Review of Information Theory

Lecture 1: Entropy, Prefix-Free Codes and Kraft's inequality

Lemma 1.1. Jensen's Inequality
Let $S \subseteq \mathbb{R}^n$ be a convex set and let X be a random variable taking values inside S. Then, for a convex function $f : S \to \mathbb{R}$, we have that:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Equivalently, for a concave function $f : S \to \mathbb{R}$, we have that:

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]).$$

Proposition 1.2. Lower and Upper Bound of Entropy
Let X be a random variable supported on a finite set $\mathcal{X}$ as above. Then:

$$0 \leq H(X) \leq \log(|\mathcal{X}|).$$

Proposition 1.3. Kraft's Inequality
Let $|\mathcal{X}| = n$. There exists a prefix-free code for $\mathcal{X}$ over {0,1} with codeword lengths $\ell_1, \ldots, \ell_n$ if and only if:

$$\sum_{i=1}^{n} \frac{1}{2^{\ell_i}} \leq 1.$$

For codes over a larger alphabet Σ, we replace 2^{ℓ_i} above by $|\Sigma|^{\ell_i}$.

Lecture 2: Conditional and Joint Entropy, Subadditivity of Entropy, Source Coding Theorem

Claim 2.1. Minimal Bits to Communicate a Random Variable
Let X be a random variable taking values in $\mathcal{X}$ and let $C : \mathcal{X} \to \{0,1\}$ be a prefix-free code.

Then the expected number of bits used by $\mathbf{C}$ to communicate the value of $\mathbf{X}$ is at least $H(X)$.

Definition 2.2. The Shannon Code
We now construct a (prefix-free) code for conveying the value of X, using at most $H(X)+1$ bits on average (over the distribution of X). For an element $x \in \mathcal{X}$ which occurs with probability $p(x)$, we will use a codeword of length $\lceil \log(1/p(x)) \rceil$. By Kraft's inequality, there exists a prefix-free code with these codeword lengths, since

$$\sum_{x \in \mathcal{X}} \frac{1}{2^{|C(x)|}} = \sum_{x \in \mathcal{X}} \frac{1}{2^{\lceil \log(1/p(x)) \rceil}} \leq \sum_{x \in \mathcal{X}} \frac{1}{2^{\log(1/p(x))}} = \sum_{x \in \mathcal{X}} p(x) = 1$$

Also, the expected number of bits used is

$$\sum_{x \in \mathcal{X}} p(x) \cdot \lceil \log(1/p(x)) \rceil \leq \sum_{x \in \mathcal{X}} p(x) \cdot (\log(1/p(x)) + 1) = H(X) + 1$$

This code is known as the Shannon code.

Proposition 2.3. Information Never Hurts

$$H(Y) \geq H(Y \mid X)$$

Proposition 2.4. Fundamental Source Coding Theorem
For all $\varepsilon > 0$ there exists a n_0 such that for all $n \geq n_0$ and given n copies of $X, X_1, \ldots, X_n$ sampled i.i.d., it is possible to communicate $(X_1, \ldots, x_n)$ using at most $H(X) + \varepsilon$ bits per copy on average.

Proposition 2.5. Cauchy-Schwarz Inequality

$$\left(\sum_{i=1}^{n} a_i \cdot b_i \right) \leq \left(\sum_{i=1}^{n} a_i^2 \right) \cdot \left(\sum_{i=1}^{n} b_i^2 \right)$$

Lecture 3: Shearer's Lemma and Mutual Information

Lemma 3.1. Shearer's Lemma
Let $\{X_1, \ldots, X_n\}$ be a set of random variables. For any $S \subset [n]$, let us denote $X_S = \{X_i : i \in S\}$. Let $\mathcal{F} \subseteq 2^{[n]}$ be a collection of subsets of $[n]$ with the property that for all $i \in [n]$, we have that $|\{S \in \mathcal{F} \mid S \ni i\}| \geq t$. Then

$$t \cdot H(X_1, \ldots, X_n) \leq \sum_{S \in \mathcal{F}} H(X_S)$$

Lemma 3.2. Shearer's Lemma: Distribution Version
Let $\{X_1, \ldots, X_n\}$ be a set of random variables. For any $S \subset [n]$, let us denote $X_S = \{X_i : i \in S\}$. Let D be an arbitrary distribution on $2^{[n]}$ (set of all subsets of $[n]$) and let μ be such that $\forall i \in [n] \mathbb{P}_{S \sim D}[i \in S] \geq \mu$. Then:

$$\mu \cdot H(X_1, \ldots, X_n) \leq \mathop{\mathbb{E}}_{S \sim D}[H(X_S)]$$

Proposition 3.3. Bounding the Volume of a Body
Let $S \subseteq \mathbb{R}^d$ be a finite set of points in dimensions, and let $S_1, \ldots, S_d$ denote the set of projections orthogonal to each of the d coordinate axes. Then we have

$$|S| \leq \left(\prod_{i=1}^{d} |S_i| \right)^{1/(d-1)}$$

Proposition 3.4. Loomis-Whitney Inequality
Let $B \subseteq \mathbb{R}^d$ be a measurable body and let $B_1, \ldots, B_d$ denote its projections orthogonal to each of the coordinate axes. Then, we have

$$\mathrm{Vol}_d(B) \leq \left(\prod_{i=1}^{d} \mathrm{Vol}_{d-1}(B_i) \right)^{1/(d-1)}$$

Lemma 3.5. Chain Rule of Mutual Information

$$I((X_1, \ldots, X_m); Y) = \sum_{i=1}^{m} I(X_i; Y \mid X_1, \ldots, X_{i-1})$$

Lemma 3.6. Data Processing Inequality
Let $X \to Y \to Z$ be a Markov chain. Then:

$$I(X; Y) \geq I(X; Z)$$

Definition 3.7. Sufficient Statistics
For random variables X and Y, a function $g(Y)$ is called a sufficient statistic $(of\ Y)$ for X if $I(X;Y) = I(X;g(Y))$ i.e., $g(Y)$ contains all the relevant information about X.

Lecture 4: Fana's inequality, Graph Entropy, KL Divergence, TV Distance and Pinsker's Inequality

Lemma 4.1. Fano's Inequality
Let $X \to Y \to \widehat{X}$ be a Markov chain, and let $p_e = \mathbb{P}[\hat{X} \neq X]$. Let $\mathrm{H}_2\left(\mathrm{p}_e\right)$ denote the binary entropy function computed at p_e. Then,

$$H_2\left(p_e\right) + p_e \cdot \log(|\mathcal{X}| - 1) \geq H(X \mid \widehat{X}) \geq H(X \mid Y)$$

Definition 4.2. Graph Entropy
Given a graph $G = (\mathcal{V}, \mathcal{E})$, we define the graph entropy $H(G)$ as

$$\min_{X,Y} I(X;Y)$$

s.t. X is uniformly distributed over $\mathcal{V}$; and Y is an independent set in G containing X.

Proposition 4.3. Subadditivity of Graph Entropy
Let $G_1 = (\mathcal{V}, \mathcal{E}_1)$ and $G_2 = (\mathcal{V}, \mathcal{E}_2)$ be two graphs, and let $G = \left(\mathcal{V}, \mathcal{E}_1 \cup \mathcal{E}_2\right)$, which we denote by $G = G_1 \cup G_2$. Then,

$$H(G) = H\left(G_1 \cup G_2\right) \leq H\left(G_1\right) + H\left(G_2\right)$$

Definition 4.4. KL Divergence

$$D(P\|Q) := \sum_{x \in U} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Lemma 4.5. Non-negativity of KL Divergence
Let P and Q be distributions on a finite universe $\mathcal{X}$. Then $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

Lemma 4.6. Chain Rule for KL Divergence
Let $P(X,Y)$ and $Q(X,Y)$ be two distributions for a pair of variables X and Y. Then,

$$D(P(X,Y)\|Q(X,Y)) = D(P(X)\|Q(X)) + \mathop{\mathbb{E}}_{x\sim P}[D(P(Y \mid X=x)\|Q(Y \mid X=x))]$$
$$= D(P(X)\|Q(X)) + D(P(Y \mid X)\|Q(Y \mid X))$$

Definition 4.7. Total Variation
Let P and Q be two distributions on a finite universe $\mathcal{X}$. Then the *total-variation distance* or *statistical distance* between P and Q is defined as

$$\delta_{TV}(P,Q) = \frac{1}{2} \cdot \|P-Q\|_1 = \frac{1}{2} \cdot \sum_{x\in\mathcal{X}} |p(x)-q(x)|$$

The quantity $\|P-Q\|_1$ is referred to as the ℓ_1*-distance* between P and Q.

Lemma 4.8. The Classifier's Behavior Over Two Distributions
Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then:

$$|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]| \le \frac{B}{2} \cdot \|P-Q\|_1 = B \cdot \delta_{TV}(P,Q)$$

Lemma 4.9. Pinsker's Inequality / TV as Lower Bound for KL Divergence
Let P and Q be two distributions defined on a universe $\mathcal{X}$. Then:

$$D(P\|Q) \ge \frac{1}{2\ln 2} \cdot \|P-Q\|_1^2.$$

Proposition 4.10. Pinsker's Inequality for $\mathcal{X} = \{0,1\}$
Consider when $\mathcal{X}$ = {0,1} and P, Q are distributions as below

$$P = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \quad \text{and} \quad Q = \begin{cases} 1 & \text{w.p. } q \\ 0 & \text{w.p. } 1-q \end{cases}$$

Then,

$$D(P\|Q) = p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right) \quad ,$$

$$\|P-Q\|_1 = 2 \mid p-q \mid,$$

$$p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right) \ge \frac{2}{\ln 2} \cdot (p-q)^2.$$

Proposition 4.10. TV and KL Divergence on $\{0,1\}$
Let P and Q be distributions on a finite set $\mathcal{X}$. Then, there exist distributions P', Q' on {0,1} such that

$$\|P' - Q'\|_1 = \|P - Q\|_1 \text{ and } D(P\|Q) \geq D(P'\|Q')$$

Lecture 5: Convexity of KL, Lower Bound for Distinguishing Coins and Bandit Problems

Proposition 5.1. Log-Sum Inequality
For $a_1, a_2, b_1, b_2 \geq 0$

$$(a_1 + a_2) \cdot \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \leq a_1 \cdot \log\left(\frac{a_1}{b_1}\right) + a_2 \cdot \log\left(\frac{a_2}{b_2}\right).$$

Proposition 5.2. Convexity of KL Divergence
Let P_1, P_2, Q_1, Q_2 be distributions on a finite universe $\mathcal{X}$, and let $\alpha \in [0,1]$. Then,

$$D(\alpha \cdot P_1 + (1-\alpha) \cdot P_2 \| \alpha \cdot Q_1 + (1-\alpha) \cdot Q_2) \leq \alpha \cdot D(P_1\|Q_1) + (1-\alpha) \cdot D(P_2\|Q_2).$$

Proposition 5.3. Upper Bound for TV

$$\min\{D(P\|Q), D(Q\|P)\} \geq \frac{1}{2\ln 2} \cdot \|P - Q\|_1^2$$

Proposition 5.4. Lower Bound for Expected Regret for Two-Armed Bandits

$$\mathbb{E}[\$regret\$] \geq \left(\frac{1}{2} + \varepsilon\right) \cdot n - \mathbb{E}\left[\sum_{t=1}^{n} X_{C_t,t}\right]$$

Proposition 5.5. Lower Bound for Expected Regret for Two-Armed Bandits

$$\mathbb{E}[\text{ regret }] \geq \varepsilon \cdot \mathbb{E}\left[|\{t \mid C_t \neq H\}|\right]$$

Proposition 5.6. Lower Bound for Expected Regret for Two-Armed Bandits (Cont'd)

$$\mathbb{E}\left[|\{t \mid C_t \neq H\}|\right] \geq \frac{n}{2} \cdot \left(1 - \frac{1}{2} \cdot \left\|P_\ell\left(Z_1, \ldots, Z_n\right) - P_r\left(Z_1, \ldots, Z_n\right)\right\|_1\right)$$

Proposition 5.7. Lower Bound for Expected Regret for Two-Armed Bandits (Cont'd)

$$\left\|P_\ell\left(Z_1, \ldots, Z_n\right) - P_r\left(Z_1, \ldots, Z_n\right)\right\|_1^2 \leq c \cdot \varepsilon^2 \cdot n$$

Proposition 5.8. Lower Bound for Expected Regret for Two-Armed Bandits (Cont'd)

$$\begin{aligned}
\mathbb{E}[\text{ regret }] &\geq \frac{\varepsilon n}{2} \cdot \left(1 - \frac{1}{2} \cdot \left\|P_\ell\left(Z_1, \ldots, Z_n\right) - P_r\left(Z_1, \ldots, Z_n\right)\right\|_1\right) \\
&\geq \frac{\varepsilon n}{2} \cdot \left(1 - \sqrt{c \cdot \varepsilon^2 \cdot n}\right) \\
&= \varepsilon n \cdot \left(\frac{1}{2} - c_0 \cdot \sqrt{\varepsilon^2 \cdot n}\right)
\end{aligned}$$

Lecture 6: Differential Entropy and Gaussian Computation

Definition 6.1. Differential Entropy

Let X be a random variable taking values in $\mathbb{R}^n$, with density p. Then the differential entropy of X is defined to be the following integral (if it exists):

$$h(X) := \int p(x) \cdot \log\left(\frac{1}{p(x)}\right) dx$$

Definition 6.2. Differential KL Divergence

$$D(P\|Q) := \int p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx$$

Proposition 6.3. Change of Variables

Exercise 2.1 (Change of variables). Let X be a random variable over $\mathbb{R}^n$ with associated density function p_X. Using the Jacobian for change of variables in integrals, check that

- If $c \in \mathbb{R}^n$ is a fixed vector, then the density function for $Y = X + c$ is given by $p_Y(y) = p_X(y - c)$
- If $A \in \mathbb{R}^{n \times n}$ is a nonsingular matrix, then the density function for $Y = AX$ is given by $p_Y(y) = \frac{p_X(A^{-1}y)}{|A|}$, where $|A|$ denotes $|\det(A)|$

Proposition 6.4. Change of Entropy
Let X be a continuous random variable over $\mathbb{R}^n$. Let $c \in \mathbb{R}^n$ and let $A \in \mathbb{R}^{n \times n}$ be a non-singular matrix. Then

- $h(X + c) = h(X)$
- $h(AX) = h(X) + \log |A|$

Fact 6.5. Entropy of Gaussians
Using the fact that $Y \sim N(\mu, \Sigma)$ can be written as $Y = \Sigma^{1/2} X + \mu$, where $X = N(0, I_n)$ (check this!) we get that

$$h(Y) = h(X) + \log\left(\left|\Sigma^{1/2}\right|\right) = \frac{n}{2} \cdot \log(2\pi \cdot e) + \frac{1}{2} \cdot \log |\Sigma|$$

Lecture 7: Type Method, Chernoff Bound and Sanov's Theorem

Definition 7.1. Type
The type $P_{\bar{x}}$ of $\overline{\mathbf{x}}$, also called the empirical distribution of $\overline{\mathbf{x}}$, is a distribution $\hat{P}$ on $\mathcal{X}$, defined as

$$\hat{P}(a) := \frac{|\{i : x_i = a\}|}{n} \quad \forall a \in \mathcal{X}$$

We use $\mathcal{T}_n$ to denote the set of all types coming from sequences of length n. We also use $\mathcal{C}_P$ to denote the set of all sequences with the type P. $\mathcal{C}_P$ is called the type class of P.

$$\mathcal{C}_P := \{\overline{\mathbf{x}} \in \mathcal{X}^n \mid P_{\overline{\mathbf{x}}} = P\}$$

Fact 7.2. Size of Type Class

$$|\mathcal{T}_n| = \binom{n + r - 1}{r - 1} \leq (n + 1)^r$$

Proposition 7.3. Size of Type Class
For any type $P \in \mathcal{T}_n$, we have

$$\frac{2^{n \cdot H(P)}}{(n+1)^r} \leq |\mathcal{C}_P| \leq 2^{n \cdot H(P)}$$

Proposition 7.4. Product Distribution
Let Q be any distribution on U and let Q^n the product distribution on $\mathcal{X}^n$. Let $\overline{\mathbf{x}}, \overline{\mathbf{y}} \in \mathcal{X}^n$ be such that $P_{\overline{\mathbf{x}}} = P_{\overline{\mathbf{y}}}$. Then, $Q^n(\overline{\mathbf{x}}) = Q^n(\overline{\mathbf{y}})$

Theorem 7.5. Bound for the Probability of Type
For any product distribution Q^n and type P on $\mathcal{X}^n$, we have

$$\frac{2^{-n \cdot D(P\|Q)}}{(n+1)^r} \leq \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim Q^n}\left[P_{\overline{\mathbf{x}}} = P\right] \leq 2^{-n \cdot D(P\|Q)}$$

Theorem 7.6. Chernoff Bound
For $\overline{\mathbf{X}} = (X_1, \ldots, X_n) \sim_{Q^n} U^n$ with Q the uniform distribution on $\mathcal{X} = \{0, 1\}$, we have

$$\mathop{\mathbb{P}}_{Q^n}\left[\sum_{i=1}^{n} X_i \geq \frac{n}{2} + \varepsilon n\right] \leq (n+1) \cdot 2^{-c \cdot n \cdot \varepsilon^2}$$

Theorem 7.7. Sanov's Theorem
Let Π be a set of distributions on $\mathcal{X}$, and $|\mathcal{X}| = r$. Then

$$\mathop{\mathbb{P}}_{Q^n}\left[P_{\overline{\mathbf{x}}} \in \Pi\right] \leq (n+1)^r \cdot 2^{-n \cdot \delta}$$

where $\delta = \inf_{P \in \Pi} D(P\|Q)$. Moreover, if Π is the closure of an open set and

$$P^* := \mathop{\arg\min}_{P \in \Pi} D(P\|Q)$$

then

$$\frac{1}{n} \cdot \log\left(\mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim Q^n}\left[P_{\overline{\mathbf{x}}} \in \Pi\right]\right) \to -D\left(P^* \| Q\right)$$

Lecture 8: Binary and Multiple Hypothesis Testing

Definition 8.1. Definition of Two Errors

$$\alpha(T) := \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_0^n}[T(\overline{\mathbf{x}}) = 1] \quad \text{(False Positive)}$$
$$\beta(T) := \mathop{\mathbb{P}}_{\frac{P}{\mathbf{x}} \sim p_n}[T(\overline{\mathbf{x}}) = 0] \quad \text{(False Negative)}$$

Proposition 8.2. Minimum of the Sum of Two Errors

$$\min_T \{\alpha(T) + \beta(T)\} = 1 - \delta_{TV}\left(P_0^n, P_1^n\right) = 1 - \frac{1}{2} \cdot \left\| P_0^n - P_1^n \right\|_1$$

Lemma 8.3. Neyman-Pearson Lemma

Let T be a test of the form

$$T(\overline{\mathbf{x}}) = \begin{cases} 1 \text{ if } P_1^n(\overline{\mathbf{x}}) / P_0^n(\overline{\mathbf{x}}) \geq \Delta \\ 0 \text{ if } P_0^n(\overline{\mathbf{x}}) / P_1^n(\overline{\mathbf{x}}) < \Delta \end{cases}$$

for some constant $\Delta > 0$. Let T' be any other test. Then,

$$\alpha\left(T'\right) \geq \alpha(T) \quad \text{or} \quad \beta\left(T'\right) \geq \beta(T)$$

Remark 8.4. Another Perspective on Test

The test $T(\overline{\mathbf{x}})$ considered above can be written in the following form

$$\frac{P_1^n(\overline{\mathbf{x}})}{P_0^n(\overline{\mathbf{x}})} \geq \Delta \quad \Leftrightarrow \quad D\left(P_{\overline{\mathbf{x}}} \| P_0\right) - D\left(P_{\overline{\mathbf{x}}} \| P_1\right) \geq \frac{1}{n} \cdot \log \Delta$$

We define the following sets of probability distributions.

$$\Pi := \left\{ P \mid D\left(P \| P_0\right) - D\left(P \| P_1\right) \geq \frac{1}{n} \cdot \log \Delta \right\}$$
$$\Pi^c := \left\{ P \mid D\left(P \| P_0\right) - D\left(P \| P_1\right) < \frac{1}{n} \cdot \log \Delta \right\}$$

Remark 8.5. Estimation of Two Errors

$$\alpha(T) = \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_0^n}\left[P_{\overline{\mathbf{x}}} \in \Pi\right] \approx 2^{-n \cdot D(P_0^* \| P_0)}$$

$$\beta(T) = \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_1^n}\left[P_{\overline{\mathbf{x}}} \in \Pi^c\right] \approx 2^{-n \cdot D(P_1^* \| P_1)}$$

$$P_0^*(x) = P_1^*(x) = P^* = \frac{P_0^\lambda(x) \cdot P_1^{1-\lambda}(x)}{\sum_{y \in \mathcal{X}} P_0^\lambda(y) \cdot P_1^{1-\lambda}(y)}$$

Lemma 8.6. Fano's Inequality
Let $Z \to Y \to \widehat{Z}$ be a Markov chain with Z taking values in a finite set $\mathcal{Z}$, and let $p_e = \mathbb{P}[\hat{Z} \neq Z]$. Let $H_2\left(p_e\right)$ denote the binary entropy function computed at p_e. Then,

$$H_2\left(p_e\right) + p_e \cdot \log(|\mathcal{Z}| - 1) \geq H(Z \mid \widehat{Z}) \geq H(Z \mid Y)$$

Proposition 8.7. Bound on the Binary Error
Let $V \to \overline{\mathbf{X}} \to \widehat{V}$ be the Markov chain as above. Then,

$$p_e = \mathbb{P}[V \neq \widehat{V}] \geq 1 - \frac{n \cdot \mathbb{E}_{v_1, v_2 \in \mathcal{V}}\left[D\left(P_{v_1} \| P_{v_2}\right)\right] + 1}{\log |\mathcal{V}|}$$

Lecture 9: Minimax's Rate and Le Cam's Method

Definition 9.1. Minimax Risk

$$\mathcal{M}_n(\Pi, \ell) := \inf_{\hat{\theta}} \sup_{P \in \Pi} \mathop{\mathbb{E}}_{\mathbf{x} \sim P^n}[\ell(\hat{\theta}(\overline{\mathbf{x}}), \theta(P))]$$

Lemma 9.2. Bound for the Minimax Risk
Let $\{P_v\}_{v \in \mathcal{V}} \subseteq \Pi$ be a finite set of distributions such that $\forall v_1, v_2 \in \mathcal{V}$ with $v_1 \neq v_2$, $\rho\left(\theta\left(P_{v_1}, P_{v_2}\right)\right) \geq 2\delta$. Let ℓ be as above. Then,

$$\mathcal{M}(\Pi, \ell) \geq \Phi(\delta) \cdot \inf_T\{\mathbb{P}[T(\overline{\mathbf{x}}) \neq V]\}$$

Note that the setting in the RHS above is exactly as considered in hypothesis testing. We think of V as uniformly distributed over the set $\mathcal{V}$ and $\overline{\mathbf{x}}$ as drawn from P_v^n

Exercise 9.3. Bound for Expectation
Let $P : \{0, 1\} \to [0, 1]$ be any distribution with $\mathbb{E}_{x \sim P}[x] = p(1) = \mu$. Show that

$$\mathop{\mathbb{E}}_{(x_1,\ldots,x_n)\sim P^n}\left[\left|\frac{1}{n}\cdot\sum_{i\in[n]}x_i-\mu\right|^2\right]=O\left(\frac{1}{n}\right)$$

Fact 9.4. Lower Bound for Classifier Error

We now consider a high-dimensional problem, where we can prove lower bounds using bounds for testing multiple hypotheses. Recall that for a random variable V uniformly distrbuted over a set of hypotheses $\mathcal{V}$, the probability of of error for any classifier $T(\overline{\mathbf{x}})$ with input $\overline{\mathbf{x}}$ coming from P_v^n for a randomly chosen $v\in\mathcal{V}$, is lower bounded as

$$\mathbb{P}[T(\overline{\mathbf{x}})\neq V]\geq 1-\frac{n\cdot\mathbb{E}_{v_1,v_2\in\mathcal{V}}\left[D\left(P_{v_1}\|P_{v_2}\right)\right]+1}{\log|\mathcal{V}|}$$

As before, we will combine the above bound with Lemma 1.1 to prove the desired lower bound on the minimax rate using

$$\mathcal{M}_n(\Pi,\ell)=\inf_{\hat{\theta}}\sup_{P\in\Pi}\mathop{\mathbb{E}}_{\overline{\mathbf{x}}\sim P^n}[\ell(\hat{\theta}(\overline{\mathbf{x}}),\theta(P))]\geq\Phi(\delta)\cdot\inf_{T}\{\mathbb{P}[T(\overline{\mathbf{x}})\neq V]\}$$

Proposition 9.5. Gaussian Mean Estimation

Let $\hat{\theta}\left(x_1,\ldots,x_n\right)=\frac{1}{n}\cdot\sum_{i\in[n]}x_i$. Then, for any $\mu\in\mathbb{R}^d$, we have that

$$\mathop{\mathbb{E}}_{\overline{\mathbf{X}}\sim(N(\mu,I_d))^n}\left[\left\|\frac{1}{n}\sum_{i\in[n]}X_i-\mu\right\|_2^2\right]=\frac{d}{n}$$

Fact 9.6. Chain Rule on KL for Gaussians

$$D\left(N\left(\mu_1,I_d\right)\|N\left(\mu_2,I_d\right)\right)=\frac{1}{2\ln 2}\cdot\|\mu_1-\mu_2\|_2^2$$

Lemma 9.7. Packing Lemma

There exists a collection of vectors $\mathcal{V}\subseteq\mathbb{R}^d$ such that $|\mathcal{V}|\geq 2^d$ and for all $v_1,v_2\in\mathcal{V},v_1\neq v_2$, we have

$$\frac{1}{2}\leq\|v_1-v_2\|_2\leq 2$$

Definition 9.8. Covering and Packing of Numbers
Let S be a set of points with a metric $\rho(\cdot,\cdot)$. A collection of points $\mathcal{C} \subseteq S$ is called a δ -covering of S (with respect to the metric ρ) if

$$\forall x \in S, \exists y \in \mathcal{C} \quad \rho(x,y) \leq \delta$$

A set of points $\mathcal{P}$ is called a δ -packing if

$$\forall x,y \in \mathcal{P}, x \neq y \quad \rho(x,y) > \delta$$

The size of the minimal δ -covering, denoted as $N(\delta, S, \rho)$, is called the δ -covering number of S and the size of the maximal δ -packing is called the δ -packing number. The quantity $\log N(\delta, S, \rho)$ is also called the metric entropy of S.

Exercise 9.9. Bound on Packing and Covering

$$M(2\delta, S, \rho) \leq N(\delta, S, \rho) \leq M(\delta, S, \rho)$$

Lemma 9.10. Proof on Packing Lemma
Lemma 3.7. There exists a collection of vectors $\mathcal{V} \subseteq \frac{1}{\sqrt{d}} \cdot \{-1,1\}^d$ such that $|\mathcal{V}| \geq 2^{d/20}$ and for all $v_1, v_2 \in \mathcal{V}, v_1 \neq v_2$, we have

$$\frac{1}{2} \leq \|v_1 - v_2\|_2 \leq 2$$

Lecture 10: Sparse Mean Estimation and I-Projection

Proposition 10.1. Distribution for the Difference
Let $\overline{\mathbf{x}} \sim (N(\mu, I_d))^n$ be a sequence of n independent samples, and let $\eta = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ be the empirical mean. Then $\eta - \mu$ is distributed according to the Gaussian distribution $N\left(0, \frac{1}{n} \cdot I_d\right)$

Corollary 10.2. Tail Bound
Let $\overline{\mathbf{x}} = (x_1, \ldots, x_n) \sim (N((\mu, I_d))^n$ as above. Then,

$$\mathbb{P}\left[\exists j \in [d] \quad |\mu_j - \eta_j| \geq t\right] \leq 2d \cdot \exp\left(-nt^2/2\right)$$

Claim 10.3. Concentration Bound for the Estimator
For the estimator $\widehat{\mu}$ as above

$$\mathbb{P}\left[\|\mu - \widehat{\mu}\|_2 \geq t\right] \leq 2d \cdot \exp\left(-nt^2/18\right)$$

Claim 10.4. Expectation on the Concentration

$$\mathop{\mathbb{E}}_{\overline{\mathbf{x}} \sim (N(\mu, I_d))^n} \left[\|\mu - \widehat{\mu}(\overline{\mathbf{x}})\|_2^2\right] = O\left(\frac{\log d}{n}\right)$$

Definition 10.5. I-Projection
Let Π be a closed convex set of distributions over U. In addition, assume that $\mathrm{Supp}(Q) = U$. Then

$$\mathrm{Proj}_{\Pi}(Q) := \mathop{\arg\min}_{P \in \Pi} D(P\|Q) = P^*$$

Theorem 10.6. Lower Bound for KL Divergence
Let $P^* = \mathrm{Proj}_{\Pi}(Q)$. Then, for all $P \in \Pi$

$$\begin{aligned} &\mathrm{Supp}(P) \subseteq \mathrm{Supp}(P^*) \\ &D(P\|Q) \geq D\left(P\|P^*\right) + D\left(P^*\|Q\right) \end{aligned}$$

Definition 10.7. Linear Family of Distributions
For any given real-valued functions $f_1, f_2, \ldots, f_k$ on $\mathcal{X}$ and $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}$, the set

$$\mathcal{L} = \left\{P \mid \sum_{x \in \mathcal{X}} p(x) \cdot f_i(x) = \mathop{\mathbb{E}}_{x \sim P}[f_i(x)] = \alpha_i, \forall i \in [k]\right\}$$

is called a linear family of distributions.

Lemma 10.8. Tight Bound for Linear Family
Let $\mathcal{L}$ be a linear family given by

$$\mathcal{L} = \left\{P : \sum_{x \in U} p(x) \cdot f_i(x) = \alpha_i, i \in [k]\right\}$$

and $\bigcup_{P \in \mathcal{L}} \mathrm{Supp}(P) = U$. Let $P^* = \mathrm{Proj}_{\mathcal{L}}(Q)$. Then, for all $P \in \mathcal{L}$

- There exists $\beta > 0$ such that for $t \in [-\beta, 0], P_t = tP + (1-t)P^* \in \mathcal{L}$.

- $D(P\|Q) = D\left(P\|P^*\right) + D\left(P^*\|Q\right)$

Then the I-Projection P^* of Q onto $\mathcal{L}$ satisfies the Pythagorean identity

$$D(P\|Q) = D\left(P\|P^*\right) + D\left(P^*\|Q\right)$$

Definition 10.9. Exponential Family
Let Q be a given distribution. For any given functions $g_1, g_2, \ldots, g_k$ on $\mathcal{X}$, the set
$\mathcal{E}_Q\left(g_1, \ldots, g_k\right) := \left\{P \mid \exists \lambda_1, \ldots, \lambda_k \in \mathbb{R} \forall x \in \mathcal{X}, \quad p(x) = c \cdot q(x) \cdot \exp\left(\sum_{i=1}^k \lambda_i g_i(x)\right)\right\}$
is called an exponential family of distributions.

Lecture 11: Matrix Scaling and Error-Correcting Codes

Definition 11.1. Rate
We define the rate of a code as above to be

$$R := \frac{\log M}{n} \text{ (bits per transmission)}.$$

We say that a rate R is achievable for a channel, if there exists a sequence of codes for $n \geq n_0$ with rates at least R and error probabilities $p_e^{(n)}$ such that $p_e^{(n)} \to 0$ as $n \to \infty$. We define the maximum achievable rate for a channel as $R^* = \sup\{R \mid R \text{ is achievable }\}$.

Definition 11.2. Channel Capacity

$$C := \max_{P(X)} I(X;Y)$$

Theorem 11.3. Channel Capacity and Best Rate
For any discrete memoryless channel, we have $R^* = C$.

Proposition 11.4. Channel Capacity as an Upper Bound
Let R be any achievable rate for a given channel with capacity C. Then, $R \leq \mathrm{C}$.

Claim 11.5. Upper Bound on Expected Error Probability
Let C be random code constructed as above in the random setting. Then

$$\mathop{\mathbb{E}}_{C}\left[p_e\right] \leq n \cdot 2^{-n \cdot D(p+\delta\|p)} + 2^{nR} \cdot n \cdot 2^{-n \cdot D\left(p+\delta\|\frac{1}{2}\right)}$$

where $D(p\|q)$ denotes $D(\mathbf{Bern}(p)\|\,\mathbf{Bern}(q))$ as usual.

Lecture 12: Linear Codes and Explicit Constructions

Definition 12.1. Parity Check Matrix
Definition 1.1 (Parity Check Matrix). Let $b_1, \ldots, b_{n-k} \in \mathbb{F}_q^n$ be a basis for the null space of G^T corresponding to a linear code C. Then $H \in \mathbb{F}_q^{(n-k)\times n}$, defined by

$$H^T = \left[\begin{array}{c|c|c|c} b_1 & b_2 & \ldots & b_{n-k} \end{array}\right]$$

is called a parity check matrix for $\mathbf{C}$.

Proposition 12.2. Decoding Algorithm with Vanishing Probability
Let $H \in \mathbb{F}_2^{m\times n}$ and Decom $: \mathbb{F}_2^m \to \mathbb{F}_2^n$ define a linear compression scheme as above. Then the linear code $\mathbf{C} = \{x \mid Hx = 0\}$ has a decoding algorithm with vanishing probability of error, for transmission through the channel $BSC(p)$.

Definition 12.3. Polarizing
An invertible matrix $P \in \mathbb{F}_2^{n\times n}$ is said to be (ε, τ) -polarizing for the random variable $Z \sim (\mathbf{Bern}(p))^n$ if for

$$W = PZ \quad \text{and} \quad S_\tau = \{i \in [n] \mid H\left(W_i \mid W_{<i}\right) \geq \tau\}$$

we have that $|S_\tau| \leq (H_2(p) + \varepsilon) \cdot n$

Theorem 12.4. Speed of Polarization
For all $\gamma > 0$, there exist constant $\alpha \in (0,1), \beta > 0$ such that for all $t \in \mathbb{N}$, we have

$$\mathbb{P}\left[X_t \in \left(\gamma^t, 1-\gamma^t\right)\right] \leq \beta \cdot \alpha^t$$

Lecture 13: Adversarial Error Models and Reed-Solomon Codes

Definition 13.1. Distance of a Code
Let $C \subseteq \mathcal{X}^n$ be a code. We define the distance of a code $\Delta(C)$ as

$$\Delta(C) := \min_{\substack{x,y\in C \\ x\neq y}} \Delta(x,y).$$

The distance can be used to understand the number of errors one can correct. Note that there are no probabilities in the error correcting model. Thus, we take the meaning of "correcting" y to finding the closest $x_0 \in C$ i.e., $x_0 = \mathbf{argmin}_{z \in C}(\Delta(y, z))$. The questions is if this correctly recovers the x that was sent (and corrupted to y by at most t errors).

Proposition 13.2. Correction of Errors
A code $C \subseteq \mathcal{X}^n$ can correct t errors if and only if $\Delta(C) \geq 2t + 1$

Proposition 13.3. Bound on Code Length
Let $C \subseteq \mathbb{F}_2^n$ be any distance-3 code, i.e., $\Delta(C) \geq 3$. Then

$$|C| \leq \frac{2^n}{n+1}$$

Proposition 13.4. Hamming Bound
Let $C \subseteq \mathcal{X}^n$ be any code with $\Delta(C) \geq d$. Then

$$|C| \leq \frac{|\mathcal{X}|^n}{\left|B\left(\cdot, \left\lfloor \frac{d-1}{2} \right\rfloor\right)\right|}$$

where $|B(\cdot, r)|$ denotes the size of a ball $B(x, r)$, which is the same for all $x \in \mathcal{X}^n$.

Remark 13.5. Hamming Bound in terms of Entropy
Remark 1.8. The Hamming bound also gives us a bound on the rate of the code in terms of entropy. Let $\mathcal{X} = \mathbb{F}_2, d = \delta n$ for $\delta \leq \frac{1}{2}$, and let $|C| = 2^k$. Since $|B(\cdot, r)| = \sum_{i=1}^{r} \binom{n}{i} \geq \frac{1}{n} \cdot 2^n \cdot H_2\left(\frac{r}{n}\right)$ for $r \leq \frac{n}{2}$ (why?), we have

$$2^k \leq n \cdot 2^{n - H_2(\delta/2)} \Rightarrow \frac{k}{n} \leq 1 - H_2(\delta/2) + o(1)$$

Theorem 13.6. Singleton Bound
Let $C \subseteq \mathcal{X}^n$ be a distance-d code, with $|C| \geq |\mathcal{X}|^k$. Then

$$d \leq n - k + 1$$

Definition 13.7. Reed-Solomon Code
Definition 2.1 (Reed-Solomon Code). Assume $q \geq n$ and fix $S = \{a_1, \ldots, a_n\} \subseteq \mathbb{F}_q$, distinct s.t. $|S| = n$. For each message $(m_0, \ldots, m_{k-1}) \in \mathbb{F}_q^k$, consider the polynomial

$f_m(X) = m_0 + m_1 X + \cdots + m_{k-1} X^{k-1}$. Then the Reed-Solomon Code is defined by its encoding as:

$$C\left(m_0, \ldots, m_{k-1}\right) = \left(f_m\left(a_1\right), \ldots, f_m\left(a_n\right)\right)$$

Alternatively, we can also define the code directly as the subspace

$$C = \left\{\left(f\left(a_1\right), \ldots, f\left(a_n\right)\right) \mid f \in \mathbb{F}_q^{\leq (k-1)}[X]\right\}$$

Let's compute the distance of the Reed-Solomon Code:

Claim 13.8. Distance of Reed-Solomon Code

$$\Delta(C) \geq n - k + 1$$

Remark 13.9. Linearity of Reed-Solomon Code
The Reed-Solomon Code is a linear code, as can be seen from the encoding map

$$C\left(m_0, \ldots, m_{k-1}\right) = \begin{bmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & a_n^2 & \ldots & a_n^{k-1} \end{bmatrix} \begin{bmatrix} m_0 \\ m_1 \\ \vdots \\ m_{k-1} \end{bmatrix}$$

Algorithm 12.10 Unique Decoding for Reed-Solomon Codes
Input: $\{(a_i, y_i)\}_{i=1,\ldots,n}$

- Find $e, g \in \mathbb{F}_q[X]$ such that $E \neq 0, \deg(e) \leq t, \deg(g) \leq k - 1 + t$

$$\forall i \in [n] \quad g\left(a_i\right) = y_i \cdot e\left(a_i\right)$$

- Output $\frac{g}{\rho}$

Lemma 12.11. Existence for Pairs
There exists (E, Q) that satisfies the conditions in Step 1 of the algorithm.

Lemma 12.12. Equality from Solutions
For any two solutions (g_1, e_1) and (g_2, e_2) that satisfy the conditions in Step 1,

$$\frac{g_1}{e_1} = \frac{g_2}{e_2}$$