly d/n, so Assouad's lemma recovers the correct minimax rate and, unlike the two-point method, the correct linear dependence on dimension. Intuitively, the hypercube forces the estimator to solve d binary testing problems simultaneously; each unresolved bit costs order $1/n$ in squared error.

10.4 Fano's packing method

Let d now denote a metric on Θ. For $T \subseteq \Theta$, write $M(T, d, \varepsilon)$ for its ε-packing number, and let

$$C(T) := \sup_{\pi} I_\pi(\Theta; X)$$

be the capacity of the experiment restricted to input space T.

Theorem 10.6 (Fano's method)**.** *For any estimator $\widehat{\theta}$, any $T \subseteq \Theta$, and any $\varepsilon > 0$,*

$$\mathbb{P}\Big[d(\Theta, \widehat{\theta}) \geq \frac{\varepsilon}{2}\Big] \geq 1 - \frac{C(T) + \log 2}{\log M(T, d, \varepsilon)}. \tag{10.22}$$

Consequently, for every $r > 0$,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta\Big[d(\theta, \widehat{\theta})^r\Big] \geq \sup_{T \subseteq \Theta, \varepsilon > 0} \left(\frac{\varepsilon}{2}\right)^r \left[1 - \frac{C(T) + \log 2}{\log M(T, d, \varepsilon)}\right]. \tag{10.23}$$

Proof. Fix T and let

$$T' = \{\theta_1, \ldots, \theta_M\} \subseteq T$$

be an ε-packing with $M = M(T, d, \varepsilon)$. Let Θ be uniform on T', and conditioned on $\Theta = \theta_j$, generate $X \sim P_{\theta_j}$. Round the estimator to the nearest packing point:

$$\widetilde{\Theta} \in \operatorname*{arg\,min}_{\theta \in T'} d(\widehat{\theta}, \theta).$$

If $\widetilde{\Theta} \neq \Theta$, then the packing property and the triangle inequality give

$$\varepsilon \le d(\Theta, \widetilde{\Theta}) \le d(\Theta, \widehat{\theta}) + d(\widehat{\theta}, \widetilde{\Theta}) \le 2d(\Theta, \widehat{\theta}).$$

Hence

$$\mathbb{P}(\widetilde{\Theta} \neq \Theta) \le \mathbb{P}\Big[d(\Theta, \widehat{\theta}) \ge \frac{\varepsilon}{2}\Big].$$

Fano's inequality yields

$$\mathbb{P}(\widetilde{\Theta} \neq \Theta) \ge 1 - \frac{I(\Theta; X) + \log 2}{\log M},$$

and $I(\Theta; X) \le C(T)$, proving (10.22). Finally, $\mathbb{E}[Z^r] \ge a^r \mathbb{P}(Z \ge a)$ with $Z = d(\Theta, \widehat{\theta})$ and $a = \varepsilon/2$ gives (10.23). □

Takeaway

Fano replaces the structured hypercube by an arbitrary packing. The quantity $\log M(T, d, \varepsilon)$ measures how many well-separated parameters must be distinguished, while $C(T)$ measures how much information the data can reveal about their identities.

10.5 Application: sparse high-dimensional linear regression

Consider the fixed-design linear model

$$Y = A\theta + W, \qquad W \sim \mathcal{N}(0, \sigma^2 I_n), \tag{10.24}$$

where $A \in \mathbb{R}^{n \times p}$ is known and

$$\Theta_{p,s} := B_0(s) \cap B_2(1) = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s,\ \|\theta\|_2 \le 1\}.$$

Under parameter θ,

$$P_\theta = \mathcal{N}(A\theta, \sigma^2 I_n).$$

We use the metric $d(\theta, \theta') = \|\theta - \theta'\|_2$.

10.5.1 Restricted upper singular value

Define

$$\gamma_{2s} := \sup_{\substack{u \neq 0 \\ \|u\|_0 \le 2s}} \frac{\|Au\|_2}{\sqrt{n}\,\|u\|_2}. \tag{10.25}$$

The difference of two s-sparse vectors is at most $2s$-sparse, so

$$\|A(\theta - \theta')\|_2 \le \sqrt{n}\,\gamma_{2s} \|\theta - \theta'\|_2. \tag{10.26}$$

10.5.2 A sparse packing

A standard sparse Varshamov–Gilbert construction gives, for all sufficiently small $\varepsilon > 0$, a set

$$T_\varepsilon = \{\theta_1, \dots, \theta_M\} \subseteq \Theta_{p,s}$$

such that

$$\varepsilon \le \|\theta_j - \theta_k\|_2 \le 2\sqrt{2}\,\varepsilon, \qquad j \neq k, \tag{10.27}$$

and

$$\log M \ge L_{p,s} := \frac{s}{2} \log\left(\frac{p - s}{s/2}\right). \tag{10.28}$$

In the sparse regime $s \ll p$,

$$L_{p,s} \asymp s \log\left(\frac{ep}{s}\right) \asymp \log\binom{p}{s}. \tag{10.29}$$

Thus $L_{p,s}$ is the logarithmic combinatorial complexity of choosing a well-separated collection of sparse supports.

10.5.3 KL diameter and Fano's bound

For $\theta_j, \theta_k \in T_\varepsilon$,

$$\begin{aligned} D(P_{\theta_j} \| P_{\theta_k}) &= \frac{1}{2\sigma^2} \|A(\theta_j - \theta_k)\|_2^2 \\ &\le \frac{n\gamma_{2s}^2}{2\sigma^2} \|\theta_j - \theta_k\|_2^2 \\ &\le \frac{4n\gamma_{2s}^2 \varepsilon^2}{\sigma^2}. \end{aligned} \tag{10.30}$$

Capacity is no larger than the KL diameter, so

$$C(T_\varepsilon) \le \frac{4n\gamma_{2s}^2 \varepsilon^2}{\sigma^2}. \tag{10.31}$$

Applying Fano with $r = 1$,

$$R_1^*(\Theta_{p,s}) \ge \frac{\varepsilon}{2}\left[1 - \frac{4n\gamma_{2s}^2\varepsilon^2/\sigma^2 + \log 2}{L_{p,s}}\right], \tag{10.32}$$

where

$$R_1^*(\Theta_{p,s}) := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_{p,s}} \mathbb{E}_\theta\left[\|\widehat{\theta} - \theta\|_2\right].$$

Choose

$$\varepsilon^2 = \frac{\sigma^2 L_{p,s}}{16 n \gamma_{2s}^2},$$

provided this is small enough for the packing construction. If $L_{p,s} \ge 4 \log 2$, the bracket in (10.32) is at least $1/2$. Consequently,

$$R_1^*(\Theta_{p,s}) \gtrsim \min\left\{1, \frac{\sigma}{\gamma_{2s}} \sqrt{\frac{s}{n} \log\left(\frac{ep}{s}\right)}\right\}. \tag{10.33}$$

Similarly, using $r = 2$ in Fano,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_{p,s}} \mathbb{E}_\theta\left[\|\widehat{\theta} - \theta\|_2^2\right] \gtrsim \min\left\{1, \frac{\sigma^2}{\gamma_{2s}^2} \frac{s}{n} \log\left(\frac{ep}{s}\right)\right\}. \tag{10.34}$$

When $\gamma_{2s} = O(1)$, the squared-risk rate is the familiar $\sigma^2 s \log(ep/s)/n$. Under suitable design conditions, the Lasso attains this rate up to constants.

10.6 From lower bounds to entropic upper bounds

Let

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

be a model class, and observe $X^n = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} P_\theta$. An estimator is now a predictive distribution

$$\widehat{P}(\cdot \mid X^n),$$

and its KL risk is

$$\mathbb{E}_\theta\Big[D\Big(P_\theta\,\Big\|\,\widehat{P}(\cdot\mid X^n)\Big)\Big].$$

Define

$$R^*_{\mathrm{KL}}(n) := \inf_{\widehat{P}}\sup_{\theta\in\Theta}\mathbb{E}_\theta\Big[D\Big(P_\theta\|\widehat{P}(\cdot\mid X^n)\Big)\Big]. \tag{10.35}$$

The estimator may be *improper*: it need not belong to $\mathcal{P}$.

For the n-sample experiment, define its capacity

$$C_n := \sup_\pi I_\pi(\Theta; X^n), \tag{10.36}$$

where the supremum is over all priors on Θ.

Definition 10.7 (KL covering number)**.** For $\varepsilon > 0$, let $N_{\mathrm{KL}}(\mathcal{P},\varepsilon)$ be the smallest N for which there exist distributions $Q_1,\dots,Q_N$ such that

$$\forall\theta\in\Theta,\qquad \min_{1\le j\le N} D(P_\theta\|Q_j)\le\varepsilon^2. \tag{10.37}$$

The centers Q_j need not belong to $\mathcal{P}$.

Theorem 10.8 (Yang–Barron)**.** *For every $n\ge 0$,*

$$R^*_{\mathrm{KL}}(n)\le\frac{C_{n+1}}{n+1}\le\inf_{\varepsilon>0}\left\{\varepsilon^2+\frac{\log N_{\mathrm{KL}}(\mathcal{P},\varepsilon)}{n+1}\right\}. \tag{10.38}$$

The proof has three conceptual steps.

10.7 A high-level proof of the Yang–Barron theorem

10.7.1 Step 1: Bayes KL risk is conditional mutual information

Fix a prior π, and introduce an unseen future observation $X_{n+1}\sim P_\Theta$, conditionally independent of X^n given Θ. For any predictive distribution $Q_{X_{n+1}|X^n}$, the conditional KL identity gives

$$\begin{aligned}&\mathbb{E}\big[D\big(P_{X_{n+1}|\Theta}\big\|Q_{X_{n+1}|X^n}\big)\big]\\&\quad= I(\Theta; X_{n+1}\mid X^n)+\mathbb{E}\big[D\big(P_{X_{n+1}|X^n}\big\|Q_{X_{n+1}|X^n}\big)\big].\end{aligned} \tag{10.39}$$

Hence the posterior predictive distribution

$$\widehat{P}^*_\pi(\cdot\mid X^n)=P_{X_{n+1}|X^n} \tag{10.40}$$

minimizes Bayes KL risk, and

$$R^*_{\mathrm{KL,Bayes}}(\pi,n)=I(\Theta;X_{n+1}\mid X^n). \tag{10.41}$$

This is the KL analogue of the conditional Pythagorean identity for squared error.

10.7.2 Step 2: one-step predictive risk is controlled by capacity

A direct worst-case argument avoids any need to invoke a minimax theorem. Use the information-radius representation

$$C_{n+1}=\inf_Q\sup_{\theta\in\Theta}D(P_\theta^{\otimes(n+1)}\|Q). \tag{10.42}$$

Fix $\eta>0$, and choose Q such that

$$\sup_\theta D(P_\theta^{\otimes(n+1)}\|Q)\le C_{n+1}+\eta.$$

Factor Q sequentially as

$$Q(\mathrm{d}x^{n+1}) = \prod_{t=0}^{n} Q_{t+1}(\mathrm{d}x_{t+1} \mid x^t).$$

Construct the batch predictor

$$\widehat{P}(\cdot \mid X^n) := \frac{1}{n+1} \sum_{t=0}^{n} Q_{t+1}(\cdot \mid X^t). \tag{10.43}$$

By convexity of KL divergence in its second argument,

$$\begin{aligned} &\mathbb{E}_\theta\Big[D\Big(P_\theta \| \widehat{P}(\cdot \mid X^n)\Big)\Big] \\ &\quad \le \frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E}_\theta\big[D\big(P_\theta \| Q_{t+1}(\cdot \mid X^t)\big)\big] \\ &\quad = \frac{1}{n+1} D(P_\theta^{\otimes(n+1)} \| Q) \le \frac{C_{n+1} + \eta}{n+1}. \end{aligned} \tag{10.44}$$

Taking the supremum over θ, the infimum over estimators, and then letting $\eta \downarrow 0$ proves

$$R^*_{\mathrm{KL}}(n) \le \frac{C_{n+1}}{n+1}. \tag{10.45}$$

10.7.3 Step 3: a KL cover bounds capacity

Fix $\varepsilon > 0$, and let $Q_1, \dots, Q_N$ be an ε-KL cover, where $N = N_{\mathrm{KL}}(\mathcal{P}, \varepsilon)$. For each θ, select an index $J(\theta)$ satisfying

$$D(P_\theta \| Q_{J(\theta)}) \le \varepsilon^2.$$

For any prior π, since $J = J(\Theta)$ is a function of Θ,

$$\begin{aligned} I(\Theta; X^n) &= I(\Theta, J; X^n) \\ &= I(J; X^n) + I(\Theta; X^n \mid J) \\ &\le H(J) + \mathbb{E}\big[D\big(P_\Theta^{\otimes n} \| Q_J^{\otimes n}\big)\big] \\ &\le \log N + n\varepsilon^2. \end{aligned} \tag{10.46}$$

The middle inequality is the variational upper bound on conditional mutual information: one may compare each conditional law to any reference law that depends only on the conditioning variable J. Taking the supremum over priors gives

$$C_n \le \inf_{\varepsilon > 0} \left\{ n\varepsilon^2 + \log N_{\mathrm{KL}}(\mathcal{P}, \varepsilon) \right\}. \tag{10.47}$$

Combining (10.45) with (10.47) proves (10.38).

Takeaway

The Yang–Barron bound is the approximation–estimation tradeoff

$$\text{KL risk} \lesssim \underbrace{\varepsilon^2}_{\text{approximation}} + \underbrace{\frac{\log N_{\mathrm{KL}}(\varepsilon)}{n}}_{\text{description/estimation cost}}.$$

A KL cover gives a coarse model index costing $\log N$ nats and a residual mismatch costing at most ε^2 per observation.

10.8 Application to a smooth finite-dimensional model

Suppose $\Theta \subset \mathbb{R}^d$ is compact and the model is locally smooth in the sense that

$$D(P_\theta \| P_{\theta'}) \asymp \|\theta - \theta'\|_2^2. \tag{10.48}$$

Then a Euclidean ε-cover of Θ induces a KL cover, and

$$N_{\mathrm{KL}}(\mathcal{P}, \varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^d. \tag{10.49}$$

Therefore,

$$C_n \lesssim \inf_{\varepsilon > 0} \left\{ n\varepsilon^2 + d \log \frac{1}{\varepsilon} \right\}.$$

The balance occurs at $\varepsilon^2 \asymp d/n$, giving

$$C_n \lesssim d \log\left(\frac{en}{d}\right), \qquad R^*_{\mathrm{KL}}(n) \lesssim \frac{d}{n} \log\left(\frac{en}{d}\right). \tag{10.50}$$

The correct regular parametric rate is d/n, so the global entropy bound loses a logarithmic factor. Chapter 32 removes this loss by using *local* rather than global entropy.

Lecture 10 summary

Method	Geometry	Key comparison	Role
Le Cam	two points	TV overlap	basic lower bound
Assouad	hypercube	adjacent TV	dimensional lower bound
Fano	packing	$C(T)$ versus $\log M$	entropy lower bound
Yang–Barron	KL cover	C_n versus $\log N$	entropy upper bound

The second half of the course can now be read as one chain:

$$\text{metric entropy} \longrightarrow \begin{cases} \text{Fano lower bounds via packings,} \\ \text{Yang–Barron upper bounds via coverings.} \end{cases}$$

When the lower and upper entropy calculations balance at the same scale, they identify the minimax rate.

11 Lecture 11: Entropic Bounds for Statistical Estimation

11.1 Recap: Yang–Barron's construction

Let $\mathcal{P}$ be a class of probability distributions. The minimax KL risk using n samples is

$$R^*_{\mathrm{KL}}(\mathcal{P}, n) := \inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[D(P \| \widehat{P}(\cdot \mid X^n))\big]. \tag{11.1}$$

For $\varepsilon > 0$, define the KL covering number by

$$N_{\mathrm{KL}}(\mathcal{P}, \varepsilon) := \min\Big\{N : \exists Q_1, \dots, Q_N \text{ such that } \forall P \in \mathcal{P},\ \exists i \in [N],\ D(P\|Q_i) \le \varepsilon^2\Big\}. \tag{11.2}$$

For the n-sample experiment, define its capacity by

$$C_n := \sup_{\pi} I_\pi(\Theta; X^n), \tag{11.3}$$

where the supremum is over all distributions π on the parameter indexing the model class. Yang–Barron's construction gives

$$R^*_{\mathrm{KL}}(\mathcal{P}, n) \le \frac{C_{n+1}}{n+1} \le \inf_{\varepsilon > 0} \left\{\varepsilon^2 + \frac{1}{n+1} \log N_{\mathrm{KL}}(\mathcal{P}, \varepsilon)\right\}. \tag{11.4}$$

Thus, up to the immaterial replacement of $n+1$ by n,

$$R^*_{\mathrm{KL}}(\mathcal{P}, n) \lesssim \inf_{\varepsilon > 0} \left\{\varepsilon^2 + \frac{1}{n} \log N_{\mathrm{KL}}(\mathcal{P}, \varepsilon)\right\}. \tag{11.5}$$

The two terms in (11.5) are the approximation and estimation costs:

$$\underbrace{\varepsilon^2}_{\text{approximation error}} \quad + \quad \underbrace{\frac{1}{n} \log N_{\mathrm{KL}}(\mathcal{P}, \varepsilon)}_{\text{estimation/description cost}}.$$

For a regular d-dimensional parametric class, $N_{\mathrm{KL}}(\mathcal{P}, \varepsilon)$ is typically of order $(1/\varepsilon)^d$, so the global Yang–Barron bound gives

$$R^*_{\mathrm{KL}}(\mathcal{P}, n) \lesssim \frac{d}{n} \log n$$

up to constants and the more precise factor $\log(n/d)$.

Contrast with Fano. Fano's method begins with an ε-packing. If the packing points cannot be reliably tested, then the parameter cannot be estimated to precision ε; this yields a minimax *lower* bound in terms of metric entropy. Le Cam–Birgé instead fixes an ε-covering in Hellinger distance and constructs an estimator, giving an *upper* bound.

11.2 Hellinger distance: reminders

For distributions P, Q dominated by a common measure μ, with densities p, q, define

$$H^2(P, Q) := \int (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\mu = 2 - 2 \int \sqrt{pq} \, \mathrm{d}\mu. \tag{11.6}$$

The Hellinger distance $H(P, Q)$ is a metric. We will use the following facts.

(i) **Triangle inequality:**

$$H(P, R) \le H(P, Q) + H(Q, R).$$

(ii) **Convexity:** H^2 is an f-divergence, hence jointly convex. In particular,

$$H^2\big((1-\alpha)P' + \alpha P'', P\big) \le (1-\alpha)H^2(P', P) + \alpha H^2(P'', P),$$

so every Hellinger ball

$$B_H(P, \varepsilon) := \{Q : H(Q, P) \le \varepsilon\}$$

is convex.

(iii) **Relations to KL and χ^2:**

$$H^2(P, Q) \le D(P\|Q) \le \log\big(1 + \chi^2(P\|Q)\big) \le \chi^2(P\|Q). \tag{11.7}$$

In the density-estimation application we will also use a two-sided comparison under upper and lower density bounds; see Lemma 11.6.

11.3 The global Le Cam–Birgé theorem

Let $N_H(\mathcal{P}, \varepsilon)$ denote the ε-covering number of $\mathcal{P}$ under the Hellinger distance. Given n observations from an unknown $P \in \mathcal{P}$, the goal is to estimate P under squared Hellinger loss

$$\ell(P, \widehat{P}) = H^2(P, \widehat{P}).$$

Theorem 11.1 (Le Cam–Birgé)**.** *Let $\varepsilon_n > 0$ satisfy*

$$n\varepsilon_n^2 \ge \log N_H(\mathcal{P}, \varepsilon_n) \vee 1. \tag{11.8}$$

Then there exists an estimator $\widehat{P} = \widehat{P}(X^n)$, taking values in $\mathcal{P}$, such that for every $t \ge 1$,

$$\sup_{P \in \mathcal{P}} P^n\big[H(P, \widehat{P}) > 4t\varepsilon_n\big] \lesssim e^{-t^2}. \tag{11.9}$$

Consequently,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\big[H^2(P, \widehat{P})\big] \lesssim \varepsilon_n^2. \tag{11.10}$$

The proof uses pairwise comparisons for the composite hypotheses

$$H_i : P \in B_H(P_i, \varepsilon) \qquad \text{versus} \qquad H_j : P \in B_H(P_j, \varepsilon).$$

The following robust testing theorem is the main testing input.

Theorem 11.2 (Robust testing)**.** *Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two families of distributions, and define*

$$H(\mathcal{P}_0, \mathcal{P}_1) := \inf_{P \in \mathcal{P}_0,\, Q \in \mathcal{P}_1} H(P, Q).$$

There exists a test $\psi = \psi(X^n) \in \{0, 1\}$ such that

$$\sup_{P \in \mathcal{P}_0} P^n(\psi = 1) + \sup_{Q \in \mathcal{P}_1} Q^n(\psi = 0) \le \exp\Big(-\frac{n}{2} H^2\big(\mathrm{co}(\mathcal{P}_0), \mathrm{co}(\mathcal{P}_1)\big)\Big). \tag{11.11}$$

If $\mathcal{P}_0$ and $\mathcal{P}_1$ are convex, then

$$\sup_{P \in \mathcal{P}_0} P^n(\psi = 1) + \sup_{Q \in \mathcal{P}_1} Q^n(\psi = 0) \le \exp\Big(-\frac{n}{2} H^2(\mathcal{P}_0, \mathcal{P}_1)\Big). \tag{11.12}$$

Here $\mathrm{co}(\mathcal{P})$ denotes the convex hull. For the Hellinger balls used in Le Cam–Birgé,

$$\mathcal{P}_0 = B_H(P_i, \varepsilon), \qquad \mathcal{P}_1 = B_H(P_j, \varepsilon),$$

the two families are convex. If $H(P_i, P_j) > 4\varepsilon$, then

$$\begin{aligned} H(\mathcal{P}_0, \mathcal{P}_1) &\ge H(P_i, P_j) - 2\varepsilon \\ &\ge \frac{1}{2} H(P_i, P_j). \end{aligned}$$

Therefore we may choose a test ψ_{ij}, where $\psi_{ij} = 1$ means that P_j defeats P_i, such that

$$\sup_{P \in B_H(P_i, \varepsilon)} P^n(\psi_{ij} = 1) \le \exp\Big(-\frac{n}{8} H^2(P_i, P_j)\Big). \tag{11.13}$$

11.4 Three-step proof of the global theorem

Set $\varepsilon = \varepsilon_n$ and $N = N_H(\mathcal{P}, \varepsilon)$.

Step 1: Approximate the model by finitely many candidates

Choose a maximal ε-packing

$$P_1, \dots, P_N \in \mathcal{P}.$$

By maximality, it also forms an ε-cover. Hence for every true distribution $P \in \mathcal{P}$, there exists an index

$$i_0 = i_0(P) \in [N]$$

such that

$$H(P, P_{i_0}) \le \varepsilon. \tag{11.14}$$

The index $i_0(P)$ depends on the unknown P, but it is only an oracle reference point in the analysis; it is not used by the estimator. After fixing P, relabel the centers so that $i_0 = 1$.

Our task is to identify, from the data, a candidate that is not too far from P_1, and hence not too far from P.

Step 2: Construct pairwise tests and select a candidate

For any two candidates P_i, P_j such that

$$H(P_i, P_j) > 4\varepsilon,$$

use the test ψ_{ij} in (11.13) to compare

$$H_i : P \in B_H(P_i, \varepsilon) \qquad \text{and} \qquad H_j : P \in B_H(P_j, \varepsilon).$$

For each candidate P_i, define

$$T_i := \begin{cases} \max\limits_{j:\, \psi_{ij}=1,\ H(P_i,P_j)>4\varepsilon} H^2(P_i, P_j), & \text{if such a } j \text{ exists}, \\ 0, & \text{otherwise.} \end{cases} \tag{11.15}$$

Thus T_i records the maximum squared distance from P_i to a far-away candidate that defeats P_i. The value $T_i = 0$ means that no candidate outside the 4ε-neighborhood of P_i defeats it.

Define

$$\widehat{P} = P_{i^*}, \qquad i^* \in \arg\min_{i \in [N]} T_i. \tag{11.16}$$

The estimator chooses the candidate whose worst defeat is least severe.

Step 3: The selected candidate is close to the truth

Recall that the true distribution P lies within ε of P_1. The key observation is:

If $\widehat{P} = P_{i^*}$ is far from P_1, then P_1 must have lost a test against a distant candidate.

Indeed, for any pair P_i, P_j satisfying $H(P_i, P_j) > 4\varepsilon$, one defeats the other, and hence

$$\max\{T_i, T_j\} \ge H^2(P_i, P_j). \tag{11.17}$$

Since i^* minimizes T_i, we have $T_{i^*} \le T_1$. Therefore, on the event $H(P_{i^*}, P_1) > 4\varepsilon$,

$$T_1 = \max\{T_{i^*}, T_1\} \ge H^2(P_{i^*}, P_1). \tag{11.18}$$

Consequently, for every $t \geq 1$,

$$\{H(\widehat{P}, P_1) > 4t\varepsilon\} \subseteq \{T_1 > 16t^2\varepsilon^2\}. \tag{11.19}$$

Since $P \in B_H(P_1, \varepsilon)$, the robust testing bound and a union bound over the N candidates give

$$\begin{aligned} P^n\big[T_1 > 16t^2\varepsilon^2\big] &\leq N \exp(-2nt^2\varepsilon^2) && (11.20)\\ &\leq \exp(n\varepsilon^2 - 2nt^2\varepsilon^2) && (11.21)\\ &\leq e^{-t^2}. && (11.22) \end{aligned}$$

Here we used

$$\log N \leq n\varepsilon^2, \qquad n\varepsilon^2 \geq 1, \qquad t \geq 1.$$

It follows that

$$P^n\big[H(\widehat{P}, P_1) > 4t\varepsilon\big] \leq e^{-t^2}. \tag{11.23}$$

Finally, by the triangle inequality and (11.14),

$$H(P, \widehat{P}) \leq H(P, P_1) + H(P_1, \widehat{P}) \leq \varepsilon + H(P_1, \widehat{P}). \tag{11.24}$$

Thus

$$P^n\big[H(P, \widehat{P}) > (4t+1)\varepsilon\big] \leq e^{-t^2}.$$

Since $4t + 1 \leq 5t$ for $t \geq 1$, this gives the theorem up to an immaterial numerical constant; the stated constant 4 follows from slightly tighter bookkeeping.

To obtain the risk bound, use the tail-integration identity

$$\mathbb{E}[Z^2] = 2\int_0^\infty s\,\mathbb{P}(Z > s)\,\mathrm{d}s \tag{11.25}$$

for any nonnegative random variable Z. Applying this to $Z = H(P, \widehat{P})$, splitting the integral at 5ε, and using the preceding tail bound yields

$$\begin{aligned} \mathbb{E}_P\big[H^2(P, \widehat{P})\big] &\leq 25\varepsilon^2 + 2\int_{5\varepsilon}^\infty s \exp\left(-\frac{s^2}{25\varepsilon^2}\right) \mathrm{d}s \\ &\lesssim \varepsilon^2. \end{aligned}$$

Taking the supremum over $P \in \mathcal{P}$ proves (11.10).

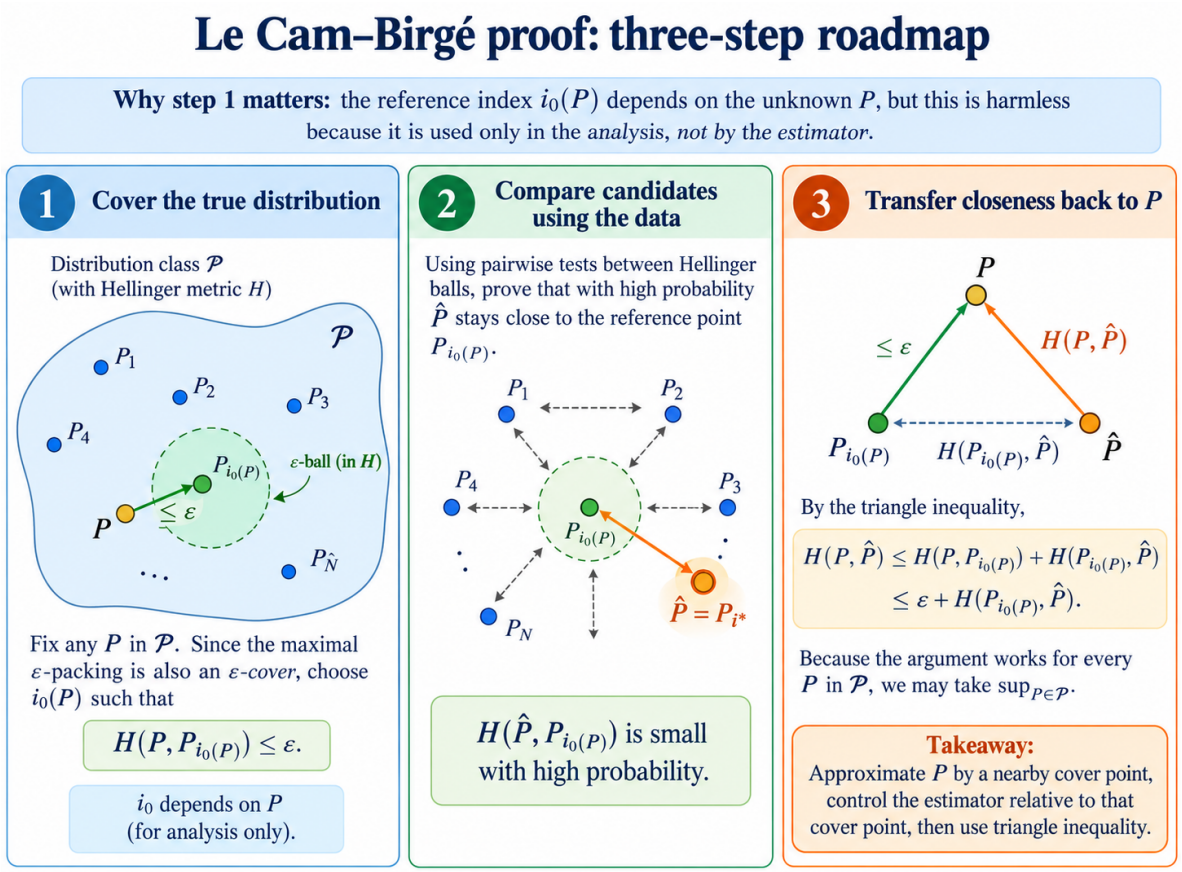


Figure 19: The three-step proof roadmap: cover the truth by an oracle center, compare candidates using robust tests, and transfer closeness back to the true distribution.

11.5 Global entropy and its local refinement

11.5.1 The global logarithmic loss

For a regular d-dimensional model,

$$N_H(\mathcal{P}, \varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^d, \qquad \log N_H(\mathcal{P}, \varepsilon) \asymp d \log \frac{1}{\varepsilon}. \tag{11.26}$$

The global theorem asks us to choose ε_n so that

$$n\varepsilon_n^2 \asymp \log N_H(\mathcal{P}, \varepsilon_n) \asymp d \log \frac{1}{\varepsilon_n}.$$

Equivalently,

$$\varepsilon_n^2 \asymp \frac{d}{n} \log \frac{1}{\varepsilon_n}.$$

When $n \gtrsim d$, the resulting scale is

$$\varepsilon_n^2 \asymp \frac{d}{n} \log \frac{n}{d}. \tag{11.27}$$

Hence the global Le Cam–Birgé theorem gives

$$R^*_{H^2}(n) \lesssim \frac{d}{n} \log \frac{n}{d}, \tag{11.28}$$

which is suboptimal by a logarithmic factor compared with the parametric rate d/n.

Why is there a loss? A global ε-cover pays for two tasks simultaneously:

(1) locating the true distribution over the whole model class;

(2) estimating it to precision ε.

This overcounts the relevant complexity: after observing the data, most of the model class should no longer be plausible, and only a local neighborhood needs fine resolution.

11.5.2 Local Hellinger entropy

Definition 11.3 (Local Hellinger entropy)**.** For $\varepsilon > 0$, define

$$N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) := \sup_{P_0 \in \mathcal{P}} \sup_{\eta \geq \varepsilon} N_H\big(B_H(P_0, \eta) \cap \mathcal{P}, \eta/2\big), \tag{11.29}$$

where

$$B_H(P_0, \eta) := \{P \in \mathcal{P} : H(P, P_0) \leq \eta\}.$$

Thus N_{loc} asks how many balls of half the radius are needed to cover an arbitrary local Hellinger ball.

Theorem 11.4 (Le Cam–Birgé: local-entropy version)**.** *Let $\varepsilon_n > 0$ satisfy*

$$n\varepsilon_n^2 \gtrsim \log N_{\mathrm{loc}}(\mathcal{P}, \varepsilon_n) \vee 1. \tag{11.30}$$

Then there exists an estimator $\widehat{P} = \widehat{P}(X^n)$, taking values in $\mathcal{P}$, such that for every $t \geq 2$,

$$\sup_{P \in \mathcal{P}} P^n\big[H(P, \widehat{P}) > 4t\varepsilon_n\big] \lesssim e^{-t^2}, \tag{11.31}$$

and consequently

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\big[H^2(P, \widehat{P})\big] \lesssim \varepsilon_n^2. \tag{11.32}$$

Brief justification of the local theorem. The estimator is the same tournament estimator as in the global theorem. The difference is only in the counting argument. Instead of union-bounding over all $N_H(\mathcal{P}, \varepsilon)$ candidates, partition the candidates around the oracle center into Hellinger shells

$$G_k := \{j : 2^k\delta \leq H(P_1, P_j) < 2^{k+1}\delta\}, \qquad \delta = 4\varepsilon.$$

Repeatedly covering a local ball by balls of half its radius gives

$$|G_k| \leq N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)^{k+c_0}$$

for a universal constant c_0. On the other hand, the robust testing error for a center in shell k is at most

$$\exp(-cn4^k\varepsilon^2).$$

Therefore, if $2^\ell \leq t < 2^{\ell+1}$,

$$\begin{aligned} P^n\big[H(P_1, \widehat{P}) > t\delta\big] &\leq \sum_{k \geq \ell} |G_k| \exp(-cn4^k\varepsilon^2) \\ &\leq \sum_{k \geq \ell} \exp\Big((k + c_0)\log N_{\mathrm{loc}} - cn4^k\varepsilon^2\Big) \\ &\lesssim e^{-c't^2}, \end{aligned}$$

provided (11.30) holds. The same triangle-inequality and tail-integration steps then yield (11.31)–(11.32).

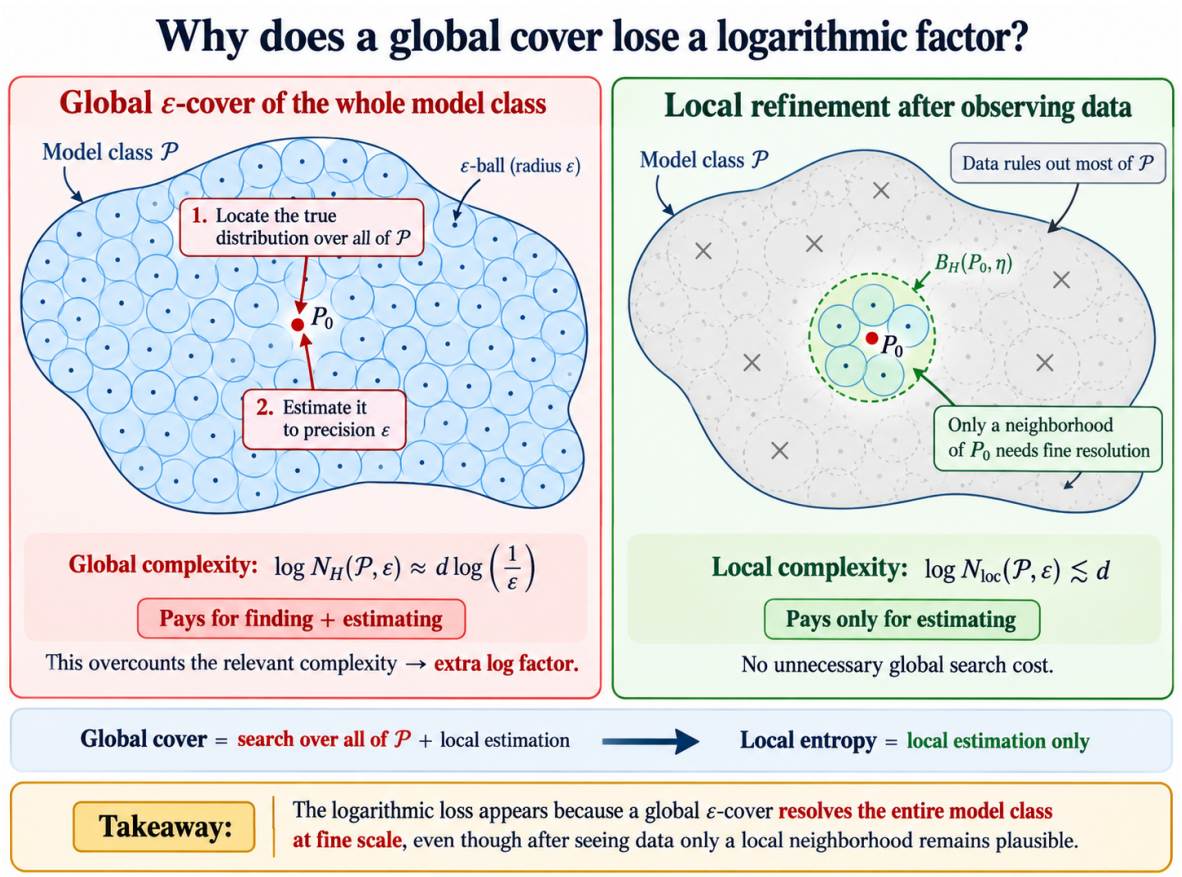


Figure 20: A global cover pays for finding and estimating, whereas local entropy pays only for resolving the neighborhood that remains plausible after observing the data.

11.6 Why local entropy is of order d in regular models

Assume that $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ and that for some norm $\|\cdot\|$ and constants $0 < c_1 \le c_2 < \infty$,

$$c_1\|\theta - \theta'\| \le H(P_\theta, P_{\theta'}) \le c_2\|\theta - \theta'\|. \tag{11.33}$$

We now prove

$$\log N_{\text{loc}}(\mathcal{P},\varepsilon) \lesssim d.$$

Fix P_{θ_0} and $\eta \ge \varepsilon$, and define

$$A := \{\theta \in \Theta : H(P_\theta, P_{\theta_0}) \le \eta\}.$$

Step 1: Convert a Hellinger ball into a parameter-space ball. If $\theta \in A$, then

$$c_1\|\theta - \theta_0\| \le H(P_\theta, P_{\theta_0}) \le \eta,$$

so

$$A \subseteq B_{\|\cdot\|}\left(\theta_0, \frac{\eta}{c_1}\right). \tag{11.34}$$

Step 2: Cover the parameter-space ball. Choose parameter-space covering radius

$$r := \frac{\eta}{2c_2}.$$

If $\|\theta - \theta_j\| \le r$, then

$$H(P_\theta, P_{\theta_j}) \le c_2\|\theta - \theta_j\| \le \frac{\eta}{2}.$$

Thus

$$N_H\big(B_H(P_{\theta_0}, \eta) \cap \mathcal{P}, \eta/2\big) \le N_{\|\cdot\|}\left(B_{\|\cdot\|}\left(\theta_0, \frac{\eta}{c_1}\right), \frac{\eta}{2c_2}\right). \tag{11.35}$$

Step 3: Apply the volume bound. For a ball of radius R in a d-dimensional normed space,

$$N_{\|\cdot\|}(B(R), r) \le \left(1 + \frac{2R}{r}\right)^d. \tag{11.36}$$

Substituting $R = \eta/c_1$ and $r = \eta/(2c_2)$ gives

$$\begin{aligned} N_H\big(B_H(P_{\theta_0}, \eta) \cap \mathcal{P}, \eta/2\big) &\le \left(1 + \frac{2(\eta/c_1)}{\eta/(2c_2)}\right)^d \\ &= \left(1 + \frac{4c_2}{c_1}\right)^d. \end{aligned}$$

The bound is uniform in θ_0 and η. Hence

$$N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) \le \left(1 + \frac{4c_2}{c_1}\right)^d, \qquad \log N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) \lesssim d. \tag{11.37}$$

Applying the local theorem gives the parametric Hellinger-risk bound

$$\inf_{\widehat{P}} \sup_{\theta \in \Theta} \mathbb{E}_\theta\big[H^2(P_\theta, \widehat{P})\big] \lesssim \frac{d}{n}. \tag{11.38}$$

11.7 Density estimation over Hölder classes

Let $\mathcal{F} = \mathcal{F}_\beta$ be a class of β-smooth densities on $[0,1]^d$. Observe

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f, \qquad f \in \mathcal{F},$$

and let $\widehat{f} = \widehat{f}(\cdot; X_1, \dots, X_n)$ be a density estimator. The squared L_2-loss is

$$\|f - \widehat{f}\|_2^2 := \int_{[0,1]^d} (f(x) - \widehat{f}(x))^2 \, \mathrm{d}x.$$

Define the minimax quadratic risk by

$$R^*_{L_2}(n; \mathcal{F}) := \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f\big[\|f - \widehat{f}\|_2^2\big]. \tag{11.39}$$

The target theorem is

$$R^*_{L_2}(n; \mathcal{F}) \asymp n^{-2\beta/(2\beta+d)}. \tag{11.40}$$

11.7.1 Reduction to densities bounded away from zero

Let

$$\mathcal{F}' := \{f \in \mathcal{F} : f(x) \ge 1/2 \text{ for every } x \in [0,1]^d\}. \tag{11.41}$$

Lemma 11.5. *The two minimax risks satisfy*

$$R^*_{L_2}(n; \mathcal{F}') \le R^*_{L_2}(n; \mathcal{F}) \le 4R^*_{L_2}(n; \mathcal{F}'). \tag{11.42}$$

Proof. The left inequality follows from $\mathcal{F}' \subseteq \mathcal{F}$.

For the right inequality, fix $f \in \mathcal{F}$ and observe $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f$. Independently sample

$$U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1]^d),$$

and define

$$Z_i := \begin{cases} U_i, & \text{with probability } 1/2, \\ X_i, & \text{with probability } 1/2, \end{cases}$$

independently across i. Then

$$Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} g, \qquad g := \frac{1+f}{2}.$$

Since $g \geq 1/2$, we have $g \in \mathcal{F}'$. Let $\widehat{g}$ be an estimator attaining, or arbitrarily closely approximating, the minimax risk on $\mathcal{F}'$, and define

$$\widehat{f} := 2\widehat{g} - 1.$$

Then

$$\|f - \widehat{f}\|_2^2 = 4\|g - \widehat{g}\|_2^2.$$

Taking the supremum over $f \in \mathcal{F}$ proves the right inequality in (11.42). □

Thus it suffices to prove the minimax rate on $\mathcal{F}'$.

11.7.2 Comparison of L_2, Hellinger, KL, and χ^2

Lemma 11.6. *Suppose P, Q have densities $f = \mathrm{d}P/\mathrm{d}\mu$ and $g = \mathrm{d}Q/\mathrm{d}\mu$, and assume*

$$0 < c_1 \leq f, g \leq c_2 < \infty.$$

Then

$$\frac{\|f-g\|_2^2}{4c_2} \leq H^2(P,Q) \leq D(P\|Q) \leq \chi^2(P\|Q) \leq \frac{\|f-g\|_2^2}{c_1}. \tag{11.43}$$

Indeed,

$$H^2(P,Q) = \int \frac{(f-g)^2}{(\sqrt{f}+\sqrt{g})^2}\,\mathrm{d}\mu, \qquad \chi^2(P\|Q) = \int \frac{(f-g)^2}{g}\,\mathrm{d}\mu,$$

and the bounds follow from $c_1 \leq f, g \leq c_2$, together with (11.7).

Consequently,

$$R^*_{L_2}(n; \mathcal{F}') \asymp R^*_{\mathrm{KL}}(n; \mathcal{F}'). \tag{11.44}$$

11.7.3 Upper bound via Yang–Barron

The metric entropy of the Hölder class satisfies

$$\log N(\mathcal{F}', \|\cdot\|_2, \varepsilon) \asymp \varepsilon^{-d/\beta}. \tag{11.45}$$

Using (11.43), an L_2-cover induces a KL cover at the same scale up to constants. Therefore, by Yang–Barron,

$$\begin{aligned} R^*_{L_2}(n; \mathcal{F}') &\asymp R^*_{\mathrm{KL}}(n; \mathcal{F}') && (11.46)\\ &\lesssim \inf_{\varepsilon > 0} \left\{ \varepsilon^2 + \frac{1}{n} \log N_{\mathrm{KL}}(\mathcal{F}', \varepsilon) \right\} && (11.47)\\ &\lesssim \inf_{\varepsilon > 0} \left\{ \varepsilon^2 + \frac{1}{n} \log N(\mathcal{F}', \|\cdot\|_2, \varepsilon) \right\} && (11.48)\\ &\asymp \inf_{\varepsilon > 0} \left\{ \varepsilon^2 + \frac{1}{n} \varepsilon^{-d/\beta} \right\} && (11.49)\\ &\asymp n^{-2\beta/(2\beta+d)}. && (11.50) \end{aligned}$$

11.7.4 Lower bound via Fano's inequality

Start from Fano's inequality. If V is uniform on $[M]$, $X^n \mid \{V = i\} \sim P_i^{\otimes n}$, and $\widehat{V} = \widehat{V}(X^n)$, then

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V; X^n) + \log 2}{\log M}. \tag{11.51}$$

By the metric-entropy lower bound,

$$\log N(\mathcal{F}, \|\cdot\|_2, \varepsilon) \asymp \log M(\mathcal{F}, \|\cdot\|_2, \varepsilon) \asymp \varepsilon^{-d/\beta}. \tag{11.52}$$

Fix $\varepsilon > 0$, and choose a 2ε-packing

$$f_1, \dots, f_M \in \mathcal{F}$$

with

$$M \geq \exp\bigl(C\varepsilon^{-d/\beta}\bigr). \tag{11.53}$$

Define

$$g_i := \frac{f_i + 1}{2}. \tag{11.54}$$

Then $g_i \in \mathcal{F}'$, and $g_1, \dots, g_M$ form an ε-packing in $\mathcal{F}'$.

Apply the estimation version of Fano's inequality under squared L_2-loss. Equivalently, given an arbitrary estimator $\widehat{f}$, round it to the nearest packing point and use (11.51). This gives

$$R^*_{L_2}(n; \mathcal{F}') \geq \left(\frac{\varepsilon}{2}\right)^2 \left(1 - \frac{C'_n + \log 2}{\log M}\right), \tag{11.55}$$

where C'_n is the capacity of the experiment

$$f \in \mathcal{F}' \longmapsto X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f.$$

In particular,

$$R^*_{L_2}(n; \mathcal{F}') \gtrsim \varepsilon^2 \left(1 - \frac{C'_n}{\log M}\right),$$

up to universal constants.

Bound the restricted capacity using Yang–Barron:

$$\begin{aligned} C'_n &\leq \inf_{\delta > 0} \bigl\{n\delta^2 + \log N_{\mathrm{KL}}(\mathcal{F}', \delta)\bigr\} && (11.56)\\ &\lesssim \inf_{\delta > 0} \bigl\{n\delta^2 + \log N(\mathcal{F}', \|\cdot\|_2, \delta)\bigr\} && (11.57)\\ &\lesssim \inf_{\delta > 0} \left\{n\delta^2 + \delta^{-d/\beta}\right\} && (11.58)\\ &\asymp n^{d/(2\beta+d)}. && (11.59) \end{aligned}$$

Choose

$$\varepsilon = c\, n^{-\beta/(2\beta+d)} \tag{11.60}$$

with $c > 0$ sufficiently small. Then (11.53) gives

$$\log M \gtrsim c^{-d/\beta} n^{d/(2\beta+d)},$$

so, for sufficiently small c and sufficiently large n,

$$C'_n + \log 2 \leq \frac{1}{2} \log M.$$

Substitution into (11.55) yields

$$R^*_{L_2}(n;\mathcal{F}') \gtrsim n^{-2\beta/(2\beta+d)} \left(1 - \frac{\frac{1}{2}\log M}{\log M}\right) \asymp n^{-2\beta/(2\beta+d)}.$$

Combining this lower bound with (11.50) and the reduction (11.42) proves

$$R^*_{L_2}(n;\mathcal{F}) \asymp n^{-2\beta/(2\beta+d)}.$$

11.7.5 Upper bound via the local Le Cam–Birgé theorem

The handwritten lecture also obtained the upper bound directly from the local Le Cam–Birgé theorem. This route connects the main theorem of the lecture to the density-estimation application without passing through predictive KL risk.

For the lower-bounded class $\mathcal{F}'$, Lemma 11.6 shows that squared Hellinger and squared L_2 distances are equivalent up to constants. In particular, an Hellinger ball of radius η is contained in an L_2-ball of radius of order η, and conversely. Using the global Hölder entropy bound (11.45), for every $\eta \geq \varepsilon$,

$$\begin{aligned} \log N_H\big(B_H(f_0,\eta) \cap \mathcal{F}', \eta/2\big) &\leq \log N_H(\mathcal{F}', \eta/2) \\ &\lesssim \eta^{-d/\beta} \\ &\leq \varepsilon^{-d/\beta}. \end{aligned} \tag{11.61}$$

Taking the supremum over $f_0 \in \mathcal{F}'$ and $\eta \geq \varepsilon$ gives

$$\log N_{\mathrm{loc}}(\mathcal{F}', \varepsilon) \lesssim \varepsilon^{-d/\beta}. \tag{11.62}$$

Choose ε_n so that

$$n\varepsilon_n^2 \asymp \varepsilon_n^{-d/\beta}.$$

Then

$$\varepsilon_n^2 \asymp n^{-2\beta/(2\beta+d)}. \tag{11.63}$$

The local Le Cam–Birgé theorem and the leftmost inequality in (11.43) therefore yield

$$\begin{aligned} R^*_{L_2}(n;\mathcal{F}') &\lesssim R^*_{H^2}(n;\mathcal{F}') \\ &\lesssim \varepsilon_n^2 \\ &\asymp n^{-2\beta/(2\beta+d)}. \end{aligned} \tag{11.64}$$

Together with the reduction (11.42), this gives the same upper bound on the full Hölder class $\mathcal{F}$.

Remark 11.7. For this nonparametric class, the local entropy bound $\log N_{\mathrm{loc}}(\mathcal{F}', \varepsilon) \lesssim \varepsilon^{-d/\beta}$ has the same order as the global entropy. The local refinement is nevertheless important conceptually: it removes the logarithmic loss for regular finite-dimensional models and provides a direct Hellinger-risk construction here.

Takeaway

For Hölder density estimation, the upper and lower bounds use the same entropy scale:

$$\log N(\varepsilon) \asymp \log M(\varepsilon) \asymp \varepsilon^{-d/\beta}.$$

The upper bound follows either from Yang–Barron or, as in the handwritten lecture, from the local Le Cam–Birgé theorem; Fano gives the lower bound. All three arguments identify the critical resolution through

$$n\varepsilon^2 \asymp \varepsilon^{-d/\beta},$$

which yields the exact minimax rate $n^{-2\beta/(2\beta+d)}$.