Foundations of Data Science

Lecture Notes — Fall 2025

by Gastpar, Telatar, Urbanke

Version of September 24, 2025

EPFL

Contents

1 Introduction **7**
- 1.1 Foreword . . . 7
- 1.2 Acknowledgments . . . 8
- 1.3 Practical Information, Fall 2024, EPFL . . . 8
- 1.4 Preliminary Lecture Schedule, Fall 2025, EPFL . . . 9

2 Some Useful Notions from Probability **11**
- 2.1 Basic Distributions . . . 11
- 2.2 Some Basic Inequalities . . . 12
  - 2.2.1 Jensen's Inequality . . . 12
  - 2.2.2 The Markov Inequality . . . 13
- 2.3 Subgaussian Random Variables and Tail Bounds . . . 13
- 2.4 Subexponential Random Variables and Tail Bounds . . . 15
- 2.5 Conditional Expectation . . . 16
- 2.6 Problems . . . 17

3 Some Useful Notions from Linear Algebra **21**
- 3.1 Definitions and Notation . . . 21
- 3.2 Symmetric Matrices : Spectral Decomposition . . . 21
- 3.3 General Matrices : Singular Value Decomposition . . . 22
- 3.4 Rank of a matrix ; Norm of a matrix . . . 22
- 3.5 Low-rank Matrix Approximation . . . 23
- 3.6 Problems . . . 24

4 Information Measures **27**
- 4.1 L_1 and Total Variation Distance . . . 27
- 4.2 KL Divergence/Relative Entropy . . . 29
  - 4.2.1 Examples and Applications . . . 30
  - 4.2.2 Total Variation Distance versus KL Divergence . . . 31
- 4.3 Entropy . . . 32
- 4.4 Variational Representations . . . 33
  - 4.4.1 L_1 / Total Variation Distance . . . 33
  - 4.4.2 Kullback-Leibler (KL) Divergence . . . 33
- 4.5 Mutual Information . . . 34
- 4.6 Extension to Continuous Alphabets . . . 36
  - 4.6.1 KL Divergence . . . 37
  - 4.6.2 Differential Entropy . . . 37

4.7 Problems . . . 38

5 Multi-Arm Bandits 43
5.1 Introduction . . . 43
5.2 Some References . . . 44
5.3 Stochastic Bandits with a Finite Number of Arms . . . 44
5.3.1 Set-Up . . . 44
5.3.2 Explore then Exploit . . . 44
5.3.3 The Upper Confidence Bound Algorithm . . . 47
5.3.4 Information-theoretic Lower Bound . . . 51
5.4 Further Topics . . . 53
5.4.1 Asymptotic Optimality . . . 53
5.4.2 Adversarial Bandits . . . 54
5.4.3 Contextual Bandits . . . 56
5.5 Problems . . . 56

6 Detection and Estimation 59
6.1 Detection . . . 59
6.1.1 Binary hypothesis testing . . . 59
6.1.2 Hypothesis testing with repeated independent observations . . . 61
6.2 Estimation . . . 61
6.2.1 MMSE Estimation . . . 61
6.2.2 Linear MMSE Estimation . . . 63
6.3 Wiener Filtering, Smoothing, Prediction . . . 64
6.4 Adaptive Filters . . . 65
6.5 Parameter Estimation . . . 67
6.5.1 Fisher Information . . . 67
6.5.2 Fisher Information — Beyond one dimension . . . 69
6.6 Problems . . . 69

7 Distribution Estimation, Property Testing and Property Estimation 75
7.1 Distribution Estimation . . . 75
7.1.1 Notation and Basic Task . . . 75
7.1.2 Empirical Estimator . . . 75
7.1.3 Loss Functions . . . 76
7.1.4 Min-Max Criterion . . . 76
7.1.5 Risk of Empirical Estimator in ℓ_2^2 . . . 77
7.1.6 Risk of "Add Constant" Estimator in ℓ_2^2 . . . 78
7.1.7 Matching lower bound for ℓ_2^2 . . . 79
7.1.8 Risk in ℓ_1 . . . 80
7.1.9 Risk in KL-Divergence . . . 81
7.1.10 The problem with the min-max formulation . . . 81
7.1.11 Competitive distribution estimation . . . 81
7.1.12 Multi-set genie estimator . . . 82
7.1.13 Natural Genie and Good-Turing Estimator . . . 82
7.2 Property Testing . . . 84
7.2.1 General Idea . . . 84
7.2.2 Testing Against a Uniform Distribution . . . 85

7.3 Property Estimation . . . 89
7.3.1 Entropy Estimation . . . 89
7.4 Problems . . . 89

8 Exponential Families and Maximum Entropy Distributions 95
8.1 Definition . . . 95
8.2 Examples . . . 96
8.3 Convexity of $A(\theta)$ . . . 97
8.4 Derivatives of $A(\theta)$ . . . 98
8.5 Application to Parameter Estimation and Machine Learning . . . 99
8.6 Conjugate Priors . . . 99
8.7 Maximum Entropy Distributions . . . 100
8.8 Application To Physics . . . 102
8.9 I-Projections . . . 104
8.10 Relationship between θ and $\mathbb{E}[\phi(x)]$ . . . 105
8.10.1 The forward map $\nabla A(\theta)$ . . . 105
8.10.2 The backward map . . . 106
8.11 Problems . . . 107

9 Signal Representations 109
9.1 Fourier Representations . . . 110
9.1.1 DFT and FFT . . . 110
9.1.2 The Other Fourier Representations . . . 111
9.2 The Hilbert Space Framework for Signal Representation . . . 112
9.3 General Bases, Frames, and Time-Frequency Analysis . . . 114
9.3.1 The General Transform . . . 114
9.3.2 The Heisenberg Box Of A Signal . . . 116
9.3.3 The Uncertainty Relation . . . 116
9.3.4 The Short-time Fourier Transform . . . 118
9.4 Problems . . . 119

10 Compression and Dimensionality Reduction 121
10.1 Data compression . . . 121
10.2 Dimensionality Reduction . . . 125
10.2.1 PCA . . . 125
10.2.2 Johnson-Lindenstrauss . . . 127
10.3 Problems . . . 129

11 Information Measures and Generalization Error 137
11.1 Setup and Problem Statement . . . 137
11.2 Bounds on the Generalization Error . . . 138
11.2.1 L_1-Distance Bound . . . 139
11.2.2 Mutual Information Bound . . . 140
11.3 Exploration Bias . . . 141
11.3.1 Mutual Information Bound . . . 143
11.4 Problems . . . 144

Chapter 1

Introduction

1.1 Foreword

This is a set of lecture notes for the MS level class called *Foundations of Data Science* (COM-406) at EPFL. In 2017, 2018, and 2019 this course was called *"Information Theory and Signal Processing (for Data Science)"* and it was taught jointly by M. Gastpar, E. Telatar, and R. Urbanke. The class was first designed for the Fall Semester 2017.

This course discusses topics that are essential for the understanding and design of modern ML algorithms but are typically not taught in the standard ML courses. These include, slightly more advanced notions of probability (e.g., useful tail bounds and exponential families), basic notions of information theory (which is one of the main tool to derive bounds on algorithms), estimation and detection (which is equivalent to regression and classification but assuming that the underlying probabilistic model is known), multi-arm bandits (a basic version of reinforcement learning), and important notions of signal processing. These topics themselves are often interconnected. E.g., we will learn that exponential families contain those probability distributions that maximize the entropy under moment constraints.

Lausanne, Switzerland, September 2025 M. Gastpar

1.2 Acknowledgments

The authors thank Dr. Ibrahim Issa for contributions to the class development as well as to the Lecture Notes.

1.3 Practical Information, Fall 2024, EPFL

Instructor:
Michael Gastpar, `michael.gastpar@epfl.ch`, Office: INR 130

Teaching Assistants:
Millen Kanabar, `millen.kanabar@epfl.ch`, Office: INR 033
Yunzhen Yao, `@epfl.ch`, Office: INR 031

Administrative Assistants:
Muriel Bardet, `muriel.bardet@epfl.ch`, Office: INR 137

Class Meetings:
Tuesdays:

- 11:15-12:30, BC 01 (Lecture)
- *12:30-13:15, Lunch Break*
- 13:15-14:30, BC 01 (Lecture)
- 14:30-15:00, BC 01 (Solve HW Problem 1 together)

Wednesdays, 13:15-15:00, GC B3 30 (Exercises)

Class Web Page: We will use Moodle. Please check frequently.

Official Prerequisites:
COM-300 "Modèles stochastiques pour les communications" (or equivalent)
COM-202 "Signal processing" (or equivalent)
CS-233 "Introduction to machine learning"

Homework: We will have weekly homework sets. A part of your homework will be graded.

Midterm Exam: Wednesday, November 12, 2025, 13:15-15:00.

Final Exam: The Final Exam for the course will take place at some point between January 13 and February 1, 2026. The precise date will be decided by EPFL some time in November 2026.

Grading:

- If you do not hand in your final exam your overall grade will be NA.
- Otherwise, your grade will be determined based on the following weighted average: 10% for the Homework, 30% for the Midterm Exam, 60% for the Final Exam.

1.4 Preliminary Lecture Schedule, Fall 2025, EPFL

Date	Topics	Reading
Sept 9 *Sept 10*	General Introduction ; Basics of Probability *Exercise: HW 1*	Chapter 2
Sept 16 *Sept 17*	Information Measures *Exercise: HW 2*	Chapter 4
Sept 23 **Sept 24**	Information Measures **Lecture (exceptionally)** Information Measures	Chapter 4
Sept 30 *Sept 30* *Oct 1*	11:15- : Information Measures: Fano method *(Millen)* *13:15-15:00 : Exceptionally: Exercise: HW2* *Exercise: HW 2*	Chapter 4
Oct 7 *Oct 8*	Multi-arm Bandits *Exercise: HW 3*	Chapter 5
Oct 14 *Oct 15*	Multi-arm Bandits *Exercise: HW 3*	Chapter 5
Oct 21&22	Fall Break	
Oct 28 *Oct 29*	Detection & Estimation *Exercise: HW 4*	Chapter 6
Nov 4 *Nov 5*	Distribution Estimation *Exercise: HW 4*	Chapter 7
Nov 11 **Nov 12**	Property Testing **Midterm Exam**	Chapter 7
Nov 18 *Nov 19*	Exponential Family Distributions *Exercise: HW 5*	Chapter 8
Nov 25 *Nov 26*	Signal Representations *Exercise: HW 6*	Chapters 3 and 9
Dec 2 *Dec 3*	Signal Representations *Exercise: HW 6*	Chapter 9
Dec 9 *Dec 10*	Compression *Exercise: HW 7*	Chapter 10
Dec 16 *Dec 17*	Information Measures and Generalization Behavior *Exercise: HW 7*	Chapter 11

Chapter 2

Some Useful Notions from Probability

Perhaps the most important prerequisite is that you are familiar and comfortable with basic notions of probability. Except for Chapter 9, all chapters heavily use probability. We collect here some slightly more advanced results that will be useful later. In particular, we discuss tail bounds for subgaussian and subexponential distributions and we recall basic properties of conditional expectation.

2.1 Basic Distributions

In the following it will often be convenient to treat continuous and discrete cases together. So we will assume that we have a space $\mathcal{X}$ and a measure ν. Let us list our most important examples:

1. Reals: Let $\mathcal{X} = \mathbb{R}$ and let ν be the Lebesgue measure on $\mathbb{R}$; recall that ν assigns to intervals $[a, b]$, $a \leq b$, the measure $\nu([a, b]) = b - a$.
2. Bernoulli: Let $\mathcal{X} = \{0, 1\}$ and let ν be the counting measure on $\{0, 1\}$; i.e., $\nu(\emptyset) = 0$, $\nu(\{0\}) = \nu(\{1\}) = 1$, $\nu(\{0, 1\}) = 2$.
3. Poisson: Let $\mathcal{X} = \mathbb{N}$ and let ν be the counting measure on $\mathbb{N}$; i.e., for $S \subseteq \mathbb{N}$, $\nu(S) = |S|$, the cardinality of the set S.

In the sequel it will hopefully be clear what the base measure is and so we will typically not include it in our notation.

Example 2.1 (Gaussian)**.** Let $\mathcal{X} = \mathbb{R}$ and let ν be the Lebesgue measure on $\mathbb{R}$. Then the density of the normal distribution with mean μ and variance σ^2 can be written as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Example 2.2 (Poisson)**.** Let $\mathcal{X} = \mathbb{N}$ and let ν be the counting measure on $\mathcal{X}$. The density of the Poisson distribution with parameter λ is

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Example 2.3 (Bernoulli)**.** Let $\mathcal{X} = \{0, 1\}$ and let ν be the couting measure on $\mathcal{X}$. The density of the Bernoulli distribution with $P(X = 1) = p$ is

$$p(x) = p^x (1 - p)^{1-x}.$$

Example 2.4 (Multinomial)**.** A generalization of the Bernoulli measure is the multinomial. Let $\mathcal{X} = \{0, \cdots, n\}^d$ and let ν be the counting measure on $\mathcal{X}$. The density of the multinomial distribution with parameter $\alpha = (\alpha_1, \cdots, \alpha_d)$ is

$$p(x_1, \cdots, x_d) = \binom{n}{x_1, \cdots, x_d} \prod_{i=1}^{d} \alpha_i^{x_i},$$

if $x_1 + x_2 + \ldots + x_d = n$, and $p(x_1, \cdots, x_d) = 0$ otherwise.

Example 2.5 (Dirichlet)**.** The Dirichlet distribution of order $d \geq 2$ with parameter $\alpha = (\alpha_1, \cdots, \alpha_d)$, $\alpha_i > 0$, has a density with respect to the Lebesgue measure on $\mathbb{R}^{d-1}$ of the form

$$p_\alpha(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}$$

where x belongs to the $(d-1)$-dimensional simplex, i.e., $\sum_{i=1}^{d} x_i = 1$, $x_i \geq 0$, and where

$$B(\alpha) = \frac{\prod_{i=1}^{d} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{d} \alpha_i)}.$$

If $d = 2$ then the Dirichlet distribution is called the Beta distribution.

Note that the $(d-1)$-dimensional simplex is the space of all probability distributions with support of size d. Hence, one of the main uses of the Dirichlet distribution is to provide a prior on such distributions. We will see an example of this usage in Section 7.1.7.

2.2 Some Basic Inequalities

2.2.1 Jensen's Inequality

A function $f(x) : \mathbb{R} \to \mathbb{R}$ is called *convex* if for all x_1 and x_2 and for all $0 \leq \lambda \leq 1$, $\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2)$. It is called *strictly convex* if the inequality is strict for all $0 < \lambda < 1$ (and for all x_1 and x_2). Examples of convex functions are $f(x) = e^{\alpha x}$ for any real-valued α and $f(x) = -\log(x)$ (for positive values of x).

Theorem 2.1 (Jensen's Inequality)**.** *If $f(\cdot)$ is a convex function and X is a random variable, then*

$$\mathbb{E}[f(X))] \geq f(\mathbb{E}[X]). \tag{2.1}$$

Moreover, if $f(\cdot)$ is strictly convex, we have equality in (2.1) if and only if $X = \mathbb{E}[X]$ with probability 1 (that is, X is a constant).

For a proof, see e.g. [1, Theorem 2.6.2 on p.27].

A function $g(\cdot)$ is called *concave* if the function $f(x) := -g(x)$ is convex. Theorem 2.1 directly implies that if $g(\cdot)$ is a concave function and X is a random variable, then

$$\mathbb{E}[g(X))] \leq g(\mathbb{E}[X]). \tag{2.2}$$

Moreover, if $g(\cdot)$ is strictly concave, we have equality in (2.2) if and only if $X = \mathbb{E}[X]$ with probability 1 (that is, X is a constant).

2.2.2 The Markov Inequality

For any non-negative random variable X, the *Markov inequality* states that

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{2.3}$$

As a direct corollary, since for *any* real-valued random variable X, we have that X^2 is non-negative, we can apply the Markov inequality to X^2 to get the *Chebyshev inequality*:

$$\mathbb{P}(X^2 \geq a) \leq \frac{\mathbb{E}[X^2]}{a}. \tag{2.4}$$

Often, you will see the Chebyshev inequality applied to the random variable $|X - \mu|$, where $\mu = \mathbb{E}[X]$. Then, we can write $\mathbb{P}(|X - \mu| \geq b) \leq \frac{\mathrm{Var}(X)}{b^2}$.

2.3 Subgaussian Random Variables and Tail Bounds

We will often need to bound the tail of random variables. This is typically done via the Chernoff bound. It will be slightly more convenient to use this bound in a "packaged" form. This leads us to the notion of *subgaussian*[1] random variables.

Definition 2.1. A random variable X with mean μ is σ^2-subgaussian if for all $\lambda \in \mathbb{R}$ it holds that $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\sigma^2/2}$.

Definition 2.2. The quantity $\mathbb{E}[e^{\lambda(X-\mu)}]$ is called the *moment-generating* function of the random variable $(X - \mu)$.

Lemma 2.2 (Basic Properties of Subgaussians)**.** *Let X_i, $i = 1, 2$, with means μ_i, be two σ_i^2-subgaussian independent random variables. Then*

(i) $\mathbb{E}[(X_i - \mu_i)^2] \leq \sigma_i^2$.

(ii) For all $\alpha \in \mathbb{R}$, αX_i is $(\alpha^2\sigma_i^2)$-subgaussian.

(iii) $X_1 + X_2$ *is* $(\sigma_1^2 + \sigma_2^2)$*-subgaussian.*

Proof. Pick λ so that $\lambda\mathbb{E}[X_i] \geq 0$. Then using our assumption in the second step,

$$1 + \frac{1}{2}\lambda^2\sigma_i^2 + O(\lambda^4) \geq e^{\lambda^2\sigma_i^2/2} \geq \mathbb{E}[e^{\lambda(X_i-\mu_i)}] \geq 1 + \lambda\mathbb{E}[X_i - \mu_i] + \frac{1}{2}\lambda^2\mathbb{E}[(X_i - \mu_i)^2] + O(\lambda^3).$$

Claim (i) follows by letting λ tend to 0. Claim (ii) is true since $\mathbb{E}[e^{\lambda(X_i-\mu_i)}] \leq e^{\lambda^2\sigma_i^2/2}$ implies $\mathbb{E}[e^{\lambda(\alpha(X_i-\mu_i))}] = \mathbb{E}[e^{(\lambda\alpha)(X_i-\mu_i)}] \leq e^{\lambda^2\alpha^2\sigma_i^2/2} = e^{\lambda^2(\alpha\sigma_i)^2/2}$. And to prove claim (iii), note that $\mathbb{E}[e^{\lambda(X_1-\mu_1+X_2-\mu_2)}] = \mathbb{E}[e^{\lambda(X_1-\mu_1)}e^{\lambda(X_2-\mu_2)}] = \mathbb{E}[e^{\lambda(X_1-\mu_1)}]\mathbb{E}[e^{\lambda(X_2-\mu_2)}] \leq e^{\lambda^2\sigma_1^2/2}e^{\lambda^2\sigma_2^2/2} = e^{\lambda^2(\sigma_1^2+\sigma_2^2)/2}$. □

[1] Unfortunately, the literature is not consistent. We will follow Definition 2.1 and thus, "2-subgaussian" means that $\sigma^2 = 2$. However, some authors use "2-subgaussian" to mean that $\sigma = 2$, and thus, $\sigma^2 = 4$. For this reason, some authors prefer the very explicit (if a bit heavy handed) terminology "subgaussian with variance proxy σ^2."

Lemma 2.3 (Tail bound for subgaussian random variables)**.** *Let X with mean μ be σ^2-subgaussian. Then, for all $\eta > 0$,*

$$\mathbb{P}(X - \mu \geq \eta) \leq e^{-\frac{\eta^2}{2\sigma^2}}, \tag{2.5}$$

$$\mathbb{P}(X - \mu \leq -\eta) \leq e^{-\frac{\eta^2}{2\sigma^2}}. \tag{2.6}$$

Proof. We have

$$\begin{aligned}
\mathbb{P}\{X - \mu \geq \eta\} &\overset{\lambda \geq 0}{=} \mathbb{P}\{e^{\lambda(X-\mu)} \geq e^{\lambda\eta}\} \\
&\overset{\text{Markov inequality}}{\leq} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda\eta}} \\
&\overset{\sigma^2-\text{subgaussian}}{\leq} e^{\frac{1}{2}\lambda^2\sigma^2 - \lambda\eta},
\end{aligned}$$

which holds for all non-negative λ. Plug in $\lambda = \eta/\sigma^2$ to obtain the claimed bound. (This can also be shown to be the optimal choice simply by taking derivatives.) Likewise,

$$\begin{aligned}
\mathbb{P}\{X - \mu \leq -\eta\} &\overset{\lambda \geq 0}{=} \mathbb{P}\{e^{-\lambda(X-\mu)} \geq e^{\lambda\eta}\} \\
&\overset{\text{Markov inequality}}{\leq} \frac{\mathbb{E}[e^{-\lambda(X-\mu)}]}{e^{\lambda\eta}} \\
&\overset{\sigma^2-\text{subgaussian}}{\leq} e^{\frac{1}{2}\lambda^2\sigma^2 - \lambda\eta},
\end{aligned}$$

Again, plug in $\lambda = \eta/\sigma^2$ to obtain the claimed bound. □

Lemma 2.4 (Zero-mean Gaussian is subgaussian)**.** *Let X be a Gaussian random variable with mean zero and variance σ^2. Then X is σ^2-subgaussian.*

Proof. By Lemma 2.2 we can assume that $\sigma^2 = 1$. We then have

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &= \frac{1}{\sqrt{2\pi}} \int e^{\lambda x} e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int e^{\lambda x - x^2/2} dx \\
&= e^{\lambda^2/2} \frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(\lambda - x)^2} dx \\
&= e^{\lambda^2/2}.
\end{aligned}$$

□

Lemma 2.5 (Zero-mean RV with finite range is subgaussian)**.** *Let X be a zero-mean random variable with $X \in [a, b]$. Then X is $(b-a)^2/4$-subgaussian.*

Proof. This is an important and very convenient example. You will do this proof in the homework. □

Lemma 2.6 (Hoeffding's Bound)**.** *Assume that $X_1 - \mu, \cdots, X_m - \mu$ are zero-mean independent σ^2-subgaussian random variables. Let $\hat{\mu}$ be the empirical mean $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m X_i$. Then $\hat{\mu}$ satisfies*

$$\mathbb{P}\{\hat{\mu} \geq \mu + \epsilon\} \leq \exp\{-\frac{m\epsilon^2}{2\sigma^2}\},$$

$$\mathbb{P}\{\hat{\mu} \leq \mu - \epsilon\} \leq \exp\{-\frac{m\epsilon^2}{2\sigma^2}\}.$$

Proof. Using Lemma 2.2, we infer that $\frac{1}{m}\sum_{t=1}^{m}(X_t-\mu)$ is $\frac{\sigma^2}{m}$-subgaussian. The lemma now follows from the standard tail bound for subgaussians given in Lemma 2.3. □

2.4 Subexponential Random Variables and Tail Bounds

Although many random variables are subgaussian there are important random variables that are not. E.g., let X be a zero-mean, unit-variance Gaussian random variable and consider $Y = X^2$. Note that $p(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Hence $\mathbb{P}\{Y \leq \alpha\} = \mathbb{P}\{X^2 \leq \alpha\} = \int_{-\sqrt{\alpha}}^{\sqrt{\alpha}} p(x)dx$. Therefore, $p(y) = \frac{d\mathbb{P}\{Y\leq y\}}{dy} = 2p(x)|_{x=\sqrt{y}} \times \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}$. Note that $\mathbb{E}[Y] = \mathbb{E}[X^2] = 1$, i.e., Y has mean 1.

The moment generating function of $Y-1$ is

$$\begin{aligned}\mathbb{E}[e^{\lambda(Y-1)}] &= \int_{y\geq 0} e^{\lambda(y-1)}\frac{e^{-y/2}}{\sqrt{2\pi y}}dy = e^{-\lambda}\int_{y\geq 0}\frac{e^{-y(1/2-\lambda)}}{\sqrt{2\pi y}}dy \\ &= \frac{1}{\sqrt{2\pi}}\int_x e^{\lambda(x^2-1)}e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}e^{-\lambda}\int_x e^{-x^2(1/2-\lambda)}dx \overset{\lambda<\frac{1}{2}}{=} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}.\end{aligned}$$

For $\lambda \geq \frac{1}{2}$ the moment generating function does not exist. Hence Y is not subgaussian.

Definition 2.3 (Subexponential RV)**.** We say that a rv X with mean μ is subexponential with parameters (ν, b) if

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}, \forall|\lambda| < \frac{1}{b}.$$

Let us go back to $Y = X^2$, where X is a zero-mean, unit-variance Gaussian. We have

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}, \forall|\lambda| < \frac{1}{4}.$$

Therefore, we see that Y is subexponential with parameters $(\nu, b) = (2, 4)$. Subexponential random variables will appear for example in Section 10.2.2 below.

For subexponential random variables we can derive concentration bounds in a similar manner as for subgaussian random variables.

Lemma 2.7 (Tail Bound for Subexponential RVs)**.**

$$\mathbb{P}\{X-\mu \geq t\} \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}}, 0 \leq t \leq \nu^2/b, \\ e^{-\frac{t}{2b}}, t > \nu^2/b. \end{cases}$$

Proof. Without essential loss of generality we can assume that X has zero mean. Using our standard approach we then have

$$\mathbb{P}\{X \geq t\} \leq \mathbb{E}[e^{\lambda X}]e^{-\lambda t} \leq e^{-\lambda t + \frac{\lambda^2\nu^2}{2}},$$

for all $0 \leq \lambda < 1/b$. In order to find the best bound it remains to determine the best parameter λ for any fixed t. This is easily done.

If we ignore at first the bound on λ then we see that the optimal value of $\lambda = t/\nu^2$. This is feasible as long as $t/\nu^2 < 1/b$, or, $t < \nu^2/b$, and we get a bound of $e^{-t^2/(2\nu^2)}$.

And for $t \geq \nu^2/b$ the best choice (the one that gives the tightest bound) is $\lambda = 1/b$ (the maximum allowed value) and we get a bound of $e^{-t/b+1/(2b)\nu^2/b}$. Using the relationships $\nu^2/b \leq t$ this can be further bounded by $e^{-t/(2b)}$. □

At several points in previous derivations have we made some simplifications and used bounds. Why are we content with potentially loose bounds? Note that bounds of the given form are particularly useful. E.g., just like in the subgaussian case, if we consider the sum of several independent rvs, each of which is subexponential, then it is easy to see that the resulting rv is again subexponential and that its parameters are easily computed from the parameters of the individual random variables. A similar thing happens if we scale a random variable. This is a considerable advantage of this formulation, well worth loosing slightly in the tightness of the final bound.

For the case of X^2 the square of a Gaussian rv we explicitly computed its moment-generating function and then found an appropriate upper bound. But this is in general difficult to do. And in some situations we might not even know the exact distribution but perhaps only have some constraints that are known (e.g., we know that the random variable is bounded in some region).

It is therefore of interest to have some other criteria at hand that can certify that a rv is subexponential and that is potentially easier to handle. One such set of conditions was discovered by Bernstein and concerns the moments of the random variable. That moments play a role here is not too surprising. If we look at the Taylor series expansion of $e^{\lambda X}$ then this involves all the moments of X. So one way to bound the moment-generating function is to bound all moments of X.

We will not pursue this avenue any further at this point. But should in the future you hear someone talk about Bernstein bounds you know that this is just a small variation of what we have talked about.

2.5 Conditional Expectation

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a random variable in this space. Let $\mathcal{G} \subseteq \mathcal{F}$. Then $\mathbb{E}[X \mid \mathcal{G}]$ is the unique random variable Z that

1. is $\mathcal{G}$-measurable,
2. for all $A \in \mathcal{G}$, $\mathbb{E}[X\mathbb{1}_{\{A\}}] = \mathbb{E}[Z\mathbb{1}_{\{A\}}]$.

If Y is another random variable in the probability space then $\mathbb{E}[X \mid Y]$ is $\mathbb{E}[X \mid \mathcal{G}]$ where $\mathcal{G}$ is the σ-algebra generated by Y.

In words, the two properties mean that $\mathbb{E}[X \mid \mathcal{G}]$ is a random variable that is constant on atoms of $\mathcal{G}$ and on each such atom it is equal to its average. One special case is if $\mathcal{G}$ is given by a random variable Y. If we think of Y as representing a certain information then $\mathbb{E}[X \mid Y]$ represents the best prediction of X given Y in the sense that X is the average of all its values that are compatible with the observation of Y (the average over all ω where Y takes on the given value). If you want to visualize this then visualize a partition that is determined by Y (the partition signifies the region where Y takes on a constant value). Then $\mathbb{E}[X \mid Y]$ is a random variable that is also constant on each of those members of the partition and takes a value that is the average of X over this region. Conditional expectation plays a fundamental role, see e.g. Section 6.2.1.

Lemma 2.8 (Conditional Expectation)**.** *Let X, Y, and Z be random variables and $a, b \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. Assuming all the following expectations exist we have*

$\mathbb{E}[a \mid Y] = a$,

$\mathbb{E}[aX + bZ \mid Y] = a\mathbb{E}[X \mid Y] + b\mathbb{E}[Z \mid Y]$,

$\mathbb{E}[X \mid Y] \geq 0$ *if* $X \geq 0$,

$\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ *if* X *and* Y *are independent,*

$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$,

$\mathbb{E}[Xg(Y) \mid Y] = g(Y)\mathbb{E}[X \mid Y]$,

$\mathbb{E}[X \mid Y, g(Y)] = \mathbb{E}[X \mid Y]$,

$\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y] = \mathbb{E}[X \mid Y]$.

2.6 Problems

Problem 2.1 (Review of Random Variables)**.** Let X and Y be discrete random variables defined on some probability space with a joint pmf $p_{XY}(x, y)$. Let $a, b \in \mathbb{R}$ be fixed.

(a) Prove that $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. Do not assume independence.

(b) Prove that if X and Y are independent random variables, then $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

(c) Assume that X and Y are not independent. Find an example where $\mathbb{E}[X \cdot Y] \neq \mathbb{E}[X] \cdot \mathbb{E}[Y]$, and another example where $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

(d) Prove that if X and Y are independent, then they are also uncorrelated, i.e.,

$$Cov(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = 0. \tag{2.7}$$

(e) Find an example where X and Y are uncorrelated but dependent.

(f) Assume that X and Y are uncorrelated and let σ_X^2 and σ_Y^2 be the variances of X and Y, respectively. Find the variance of $aX + bY$ and express it in terms of $\sigma_X^2, \sigma_Y^2, a, b$. **Hint:** First show that $Cov(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

Problem 2.2 (Review of Gaussian Random Variables)**.** A random variable X with probability density function

$$p_X(x) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{2.8}$$

is called a *Gaussian* random variable.

(a) Explicitly calculate the mean $\mathbb{E}[X]$, the second moment $\mathbb{E}[X^2]$, and the variance $Var[X]$ of the random variable X.

(b) Let us now consider events of the following kind:

$$\Pr(X < \alpha). \tag{2.9}$$

Unfortunately for Gaussian random variables this cannot be calculated in closed form. Instead, we will rewrite it in terms of the standard Q-function:

$$Q(x) \quad = \quad \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du \tag{2.10}$$

Express $\Pr(X < \alpha)$ in terms of the Q-function and the parameters m and σ^2 of the Gaussian pdf.

Like we said, the Q-function cannot be calculated in closed form. Therefore, it is important to have *bounds* on the Q-function. In the next 3 subproblems, you derive the

most important of these bounds, learning some very general and powerful tools along the way:

(c) Derive the Markov inequality, which says that for any non-negative random variable X and positive a, we have

$$\Pr(X \geq a) \quad \leq \quad \frac{\mathbb{E}[X]}{a}. \tag{2.11}$$

(d) Use the Markov inequality to derive the Chernoff bound: the probability that a real random variable Z exceeds b is given by

$$\Pr(Z \geq b) \leq \mathbb{E}\big[e^{s(Z-b)}\big], \qquad s \geq 0. \tag{2.12}$$

(e) Use the Chernoff bound to show that

$$Q(x) \leq e^{-\frac{x^2}{2}} \quad \text{for } x \geq 0. \tag{2.13}$$

Problem 2.3 (Moment Generating Function)**.** In the class we had considered the logarithmic moment generating function

$$\phi(s) := \ln \mathbb{E}[\exp(sX)] = \ln \sum_x p(x) \exp(sx)$$

of a real-valued random variable X taking values on a finite set, and showed that $\phi'(s) = \mathbb{E}[X_s]$ where X_s is a random variable taking the same values as X but with probabilities $p_s(x) := p(x) \exp(sx) \exp(-\phi(s))$.

(a) Show that

$$\phi''(s) = \mathrm{Var}(X_s) := \mathbb{E}[X_s^2] - \mathbb{E}[X_s]^2$$

and conclude that $\phi''(s) \geq 0$ and the inequality is strict except when X is deterministic.

(b) Let $x_{\min} := \min\{x : p(x) > 0\}$ and $x_{\max} := \max\{x : p(x) > 0\}$ be the smallest and largest values X takes. Show that

$$\lim_{s \to -\infty} \phi'(s) = x_{\min}, \quad \text{and} \quad \lim_{s \to \infty} \phi'(s) = x_{\max}.$$

Problem 2.4 (Bounded random variables are subgaussian)**.** This problem is a guided proof of a slightly weakened version of Lemma 2.4.

(a) Prove the following inequality:

$$\cosh(x) = (e^x + e^{-x})/2 \leq e^{x^2/2}. \tag{2.14}$$

(b) Using the previous inequality, give an upper bound on the moment generating function of a random variable S that only takes the values $+1$ and -1, with equal probability.

Hint: The upper bound should depend on the parameter of the moment generating function.

(c) Consider any random variable X and let X' be a random variable *independent of* X, but with exactly the same distribution. Show that

$$\mathbb{E}_X[e^{\lambda(X - \mathbb{E}[X])}] \leq \mathbb{E}_{X,X'}[e^{\lambda(X - X')}]. \tag{2.15}$$

(d) Show that the random variables $(X - X')$ and $S(X - X')$, where S is as in Part (b) and assumed independent of X and X', have the same distribution.

(e) From the previous part, we thus know that

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}]. \tag{2.16}$$

Now assume that X is a bounded random variable, $X \in [a, b]$. Condition on $X = x$ and $X' = x'$, and take expectation over S. Observe that $(x - x')^2 \leq (b - a)^2$. Use this and your result from Part (b) to further upper bound $\mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}]$.

(f) Combine your results to give an upper bound on the moment generating function of a centered bounded random variable $X - \mathbb{E}[X]$, where $X \in [a, b]$.

Hint: The upper bound should depend on the parameter of the moment generating function as well as a and b.

(g) Compare your result to Lemma 2.4. Discuss the differences.

Problem 2.5 (Hoeffding's Lemma)**.** Prove Lemma 2.5 in the lecture notes. In other words, prove that if X is a zero-mean random variable taking values in $[a, b]$ then

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2}{2}[(a-b)^2/4]}.$$

Expressed differently, X is $[(a - b)^2/4]$-subgaussian.

Problem 2.6 (Gaussian Variance Estimation)**.** Consider estimating the mean μ and variance σ^2 from n independent samples $(X_1, \ldots, X_n)$ of a Gaussian with this mean and variance.

(a) Show that $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator of μ.

(b) Show that

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

is a biased estimator of σ^2 whereas

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

is an unbiased estimator of σ^2.

(c) Show that S_n^2 has a lower mean squared error than S_{n-1}^2. Thus it is possible that a biased estimator may be better than an unbiased one.

Problem 2.7 (Expected Maximum of Subgaussians)**.** Let $\{X_i\}_{i=1}^n$ be a collection of n σ^2-subgaussian random variables, not necessarily independent of each other. Let $Y = \max_{i\in\{1,2,\cdots,n\}} X_i$. Prove that $\mathbb{E}[Y] \leq \sqrt{2\sigma^2 \log n}$. *Hint:* Recall that by Jensen, $e^{\lambda\mathbb{E}[X]} \leq \mathbb{E}[e^{\lambda X}]$.

Chapter 3

Some Useful Notions from Linear Algebra

A key set of tools here (and throughout engineering and computer science) is linear algebra.

3.1 Definitions and Notation

We will denote (column) vectors by bold symbols $\mathbf{x}$. The transpose of a vector $\mathbf{x}$ is the row vector $\mathbf{x}^T$. The Hermitian transpose (transpose and complex conjugate) of a complex-valued vector $\mathbf{x}$ is the row vector $\mathbf{x}^H$. The inner product (or dot product) of two vectors (of equal length) will be denoted as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^H \mathbf{x}$. (That is, following standard notational conventions, in the $\langle \cdot, \cdot \rangle$ notation, it is the *second* argument that is complex-conjugated.) The 2-norm of a vector is $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. More generally, the p-norm of a vector is $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. These are genuine norms (triangle inequality, scaling, zero only for the zero vector) for all real numbers $p \geq 1$. For $0 \leq p < 1$, they are not norms since they violate the triangle inequality, but they are nonetheless of interest in applications.

Matrices are denoted by upper-case symbols A. They are of dimensions $m \times n$, meaning that they have m rows and n columns. The entry in row i, column j is denoted by $\{A\}_{ij}$, or simply A_{ij} when there is no confusion possible. The identity matrix is denoted by I. The matrix-vector product $\mathbf{y} = A\mathbf{x}$, where $\mathbf{x}$ is of length n, is the vector $\mathbf{y}$ of length m with entries $y_i = \sum_{j=1}^n A_{ij} x_j$. We use A^T to denote the transpose and A^H to denote the Hermitian transpose of the matrix A. Consider two matrices A and B with columns denoted by $\mathbf{a}_i$ and $\mathbf{b}_i$, respectively. Then, the matrix product $B^H A$ (if dimension-compatible) has as its entry in row i, column j the inner product $\langle \mathbf{a}_j, \mathbf{b}_i \rangle$. An alternative and equally useful expression for matrix multiplication is that $AB^H = \sum_i \mathbf{a}_i \mathbf{b}_i^H$ (if the matrices are dimension-compatible). A *unitary* matrix is a matrix U satisfying $UU^H = U^H U = I$. The *trace* of a square matrix, $\mathrm{trace}(A)$, is the sum of its diagonal entries. A particularly useful property is that for any two (dimension-compatible) matrices A and B, we have $\mathrm{trace}(AB) = \mathrm{trace}(BA)$. Another useful property is that $(AB)^H = B^H A^H$.

3.2 Symmetric Matrices : Spectral Decomposition

Perhaps the most important class of matrices are the symmetric (hence square) matrices. That is, matrices A for which we have $A = A^H$ (and thus *a fortiori*, $m = n$). Such matrices

always admit a *spectral decomposition*,[1] i.e., they can be written as

$$A = U\Lambda U^H, \tag{3.1}$$

where Λ is a real-valued diagonal matrix and U is a unitary matrix. The n columns of U are called the *eigenvectors* of A, denoted by $\mathbf{u}_1, \ldots, \mathbf{u}_n$. The n diagonal entries of Λ, usually denoted by λ_i, are called the corresponding *eigenvalues*. By inspection, Formula (3.1) can also be expressed as

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^H, \tag{3.2}$$

a shape that will be of interest to us in this class. A property with many uses is the fact that $\text{trace}(A) = \sum_{i=1}^{n} \lambda_i$, which follows simply from $\text{trace}(A) = \text{trace}(U\Lambda U^H) = \text{trace}(\Lambda U^H U)$.

3.3 General Matrices : Singular Value Decomposition

For general matrices A, one can construct two instructive symmetric matrices, namely, AA^H and $A^H A$. Both of these admit spectral decompositions:

$$AA^H = U\Lambda' U^H \quad \text{and} \quad A^H A = V\Lambda'' V^H, \tag{3.3}$$

and it is straightforward to show that, *(i),* the non-zero entries in Λ' and Λ'' are the same, i.e., AA^H and $A^H A$ have the same eigenvalues, and *(ii),* all eigenvalues are non-negative. From these one can construct the *singular value decomposition*

$$A = U\Sigma V^H = \sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \tag{3.4}$$

where Σ is an $m \times n$ diagonal matrix whose entries σ_i are simply the square roots of the eigenvalues of AA^H (or, equivalently, $A^H A$). The values σ_i are thus non-negative and are referred to as the *singular values* of the matrix A.

3.4 Rank of a matrix ; Norm of a matrix

The *rank* of a matrix A is the number of non-zero singular values it has in its singular value decomposition and will be denoted by $\text{rank}(A)$. An important relationship is that for any two (dimension-compatible) matrices, we have $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$. Note that no similarly useful and non-trivial relationship can be given for the rank of the *sum* of two matrices.

We will also find it useful to define *norms* for matrices. First, let us introduce the so-called *operator norms* that are derived from standard vector norms as follows:

$$\|A\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \tag{3.5}$$

An interesting special case is when $p = 2$, which is often called the *spectral norm* of the matrix A, and is easily seen to be equal to the largest singular value of the matrix A.

[1] In general, the eigendecomposition is expressed as $A = Q\Lambda Q^{-1}$. When Q turns out to be a unitary matrix (thus, $Q^{-1} = Q^H$), then one often refers to the eigendecomposition as a *spectral decomposition.*

Of equal importance is the Frobenius norm

$$\|A\|_F \;=\; \sqrt{\sum_{i,j} |A_{ij}|^2}. \tag{3.6}$$

A first interesting observation (which can be proved by elementary manipulations) is that $\|A\|_F^2 = \mathrm{trace}(A^H A)$. This also implies that $\|A\|_F^2 = \sum_{i=1}^r \sigma_i^2$. Another interesting property is that for any two (dimension-compatible) matrices A and B, we have that $\|AB\|_F \leq \|A\|_F \|B\|_F$, a consequence of the Cauchy-Schwarz inequality.

3.5 Low-rank Matrix Approximation

Let us consider the following intuitively pleasing problem: Given a matrix $A \in \mathbb{R}^{m\times n}$, we seek to find a matrix $B \in \mathbb{R}^{m\times n}$ of rank no larger than p such that B is as close as possible to A, i.e., such that the norm $\|A - B\|$ is as small as possible. If we use as the norm the spectral or the Frobenius norm, then this problem has an intuitively pleasing solution, given in the following theorem.

Theorem 3.1 (Eckart-Young)**.** *Let the SVD of the rank-r matrix A be*

$$A \;=\; \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \quad \textit{with } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r. \tag{3.7}$$

For integers p between 1 and $r-1$, let $\hat{A}_p$ denote the truncated sum

$$\hat{A}_p \;=\; \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^H. \tag{3.8}$$

Then, we have

$$\min_{B:\mathrm{rank}(B)\leq p} \|A - B\|_2 \;=\; \sigma_{p+1} \tag{3.9}$$

$$\min_{B:\mathrm{rank}(B)\leq p} \|A - B\|_F \;=\; \sqrt{\sum_{k=p+1}^{r} \sigma_k^2}, \tag{3.10}$$

and a minimizer of each of the two is $B = \hat{A}_p$. For the Frobenius norm, $\hat{A}_p$ is the unique minimizer if and only if $\sigma_p > \sigma_{p+1}$ (strict inequality).

Proof. First, observe that $\mathrm{rank}(\hat{A}_p) \leq p$ and that we can write $A - \hat{A}_p = \sum_{k=p+1}^r \sigma_k \mathbf{u}_k \mathbf{v}_k^H$. Therefore :

- $\|A - \hat{A}_p\|_2 = \sigma_{p+1}$, thus $\min_{B:\mathrm{rank}(B)\leq p} \|A - B\|_2 \leq \sigma_{p+1}$, and
- $\|A - \hat{A}_p\|_F = \sqrt{\sum_{k=p+1}^r \sigma_k^2}$, thus $\min_{B:\mathrm{rank}(B)\leq p} \|A - B\|_F \leq \sqrt{\sum_{k=p+1}^r \sigma_k^2}$.

The more interesting part is the converse. We provide the converse proof only for the spectral norm in these notes. Consider any matrix $B \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(B) \leq p$. Its null space has dimension no smaller than $n - p$, and thus, the dimension of the intersection

null$(B) \cap$ span$\{\mathbf{v}_1, \cdots, \mathbf{v}_{p+1}\}$ is at least one. For all vectors $\mathbf{x} \in \mathbb{R}^n$ of norm one in this intersection, we have

$$(A-B)\mathbf{x} \;=\; A\mathbf{x} = \sum_{k=1}^{p+1} \sigma_k (\mathbf{v_k}^H \mathbf{x}) \mathbf{u}_k, \tag{3.11}$$

thus, using the fact that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{p+1}$,

$$\|(A-B)\mathbf{x}\|^2 \;=\; \sum_{k=1}^{p+1} \sigma_k^2 |\mathbf{v_k}^H \mathbf{x}|^2 \geq \sum_{k=1}^{p+1} \sigma_{p+1}^2 |\mathbf{v_k}^H \mathbf{x}|^2 = \sigma_{p+1}^2 \sum_{k=1}^{p+1} |\mathbf{v_k}^H \mathbf{x}|^2 \tag{3.12}$$

But since the unit-norm vector $\mathbf{x}$ lies in the span of $\{\mathbf{v}_1, \cdots, \mathbf{v}_{p+1}\}$, we must have $\sum_{k=1}^{p+1} |\mathbf{v_k}^H \mathbf{x}|^2 = 1$. Therefore, for every matrix B with rank$(B) \leq p$, a unit-norm vector $\mathbf{x}$ can be found such that $\|(A-B)\mathbf{x}\|^2 \geq \sigma_{p+1}^2$. Thus, $\|A-B\|_2 \geq \sigma_{p+1}$. □

3.6 Problems

Problem 3.1 (Some review problems on linear algebra)**.** *(a) (Frobenius norm)* Prove that $\|A\|_F^2 = \text{trace}(A^H A)$.

(b) (Singular Value Decomposition) Let $\sigma_i(A)$ denote the i^{th} singular value of an $m \times n$ matrix A. Prove that $\|A\|_F^2 = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)$

(c) (Projection Matrices) Consider a set of k orthonormal vectors in $\mathbb{C}^n$, denoted by $\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_k}$. The projection matrix (that projects an arbitrary vector into the subspace spanned by these orthonormal vectors) is given by

$$P \;=\; \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^H. \tag{3.13}$$

- Prove that this matrix is *Hermitian,* i.e., $P^H = P$.
- Prove that this matrix is *idempotent,* i.e., $P^2 = P$. (In words, projecting twice into the same subspace is the same as projecting only once.)
- Prove that trace$(P) = k$, i.e., equal to the dimension of the subspace.
- Prove that the diagonal entries of P must be real-valued and non-negative. Then, prove that the diagonal entries of P cannot be larger than 1 (this is a little more tricky).

Problem 3.2 (Eckart–Young Theorem)**.** In class, we proved the converse part of the Eckart–Young theorem for the spectral norm. Here, you do the same for the case of the Frobenius norm.

(a) For any matrix A of dimension $m \times n$ and an arbitrary orthonormal basis $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ of $\mathbb{C}^n$, prove that

$$\|A\|_F^2 \;=\; \sum_{k=1}^{n} \|A\mathbf{x}_k\|^2. \tag{3.14}$$

(b) Consider any $m \times n$ matrix B with rank$(B) \leq p$. Clearly, its null space has dimension no smaller than $n-p$. Therefore, we can find an orthonormal set $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-p}\}$ in the null

space of B. Prove that for such vectors, we have

$$\|A - B\|_F^2 \quad \geq \quad \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2. \tag{3.15}$$

(c) (This requires slightly more subtle manipulations.) For any matrix A of dimension $m \times n$ and any *orthonormal* set of $n - p$ vectors in $\mathbb{C}^n$, denoted by $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-p}\}$, prove that

$$\sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \quad \geq \quad \sum_{j=p+1}^{r} \sigma_j^2. \tag{3.16}$$

Hint: Consider the case $m \geq n$ and the set of vectors $\{\mathbf{z}_1, \cdots, \mathbf{z}_{n-p}\}$, where $\mathbf{z}_k = V^H \mathbf{x}_k$. Express your formulas in terms of these and the SVD representation $A = U\Sigma V^H$.

(d) Briefly explain how *(a)-(c)* imply the desired statement.

Chapter 4

Information Measures

Whenever you see a lower bound in data science, chances are that it is based on information-theoretic ideas. This chapter collects basic information-theoretic notions and inequalities. It is important not to over-interpret those definitions. E.g., entropy is a measure of "information." But it is only through specific operational settings that these quantities acquire their meaning: E.g., if we are compressing a "source" then entropy gives us a lower bound on the number of bits that we need in order to describe the source. But be aware that not every time you see the word "information" will entropy be the correct quantity to measure it.

Let $\mathcal{X}$ denote a discrete alphabet and let $\Pi := \Pi(\mathcal{X})$ denote the set of all probability distributions on $\mathcal{X}$. With $K = |\mathcal{X}|$, we can identify Π with the *simplex* in $\mathbb{R}^K$: the set of all $(p_1, \ldots, p_K) \in \mathbb{R}^K$ with $\sum_k p_k = 1$, and $p_k \geq 0$. In the sequel we will say that $P \in \Pi$ is a distribution and we will typically work with the probability mass function (pmf) $p(x) = P(X = x)$, $x \in \mathcal{X}$. The *support* of a probability distribution is the set of values $x \in \mathcal{X}$ for which $p(x) > 0$, strictly positive.

4.1 L_1 and Total Variation Distance

How should we compare two probability distributions? Since they are vectors, we might simply start with the Euclidean distance between them. This choice is not invalid, but it does not appear to lead to a fruitful perspective. There are many reasons for this. We will see some of them later in the class. Instead, we start with the L_1 distance between the two probability distributions (thought of as vectors).

Definition 4.1. Let P and Q be two probability mass functions on a finite set $\mathcal{X}$. Then, the L_1 distance between P and Q is defined as

$$\|P - Q\|_1 = \sum_x |p(x) - q(x)|. \tag{4.1}$$

Lemma 4.1. *We have $0 \leq \|P - Q\|_1 \leq 2$. There is equality on the left if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$. There is equality on the right if and only if the supports of P and Q are disjoint.*

Total Variation distance, and an interpretation

A second, seemingly unrelated way of measuring the distance between two distributions is the Total Variation distance.

Definition 4.2. Let P and Q be two probability mass functions on a finite set $\mathcal{X}$. Then, the Total Variation distance between P and Q is defined as

$$\delta(P,Q) = \max_{S\subseteq\mathcal{X}} P(S) - Q(S). \tag{4.2}$$

The total variation distance has several interesting interpretations. Here is one of them. Consider the following way of rewriting:

$$1-\delta(P,Q) = 1 - \left\{\max_{S\subseteq\mathcal{X}} P(S) - Q(S)\right\} \tag{4.3}$$

$$= \min_{S\subseteq\mathcal{X}} \{(1-P(S)) + Q(S)\} \tag{4.4}$$

$$= \min_{S\subseteq\mathcal{X}} P(S^c) + Q(S), \tag{4.5}$$

where S^c denotes the complement of the set S in $\mathcal{X}$. This last quantity can be interpreted as follows:

1. We receive a (random) sample $X \in \mathcal{X}$. Our task is to decide if the sample came from P or from Q.
2. That is, for every possible outcome $x \in \mathcal{X}$, we need to say if it came from P or from Q. Denote by $S \subseteq \mathcal{X}$ the subset of all values $x \in \mathcal{X}$ for which we decide P.
3. Now, suppose that the sample X came from P. Then, the probability that our rule gets it wrong (*i.e.,* outputs Q) is precisely the probability that the sample $X \in S^c$, which is $P(S^c)$. (Sometimes called a *Type-I error.*)
4. Conversely, suppose that the sample X came from Q. Then, the probability that our rule gets it wrong (*i.e.,* outputs P) is precisely the probability that the sample $X \in S$, which is $Q(S)$. (Sometimes called a *Type-II error.*)
5. Goal: Select S such as to minimize the *sum* of the two error probabilities (*Type-I error* plus *Type-II error*).
6. The performance of this (optimal) S is exactly equal to $1-\delta(P,Q)$.

Total Variation distance is half of L_1

Perhaps initially to some surprise, L_1 distance and Total Variation distance are the same (up to a factor of two), which we record in the following lemma.

Lemma 4.2. *Let P and Q be two probability mass functions on a finite set $\mathcal{X}$. Then,*

$$\|P-Q\|_1 = 2\delta(P,Q).$$

Proof. Let $A = \{z \in \mathcal{Z} : P(z) \geq Q(z)\}$, and A^c the complement of A in $\mathcal{X}$. Then,

$$\begin{aligned}\|P-Q\|_1 &= \sum_{x\in A} p(x) - q(x) + \sum_{x\in A^c} q(x) - p(x) = P(A) - Q(A) + Q(A^c) - P(A^c) \\ &= P(A) - Q(A) + 1 - Q(A) - 1 + P(A) = 2(P(A) - Q(A)).\end{aligned}$$

To complete the proof, start by considering any other set S by adding elements to A to observe that for such S, we have $P(S) - Q(S) \leq P(A) - Q(A)$. Then, consider any other set S by removing elements from A, and again observe that for such S, we have $P(S) - Q(S) \leq P(A) - Q(A)$. □

4.2 KL Divergence/Relative Entropy

The most fundamental measure in information theory is the so-called *KL divergence*. It is a measure of how different two distributions are.

Definition 4.3 (KL Divergence)**.** For P and Q in Π, we call $D(P\|Q) = \sum_x p(x) \log[p(x)/q(x)]$ the *divergence of P from Q*.

In the sum above, we skip the terms with $p(x) = 0$, and we set $D(P\|Q) = +\infty$ if there is an x for which $p(x) > 0$ and $q(x) = 0$. This divergence is also called the KL divergence (there are other divergences as we will discuss in the homeworks), where KL stands for Kullback-Leibler.

Definition 4.4. Given two alphabets $\mathcal{X}$ and $\mathcal{Y}$, a *probability kernel from $\mathcal{X}$ to $\mathcal{Y}$* is a matrix $W = \big[W(y|x) : x \in \mathcal{X},\ y \in \mathcal{Y}\big]$ such that $W(y|x) \geq 0$, and for each $x \in \mathcal{X}$, $\sum_y W(y|x) = 1$. We will write $W : \mathcal{X} \to \mathcal{Y}$ to indicate that W is such a kernel. The set of probability kernels describes all possible conditional probabilities on $\mathcal{Y}$, conditional on elements of $\mathcal{X}$.

Lemma 4.3 (Data Processing Inequality for Divergence)**.** *Given P and Q in $\Pi(\mathcal{X})$ and $W : \mathcal{X} \to \mathcal{Y}$, let $\tilde{p}(y) = \sum_x P(x)W(y|x)$ and $\tilde{q}(y) = \sum_x q(x)W(y|x)$. Then $\tilde{P}$ and $\tilde{Q}$ are in $\Pi(\mathcal{Y})$, and*

$$D(\tilde{P}\|\tilde{Q}) \leq D(P\|Q).$$

The inequality is strict, unless $q(x)/p(x) = \tilde{q}(y)/\tilde{p}(y)$ for all x, y with $p(x)W(y|x) > 0$.

Proof. That $\tilde{P}$ and $\tilde{Q}$ are probability distributions is an easy consequence of W being a probability kernel. To prove the claimed inequality between the divergences let us first show that log is a strictly concave function. I.e., for any non-negative $\lambda_1, \ldots, \lambda_K$ for which $\sum_k \lambda_k = 1$, and any positive $x_1, \ldots, x_K$, we have, with $\bar{x} = \sum_k \lambda_k x_k$,

$$\sum_k \lambda_k \log x_k \leq \log \bar{x},$$

and equality happens if and only if for all k with $\lambda_k > 0$, we have $x_k = \bar{x}$. It suffices to prove this statement with ln instead of log. To that end, first note that with $f(x) = \ln x$ we have $f'(x) = 1/x$ and $f''(x) = -1/x^2 < 0$. Thus, Taylor expansion of $\ln x$ around 1 yields $\ln x = (x-1) - (x-1)^2/(2\xi^2)$ for some ξ between 1 and x, and we see that $\ln x \leq x - 1$, with equality if and only if $x = 1$. Consequently

$$\sum_k \lambda_k \ln x_k - \ln \bar{x} = \sum_k \lambda_k \ln[x_k/\bar{x}] \leq \sum_k \lambda_k [x_k/\bar{x} - 1] = 1 - 1 = 0,$$

with the inequality being strict if there is a k for which $\lambda_k > 0$ and $x_k \neq \bar{x}$.

Having thus proved the strict concavity of log, now observe (with $p(x)W(y|x)$'s cast in

the role of λ_k's) that

$$\begin{aligned} D(\tilde{P}\|\tilde{Q}) - D(P\|Q) &= \sum_y \tilde{p}(y) \log \frac{\tilde{p}(y)}{\tilde{q}(y)} - \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x,y} W(y|x)p(x) \log \frac{\tilde{p}(y)}{\tilde{q}(y)} - \sum_{x,y} W(y|x)p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x,y} W(y|x)p(x) \log \frac{\tilde{p}(y)q(x)}{\tilde{q}(y)p(x)} \\ &\leq \log\left[\sum_{x,y} W(y|x) \frac{\tilde{p}(y)q(x)}{\tilde{q}(y)}\right] = \log\left[\sum_y \tilde{p}(y)\right] = 0. \end{aligned}$$ □

Corollary 4.4 (Non-Negativity of Divergence)**.** $D(P\|Q) \geq 0$ *with equality if and only if* $P = Q$*.*

Proof. Take $\mathcal{Y} = \{0\}$ and set $W(0|x) = 1$. Then $\tilde{p}(0) = \tilde{q}(0) = 1$ and $D(\tilde{P}\|\tilde{Q}) = 0$. □

Corollary 4.5. $D(P\|Q)$ *is a convex function of the pair* (P, Q)*.*

Proof. Suppose P_0, Q_0, P_1, Q_1 are in $\Pi(\mathcal{X})$ and suppose $0 \leq \lambda \leq 1$. We need to show that

$$D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1) \leq (1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1).$$

To that end consider the distributions P and Q on the set $\{0,1\} \times \mathcal{X}$ with

$$p(z,x) = \begin{cases} (1-\lambda)p_0(x) & \text{if } z = 0 \\ \lambda p_1(x) & \text{if } z = 1, \end{cases} \quad \text{and} \quad q(z,x) = \begin{cases} (1-\lambda)q_0(x) & \text{if } z = 0 \\ \lambda q_1(x) & \text{if } z = 1, \end{cases}$$

Consider also the channel $W : \{0,1\} \times \mathcal{X} \to \mathcal{X}$ with $W(x'|(z,x)) = \mathbb{1}\{x' = x\}$. It is easily checked that $D(P\|Q) = (1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1)$ and also that

$$\tilde{P} = (1-\lambda)P_0 + \lambda P_1 \quad \text{and} \quad \tilde{Q} = (1-\lambda)Q_0 + \lambda Q_1.$$

The conclusion now follows from Lemma 4.3. □

4.2.1 Examples and Applications

The Gaussian case

Example 4.1 (KL Distance between two Gaussians)**.** If P_i, $i = 1, 2$, are two Gaussians with means μ_i and variances σ_i^2, then

$$D_{KL}(P_1||P_2) = \ln(\sigma_2/\sigma_1) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

A Probability Bound

Let A be an event and P be any distribution. Then trivially $P(A) + P(A^c) = 1$. What happens if instead we consider $P(A) + Q(A^c)$ where P and Q are close but not identical? The following lemma gives a handy bound (and will be useful to us below in the proof of Lemma 5.2).

Lemma 4.6. *Let A be an event and P and Q be any distributions. Then*

$$P(A) + Q(A^c) \geq \frac{1}{2} e^{-D(P||Q)}.$$

Proof.

$$\begin{aligned}
P(A) + Q(A^c) &= \int_{x \in A} p(x) + \int_{x \in A^c} q(x) \\
&\geq \int_{x \in A} \min\{p(x), q(x)\} + \int_{x \in A^c} \min\{p(x), q(x)\} \\
&= \int \min\{p(x), q(x)\} \\
&\overset{(a)}{\geq} \frac{1}{2} \left(\int \min\{p(x), q(x)\} \right) \left(\int \max\{p(x), q(x)\} \right) \\
&\overset{(b)}{\geq} \frac{1}{2} \left(\int \sqrt{\min\{p(x), q(x)\} \max\{p(x), q(x)\}} \right)^2 \\
&= \frac{1}{2} \left(\int \sqrt{p(x) q(x)} \right)^2 \\
&= \frac{1}{2} e^{2 \ln \int \sqrt{p(x) q(x)}} \\
&= \frac{1}{2} e^{2 \ln \int p \sqrt{\frac{q(x)}{p(x)}}} \\
&\overset{\text{Jensen}}{\geq} \frac{1}{2} e^{2 \int \frac{1}{2} p \ln \frac{q(x)}{p(x)}} \\
&= \frac{1}{2} e^{\int p \ln \frac{q(x)}{p(x)}} \\
&= \frac{1}{2} e^{-D(P||Q)}.
\end{aligned}$$

For step (a) note that $\frac{1}{2} \left(\int \max\{p(x), q(x)\} \right) \leq 1$ with equality if the two distributions have disjoint support. Step (b) follows by Cauchy-Schwartz.[1] □

4.2.2 Total Variation Distance versus KL Divergence

The KL divergence can be used to furnish a bound on the L_1 distance of two distributions. This is called the *Pinsker* inequality. A guided proof is provided in Problem 4.5.

Lemma 4.7 (Pinsker Inequality)**.** *Let P and Q be two distributions. Then* $\|p - q\|_1 \leq \sqrt{\frac{2}{\log(e)} D(p\|q)}$.

[1] $|fg|_1^2 \leq |f|_2^2 |g|_2^2$ with $f = \sqrt{\min\{p(x), q(x)\}}$ and $g = \sqrt{\max\{p(x), q(x)\}}$.

4.3 Entropy

Definition 4.5 (Entropy)**.** For P in Π, we call $H(P) = -\sum_x p(x)\log(p(x))$ the *entropy* of P.

The formula for entropy bears a close resemblance to KL divergence. Indeed, if we let Q be the uniform distribution over $\mathcal{X}$, then

$$D(P\|\text{unif}_{\mathcal{X}}) = \sum_x p(x)\log\frac{p(x)}{1/|\mathcal{X}|} = \log|\mathcal{X}| - H(P). \tag{4.6}$$

So we may think of the entropy as the comparison of a distribution to the uniform distribution.

Lemma 4.8. *$0 \le H(P) \le \log|\mathcal{X}|$, with equality on the left if and only if there is an $x \in \mathcal{X}$ with $P(X = x) = 1$, and equality on the right if and only if P is the uniform distribution on $\mathcal{X}$.*

Proof. The non-negativity of $H(P)$ follows from $p(x) \ge 0$ and $-\log(p(x)) \ge 0$, so that each term in the sum defining $H(P)$ is non-negative. Moreover, the sum equals zero only if each term is zero, which yields the condition for $H(P)$ to equal 0.

The upper bound in Lemma 4.8 and the condition for equality follow from noting that $\log|\mathcal{X}| - H(P) = D(P\|\text{unif}_{\mathcal{X}})$ where $\text{unif}_{\mathcal{X}}$ is the uniform distribution on $\mathcal{X}$ with $\text{unif}_{\mathcal{X}}(x) = 1/|\mathcal{X}|$. As we have seen in Corollary 4.4, $D(P\|\text{unif}_{\mathcal{X}}) = 0$ if and only if P is the uniform distribution. □

If the logarithm is in base two then we call the corresponding quantity "bits," whereas if we choose the natural quantity then we speak of "nats."

If a random variable X is distributed according to P then we will also write $H(X)$ to denote $H(P)$.

Definition 4.6 (Conditional Entropy)**.** Given two random variables X and Y we define $H(X \mid Y = y)$ and $H(X \mid Y)$ (conditional entropy) to be

$$H(X \mid Y = y) = -\sum_x p(x \mid y)\log(p(x \mid y)),$$
$$H(X \mid Y) = \sum_y H(X \mid Y = y)p(y) = -\sum_{x,y} p(x,y)\log(p(x \mid y)).$$

Note, that $H(X \mid Y)$ is the average entropy if we reveal the value of the Y variable. If we have given a joint distribution with pmf $p(x_1, x_2, \cdots, x_n)$ then we have the *chain rule*

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2 \mid x_1)\cdots p(x_n \mid x_1, \cdots, x_{n-1}).$$

If we plug this representation into the formula for $H(X_1, \cdots, X_n)$ we see that we get the corresponding *chain rule* for entropies,

Lemma 4.9 (Chain Rule of Entropy)**.**

$$H(X_1, \cdots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \cdots X_{n-1}).$$

Finally, a fundamental and intuitively pleasing inequality shows that conditioning cannot increase entropy.

Lemma 4.10 (Conditioning Decreases Entropy)**.**

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

with equality if and only if X and Y are independent.

Proof.

$$H(X) - H(X \mid Y) = -\sum_x p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y) \right) = \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$
$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)p(y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y)) \geq 0.$$

□

4.4 Variational Representations

4.4.1 L_1 / Total Variation Distance

Lemma 4.11.

$$\|P - Q\|_1 = 2 \max_{f: \mathcal{Z} \to [0,1]} \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)].$$

Proof. Let $A = \{z \in \mathcal{Z} : P(z) \geq Q(z)\}$. Then,

$$\begin{aligned} \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] &= \sum_{z \in A} f(z) \left(P(z) - Q(z)\right) + \sum_{z \notin A} f(z) \left(P(z) - Q(z)\right) \\ &\leq \sum_{z \in A} \left(P(z) - Q(z)\right) \\ &= \frac{\|P - Q\|_1}{2}. \end{aligned}$$

Equality can be achieved if we choose $f(z) = \begin{cases} 1, & z \in A, \\ 0, & z \notin A \end{cases}$. □

4.4.2 Kullback-Leibler (KL) Divergence

The following lemma presents an interesting and important alternative characterization of the KL divergence. It will prove to be useful for example in Lemma 11.4.

Lemma 4.12 (Variational Form of KL Divergence - Donsker-Varadhan)**.** *Let P and Q be two distributions. Then,*

$$D(P||Q) = \sup_{\substack{f: \mathbb{R} \to \mathbb{R} \\ \mathbb{E}_Q[e^f] < +\infty}} \left\{ \mathbb{E}_P[f(Z)] - \log \mathbb{E}_Q\left[e^{f(Z)}\right] \right\}$$

Remark. The log is taken to the base e (both in the computation of $D(P||Q)$ and in the right-hand side).

Proof. 1) Suppose $D(P||Q) < +\infty$ and consider any function f. Then,

$$\begin{aligned}
\mathbb{E}_P[f(Z)] - \log \mathbb{E}_Q\left[e^{f(Z)}\right] &= \mathbb{E}_P\left[\log e^{f(Z)}\right] - \log \mathbb{E}_Q\left[e^{f(Z)}\right] \\
&= \mathbb{E}_P\left[\log \frac{e^{f(Z)}}{\mathbb{E}_Q\left[e^{f(Z)}\right]}\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{e^{f(Z)}}{\mathbb{E}_Q\left[e^{f(Z)}\right]}\frac{p}{q}\frac{q}{p}\right)\right] \\
&= D(P||Q) + \mathbb{E}_P\left[\log\left(\frac{qe^{f(Z)}}{p\mathbb{E}_Q\left[e^{f(Z)}\right]}\right)\right] \\
&= D(P||Q) - \mathbb{E}_P\left[\log \frac{p}{q\frac{e^{f(Z)}}{\mathbb{E}_Q\left[e^{f(Z)}\right]}}\right]
\end{aligned}$$

Now note that $\int q\frac{e^{f(Z)}}{\mathbb{E}_Q\left[e^{f(Z)}\right]}dz = \frac{1}{\mathbb{E}_Q\left[e^{f(Z)}\right]}\int qe^{f(Z)}dz = 1$. Hence, the second term is also a KL divergence. Then,

$$\mathbb{E}_P[f(Z)] - \log \mathbb{E}_Q\left[e^{f(Z)}\right] = D(P||Q) - D(P||\tilde{Q}) \leq D(P||Q).$$

Equality is achieved if we set $f = \log \frac{p}{q}$.
2) Suppose $D(P||Q) = +\infty$. Then, we need to show that the supremum is also $+\infty$. $D(p||q) = +\infty$ implies that there exists a set A such that $p(A) > 0$ and $q(A) = 0$. Choose $f = \lambda \mathbb{1}\{z \in A\}$. Then, $\mathbb{E}_p[f(Z)] - \log \mathbb{E}_q\left[e^{f(Z)}\right] = \lambda p(A)$. Taking $\lambda \to +\infty$ yields the result. □

4.5 Mutual Information

Entropy measures the information content of a random variable. KL divergences measure how much one distributions diverges from another one. And *mutual information* measures how much information one random variable contains about another one.

Definition 4.7 (Mutual Information)**.** Let X and Y be two random variables. Then their *mutual information* is defined as

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

Lemma 4.13 (Non-Negativity of Mutual Information)**.**

$$I(X;Y) \geq 0,$$

with equality if and only if X and Y are independent.

Proof.

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= -\sum_x p(x)\log(p(x)) + \sum_{x,y} p(x,y)\log(p(x \mid y) \\ &= \sum_{x,y} p(x,y)\log(\frac{p(x \mid y)}{p(x)}) \\ &= \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \\ &= D(p(x,y)||p(x)p(y)) \geq 0. \end{aligned}$$

□

Definition 4.8 (Conditional Mutual Information)**.**

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

Lemma 4.14 (Non-Negativity of Conditional Mutual Information)**.**

$$I(X;Y \mid Z) \geq 0,$$

with equality if and only if X and Y are condiitonally independent given Z.

Proof.

$$I(X;Y \mid Z) = \sum_z p(z) D(p(x,y|z)||p(x|z)p(y|z)) \geq 0.$$

□

Lemma 4.15 (Chain Rule for Mutual Information)**.**

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \cdots, X_{i-1}).$$

Proof.

$$\begin{aligned} I(X_1, X_2, \cdots, X_n; Y) &\overset{\text{Definition 4.7}}{=} H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n \mid Y) \\ &\overset{\text{Lemma 4.9}}{=} \sum_{i=1}^{n} H(X_i \mid X_1, \cdots, X_{i-1}) - \sum_{i=1}^{n} H(X_i \mid Y, X_1, \cdots, X_{i-1}) \\ &\overset{\text{Definition 4.8}}{=} \sum_{i=1}^{n} I(X_i; Y \mid X_1, \cdots, X_{i-1}). \end{aligned}$$

□

Lemma 4.16 (Data Processing Inequality for Mutual Information)**.** *Let $X - Y - Z$ form a Markov chain, i.e., $p(z \mid x,y) = p(z \mid y)$. Then*

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

Proof. Expanding $I(X;Y,Z)$ in two ways we get

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

The last equality holds since $I(X;Z \mid Y) = 0$ due to the assumption that X and Z are conditionally independent given Y. The result follows since $I(X;Y \mid Z) \geq 0$. □

Lemma 4.17 (Fano's Inequality)**.** *Let X and Y be random variables. Let $\hat{X} = g(Y)$ be a prediction of X based on Y (so that $X - Y - \hat{X}$ is a Markov chain) and let $P_e = \mathbb{P}(X \neq \hat{X})$ be the related probability of error. Then we have*

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y).$$

Proof. Define the random variable $E = \mathbb{1}\{X \neq \hat{X}\}$, so that $\mathbb{P}(E = 1) = \mathbb{P}(X \neq \hat{X})$. Expanding $H(X, E|\hat{X})$ in two ways using the chain-rule we get

$$H(X, E|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \overset{\text{Lemma } 4.10}{\leq} H(E) + H(X|E, \hat{X})$$

and

$$H(X, E|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(X|\hat{X}),$$

where we used that $H(E|X, \hat{X}) = 0$ since we can determine the value of E when given X and $\hat{X}$. Combining both yields

$$\begin{aligned} H(X|\hat{X}) &\leq H(E) + H(X|E, \hat{X}) \\ &\overset{\text{Definition } 4.6}{=} H(E) + \mathbb{P}(E = 1)H(X|E = 1, \hat{X}) + \mathbb{P}(E = 0)H(X|E = 0, \hat{X}) \\ &= H(E) + \mathbb{P}(E = 1)H(X|E = 1, \hat{X}) \\ &= H(P_e) + P_e H(X|E = 1, \hat{X}) \\ &\overset{\text{Lemma } 4.8}{\leq} H(P_e) + P_e \log(|\mathcal{X}| - 1), \end{aligned}$$

where in the third step we used that $H(X|E = 0, \hat{X}) = 0$ since $E = 0$ so $X = \hat{X}$ and we are given $\hat{X}$.

Finally, using lemma 4.16, we have $I(X;Y) \geq I(X;\hat{X})$ so that $H(X|Y) \leq H(X|\hat{X})$, which concludes the proof. □

4.6 Extension to Continuous Alphabets

So far, in this chapter, we have restricted attention to the case of discrete-valued random variables. For some of the arguments made in this course, it is also of interest to consider continuous random variables X, including in Chapters 8 and 11. In this section, we provide the necessary definitions and review some key properties.

4.6.1 KL Divergence

For probability density functions $f(x)$ and $g(x)$, the definition of KL divergence directly extends as

$$D(f\|g) = \int_S f(x) \log \frac{f(x)}{g(x)} dx. \tag{4.7}$$

Most importantly, we have

$$D(f\|g) \geq 0, \tag{4.8}$$

with equality if and only if $f(x) = g(x)$ almost everywhere. We point to [1, Chapter 8, Section 8.5] for a more detailed treatment of this quantity.

We may define the mutual information between two random variables X and Y with joint probability density function $f(x, y)$ as

$$I(X;Y) = D(f(x,y)\|f(x)f(y)), \tag{4.9}$$

where $f(x)$ and $f(y)$ denote the marginal probability density functions of X and Y, respectively.

4.6.2 Differential Entropy

In the continuous case, entropy is usually referred to as *differential entropy*. Following [1, Chapter 8], we define the differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ as

$$h(X) = -\int_S f(x) \log f(x) dx, \tag{4.10}$$

where S is the support set of the random variable. Most importantly, it must be noted that this quantity may be *negative*.

An interesting special case is when X is a Gaussian (normal) random variable with variance σ^2. In this case, solving the above integral gives $h(X) = \frac{1}{2} \log_2 \left((2\pi e)\sigma^2\right)$.

Moreover, one can also easily convince oneself that we can write $I(X;Y) = h(X) + h(Y) - h(X,Y)$. But we note that any or all of the differential entropies involved in this expression may be negative.

We point to [1, Chapter 8] for a detailed treatment of this quantity, but point out one small yet important and useful lemma.

Lemma 4.18. *Let the random variable X have zero mean and variance σ^2. Then,*

$$h(X) \leq \frac{1}{2} \log_2 \left((2\pi e)\sigma^2\right), \tag{4.11}$$

with equality if and only if X is Gaussian (normal).

We return to this observation in much more detail in Chapter 8.

4.7 Problems

Problem 4.1 (Entropy and pairwise independence)**.** Suppose X, Y, Z are pairwise independent fair flips, i.e., $I(X;Y) = I(Y;Z) = I(Z;X) = 0$.

(a) What is $H(X,Y)$?

(b) Give a lower bound to the value of $H(X,Y,Z)$.

(c) Give an example that achieves this bound.

Problem 4.2 (Entropy and Geometry)**.** Suppose X, Y and Z are random variables.

(a) Show that $H(X) + H(Y) + H(Z) \geq \frac{1}{2}\big[H(X,Y) + H(Y,Z) + H(Z,X)\big]$.

(b) Show that $H(X,Y) + H(Y,Z) \geq H(X,Y,Z) + H(Y)$.

(c) Show that

$$2\big[H(X,Y) + H(Y,Z) + H(Z,X)\big] \geq 3H(X,Y,Z) + H(X) + H(Y) + H(Z).$$

(d) Show that $H(X,Y) + H(Y,Z) + H(Z,X) \geq 2H(X,Y,Z)$.

(e) Suppose n points in three dimensions are arranged so that their their projections to the xy, yz and zx planes give n_{xy}, n_{yz} and n_{zx} points. Clearly $n_{xy} \leq n$, $n_{yz} \leq n$, $n_{zx} \leq n$. Use part (d) show that

$$n_{xy} n_{yz} n_{zx} \geq n^2.$$

Problem 4.3. *(Geometrical interpretation of mutual information)*
In Homework 2 we introduced the conditional KL divergence between two probability kernels $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \to \mathcal{Y}$ given a distribution P_X over $\mathcal{X}$ as

$$D(P_{Y|X} \| Q_{Y|X} | P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x)),$$

where for every $x \in \mathcal{X}$, $D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x))$ is the standard KL divergence between the two distributions $P_{Y|X}(\cdot|x)$ and $Q_{Y|X}(\cdot|x)$ over $\mathcal{Y}$.

(a) Let X and Y be two random variables with joint distribution $P_{XY} = P_X P_{Y|X}$. Show that

$$I(X;Y) = \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| P_Y),$$

where P_Y is the marginal distribution of Y. This formula shows that the mutual information can be interpreted as a weighted average of the distances between the conditional distributions $P_{Y|X}(\cdot|x)$ and the marginal distribution P_Y.

(b) Show that for any distribution Q_Y on $\mathcal{Y}$,

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

You can think of this formula as a KL equivalent of the classical $I(X;Y) = H(Y) - H(Y|X)$.

(c) Show that

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

According to this formula, the minimizing Q_Y can be interpreted as the "center of gravity" of the conditional distributions $P_{Y|X}(\cdot|x)$, and the mutual information as its radius.

Problem 4.4. *(Entropy and variance)*

Let X be a continuous random variable with density $p(x)$ and support in $\mathbb{R}$. Let $h(X)$ denote the (differential) entropy of X, where (in nats)

$$h(x) = \int -p(x)\ln p(x)dx.$$

Let $\mathrm{Var}(X)$ denote the variance of X, where

$$\mathrm{Var}(X) = \int p(x)(x - \mathbb{E}X)^2 dx.$$

Assume that you are told that $h(X) \geq h > 0$. What bound can you conclude on $\mathrm{Var}(X)$ in terms of h?

Problem 4.5 (Divergence and L_1)**.** Suppose p and q are two probability mass functions on a finite set $\mathcal{U}$. (I.e., for all $u \in \mathcal{U}$, $p(u) \geq 0$ and $\sum_{u\in\mathcal{U}} p(u) = 1$; similarly for q.)

(a) Show that the L_1 distance $\|p - q\|_1 := \sum_{u\in\mathcal{U}} |p(u) - q(u)|$ between p and q satisfies

$$\|p - q\|_1 = 2 \max_{\mathcal{S}:\mathcal{S}\subset\mathcal{U}} p(\mathcal{S}) - q(\mathcal{S})$$

with $p(\mathcal{S}) = \sum_{u\in\mathcal{S}} p(u)$ (and similarly for q), and the maximum is taken over all subsets $\mathcal{S}$ of $\mathcal{U}$.

For α and β in $[0,1]$, define the function $d_2(\alpha\|\beta) := \alpha\log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}$. Note that $d_2(\alpha\|\beta)$ is the divergence of the distribution $(\alpha, 1-\alpha)$ from the distribution $(\beta, 1-\beta)$.

(b) Show that the first and second derivatives of d_2 with respect to its first argument α satisfy $d_2'(\beta\|\beta) = 0$ and $d_2''(\alpha\|\beta) = \frac{\log e}{\alpha(1-\alpha)} \geq 4\log e$.

(c) By Taylor's theorem conclude that

$$d_2(\alpha\|\beta) \geq 2(\log e)(\alpha - \beta)^2.$$

(d) Show that for any $\mathcal{S} \subset \mathcal{U}$

$$D(p\|q) \geq d_2(p(\mathcal{S})\|q(\mathcal{S}))$$

[Hint: use the data processing theorem for divergence.]

(e) Combine (a), (c) and (d) to conclude that

$$D(p\|q) \geq \frac{\log e}{2}\|p - q\|_1^2.$$

(f) Show, by example, that $D(p\|q)$ can be $+\infty$ even when $\|p - q\|_1$ is arbitrarily small. [Hint: considering $\mathcal{U} = \{0,1\}$ is sufficient.] Consequently, there is no generally valid inequality that upper bounds $D(p\|q)$ in terms of $\|p - q\|_1$.

Problem 4.6 (Generating fair coin flips from biased coins)**.** Suppose $X_1, X_2, \ldots$ are the outcomes of independent flips of a biased coin. Let $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$, with p unknown. By processing this sequence we would like to obtain a sequence $Z_1, Z_2, \ldots$ of *fair* coin flips.

Consider the following method: We process the X sequence in sucssive pairs, (X_1X_2), (X_3X_4), (X_5X_6), mapping (01) to 0, (10) to 1, and the other outcomes (00) and (11) to the empty string. After processing X_1, X_2, we will obtain either nothing, or a bit Z_1.

(a) Show that, if a bit is obtained, it is fair, i.e., $\mathbb{P}(Z_1 = 0) = \mathbb{P}(Z_1 = 1) = 1/2$.

In general we can process the X sequence in successive n-tuples via a function $f : \{0,1\}^n \to \{0,1\}^*$ where $\{0,1\}^*$ denote the set of all finite length binary sequences (including the empty string λ). [The case in (a) is the function $f(00) = f(11) = \lambda$, $f(01) = 0$, $f(10) = 1$. The function f is chosen such that $(Z_1, \ldots, Z_K) = f(X_1, \ldots, X_n)$ are i.i.d., and fair (here K may depend on $(X_1, \ldots, X_K)$.

(b) With $h_2(p) = -p \log p - (1-p)\log(1-p)$, prove the following chain of (in)equalities.

$$\begin{aligned} nh_2(p) &= H(X_1, \ldots, X_n) \\ &\geq H(Z_1, \ldots, Z_K, K) \\ &= H(K) + H(Z_1 \ldots, Z_K|K) \\ &= H(K) + \mathbb{E}[K] \\ &\geq \mathbb{E}[K]. \end{aligned}$$

Consequently, on the average no more than $nh_2(p)$ fair bits can be obtained from $(X_1, \ldots, X_n)$.

(c) Find a good f for $n = 4$.

Problem 4.7 (Other Divergences)**.** Suppose f is a convex function defined on $(0, \infty)$ with $f(1) = 0$. Define the f-divergence of a distribution p from a distribution q as

$$D_f(p\|q) := \sum_u q(u) f(p(u)/q(u)).$$

In the sum above we take $f(0) := \lim_{t\to 0} f(t)$, $0f(0/0) := 0$, and $0f(a/0) := \lim_{t\to 0} tf(a/t) = a \lim_{t\to 0} tf(1/t)$.

(a) Show that for any non-negative a_1, a_2, b_1, b_2 and with $A = a_1 + a_2$, $B = b_1 + b_2$,

$$b_1 f(a_1/b_1) + b_2 f(a_2/b_2) \geq Bf(A/B);$$

and that in general, for any non-negative $a_1, \ldots, a_k$, $b_1, \ldots, b_k$, and $A = \sum_i a_i$, $B = \sum_i b_i$, we have

$$\sum_i b_i f(a_i/b_i) \geq Bf(A/B).$$

[Hint: since f is convex, for any $\lambda \in [0,1]$ and any $x_1, x_2 > 0$ $\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$; consider $\lambda = b_1/B$.]

(b) Show that $D_f(p\|q) \geq 0$.

(c) Show that D_f satisfies the data processing inequality: for any transition probability kernel $W(v|u)$ from $\mathcal{U}$ to $\mathcal{V}$, and any two distributions p and q on $\mathcal{U}$

$$D_f(p\|q) \geq D_f(\tilde{p}\|\tilde{q})$$

where $\tilde{p}$ and $\tilde{q}$ are probability distributions on $\mathcal{V}$ defined via $\tilde{p}(v) := \sum_u W(v|u)p(u)$, and $\tilde{q}(v) := \sum_u W(v|u)q(u)$,

(d) Show that each of the following are f-divergences.

i. $D(p\|q) := \sum_u p(u)\log(p(u)/q(u))$. [Warning: log is not the right choice for f.]
ii. $R(p\|q) := D(q\|p)$.
iii. $1 - \sum_u \sqrt{p(u)q(u)}$
iv. $\|p - q\|_1$.
v. $\sum_u (p(u) - q(u))^2/q(u)$

Problem 4.8 (Growth of Expected Capital vs Expected Growth of Capital)**.** Suppose $U_1, U_2, \ldots$ are i.i.d. random variables taking values on a finite alphabet $\mathcal{U}$; let $P(u) = \mathbb{P}(U_1 = u)$ denote their common distribution. As in class let $\hat{P}_n$ denote the empirical distribution of U^n.

Suppose $f : \mathcal{U} \to [0, \infty)$ is a non-negative real valued function defined on $\mathcal{U}$. Define now the random variables $X_0, X_1, \ldots$ as $X_0 = 1$, $X_n = f(U_n)X_{n-1}$, $\forall n \geq 1$. In other words

$$X_n = \prod_{i=1}^{n} f(U_i).$$

One refers to the value $R_n = \frac{1}{n}\log X_n$ as the (exponential) *rate of growth* of X_n. (The terminology is motivated by the relationship $X_n = \exp(nR_n)$).

Fix $\alpha = \sum_u P(u)\log f(u) = \mathbb{E}[\log f(U)]$, and for a given $\epsilon > 0$, let

$$A = \left\{Q \in \Pi : \left|\sum_u Q(u)\log f(u) - \alpha\right| < \epsilon\right\}.$$

Let $D^* = \min_{Q \notin A} D(Q\|P)$. Observe that $D^* > 0$.

(a) What can you say about $\mathbb{P}(|R_n - \alpha| \geq \epsilon)$ as n gets large? *Hint:* How are the events $\{|R_n - \alpha| \geq \epsilon\}$ and $\{\hat{P}_n \notin A\}$ related?

(b) Let $\beta = \log \mathbb{E}[f(U)]$. What is the relationship between $e_n = \frac{1}{n}\log \mathbb{E}[X_n]$ and β? Which one of α and β is larger?

In a casino a game of chance is played. The outcome of the game is a random variable U, and if the outcome is u, the money bet on that outcome is multipled by a factor $\phi(u)$. The money bet on other outcomes is lost. The game can be played successively with independent, identically distributed outcomes.

We allocate our capital among the outcomes by placing a fraction $q(u)$ of it on outcome u. Clearly $q(u) \geq 0$ and $Q = \sum_u q(u) \leq 1$. (The fraction $1 - Q$ is the fraction of our capital not bet on the game and kept in reserve.) Observe that $f(u) = (1 - Q) + q(u)\phi(u)$ is the factor our capital is multipled by if the outcome of the game is u.

Let $X_0 = 1$ be our initial capital, and let X_n, $n = 1, 2, \ldots$ denote our capital as we play the game repeatedly with a fixed allocation strategy q.

(c) Suppose $\mathcal{U} = \{0, 1\}$, $P(0) = 1/4$, $P(1) = 3/4$, $\phi(0) = \phi(1) = 2$. What is the allocation q that maximizes the value of β in (b)?

(d) Continuing with (c) and the allocation you just found, what is the value of α? What will happen to our capital X_n in the long run if we repeatedly play the game?

Chapter 5

Multi-Arm Bandits

You are likely already familiar with *supervised* and *unsupervised* learning. In *supervised learning* we are given samples of input-output pairs and are asked to learn from those. In *unsupervised* learning we only have access to input samples. Given those samples we hope to learn about the *structure* of the input. E.g., the perhaps simplest such case is clustering. But there is a third fundamental concept in ML typically referred to as *reinforcement* learning. The main new component here is that we are allowed to *interact* with our *environment* and are supposed to learn from these interactions while at the same time our interactions should also serve the purpose of maximizing some objective. This leads to a fundamental tension between *exploration* versus *exploitation*.

Reinforcement learning is a very large topic. We will explore the simplest such setting, known as bandits. Even on the topic of bandits there is much more to say than what we can cover in this short time. If you want to know more (also about the historical development) we highly recommend the book Bandit Algorithms by Tor Lattimore and Csaba Szepesvári, [2].

5.1 Introduction

The basic model is the following. For each *round* $t = 1, 2, \cdots, n$, the *learner* chooses an *action* A_t from a set of available actions $\mathcal{A}$. To each action $a \in \mathcal{A}$ corresponds a probability distribution P_a. The *environment* receives the chosen action A_t from the learner and in response generates the random variable X_t that is distributed according to P_{A_t} and has mean μ_A.

The *reward* up to and including time n is $\sum_{t=1}^{n} X_t$. The decision by the learner at time t is in general a function of the history $H_{t-1} = \{A_1, X_1, \cdots, A_{t-1}, X_{t-1}\}$. Our aim is to maximize the reward by employing an appropriate learning algorithm.

More precisely, we typically try to *minimize* the *regret* rather than maximize the reward. The regret with respect to a particular action $a \in \mathcal{A}$ is the difference of what we could have gotten if we had used action a in all n rounds versus the actual reward we got. The advantage of this *competitive* view (comparing to some other action) is that this measure is invariant to e.g. shifting all rewards by a constant amount. We typically compute the *worst case* regret, i.e., the regret with respect to the best action we could have taken, and since the reward is a random variable it is common to first average over the reward for each

action. This means we compute

$$R_n = \max_{A \in \mathcal{A}} n\mu_A - \mathbb{E}[\sum_{t=1}^{n} X_t].$$

It is probably not surprising that a good learner will be able to achieve a sublinear worst-case regret, i.e., $R_n = o(n)$. Let us quickly go over the argument. We will do a much more thorough analysis later on. Assume that $|\mathcal{A}| = K$. If we take m samples from each of these K distributions we can compute each mean with an additive error bounded by $c/\sqrt{m}$ with high probability.

Assume that we spend a fraction ϵ of the total time n on learning the K actions and afterwards always play the "best" one according to the derived estimates. In this way we will achieve a regret that behaves like $n\mu^*(\epsilon + cK^{3/2}/\sqrt{\epsilon n})$. The term $n\mu^*\epsilon$ is an upper bound on the regret that we get since for a fraction ϵ of the time (when we are learning) we might have a regret as large as μ^*. The second term, namely $n\mu^* cK^{3/2}/\sqrt{\epsilon n}$ accounts for the fact that during the remaining fraction $1 - \epsilon \leq 1$ of the time, we always play the "best" arm according to our estimates but for each arm the estimate can be off by $c\sqrt{K}/\sqrt{\epsilon n}$ with a fixed probability and so each arm in expectation will contribute a term of this order to the expected regret and we have K arms. We are still free to optimize over the choice of ϵ. If we choose $\epsilon = \frac{\sqrt{cK}}{(2\sqrt{n})^{\frac{2}{3}}}$ then we get $3\sqrt{K}(\frac{cn}{2})^{\frac{2}{3}}$, which vanishes as a function of n. So the interesting question is how *fast* we can make the normalized regret converge to 0.

In the above paragraph we have assumed that we know the *time horizon* n. This is often the case. But also the setting where the time horizon is not known a priori is of interest.

The setting we described, where the rewards come from a distribution that only depends on the chosen action and this distribution is fixed over time is called the *stochastic stationary bandit* problem. We will limit ourselves to this setting.

5.2 Some References

Besides the web page pointed out at the beginning there are many very good references for this topic.

The area goes back to a paper by William R. Thompson [3]. A good recent survey is [4]. If you are looking for a book, we can recommend [5, 6].

5.3 Stochastic Bandits with a Finite Number of Arms

5.3.1 Set-Up

Let us analyze the simplest strategy that we already mentioned in a little bit more detail. It is somewhat easier to think of problems with infinite horizons, i.e., there is no fixed n, but we assume that the game goes on forever.

5.3.2 Explore then Exploit

A sub-optimal but very natural strategy is the following. First, get sufficiently many samples from every bandit in order to determine its mean sufficiently accurately. This is the exploring stage. Then exploit the so gained knowledge and play according to these empirical means.

If we have a bandit with a finite number of arms then it is not very surprising that this strategy achieves a sub-linear regret.

Let us do the calculations. Let $X_1, \cdots, X_m$ be a sequence of iid random variables with mean $\mu = \mathbb{E}[X_i]$. Given the sequence $X_1, \cdots, X_m$, the empirical estimator for μ, call it $\hat{\mu}(X_1, \cdots, X_m)$ is

$$\hat{\mu}(X_1, \cdots, X_m) = \frac{1}{m} \sum_{t=1}^{m} X_t.$$

The above estimate is itself a random variable. Its mean is unbiased, i.e.,

$$\mathbb{E}[\hat{\mu}(X_1, \cdots, X_m)] = \frac{1}{m} \mathbb{E}[\sum_{t=1}^{m} X_t] = \mu.$$

But of course we have a variance. We will use the tail bounds discussed in Section 2.3 for this purpose.

Assume that at the start we get m samples from each of the K bandit arms. Let the expected gain from arm k be μ_k and let $\mu* = \max_{1 \le k \le K} \mu_k$. To simplify notation, let us assume that $\mu* = \mu_1$. Define $\Delta_k = \mu^* - \mu_k$. Finally, assume that each of the K arms corresponds to a random variable that is 1-subgaussian.

After the initial exploration stage we choose the bandit with the largest empirical pay-off for the remaining $n - Km$ steps. This gives us an expected regret of

$$R_n = m \sum_{k=1}^{K} \Delta_k + (n - mK) \sum_{k=1}^{K} \Delta_k \mathbb{P}\{k = \text{argmax}_j \hat{\mu}_j\}.$$

This expression is easy to explain. The first sum on the right is the expected regret due to the exploration stage – we get m samples from each arm, and in so doing, accumulate for each arm an expected regret of $m\Delta_k$.

The second sum on the right accounts for the regret that we accumulate over the remaining $n - mK$ steps in case we choose a sub-optimal arm in the exploitation stage. This second term we can now bound using our tail-bound inequalities. Recall that Δ_k is the regret if we use arm k, instead of the optimum arm 1. We get m samples. What is the probability that the average of the m samples of arm k look better than the average of the m samples of arm 1? This is equivalent to asking for the probability that

$$\mathbb{P}\{\frac{1}{m} \sum_{t=1}^{m} (X_t^{(1)} - X_t^{(k)}) \le 0\},$$

where $X_t^{(1)}$ denotes the m independent samples from arm 1 and $X_t^{(k)}$ denotes the m independent samples from arm k. Now note that by assumption $X_t^{(1)} - \mu_1$ is 1-subgaussian and so is $X_t^{(k)} - \mu_k$. Therefore $X_t^{(1)} - \mu_1 + X_t^{(k)} - \mu_k$ is 2-subgaussian by Lemma 2.2. We therefore have

$$\begin{aligned}
\mathbb{P}\{\frac{1}{m} \sum_{t=1}^{m} (X_t^{(1)} - X_t^{(k)}) \le 0\} &= \mathbb{P}\{\frac{1}{m} \sum_{t=1}^{m} (X_t^{(1)} - \mu_1 - X_t^{(k)} + \mu_k) \le \mu_k - \mu_1\} \\
&= \mathbb{P}\{\frac{1}{m} \sum_{t=1}^{m} (X_t^{(1)} - \mu_1 - X_t^{(k)} + \mu_k) \le -\Delta_k\} \\
&\le e^{-m\Delta_k^2/4}.
\end{aligned}$$

Therefore, our regret can be upper bounded as

$$\begin{aligned} R_n &= m\sum_{k=1}^{K}\Delta_k + (n-mK)\sum_{k=1}^{K}\Delta_k \mathbb{P}\{k = \text{argmax}_j \hat{\mu}_j\} \\ &\leq m\sum_{k=1}^{K}\Delta_k + (n-mK)\sum_{k=1}^{K}\Delta_k \exp\{-\frac{m\Delta_k^2}{4}\}. \end{aligned}$$

It is instructive to consider the special case of $K = 2$. Let Δ be the regret of the second best arm (compared to the best one). Our expression for the regret is then

$$R_n \leq m\Delta + (n-2m)\Delta\exp\{-\frac{m\Delta^2}{4}\} \leq \underbrace{m\Delta + n\Delta\exp\{-\frac{m\Delta^2}{4}\}}_{=\overline{R}_n}.$$

If we assume that we know n and Δ a priori we can find the optimal value of m for the bound $\overline{R}_n$. This leads to the equation

$$\begin{aligned} &\frac{d\overline{R}_n}{dm} = \Delta(1 - ne^{-\frac{m\Delta^2}{4}}\Delta^2/4) = 0, \\ &e^{-\frac{m\Delta^2}{4}}\frac{\Delta^2}{4} = \frac{1}{n}. \end{aligned}$$

We see form this equation that the optimum choice (ignoring integer constraints) is

$$m \sim \frac{4}{\Delta^2}\ln(\frac{n\Delta^2}{4}).$$

This gives us

$$R_n \sim \frac{4}{\Delta}\left(1 + \ln(\frac{n\Delta^2}{4})\right).$$

At first this looks pretty promising. The bound on the right is only logarithmic in n. But there is a slight problem with our bound. So far we have implicitly assumed that Δ is relatively large. But what if Δ is small? E.g., assume that $\Delta = \frac{1}{\sqrt{n}}$. Then the term $\frac{1}{\Delta}$ is equal to $\sqrt{n}$ and so our regret is now much larger. Indeed, what if Δ is even smaller? It seems that the regret has no bound – the smaller the gap the larger the regret. This seems counter-intuitive. Should a small gap not be good for us?

We can easily fix this bound by noting that the regret can never be larger than $n\Delta$. Hence, we have a bound of the form

$$R_n \leq \min\{n\Delta, \frac{4}{\Delta}\left(1 + \ln(\frac{n\Delta^2}{4})\right)\}.$$

Now it is easy to see that the worst case is to have a gap of order $1/\sqrt{n}$. In this case the regret is of order $\sqrt{n}$.

To summarize. If the gap is large (a constant) then we only need $\ln(n)$ samples to figure out which of the arms is best with high probability. After that we will always use the best arm. This gives us a regret of order $\ln(n)$. But if the gap is small then even though pulling the wrong trigger is less costly we need a considerably larger exploration phase. And once

the gap becomes of size $1/\sqrt{n}$, we will spend all the time in exploring and never reach the exploitation phase.

There is another issue. All our previous discussion is based on the assumption that we *know* the horizon n and the gap Δ (just look at the expression for the optimum m – it depends both on n and Δ. Perhaps it is realistic to assume that n is known. But it is not realistic to assume that we know Δ. So is there a way to choose m that is *universal*? We will explore this in the exercises. We will see that there is. But in this case the worst-case regret is of order $n^{\frac{2}{3}}$. Note that this is the same order that we got in our very first back of the envelope calculation.

5.3.3 The Upper Confidence Bound Algorithm

The upper confidence bound (UCB) algorithm is a celebrated algorithm that overcomes the shortcomings of the explore-then-exploit algorithm. Rather than separating the exploring phase from the exploiting phase these two phases are mixed and the algorithm learns continously. The idea is simple: At any point the algorithm gets a sample from that arm that, according to optimistic estimates, looks best.

Recall our upper bound of

$$\mathbb{P}\{\hat{\mu}(X_1, \cdots, X_m) - \mu \geq \epsilon\} \leq \exp(-m\epsilon^2/2).$$

If we set the right-hand side to $\delta > 0$ and then solve for δ we get

$$\mathbb{P}\{\hat{\mu}(X_1, \cdots, X_m) - \mu \geq \sqrt{\frac{2}{m}\ln(\frac{1}{\delta})}\} \leq \delta.$$

If we think of δ as small then this suggests that, at time $t-1$, it is unlikely that our empirical estimator $\hat{\mu}_{k,t-1}$ of the k-th bandit arm overestimates its mean by more than $\sqrt{\frac{2}{T_k(t-1)}\ln(\frac{1}{\delta})}$. Here $T_k(t-1)$ denotes the number of times we have chosen arm k in the first $t-1$ steps.

The idea of the UCB algorithm is to take these upper bounds on the individual confidence intervals as our estimates and to choose as an action A_t at time t that arm i that maximizes this upper bound.

To specify the algorithm it remains to specify the *confidence* level δ_t that is used at time t. We will choose

$$\delta_t = \frac{1}{f(t)} = \frac{1}{1 + t\ln^2(t)}. \tag{5.1}$$

Note that the above algorithm has the following property. Once all arms have been explored at depth all the upper bounds on the confidence intervals will be very close to the true means and so we will likely explore further only arms whose mean is very close to the maximum mean.

Let us now formally specify the algorithm. We have

$$A_t = \begin{cases} t, & t \leq K, \\ \text{argmax}_k \hat{\mu}_k(t-1) + \sqrt{\frac{2\ln f(t)}{T_k(t-1)}}, & t > K. \end{cases}$$

This algorithm is pretty intuitive. Even if a genie had given us the correct mean of the best arm for free, in order to verify that indeed this is the best arm, what we would do is

to compute its confidence interval. And how confident should we be, how should we choose δ? If we make a mistake we will pay linear regret for the remainder of the running time. Therefore δ should be smaller than $\frac{1}{n}$. If we think now of n as t then (5.1) makes sense.

Lemma 5.1. *The regret of the UCB algorithm is bounded by*

$$R_n \leq \sum_{k:\Delta_k>0} inf_{\epsilon\in(0,\Delta_k)}\Delta_k(1+\frac{7}{\epsilon^2}+\frac{2}{(\Delta_k-\epsilon)^2}(\ln f(n)+\sqrt{\pi \ln f(n)}+1)).$$

Let us compare this result to what we have seen for the explore-then-exploit algorithm. If we pick ϵ small but not too small then the dominant terms in this expression are of the form $\frac{2\ln f(n)}{\Delta_k} \sim \frac{2\ln(n)}{\Delta_k}$. This is essentially the same as what we derived for the explore-then-exploit algorithm. But this time we neither required the knowledge of n nor of Δ_k. Of course, we have the same issue when one of the Δ_k becomes small. The worst case is again when one of these gaps is of order $1/\sqrt{n}$. This will, as before, result in a regret of order $\sqrt{n}\ln(n)$. In summary, we have

$$R_n \leq \sum_{k:\Delta_k>0} \min\{n\Delta_k, \inf_{\epsilon\in(0,\Delta_k)}\Delta_k(1+\frac{7}{\epsilon^2}+\frac{2}{(\Delta_k-\epsilon)^2}(\ln f(n)+\sqrt{\pi \ln f(n)}+1))\}.$$

Proof. Let $\hat{\mu}_t$ be the empirical (natural) estimator of the mean of a 1-subgaussian random variable based on t independent observations. Let $a \in \mathbb{R}^+$ and $\epsilon > 0$. Consider the quantity

$$\mathbb{P}\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\}.$$

For t no more than $\frac{2a}{\epsilon^2}$ this probability is very close to 1. But for larger t we can use our tail bound to conclude that

$$\mathbb{P}\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\} \leq e^{-\frac{1}{2}t(\epsilon-\sqrt{\frac{2a}{t}})^2}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_t+\sqrt{\frac{2a}{t}}\geq\epsilon\}}] &= \sum_{t=1}^{n} \mathbb{P}\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\} \\
&\leq \frac{2a}{\epsilon^2} + \sum_{t\geq\frac{2a}{\epsilon^2}}^{n} \mathbb{P}\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\} \\
&\leq \frac{2a}{\epsilon^2} + \sum_{t\geq\frac{2a}{\epsilon^2}}^{n} e^{-\frac{1}{2}t(\epsilon-\sqrt{\frac{2a}{t}})^2} \\
&\overset{(a)}{\leq} \frac{2a}{\epsilon^2} + 1 + \int_{\frac{2a}{\epsilon^2}}^{\infty} e^{-\frac{1}{2}t(\epsilon-\sqrt{\frac{2a}{t}})^2} dt \\
&\overset{(b)}{=} \frac{2a}{\epsilon^2} + 1 + \frac{2}{\epsilon^2}\int_{0}^{\infty} e^{-\frac{1}{2}x^2}(x+\sqrt{2a})dx \\
&= 1 + \frac{2}{\epsilon^2}(a+\sqrt{\pi a}+1). \qquad (5.2)
\end{aligned}$$

In step (a) we note that the terms in the sum are decreasing and that each term is less than 1. We can hence bound the sum by the corresponding integral plus the constant 1 (the maximum value that the function can take on at the left boundary). In step (b) we made two substitutions. First we set $z = \epsilon\sqrt{t}$ so that $dt = 2z/\epsilon^2 dz$. This will change the lower bound to $\sqrt{2a}$ and the argument in the exponent to $-\frac{1}{2}(z - \sqrt{2a})^2$. Then we shift the integration boundaries by defining $x = z - \sqrt{2a}$. This gives us the indicated integral.

Let us now bound the regret, which has the form $R_n = \sum_{k:\Delta_k>0} \Delta_k \mathbb{E}[T_k(n)]$. The key is to find a good bound on $\mathbb{E}[T_k(n)]$. Note that

$$T_k(n) = \sum_{t=1}^{n} \mathbb{1}_{\{A_t=k\}} \leq \sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\ln f(t)}{T_1(t-1)}} \leq \mu_1-\epsilon\}} + \sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_k(t-1)+\sqrt{\frac{2\ln f(t)}{T_k(t-1)}} \geq \mu_1-\epsilon \,\wedge\, A_t=k\}} \tag{5.3}$$

The idea of this bound is the following. Rather than counting how often the upper confidence bound of arm k is larger than the upper confidence bounds of all other arms, we count how often it is larger than the upper confidence bound of arm 1.

Clearly, this count is an upper bound. Further, rather than comparing the upper confidence bound of arm k and arm 1 directly, we compare each individuall to a third quantity. This quantity is chosen to be slightly below the true mean of arm 1. We increase our count if either the upper confidence bound of arm k is above this threshold, or if the upper confidence bound of arm 1 is below this threshold. Again, this leads to an upper bound. This explains the two terms on the right in (5.3).

We start with the first one,

$$
\begin{aligned}
\mathbb{E}[\sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\ln f(t)}{T_1(t-1)}} \leq \mu_1 - \epsilon\}}] &= \sum_{t=1}^{n} \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2\ln f(t)}{T_1(t-1)}} \leq \mu_1 - \epsilon\right) \\
&\overset{(a)}{\leq} \sum_{t=1}^{n}\sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{1,s} + \sqrt{\frac{2\ln f(t)}{s}} \leq \mu_1 - \epsilon\right) \\
&\leq \sum_{t=1}^{n}\sum_{s=1}^{t} e^{-\frac{s}{2}(\sqrt{\frac{2\ln f(t)}{s}}+\epsilon)^2} \\
&= \sum_{t=1}^{n}\sum_{s=1}^{t} e^{-\ln(f(t))-\sqrt{2s\ln f(t)}-\frac{s}{2}\epsilon^2} \\
&\leq \sum_{t=1}^{n} \frac{1}{f(t)} \sum_{s=1}^{t} e^{-\frac{s}{2}\epsilon^2} \\
&= \sum_{t=1}^{n} \frac{1}{f(t)} \frac{e^{-\frac{\epsilon^2}{2}}}{1-e^{-\frac{\epsilon^2}{2}}} \\
&= \sum_{t=1}^{n} \frac{1}{f(t)} \underbrace{\frac{1}{e^{\frac{\epsilon^2}{2}}-1}}_{\text{take Taylor series}} \\
&\qquad\qquad \text{all terms are positive; keep only first two} \\
&\leq \sum_{t=1}^{n} \frac{1}{f(t)} \frac{2}{\epsilon^2} \\
&\leq \frac{2}{\epsilon^2} \sum_{t=1}^{n} \frac{1}{1+t\ln(t)^2} \\
&\overset{(b)}{\leq} \frac{2}{\epsilon^2}(2 + \int_2^{\infty} \frac{1}{x\ln(x)^2} dx) \\
&= \frac{2}{\epsilon^2}(2 + \frac{1}{\ln(2)}) \\
&\leq \frac{7}{\epsilon^2}
\end{aligned}
$$

In step (a) we argue as follows. We do not know how many samples of arm 1 we have taken at time t. Hence we bound this probability via a union bound, where we sum over all possibilities. In step (b) we used the fact that $\frac{1}{1+t\ln(t)^2}$ is a decreasing function so that the sum can be bounded by an appropriately chosen integral. In particular, we note that each term us upper bounded by 1. We hence bound the first two terms by 2 and then we bound the remainder of the sum by the corresponding integral starting at 2 (the sum starts at 3). Finally, we drop the 1 from the denominator and we extend the integral to infinity. This further upper bounds the sum and leads to a simple expression.

It remains to bound the second term in (5.3),

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_k(t-1)+\sqrt{\frac{2\ln f(t)}{T_k(t-1)}} \geq \mu_1 - \epsilon \,\wedge\, A_t = k\}}] &\overset{(a)}{\leq} \mathbb{E}[\sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_k(t-1)+\sqrt{\frac{2\ln f(n)}{T_k(t-1)}} \geq \mu_1 - \epsilon \,\wedge\, A_t = k\}}] \\
&\overset{(b)}{\leq} \mathbb{E}[\sum_{s=1}^{n} \mathbb{1}_{\{\hat{\mu}_{k,s}+\sqrt{\frac{2\ln f(n)}{s}} \geq \mu_1 - \epsilon\}}] \\
&\leq \mathbb{E}[\sum_{s=1}^{n} \mathbb{1}_{\{\hat{\mu}_{k,s}-\mu_k+\sqrt{\frac{2\ln f(n)}{s}} \geq \Delta_k - \epsilon\}}] \\
&\overset{(c)}{\leq} 1 + \frac{2}{(\Delta_k - \epsilon)^2}(\ln f(n)) + \sqrt{\pi \ln f(n)} + 1).
\end{aligned}$$

In step (a) we replaced $f(t)$ by the larger quantity $f(n)$. This gives us an upper bound.

Step (b) also warrants some explanation. As in the previous case, we do not know what $T_k(t-1)$ is, other than that it must be in the range from 1 to $t-1$. When we derived a bound on the first term in (5.3) we got around this problem by taking a union bound over all possible such values.

We could do the same thing here, but this bound would be loose. The trick is to realize that whatever value of s we have for a particular step t, the same value cannot appear again in a later step due to the condition that $A_t = k$. Therefore, it suffices to take the sum over all possible values of s *once*, i.e., *instead* of the sum over t.

Finally, in step (c) we have used (5.2) with $a = \ln(f(n))$ and ϵ replaced by $\Delta_k = \epsilon$. □

5.3.4 Information-theoretic Lower Bound

We have seen that the UCB algorithm has a worst-case (worst-case over the choice of gaps) regret of order $\sqrt{n}\ln(n)$. Could there be an algorithm that is much better than that. We will now see that we cannot hope to do better than $\sqrt{n}$.

So far we have discussed two concrete algorithms. In general, an algorithm is specified by a *policy* π. A policy is a sequence of conditional probabilities that specify the probability of the action at time t given the history $H_{t-1} = \{A_1, X_1, \cdots, A_{t-1}, X_{t-1}\}$. The policies of the explore-then-exploit as well as the UCB algorithm were deterministic (other than perhaps when breaking ties). But in general a policy might be randomized. Recall also our notion of *environment* ν. The environment is the set of K probability distributions $\nu = (\mathbb{P}_1, \cdots, \mathbb{P}_K)$.

Lemma 5.2 (Lower Bound on Worst-Case Regret)**.** *Let $K > 1$ and $n \geq K-1$. Then for any policy π there exists an environment ν so that the regret $R_n(\pi, \nu) \geq \frac{1}{27}\sqrt{(K-1)n}$. Further, this environment can be chosen to be a Gaussian environment, where all distributions are unit-variance Gaussians.*

Proof. The idea of the proof is the following. We are given a policy π. Based on this policy we construct two Gaussian environments that are quite similar and differ only in a single distribution. We then show that the given policy π cannot do well on *both* of these environments. Note that the "bad" enviroment that we prove to exist depends in general not only on the policy but also on n.

Let K be the number of arms, $K > 1$, and let π be given policy. Our first enviroment is Gaussian with unit-variance distributions and a mean vector of the form $(\Delta, 0, \cdots, 0)$, where $\Delta > 0$ is a parameter. We will chose it to be $\sqrt{(K-1)/(4n)}$.

Let $\mathbb{E}_\nu[T_k(n)]$ denote the expected number of times we choose arm k for this environment ν under the policy π (since the policy π is fixed we do not explicitly denote it). Let i, $1 \leq i \leq K$, be an arm that we choose the least often under this policy. More formally, $i = \text{argmin}_k \ \mathbb{E}_\nu[T_k(n)]$. If $i = 1$ then at least $(1 - 1/K)n$ times we do not choose arm 1 and so our regret is at least $(1 - 1/K)n\Delta$, which, for our choice of Δ gives a regret of $(1 - 1/K)n\sqrt{(K-1)/(4n)} \geq \frac{1}{4}\sqrt{(K-1)n} \geq \frac{1}{27}\sqrt{(K-1)n}$.

So let us assume that $i \neq 1$. In this case the second environment ν' is again Gaussian with unit-variance distributions and a mean vector of

$$(\Delta, 0, \cdots, 0, \underbrace{2\Delta}_{i\text{-th component}}, 0, \cdots, 0).$$

Let $p_\nu(A_1, X_1, \cdots, A_n, X_n)$ denote the joint distribution under policy π in environment ν and let $p_{\nu'}(A_1, X_1, \cdots, A_n, X_n)$ denote the joint distribution under policy π in environment ν'. How different are these distributions? Let us compute their KL divergence,

$$D(p_\nu || p_{\nu'}) = \int p_\nu(A_1, X_1, \cdots, A_n, X_n) \ln \frac{p_\nu(A_1, X_1, \cdots, A_n, X_n)}{p_{\nu'}(A_1, X_1, \cdots, A_n, X_n)}.$$

Note that these distributions can be factorized in the following form

$$p_\nu(A_1, X_1, \cdots, A_n, X_n) = \pi(A_1)\pi_\nu(X_1 \mid A_1) \cdots \pi_{H_{n-1}}(A_n)\pi_\nu(X_n \mid A_n).$$

Therefore

$$\begin{aligned}
D(p_\nu || p_{\nu'}) &= \int p_\nu(A_1, X_1, \cdots, A_n, X_n) \ln \frac{p_\nu(A_1, X_1, \cdots, A_n, X_n)}{p_{\nu'}(A_1, X_1, \cdots, A_n, X_n)} \\
&= \int p_\nu(A_1, X_1, \cdots, A_n, X_n) \ln \frac{\pi(A_1)\pi_\nu(X_1 \mid A_1) \cdots \pi_{H_{n-1}}(A_n)\pi_\nu(X_n \mid A_n)}{\pi(A_1)\pi_{\nu'}(X_1 \mid A_1) \cdots \pi_{H_{n-1}}(A_n)\pi_{\nu'}(X_n \mid A_n)} \\
&= \int p_\nu(A_1, X_1, \cdots, A_n, X_n) \ln \frac{\pi_\nu(X_1 \mid A_1) \cdots \pi_\nu(X_n \mid A_n)}{\pi_{\nu'}(X_1 \mid A_1) \cdots \pi_{\nu'}(X_n \mid A_n)} \\
&= \sum_{t=1}^{n} \int p_\nu(A_1, X_1, \cdots, A_t, X_t) \ln \frac{\pi_\nu(X_t \mid A_t)}{\pi_{\nu'}(X_t \mid A_t)} \\
&= \sum_{t=1}^{n} \int \sum_{k=1}^{K} p_\nu(A_t = k) p_\nu(X_t \mid A_t = k) \ln \frac{\pi_\nu(X_t \mid A_t = k)}{\pi_{\nu'}(X_t \mid A_t = k)} \\
&= \sum_{t=1}^{n} \sum_{k=1}^{K} p_\nu(A_t = k) D(P_k, P'_k) \\
&= \sum_{k=1}^{K} \mathbb{E}_\nu[T_k(n)] D(P_k, P'_k) \\
&\overset{(a)}{=} \mathbb{E}_\nu[T_i(n)] \frac{4\Delta^2}{2} \\
&\overset{(b)}{\leq} \frac{n}{K-1} \frac{4\Delta^2}{2} = \frac{2n\Delta^2}{K-1}.
\end{aligned}$$

In step (a) we have used the fact that the two enviroments only differ in position i and that in this position we have two unit-variance Gaussians, one with mean 0 and one with mean

2Δ. As you will show in your homework, if P_i, $i = 1, 2$, are two Gaussians with means μ_i and variances σ_i^2, then

$$D_{KL}(P_1||P_2) = \ln(\sigma_2/\sigma_1) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

To see step (b) note that amongst the $K-1$ arms $2, \cdots, K$, the one that is chosen the least cannot be chosen more than $n/(K-1)$ times. We now have

$$\begin{aligned} R_n(\pi, \nu) + R_n(\pi, \nu') &\overset{(a)}{\geq} \mathbb{P}_\nu\{T_1(n) \leq n/2\}\frac{n\Delta}{2} + \mathbb{P}_{\nu'}\{T_1(n) > n/2\}\frac{n\Delta}{2} \\ &\overset{(b)}{\geq} \frac{n\Delta}{4} e^{-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'})} \\ &\geq \frac{n\Delta}{4} e^{-\frac{2n\Delta^2}{K-1}} \\ &\overset{(c)}{=} \frac{\sqrt{n(K-1)}}{8} e^{-\frac{1}{2}} \\ &\overset{(d)}{=} \frac{2\sqrt{n(K-1)}}{27}. \end{aligned}$$

Before we justify any of these steps note that this inequality completes our proof: We have shown that for a given policy there are two environments so that the sum of their regrets at time n is at least $2\frac{\sqrt{n(K-1)}}{27}$. So at least one of these environments must have a regret of at least $\frac{\sqrt{n(K-1)}}{27}$.

Step (a) is easy to explain. If we choose the arm 1 in environment μ at most half of the time then for $n/2$ time steps we have a regret of Δ for each step. And if we choose the arm 1 in environment ν' more than half of the time then for $n/2$ time steps we have a regret of at least Δ.

In step (c) we made the choice $\Delta = \sqrt{(K-1)/(4n)}$ and in step (d) we lower bounded $e^{-\frac{1}{2}}/8 \sim 0.0758$ by $2/27 \sim 0.07407$.

Step (b) follows from Lemma 4.6. □

5.4 Further Topics

There are many extensions and variations of this topic. Let us quickly mention a few without proofs or details.

5.4.1 Asymptotic Optimality

Assume we are given a fixed policy π.

In Section 5.3.4 proved that if we first fix the time horizon n and then choose an environment we can make the regret as large as $\sqrt{n}$.

But what if we first fix the environment and then let n tend to infinity. We have seen that for the UCB algorithm the asymptotic regret scales logarithmically. Can we do better? It turns out that we cannot as long as we stick to policies that have an asymptotic regret that is upper bounded by n^α, for every $\alpha > 0$, for all environments (E.g., the UCB fulfills this condition). So the UCB algorithm is optimal also in this sense.

5.4.2 Adversarial Bandits

So far we have assumed that the environment consists of K distributions that are unknown but fixed. This is a relatively strong assumption. In the *adversarial bandit* setting we do not assume that the rewards are iid samples from a distribution. Rather, we allow them to be arbitrary numbers $x_{t,k}$ in $[0,1]$.

Assume at first that the policy is deterministic. Then it is clear that we can make the regret equal to n. For any time t, once A_t has been chosen by the policy, lets say it has value k, pick $j \neq k$, and set $x_{t,j} = 1$, and $x_{t,i} = 0$, $i \neq j$.

We can remedy this problem by making the following two changes. First, clearly we need a randomized strategy. Second, we will compare ourselves to a genie who knows all rewards $x_{t,k}$ and who picks that arm whose average reward is largest up to time n (so we do NOT compare to a genie who is allowed to pick at every time t that arm that contains the highest reward at this time).

Perhaps suprisingly, for a randomized stratety and this proper choice of genie we can make the regret almost as small as for the stochastic case. Here, the regret is the expected regret, where the expectation is over the randomness of the algorithm.

Exponential-Weight Algorithm for Exploration and Exploitation

The most common algorithm in this setting is the *Exponential-weight algorithm for Exploration and Exploitation* (Exp3 for short). It is defined as follows.

We start with a uniform distribution on the set of actions, $\mathbb{P}_{t=1,k} = 1/K$, $k = 1, \cdots, K$. At time t, we have computed the distribution $\mathbb{P}_{t,k}$. Sample an action A_t from this distribution, assume it is k. Reveal the sample. It is the number $x_{t,k}$ and we call it X_t. *Estimate* the rewards for all arms based on X_t and then compute $P_{t+1,k}$ by *updating* $P_{t,k}$.

Reward Estimation

Recall that in round t we chose some action A_t according to the distribution $\mathbb{P}_{t,k}$ and then we observed the reward X_t which is the t-reward of arm A_t. Based on this number we would like to estimate the reward of all arms. We use the estimator

$$\hat{x}_{t,k} = \frac{\mathbb{1}_{\{A_t=k\}}}{\mathbb{P}_{t,k}} X_t.$$

This makes sense. We scale each number by one over the probability that we sample it. We have[1]

$$\begin{aligned}\mathbb{E}[\hat{x}_{t,k} \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] &= \mathbb{E}[\frac{\mathbb{1}_{\{A_t=k\}}}{\mathbb{P}_{t,k}} X_t \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] \\ &= \mathbb{E}[\frac{\mathbb{1}_{\{A_t=k\}}}{\mathbb{P}_{t,k}} x_{t,k} \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] \\ &= \frac{x_{t,k}}{\mathbb{P}_{t,k}} \mathbb{E}[\mathbb{1}_{\{A_t=k\}} \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] \\ &= x_{t,k}.\end{aligned}$$

Note that the conditional expectation $\mathbb{E}[\hat{x}_{t,k} \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}]$ is a random variable (it depends on the history) that is constant on the "partitions" given by the conditioning. What the above says is that independent of the history this random variable is in fact equal to $x_{t,k}$. So irrespective of the history, if we repeated the same history many times, the expected regret we would see is always $x_{t,k}$. This makes of course sense since the history only changes $\mathbb{P}_{t,k}$ but by definition we sampled exactly according to this distribution. (In order for things to be well-defined we need to ensure that the probability of sampling a particular k is never 0.)

In the same way we can compute the variance of this estimator. We have

$$\begin{aligned}&\mathbb{E}[\hat{x}_{t,k}^2 \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] - \mathbb{E}[\hat{x}_{t,k} \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}]^2 \\ &= \mathbb{E}[\left(\frac{\mathbb{1}_{\{A_t=k\}}}{\mathbb{P}_{t,k}} X_t\right)^2 \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] - x_{t,k}^2 \\ &= \mathbb{E}[\frac{\mathbb{1}_{\{A_t=k\}}}{\mathbb{P}_{t,k}^2} x_{t,k}^2 \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] - x_{t,k}^2 \\ &= \frac{x_{t,k}^2}{\mathbb{P}_{t,k}^2} \mathbb{E}[\mathbb{1}_{\{A_t=k\}} \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}] - x_{t,k}^2 \\ &= x_{t,k}^2 \frac{1 - \mathbb{P}_{t,k}}{\mathbb{P}_{t,k}}.\end{aligned}$$

From this we see that the variance can be substantial if $\mathbb{P}_{t,k}$ is very small. This can of course cause trouble.

Updating the Probability Distribution

Now that we have discussed how we can estimate the total reward of each arm up to time $t-1$, call this quantity $\hat{S}_{t-1,k}$, we still need to discuss how we convert this estimate into the probability distribution $\mathbb{P}_{t,k}$. One standard way of doing this is to set

$$\mathbb{P}_{t,k} = \frac{e^{\eta \hat{S}_{t-1,k}}}{\sum_j e^{\eta \hat{S}_{t-1,j}}},$$

[1] Recall the definition of conditional expectation. Let X be a random variable defined on a σ-algebra $\mathcal{F}$. Let $\mathcal{G}$ be a *sub σ-algebra*. Then the conditional $\mathbb{E}[X \mid \mathcal{G}]$ is the unique random variable that is $\mathcal{G}$-measureable and so that for all $G \in \mathcal{G}$

$$\int_G \mathbb{E}[X \mid \mathcal{G}] d\mathbb{P} = \int_G X d\mathbb{P}$$

where η is a parameter that we can choose freely.

Lemma 5.3 (Regret of Exp3)**.** *For any assignment of the rewards $x_{t,k} \in [0,1]$ the expected regret of the Exp3 algorithm is bounded as*

$$R_n \leq 2\sqrt{nK\ln(K)}.$$

5.4.3 Contextual Bandits

In many scenarios we have some side information available. How can we use this information to improve our choice. One idea is to define a *context*. E.g., perhaps we built a movie recommedation site. In this case we might have a prior classification of various user "types." This could be the context.

Assume that there are a finite number of contexts. Then we could define one bandit algorithm for each of the finite number of contexts and run them independently. But we pay a price. Now each algorithm only sees a fraction of the examples!

5.5 Problems

Problem 5.1 (KL Divergence)**.** Compute the KL Divergence of two scalar Gaussians $p(x) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) = \mathcal{N}(\mu_2, \sigma_2^2)$.

Problem 5.2 (Epsilon-Greedy Algorithm)**.** Recall our original *explore-then-exploit* strategy. We had a fixed time horizon n. For some m, a function of n and the gaps $\{\Delta_k\}$, we explore each of the K arms m times initially. Then we pick the best arm according to their empirical gains and play this arm until we reach round n. We have seen that this strategy achieves an asymptotic regret of order $\ln(n)$ if the environment is fixed and we think of n tending to infinity but a worst-case regret of order $\sqrt{n}$ if we use the gaps when determining m and of order $n^{\frac{2}{3}}$ if we do not use the gaps in order to determine m.

Here is a slightly different algorithm. Let $\epsilon_t = t^{-\frac{1}{3}}$. For each round $t = 1, \ldots$, toss a coin with success probability ϵ_t. If success, then explore arms uniformly at random. If not success, then pick in this round the arm that currently has the highest empirical average.

Show that for this algorithm the expected regret at *any* time t is upper bounded by $t^{\frac{2}{3}}$ times terms in t and K of lower order. This is a similar to the worst-case of the explore-then-exploit strategy but here we do not need to know the horizon a priori. Assume that the rewards are in $[0,1]$.

Problem 5.3 (Upper Confidence Bound Algorithm)**.** In the course we analyzed the Upper Confidence Bound algorithm. As was suggested in the course, we should get something similar if instead we use the Lower Confidence Bound algorithm. It is formally defined as follows.

$$A_t = \begin{cases} t, & t \leq K, \\ \arg\max_k \hat{\mu}_k(t-1) - \sqrt{\frac{2\ln f(t)}{T_k(t-1)}}, & t > K. \end{cases}$$

Analyze the performance of this algorithm in the same way as we did this in the course for the UCB algorithm.

Hint: Is this algorithm well designed?

Problem 5.4 (Thompson Sampling with Bernoulli Losses)**.** This problem deals with a Bayesian approach to multi-arm bandits. Although we will not pursue this facet in the current problem, the Bayesian approach is useful since within this framework it is relatively easy to incorporate prior information into the algorithm.

Assume that we have K bandits, and that bandit k outputs a $\{0,1\}$-valued Bernoulli random variable with parameter $\theta_k \in [0,1]$. Let π be the uniform prior on $[0,1]^K$, i.e., the uniform prior on the set of all parameters $\theta = (\theta_1, \cdots, \theta_K)$. Let

$$T_k^1(t) = |\{\tau \leq t : A_\tau = k; Y_\tau = 1\}|,$$
$$T_k^0(t) = |\{\tau \leq t : A_\tau = k; Y_\tau = 0\}|.$$

In words, $T_k^1(t)$ is the number of times up to and including time t that we have chosen action k and the output of arm k was 1 and similarly $T_k^0(t)$ is the number of times up to and including time t that we have choses action k and the output of the arm k was 0.

The goal is to find the arm with the highest parameter, i.e., the goal is to determine

$$k^* = \text{argmax}_k \theta_k.$$

In the Bayesian approach we proceed as follows. At time time t:

1. Compute for each arm k the distribution $p(\theta_k(t)|T_k^1(t-1), T_k^0(t-1))$.
2. Generate samples of these parameters according to their distributions.
3. Pick the arm j with the largest sample.
4. Observe the output of the j-th arm, call it $Y_j(t)$, and update the counters T_j^1 and T_j^0 accordingly.

Show that this algorithm "works" in the sense that eventually it will pick the best arm. More precisely, show the following two claims.

1. Show that $p(\theta_k(t)|T_k^1(t-1), T_k^0(t-1))$ is a Beta distributed and determine α and β.
2. Show that as t tends to infinity the probability that we choose the correct arm tends to 1. [HINT: To simplify your life, you can assume that for every arm k, $T_k^1(t-1) + T_k^0(t-1) \overset{t\to\infty}{\to} \infty$.]

NOTE: Recall that the density of the Beta distribution on $[0,1]$ with parameters α and β is equal to

$$f(x;\alpha,\beta) = \text{constant } x^{\alpha-1}(1-x)^{\beta-1}.$$

Further, the expected value of $f(x;\alpha,\beta)$ is $\frac{\alpha}{\alpha+\beta}$ and its variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Problem 5.5 (Bandits with Infinitely Many Arms)**.** In the course we considered bandits with a finite number of K arms. In this problem we will see that the same ideas apply if we have infinitely many arms as long as there is some additional structure.

Assume that there is an unknown unit-norm vector $\theta \in \mathbb{R}^d$. For every unit-norm vector $u \in \mathbb{R}^d$, there is a bandit. It gives the reward $X_u = \langle u, \theta \rangle + Z_u$, where Z_u is a zero-mean unit-variance Gaussian that is independent over time and independent with respect to different bandits. The nature of the reward is known to the player.

Find a policy, i.e., a strategy of what bandit to probe at any given point in time given a specific history, that has a sublinear regret as time tends to infinity. You can assume that you know the horizon, i.e., we are looking for fixed-horizon policies.

Problem 5.6 (Lipschitz Bandits)**.** Assume for the following that you have a bandit algorithm at your disposal that has an expected regret, call it R_n, bounded by $c\sqrt{Kn\log(n)}$, where K is the number of arms and n is the time horizon.

You have to design an algorithm for the following scenario. There are infinitely many bandits. More precisely the bandits are indexed by x, $x \in [0,1]$. Bandit x has mean $\mu(x)$ (which is unknown). But you do know that the various bandits are related in the sense that

$$|\mu(x) - \mu(y)| \leq L|x - y|, \tag{5.4}$$

where L is a known constant. This is known as the Lipschitz bandit problem due to the Lipschitz condition (5.4).

A natural approach to such a bandit problem is to discretize the space of bandits. I.e., assume that you pick K positions $0 \leq x_1 < x_2 < \cdots < x_K \leq 1$ and run your given bandit problem on these K bandits.

a) Bound the expected regret as a function of K, n, L and the placement of points.

b) For n and L fixed, minimize your expression with respect to K and the placement of points.

Hint: In order to simplify your computation, you might want to slightly loosen your bound.

Chapter 6

Detection and Estimation

There are two basic tasks that we will review in this chapter. The first one is known as *detection*. Synonyms are *decision making* and *hypothesis testing*. The second topic is *estimation*.

These two tasks are the equivalent of *classification* and *regression* in ML, except that we know the underlying distribution. Therefore, detection and estimation give us a guide of what we might hope to achieve in a particular setting (assuming that we have a large amount of data). Of course, in a typical setting in ML we do not know the distribution.

In particular estimation is very much at the intersection of all the topics we discuss in this course. It is a form of statistical signal processing. Its fundamental bounds are given by information theory. And, as we just mentioned above, it relates to the basic regression task of data science. And in the typical case where we do not know the underlying distribution one possible strategy is to estimate this distribution (another topic of this course) given the data and then to predict the value according to the rules of estimation theory.

6.1 Detection

6.1.1 Binary hypothesis testing

We start with the simplest set-up, namely *binary* hypothesis testing. Consider the problem of deciding which of two hypotheses, hypothesis 0 or hypothesis 1, is true, based on an observation U. The observation U is a random variable taking values in an alphabet $\mathcal{U}$ — a finite set of $K = |\mathcal{U}|$ letters — and under hypothesis j it has distribution P_j. To avoid trivial cases we will assume that for each $u \in \mathcal{U}$ both $P_0(u)$ and $P_1(u)$ are strictly positive. Otherwise, if we observe a u with, say, $P_0(u) = 0$, we would know for sure that hypothesis 1 is true.

A deterministic decision rule associates to each $u \in \mathcal{U}$ a binary value — i.e., the rule is a function $\phi : \mathcal{U} \to \{0, 1\}$ — and we decide in favor of hypothesis $\phi(u)$ if the observation U equals u. In general, we will allow for randomized decision rules: such a rule is characterized by a function $\phi : \mathcal{U} \to [0, 1]$ that associates to each $u \in \mathcal{U}$ a value in the *interval* $[0, 1]$, that gives the probability of deciding in favor of hypothesis 1. If our observation U equals u, we flip a coin that comes heads with probability $\phi(u)$ and tails with probability $1 - \phi(u)$, and decide accordingly: 1 if heads, 0 if tails. We will identify a decision rule with the function ϕ.

In this set up there are two kinds of error: deciding 1 when the true hypothesis is 0, and deciding 0 when the true hypothesis is 1. For a fixed rule ϕ let $\pi_\phi(i|j)$ denote the

probability of deciding i when the truth is j. We have

$$\pi_\phi(0|1) = \sum_u P_1(u)[1-\phi(u)], \quad \pi_\phi(1|0) = \sum_u P_0(u)\phi(u).$$

Given P_0 and P_1 and a positive real number $\eta > 0$, let Φ_η to be the set of decision rules ϕ of the form

$$\phi(u) = \begin{cases} 1 & \text{if } P_1(u) > \eta P_0(u) \\ 0 & \text{if } P_1(u) < \eta P_0(u). \end{cases} \tag{6.1}$$

Note that if there is no u for which $P_1(u) = \eta P_0(u)$, the test ϕ is uniquely specified and Φ_η contains only this test.

Lemma 6.1. *The rules in Φ_η are minimizers of $\pi(0|1) + \eta\pi(1|0)$.*

Proof. For any rule $\phi \in \Phi_\eta$, as a consequence of (6.1), for every $u \in \mathcal{U}$

$$P_1(u)[1-\phi(u)] + \eta P_0(u)\phi(u) = \min\{P_1(u), \eta P_0(u)\}.$$

Thus for any rule $\phi \in \Phi_\eta$

$$\pi_\phi(0|1) + \eta\pi_\phi(1|0) = \sum_u P_1(u)[1-\phi(u)] + \eta P_0(u)\phi(u) = \sum_u \min\{P_1(u), \eta P_0(u)\}.$$

Suppose now ψ is any decision rule. The lemma follows by noting that

$$\pi_\psi(0|1) + \eta\pi_\psi(1|0) = \sum_u P_1(u)[1-\psi(u)] + \eta P_0(u)\psi(u) \geq \sum_u \min\{P_1(u), \eta P_0(u)\}. \quad \square$$

Theorem 6.2. *For any $\alpha \in [0,1]$, (i) there is a rule ϕ of the form (6.1) such that $\pi_\phi(0|1) = \alpha$, and (ii) for any decision rule ψ either $\pi_\psi(0|1) \geq \pi_\phi(0|1)$ or $\pi_\psi(1|0) \geq \pi_\phi(1|0)$.*

Proof. Assertion (ii) follows from the lemma above: a ψ that violates both the inequalities would contradict the lemma. It thus suffices to prove (i), the existence of a rule ϕ of the form (6.1) with $\pi_\phi(0|1) = \alpha$. To that end, define $L(u) = P_1(u)/P_0(u)$, and label the elements of $\mathcal{U}$ as $\mathcal{U} = \{u_1, \ldots, u_K\}$ such that $L(u_1) \geq L(u_2) \geq \cdots \geq L(u_K)$. Now define, $a_i = \sum_{j=1}^{i} P_1(u_j)$ for $i = 0, \ldots, K$. We then have $0 = a_0 < a_1 < \cdots < a_K = 1$. Given $0 \leq \alpha \leq 1$, we can find $1 \leq i \leq K$ for which $a_{i-1} \leq 1-\alpha \leq a_i$, so that $1-\alpha = (1-\rho)a_{i-1} + \rho a_i$ for some $\rho \in [0,1]$. Then, the rule

$$\phi(u) = \begin{cases} 1 & u \in \{u_1, \ldots, u_{i-1}\} \\ \rho & u = u_i \\ 0 & u \in \{u_{i+1}, \ldots, u_K\} \end{cases}$$

is of the form (6.1) with $\eta = L(u_i)$, and $\pi_\phi(0|1) = \alpha$. $\square$

Rules of the form (6.1) are based on a *likelihood ratio test*: they compare the likelihood ratio $P_1(u)/P_0(u)$ to a threshold η to make a decision. If the likelihood ratio is larger than the threshold, decide 1; if less, decide 0. Equivalently one may compare the *log likelihood ratio*, $\log(P_1(u)/P_0(u))$ to the threshold $\log\eta$.

The theorem stated just above shows the dominant nature of likelihood ratio tests in making decisions: given any decision rule ψ, we can find a (log) likelihood ratio test ϕ which is 'as good or better' — in the sense that the two error probabilities satisfy $\pi_\phi(0|1) \leq \pi_\psi(0|1)$ and $\pi_\phi(1|0) \leq \pi_\psi(1|0)$.

6.1.2 Hypothesis testing with repeated independent observations

Suppose now that we make repeated independent observations of U. That is, we observe a sequence $U_1, \ldots, U_n$ of independent and identically distributed (i.i.d.) random variables, with common distribution P_i under hypothesis i, for $i = 0, 1$.

The log likelihood ratio tests for this scenario are of the form

$$\phi(u_1, \ldots, u_n) = \begin{cases} 1 & \Lambda_n(u_1, \ldots, u_n) > t \\ 0 & \Lambda_n(u_1, \ldots, u_n) < t \end{cases}$$

where

$$\Lambda_n(u_1, \ldots, u_n) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_1(u_i)}{P_0(u_i)}$$

is the normalized log likelihood ratio for the observation $u_1, \ldots, u_n$.

If hypothesis 0 is true, then $U_1, \ldots, U_n$ are i.i.d. random variables with distribution P_0, and, by the law of large numbers

$$\Lambda_n(U_1, \ldots, U_n) \to E_0\left[\log \frac{P_1(U_1)}{P_0(U_1)}\right] = \sum_u P_0(u) \log \frac{P_1(u)}{P_0(u)}$$

as n gets large. In the expression above, the subscript 0 to the expectation operator indicates that the expectation is taken with the distribution of the U_i's given by P_0. Similarly, if hypothesis 1 is true,

$$\Lambda_n(U_1, \ldots, U_n) \to E_1\left[\log \frac{P_1(U_1)}{P_0(U_1)}\right] = \sum_u P_1(u) \log \frac{P_1(u)}{P_0(u)}$$

as n gets large.

We of course recognize the above two quantities as $D(P_0 \| P_1)$ and $D(P_1 \| P_0)$.

Thus, as n gets large $\Lambda_n(U_1, \ldots, U_n)$ concentrates around $-D(P_0 \| P_1) \leq 0$ under hypothesis 0 and, concentrates around $D(P_1 \| P_0) \geq 0$ under hypothesis 1. One expects that the threshold t will be chosen to lie between $-D(P_0 \| P_1)$ and $D(P_1 \| P_0)$ so that under either hypothesis, making a wrong decision becomes a large deviations event — an event that the empirical average of a collection of i.i.d. random variables deviates significantly from its expected value.

6.2 Estimation

6.2.1 MMSE Estimation

Consider two (real- or complex-valued) random vectors $\mathbf{D}$ and $\mathbf{X}$ with known joint probability density function $p_{\mathbf{D},\mathbf{X}}$. Suppose that using only $\mathbf{X}$, we are tasked to construct an estimate of $\mathbf{D}$. This estimate is thus a function $g(\mathbf{x})$, to be selected optimally. A natural criterion for choosing the estimator is to find that function $g(\cdot)$ that minimizes the so-called standard mean-squared error

$$\mathbb{E}\left[\|\mathbf{D} - g(\mathbf{X})\|^2 \middle| \mathbf{X} = \mathbf{x} \right]. \tag{6.2}$$

In words, for a given input $\mathbf{X} = \mathbf{x}$, we choose the estimate so that the average of the squared error is as small as possible.

This problem has a nice and intuitively pleasing solution, pick $g(\mathbf{x})$ to be the mean of $\mathbf{D}$ given the observation $\mathbf{x}$, i.e.,

$$g(\mathbf{x}) \quad = \quad \mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right]. \tag{6.3}$$

We will use the shorthand notation $\hat{\mathbf{D}}_{MMSE}(\mathbf{X}=\mathbf{x})$ for this optimal estimator (optimal in the mean-squared error sense). To see that this estimator is optimal write

$$\mathbb{E}\left[\left\|\mathbf{D}-g(\mathbf{X})\right\|^2\middle|\,\mathbf{X}=\mathbf{x}\right] = \mathbb{E}\left[\left\|(\mathbf{D}-\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right])+(\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right]-g(\mathbf{X}))\right\|^2\middle|\,\mathbf{X}=\mathbf{x}\right] \tag{6.4}$$

Expand now the right hand term into three parts as

$$\mathbb{E}\left[\left\|\mathbf{D}-\mathbb{E}\left[\mathbf{D}\middle\|\,\mathbf{X}=\mathbf{x}\right]\right\|^2\middle|\,\mathbf{X}=\mathbf{x}\right]+ \tag{6.5}$$

$$2\mathrm{Re}\quad\left\{\mathbb{E}\left[\left\langle\mathbf{D}-\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right],\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right]-g(\mathbf{X})\right\rangle\middle|\,\mathbf{X}=\mathbf{x}\right]\right\}+ \tag{6.6}$$

$$\mathbb{E}\left[\left\|\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right]-g(\mathbf{X})\right\|^2\middle|\,\mathbf{X}=\mathbf{x}\right]. \tag{6.7}$$

To prove the claim observe that the middle part is zero and the third terms is non-negative. To see that the middle part is zero, note that the whole left side $\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right]-g(\mathbf{X})$ is X-measureable and hence we can take this part out of the conditional expectation. The remaining expression on the left evaluates to zero if we take the expectation and hence the whole inner product is zero.

The optimality of this estimator is used for example in Section 7.1.7.

Example 6.1 (Gaussian signal and noise)**.** Let D be a real-valued zero-mean unit-variance Gaussian random variable. Let $X=D+Z$, where Z is a zero-mean Gaussian random variable of variance σ^2. Then,

$$\begin{aligned}\hat{D}_{MMSE}(X=x) &= \mathbb{E}\left[D|X=x\right]=\int_{-\infty}^{\infty}\mathrm{d}p_{D|X}(\mathrm{d}|x)d\mathrm{d}=\int_{-\infty}^{\infty}\mathrm{d}\frac{p_{X|D}(x|\mathrm{d})p_D(\mathrm{d})}{p_X(x)}d\mathrm{d}\\ &= \int_{-\infty}^{\infty}\mathrm{d}\frac{\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mathrm{d})^2}{2\sigma^2})\frac{1}{\sqrt{2\pi}}\exp(-\frac{\mathrm{d}^2}{2})}{\frac{1}{\sqrt{2\pi(1+\sigma^2)}}\exp(-\frac{x^2}{2(1+\sigma^2)})}d\mathrm{d} \qquad (6.8)\\ &= \int_{-\infty}^{\infty}\mathrm{d}\frac{1}{\sqrt{2\pi\rho^2}}e^{-\frac{(\mathrm{d}-\mu)^2}{2\rho^2}}d\mathrm{d} \quad = \quad \mu, \qquad (6.9)\end{aligned}$$

where $\mu=\frac{1}{1+\sigma^2}x$ and $\rho^2=\frac{\sigma^2}{1+\sigma^2}$. Hence, the mean-squared error incurred by this optimum estimator is

$$\mathbb{E}\left[\left|D-\hat{D}_{MMSE}(X)\right|^2\right] \quad = \quad \rho^2=\frac{\sigma^2}{1+\sigma^2}. \tag{6.10}$$

In this example, the conditional expectation $\mathbb{E}\left[D|X=x\right]=\frac{1}{1+\sigma^2}x$ is a *linear* function of the observation x. Extending the calculation performed in this example, one can establish that whenever $\mathbf{D}$ and $\mathbf{X}$ are jointly Gaussian random vectors, then the conditional expectation can be written as $\mathbb{E}\left[\mathbf{D}\middle|\,\mathbf{X}=\mathbf{x}\right]=A\mathbf{x}$ for a matrix A. That is, in this case, the conditional expectation is again a *linear* function of the observation $\mathbf{x}$. Once we know this fact, then finding the optimal matrix A is not difficult. We do this in the following section.

6.2.2 Linear MMSE Estimation

In a slight variation of the consideration, let us now assume that the estimator must be linear (with fixed coefficients, independent of the data). Let us first consider the case where the desired data D is scalar. That is, we seek to find

$$\hat{D}_{LMMSE}(\mathbf{X}) \quad = \quad \mathbf{w}^T\mathbf{X}, \tag{6.11}$$

where $\mathbf{w}$ is a fixed vector of coefficients. In the MMSE perspective, we strive to select this vector such as to minimize

$$\mathbb{E}\left[\left|D - \hat{D}_{LMMSE}(\mathbf{X})\right|^2\right]. \tag{6.12}$$

To express the solution, it is convenient to introduce the notation

$$R_{\mathbf{X}} \quad = \quad \mathbb{E}[\mathbf{X}\mathbf{X}^H] \tag{6.13}$$

for the *covariance matrix* of the data (without essential loss in generality we assume zero-mean signals throughout this section), and

$$\mathbf{r}_{D\mathbf{X}} \quad = \quad \mathbb{E}[D\mathbf{X}^*] \tag{6.14}$$

for the covariance between the desired and the observed data. With this, the optimal coefficients, called the *Wiener coefficients* are (assuming that the matrix $R_{\mathbf{X}}$ is invertible — for the more general case, see the homework)

$$\mathbf{w} \quad = \quad R_{\mathbf{X}}^{-1}\mathbf{r}_{D\mathbf{X}}, \tag{6.15}$$

and the corresponding mean-squared error can be expressed as

$$\mathbb{E}\left[\left|D - \hat{D}_{LMMSE}(\mathbf{X})\right|^2\right] \quad = \quad \sigma_D^2 - \mathbf{r}_{D\mathbf{X}}^H R_{\mathbf{X}}^{-1}\mathbf{r}_{D\mathbf{X}}, \tag{6.16}$$

where σ_D^2 denotes the variance of the desired data D.

To prove this, it is instructive to observe that with the optimal coefficient vector $\mathbf{w}$, the error must be orthogonal to the observed data, which (here) means that

$$\mathbb{E}[(D - \mathbf{w}^H\mathbf{X})\mathbf{X}^H] \quad = \quad \mathbf{0}^H. \tag{6.17}$$

This can be established in various ways: (i) we can observe that the objective function is convex and find the gradient with respect to the coefficient vector; (ii) we can observe that random variables are just functions and that the expectation $\mathbb{E}[XY^*]$ gives rise to a valid inner product; hence we are working in a Hilbert space (see Section 9.2) and we can invoke the orthogonality principle from this setting; or, (iii), we can proceed with our standard trick and expand

$$\begin{aligned}\mathbb{E}[\left|D - \mathbf{w}^T\mathbf{X}\right|^2] \quad &= \quad \mathbb{E}[\left|D - \mathbf{w}_\perp^T\mathbf{X} + \mathbf{w}_\perp^T\mathbf{X} - \mathbf{w}^T\mathbf{X}\right|^2] \\ &= \quad \mathbb{E}[\left|D - \mathbf{w}_\perp^T\mathbf{X}\right|^2] + \underbrace{2\mathrm{Re}\,\mathbb{E}[\langle D - \mathbf{w}_\perp^T\mathbf{X}, (\mathbf{w}_\perp - \mathbf{w})^T\mathbf{X}\rangle]}_{=0} + \underbrace{\mathbb{E}[\left|(\mathbf{w}_\perp - \mathbf{w})^T\mathbf{X}\right|^2]}_{\geq 0}\end{aligned} \tag{6.18}$$

where $\mathbf{w}_\perp^T$ is the vector chosen according to this orthogonality principle.

The orthogonality condition can be rewritten as

$$\mathbb{E}[D\mathbf{X}^H] - \mathbf{w}^T\mathbb{E}[\mathbf{X}\mathbf{X}^H] \;=\; \mathbf{0}^H. \tag{6.19}$$

Assuming that the matrix $\mathbb{E}[\mathbf{X}\mathbf{X}^H]$ is invertible, this implies the claimed formula.

The corresponding incurred mean-squared error can be calculated as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|D - \hat{D}_{LMMSE}(\mathbf{X})\right\|^2\right] &= \mathbb{E}\left[\left(D - \mathbf{w}^T\mathbf{X}\right)^*\left(D - \mathbf{w}^T\mathbf{X}\right)\right] \\ &= \mathbb{E}\left[\left(D - \mathbf{w}^T\mathbf{X}\right)^* D\right] - \mathbf{w}^T \underbrace{\mathbb{E}\left[\left(D - \mathbf{w}^T\mathbf{X}\right)^*\mathbf{X}\right]}_{=\mathbf{0}\text{, due to orthogonality}} \\ &= \mathbb{E}\left[|D|^2\right] - \mathbf{w}^H\mathbb{E}\left[\mathbf{X}^*D\right], \end{aligned} \tag{6.20}$$

and if we plug in the formula for the optimal Wiener solution for $\mathbf{w}$, we obtain the claimed formula.

Recall that estimation is equivalent to regression. Hence the equivalent setting to the current one in data science least squares. If you go back to your ML notes you will see that we encountered the equivalent of orthogonality condition (6.17) when we derived the solution, except that in ML we take the empirical quantities since we do not have access to the actual distribution.

More precisely, and adapting to the notation that we use here our cost function was given by

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2N}\sum_{n=1}^{N}(D_n - \mathbf{w}^H\mathbf{X}_n)^2,$$

where $\{(D_n, \mathbf{X}_n\}$ are the samples and where we typically assumed that all quantities are real-valued. If we take the gradient of this expression wrt $\mathbf{w}$ we get

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N}(D_n - \mathbf{w}^H\mathbf{X}_n)\mathbf{X}_n = 0.$$

We see that this is the equivalent to the orthogonality condition where we take empirical quantities rather than expectations.

6.3 Wiener Filtering, Smoothing, Prediction

The tools of signal processing are often most powerful if we consider (long) sequences of data. That is, we now suppose that we have a time-domain signal $D[n]$ (where n ranges over integers), and the observed data is $X[n]$. In the world view of signal processing, we would then form an estimate of the form

$$\hat{D}[n] \;=\; \sum_{k=-p_0}^{p_1} w[k]X[n-k], \tag{6.21}$$

where p_0 and p_1 are non-negative integers. Defining the vector (of length $p_0 + p_1 + 1$)

$$\mathbf{X}[n] \;=\; (X[n+p_0], X[n+p_0-1], \ldots, X[n], \ldots, X[n-p_1+1], X[n-p_1])^T \tag{6.22}$$

and the vector $\mathbf{w}$ containing the corresponding $p_0 + p_1 + 1$ filter coefficients, namely,

$$\mathbf{w} = (w[-p_0], w[-p_0+1], \ldots, w[0], \ldots, w[p_1-1], w[p_1])^T, \tag{6.23}$$

we can express the optimal coefficients as

$$\mathbf{w}^T = \mathbb{E}[D[n]\mathbf{X}[n]^H] \left(\mathbb{E}[\mathbf{X}[n]\mathbf{X}[n]^H]\right)^{-1}, \tag{6.24}$$

which, in general, depends on n. This motivates the definition of *wide-sense stationary* random processes that you have encountered in earlier classes. For such processes, we have that

$$\mathbb{E}[X[n]X^*[n-k]] = R_X[k], \tag{6.25}$$
$$\mathbb{E}[D[n]X^*[n-k]] = R_{DX}[k], \tag{6.26}$$

that is, these expectations do not depend on n, but only on the "lag" k between the two arguments. With this, it can easily be verified that the above formula for the optimal coefficient vector $\mathbf{w}$ does not depend on n.

An equally enlightening but alternative view is to allow p_0 and p_1 to be infinite. In this case, the orthogonality principle stipulates that the optimum filter coefficients must satisfy

$$\mathbb{E}\left[\left(D[n] - \sum_{k=-\infty}^{\infty} w[k]X[n-k]\right)X^*[n-\ell]\right] = 0, \tag{6.27}$$

for all integers ℓ. Rewriting,

$$\mathbb{E}\left[D[n]X^*[n-\ell]\right] - \sum_{k=-\infty}^{\infty} w[k]\mathbb{E}\left[X[n-k]X^*[n-\ell]\right] = 0, \tag{6.28}$$

or

$$R_{DX}[\ell] - \sum_{k=-\infty}^{\infty} w[k]R_X[\ell-k] = 0, \tag{6.29}$$

where we observe that the sum is a convolution. This suggests that it may be instructive to take Fourier transforms:

$$S_{DX}(e^{j\omega}) - W(e^{j\omega})S_{XX}(e^{j\omega}) = 0. \tag{6.30}$$

6.4 Adaptive Filters

Let us consider the scenario where

$$\hat{D}[n] = \sum_{k=0}^{p} w[k]X[n-k]. \tag{6.31}$$

Suppose that we pick an arbitrary initial choice of filter coefficients $\mathbf{w}_0$. Let us take the perspective that we gradually update these filter coefficients so as to make them better. A

classical choice is called *gradient descent.* Here, we consider the gradient of the error (with respect to the filter coefficients), which is easily found to be

$$\nabla_{\mathbf{w}_n} \mathbb{E}\left[\left|D[n] - \mathbf{w}_n^T \mathbf{X}[n]\right|^2\right] = -2\mathbb{E}\left[\left(D[n] - \mathbf{w}_n^T \mathbf{X}[n]\right) \mathbf{X}^*[n]\right] \tag{6.32}$$

The idea is to take a ("small") step *against* the gradient, i.e.,

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu\mathbb{E}\left[\left(D[n] - \mathbf{w}_n^T \mathbf{X}[n]\right) \mathbf{X}^*[n]\right], \tag{6.33}$$

where the step-size parameter μ is to be chosen wisely.

To gain some understanding of what this algorithm does, let us consider the special case where the signals $D[n]$ and $X[n]$ are jointly wide-sense stationary, and hence, all expected values above do not depend on n. For this special case, the update equation becomes

$$\begin{aligned} \mathbf{w}_{n+1} &= \mathbf{w}_n + \mu\mathbb{E}\left[\left(D - \mathbf{w}_n^T \mathbf{X}\right) \mathbf{X}^*\right] && (6.34)\\ &= \mathbf{w}_n + \mu\left(\mathbf{r}_{D\mathbf{X}} - R_{\mathbf{X}}\mathbf{w}_n\right) && (6.35) \end{aligned}$$

If we plug in $\mathbf{w}_n = R_{\mathbf{X}}^{-1}\mathbf{r}_{D\mathbf{X}}$ (which is the optimal Wiener solution), then the algorithm will not move any further, and thus, will stay at the globally optimal solution, which is a first important sanity check. In more detail, we can also suppose that we start the algorithm with an arbitrary $\mathbf{w}_0$. Let us denote the optimal Wiener solution by $\overline{\mathbf{w}}$. Then, we can express

$$\begin{aligned} \mathbf{w}_{n+1} - \overline{\mathbf{w}} &= \mathbf{w}_n - \overline{\mathbf{w}} + \mu\mathbf{r}_{D\mathbf{X}} - R_{\mathbf{X}}\mathbf{w}_n \\ &= \left(I - \mu R_{\mathbf{X}}\right)\left(\mathbf{w}_n - \overline{\mathbf{w}}\right), && (6.36) \end{aligned}$$

where, for the last step, we have used the fact that the Wiener solution satisfies $\mathbf{r}_{D\mathbf{X}} = R_{\mathbf{X}}\overline{\mathbf{w}}$. It is then instructive to express the matrix in terms of its spectral decomposition $R_{\mathbf{X}} = U\Lambda U^H$, leading to $p+1$ independent recursions. Specifically, defining the notation $\mathbf{u}_n = U^H\left(\mathbf{w}_n - \overline{\mathbf{w}}\right)$, we obtain

$$\begin{aligned} \mathbf{u}_{n+1} &= U^H\left(\mathbf{w}_{n+1} - \overline{\mathbf{w}}\right) \\ &= U^H\left(I - \mu R_{\mathbf{X}}\right)\left(\mathbf{w}_n - \overline{\mathbf{w}}\right) \\ &= U^H\left(U\left(I - \mu\Lambda\right)U^H\right)\left(\mathbf{w}_n - \overline{\mathbf{w}}\right) \\ &= \left(I - \mu\Lambda\right)\mathbf{u}_n, && (6.37) \end{aligned}$$

and since Λ is a diagonal matrix, each of the $p+1$ components of the vector $\mathbf{u}_n$ follows a separate recursion, independently of the others. Clearly, the overall sequence converges if and only if all $p+1$ so-called *modes* converge individually. But each one of them is simply an exponential series governed by $(1-\mu\lambda_i(R_{\mathbf{X}}))^n$ (times the initial value). Such an exponential series converges (to zero) if and only if $|1-\mu\lambda_i(R_{\mathbf{X}})| < 1$, or, equivalently, $\lambda_{max}(R_{\mathbf{X}}) < 2/\mu$. Here, we are also using the fact that for covariance matrices $R_{\mathbf{X}}$, eigenvalues must be non-negative.

To make the algorithm useful, we cannot use the shape given in Equation (6.33) since in any realistic scenario, we would not know the involved expected values. Instead, we may estimate them from the data. In the extreme case, we could estimate the expectation from just a single sample, hence use $\mathbb{E}\left[\left(D[n] - \mathbf{w}_n^T\mathbf{X}[n]\right)\mathbf{X}^*[n]\right] \approx \left(D[n] - \mathbf{w}_n^T\mathbf{X}[n]\right)\mathbf{X}^*[n]$. Of course, this should not be expected to be a particularly good estimate of said expectation, but in exchange, we can actually calculate this simply based on the data at hand (at least for all times n for which we have "training data," i.e., where we know the true desired

outcome $D[n]$). This is called the LMS adaptive algorithm and was discovered in 1960 [7]. It is thus characterized by the update equation

$$\mathbf{w}_{n+1} \quad = \quad \mathbf{w}_n + \mu \left(D[n] - \mathbf{w}_n^T \mathbf{X}[n] \right) \mathbf{X}^*[n]. \tag{6.38}$$

The full analysis of the convergence of this algorithm, even for the special case of wide-sense stationary data, is not feasible in closed form. It is important to observe that here, the filter coefficients $\mathbf{w}_n$ are random vectors, induced by the data. Hence, a first order of business would be to prove convergence of the mean $\mathbb{E}[\mathbf{w}_n]$, perhaps starting with the case of wide-sense stationary data. Unfortunately, it is quickly seen that the resulting vector sequence $\{\mathbb{E}[\mathbf{w}_n]\}_{n \geq 0}$ depends on higher-order statistics of the data and is thus out of reach. A common alternative (approximate) consideration is to assume that the filter taps $\mathbf{w}_n$ are *(statistically) independent* of the data vector corresponding to the same time slot, $\mathbf{X}[n]$. While this is not exactly true, it may hold approximately if μ is sufficiently small. Under this so-called *independence assumption* for the LMS, one easily finds that convergence is again determined by the vector sequence from Equation (6.36), and thus, by the eigenvalues of the covariance matrix of the observed data. A well written account on adaptive filters can be found, e.g., in [8], and an exhaustive compendium in [9].

6.5 Parameter Estimation

In Statistical Signal Processing, a well-studied alternative consideration is to only assume the noise to be random, but the signal to be arbitrary and unknown. This is often referred to as *parameter estimation.* In this view, there is an underlying unknown parameter θ, and the observations are distributed according to a known family of distributions $p_\theta(x)$. Upon observing data $X = x$, we have to produce an estimate $T(x)$ of the unknown parameter θ. The most common figure of merit is the *mean-squared error* $\mathbb{E}[(\theta - T(X))^2]$, where the expectation is over the distribution $p_\theta(x)$, for every fixed θ. That is, every possible estimator $T(x)$ is characterized by a corresponding mean-squared error *function,* i.e., as a function of θ. One important result is the Cramér-Rao lower bound. For all unbiased estimators, this bound gives a lower limit to the mean-squared error function in terms of the *Fisher information.* We also note that this lower limit is not generally attainable, i.e., there does not always exist a $T(x)$ attaining this lower limit.

6.5.1 Fisher Information

In this perspective, there is a family of distributions $\{p_\theta(x) : \theta \in \mathbb{R}\}$, indexed by a real-valued parameter θ. Note that the theory can be extended in a direct fashion to the case where θ is a real-valued vector of some dimension d. In the sequel, we will use the following notational convention:

$$\mathbb{E}_\theta[g(X)] = \int_{-\infty}^{\infty} g(x) p_\theta(x) dx. \tag{6.39}$$

We study the performance of an estimator $T(x)$ in the mean-squared error sense, that is,

$$\mathbb{E}_\theta[(T(X) - \theta)^2], \tag{6.40}$$

which is a function of θ. We are looking for estimators $T(x)$ that are simultaneously good for all values of θ.

Definition 6.1. The *bias* of $T(X)$ is $B_T(\theta) = \mathbb{E}_\theta[T(X) - \theta]$.

Definition 6.2. $T(X)$ is called *unbiased* if $B_T(\theta) = 0$ for all values of θ.

Definition 6.3. For a collection of distributions $\{p_\theta(x) : \theta \in \mathbb{R}\}$, the *score* is $\ell(\theta) = \frac{d}{d\theta} \log p_\theta(x)$.

Lemma 6.3. $\mathbb{E}_\theta[\ell(\theta)] = 0$.

Definition 6.4. The *Fisher information* associated to the class of distributions $\{p_\theta(x) : \theta \in \mathbb{R}\}$ is defined as

$$I_\theta = \mathbb{E}_\theta[\ell^2(\theta)]. \tag{6.41}$$

Theorem 6.4 (Cramér-Rao lower bound)**.** *For any unbiased estimator $T(x)$ for the class of distributions $\{p_\theta(x) : \theta \in \mathbb{R}\}$, it holds that*

$$\mathbb{E}_\theta[(T(X) - \theta)^2] \geq \frac{1}{I_\theta}. \tag{6.42}$$

Proof. From the Cauchy-Schwarz inequality, we can write

$$\begin{aligned}
(\mathbb{E}_\theta[\ell(\theta)(T(X) - \theta)])^2 &\leq \mathbb{E}_\theta[\ell^2(\theta)]\mathbb{E}_\theta[(T(X) - \theta)^2] && (6.43)\\
&= I_\theta \mathbb{E}_\theta[(T(X) - \theta)^2]. && (6.44)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}_\theta[\ell(\theta)(T(X) - \theta)] &= \mathbb{E}_\theta[\ell(\theta)T(X)] - \underbrace{\mathbb{E}_\theta[\ell(\theta)\theta]}_{=0} && (6.45)\\
&= \mathbb{E}_\theta\left[\frac{d}{d\theta} \log p_\theta(X) T(X)\right] && (6.46)\\
&= \int_{-\infty}^{\infty} \frac{\frac{d}{d\theta} p_\theta(x)}{p_\theta(x)} T(x) p_\theta(x) dx && (6.47)\\
&= \int_{-\infty}^{\infty} \frac{d}{d\theta} p_\theta(x) T(x) dx && (6.48)\\
&= \frac{d}{d\theta} \underbrace{\int_{-\infty}^{\infty} p_\theta(x) T(x) dx}_{=\theta} && (6.49)\\
&= 1, && (6.50)
\end{aligned}$$

which completes the proof. □

Example 6.2 (Fisher Information of Bernoulli)**.** For $x \in \{0, 1\}$, let $P_\theta(x) = \theta^x (1 - \theta)^{1-x}$, where $\theta \in [0, 1]$. Then

$$I_\theta = -\sum_{x \in \{0,1\}} P_\theta(x) \frac{d \log(P_\theta(x))}{d\theta^2} = \frac{1}{\theta} + \frac{1}{1 - \theta} = \frac{1}{\theta(1 - \theta)}.$$

Example 6.3 (Fisher Information and Estimation of Bernoulli)**.** Assume that we want to estimate the mean of a Bernoulli random variable with parameter p from n iid samples. Let $\hat{p}$ denote the empirical sample mean. Then we have

$$\mathbb{E}[(\hat{p} - p)^2] = \frac{1}{n} \mathrm{Var}(X) = \frac{p(1 - p)}{n} = \frac{1}{I_p} \frac{1}{n}.$$

6.5.2 Fisher Information — Beyond one dimension

Let $p_\theta(x)$ denote the densities associated to a family of distributions parametrized by θ. The *Fisher* information associated to this family is defined as

$$I_\theta = \mathbb{E}_\theta[(\underbrace{\nabla_\theta \log p_\theta(X)}_{\ell(\theta)})(\nabla_\theta \log p_\theta(X))^T].$$

First note that

$$\begin{aligned}\mathbb{E}_\theta[\ell_\theta] &= \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx \\ &= \int \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} p_\theta(x)dx \\ &= \int \nabla_\theta p_\theta(x)dx = \nabla_\theta \int p_\theta(x)dx = 0.\end{aligned}$$

Further,

$$\begin{aligned}\nabla_\theta^2 \log p_\theta(x) &= \frac{\nabla_\theta^2 p_\theta(x)}{p_\theta(x)} - \frac{\nabla_\theta p_\theta(x)\nabla_\theta p_\theta(x)^T}{p_\theta(x)^2} \\ &= \frac{\nabla_\theta^2 p_\theta(x)}{p_\theta(x)} - \ell(\theta)\ell(\theta)^T.\end{aligned}$$

Lemma 6.5 (Alternative Characterization)**.** *Unders suitable smoothness conditions on the density (hence justifying the exchange of taking integration with taking derivatives) the Fisher information can also be written as*

$$I_\theta = -\mathbb{E}_\theta[\nabla_\theta^2 \log p_\theta(X)].$$

Proof.

$$\begin{aligned}I_\theta &= \mathbb{E}_\theta[\ell(\theta)\ell(\theta)^T] \\ &= -\int p_\theta(x)\nabla^2 \log p_\theta(x)dx + \int \nabla^2 p_\theta(x)dx \\ &= -\mathbb{E}[\nabla^2 \log p_\theta(x)] + \nabla^2 \int p_\theta(x)dx = -\mathbb{E}[\nabla^2 \log p_\theta(x)].\end{aligned}$$

□

Lemma 6.6 (Cramer Rao Bound)**.** *Let T be an unbiased (under P_θ) estimator of a function $\phi(\theta)$. Then*

$$Var(T) \geq \nabla_\theta \phi(\theta)^T I_\theta^{-1} \nabla_\theta \phi(\theta)$$

6.6 Problems

Problem 6.1 (MMSE Estimation)**.** Consider the scenario where $p(x|d) = de^{-dx}$, for $x \geq 0$ (and zero otherwise), that is, the observed data x is distributed according to an exponential with mean $1/d$. Moreover, the desired variable d itself is also exponentially distributed, with parameter λ, that is, $p(d) = \lambda e^{-\lambda d}$.

(a) Find the MMSE estimator of d given x, and calculate the corresponding mean-squared error incurred by this estimator.

(b) Find the MAP estimator of d given x.

Problem 6.2 (Tweedie's Formula)**.** For the special case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, *Tweedie's formula* says that the conditional mean (that is, the MMSE estimator) can be expressed as

$$\mathbb{E}[D|\, X = x] = x + \sigma^2 \ell'(x), \tag{6.51}$$

where

$$\ell'(x) = \frac{d}{dx} \log f_X(x), \tag{6.52}$$

where $f_X(x)$ denotes the marginal PDF of X. In this exercise, we derive this formula.

(a) Assume that $f_{X|D}(x|d) = e^{\alpha d x - \psi(d)} f_0(x)$ for some functions $\psi(d)$ and $f_0(x)$ and some constant α (such that $f_{X|D}(x|d)$ is a valid PDF for every value of d). Define

$$\lambda(x) = \log \frac{f_X(x)}{f_0(x)}, \tag{6.53}$$

where $f_X(x)$ is the marginal PDF of X, i.e., $f_X(x) = \int f_{X|D}(x|\delta) f_D(\delta) d\delta$. With this, establish that

$$\mathbb{E}[D|\, X = x] = \frac{1}{\alpha} \frac{d}{dx} \lambda(x). \tag{6.54}$$

(b) Show that the case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, is indeed of the form required in Part *(a)* by finding the corresponding $\psi(d)$, $f_0(x)$, and α. Show that in this case, we have

$$\frac{f_0'(x)}{f_0(x)} = -\frac{x}{\sigma^2}, \tag{6.55}$$

and use this fact in combination with Part *(a)* to establish Tweedie's formula.

Problem 6.3 (FIR Wiener Filter)**.** Consider a (discrete-time) signal that satisfies the difference equation $d[n] = 0.5 d[n-1] + v[n]$, where $v[n]$ is a sequence of uncorrelated zero-mean unit-variance random variables. We observe $x[n] = d[n] + w[n]$, where $w[n]$ is a sequence of uncorrelated zero-mean random variables with variance 0.5.

(a) (you may skip this at first and do it later — it is conceptually straightforward) Show that for this signal model, the autocorrelation function of the signal $d[n]$ is

$$\mathbb{E}[d[n]d[n+k]] \;=\; \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, \tag{6.56}$$

and thus the autocorrelation function of the signal $x[n]$ is

$$\mathbb{E}[x[n]x[n+k]] \;=\; \begin{cases} \frac{11}{6}, & \text{for } k = 0, \\ \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, & \text{otherwise.} \end{cases} \tag{6.57}$$

(b) We would like to find an (approximate) linear predictor $\hat{d}[n+3]$ using only the observations $x[n], x[n-1], x[n-2], \ldots, x[n-p]$. Using the Wiener Filter framework, determine the optimal coefficients for the linear predictor. Find the corresponding mean-squared error for your predictor.

(c) We would like to find a linear denoiser $\hat{d}[n]$ using *all* of the samples $\{x[k]\}_{k=-\infty}^{\infty}$. Find the filter coefficients and give a formula for the incurred mean-squared error.

Problem 6.4 (Wiener Filter and Irrelevant Data)**.** As we have seen in class, the (FIR) Wiener filter is given by

$$\mathbf{w} \;=\; R_x^{-1}\mathbf{r}_{dx}, \tag{6.58}$$

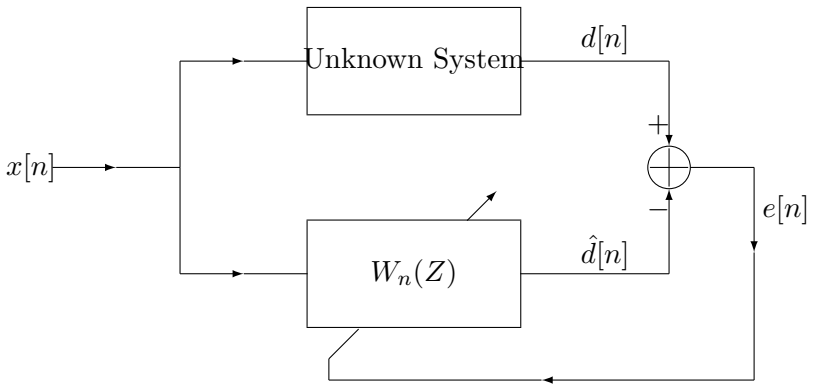
where R_x is the autocorrelation matrix of the data that's being used, and $\mathbf{r}_{dx}$ is the cross-correlation between the data and the desired output. For this to be well defined, R_x should be full rank. In this problem, we study this question in more detail.

(a) In many applications, the signal acquisition process is noisy. That is, the data $x[n] = s[n] + w[n]$, where $s[n]$ is an *arbitrary* signal, and $w[n]$ is white noise. Prove that in this case, the $p-$dimensional autocorrelation matrix R_x is full rank (i.e., invertible) for any p. (Note: Be careful not to make *any* assumptions about the signal $s[n]$.)

(b) In some other cases, R_x could be rank-deficient. To study this, prove first that if the (FIR) Wiener filter based on the data $\mathbf{x} = \{x[n]\}_{n=0}^{p-1}$ is $\mathbf{w}$, then the (FIR) Wiener filter based on the modified data $A\mathbf{x}$ (where A is an invertible matrix) is $A^{-H}\mathbf{w}$, (where we use the relatively common notation $A^{-H} = (A^{-1})^H$).

(c) Explain how to find the (FIR) Wiener filter when R_x is rank-deficient. Discuss existence and uniqueness. *Hint:* Use Part *(b)* to transform your data to a more convenient basis.

Problem 6.5 (Adaptive Filters)**.** One of the many uses of adaptive filters is for system identification as shown in the figure blow. In this configuration, the same input is applied to an adaptive filter and to an unknown system, and the coefficients of the adaptive filter are adjusted until the difference between the outputs of the two systems is as small as possible.



Let the unknown system that is to be characterized by

$$d[n] = x[n] + 1.8x[n-1] + 0.81x[n-2] \tag{6.59}$$

With an input $x[n]$ consisting of 1000 samples of unit variance white Gaussian noise, create the reference signal $d[n]$.

(a) Determine the range of values for the step size μ in the LMS algorithm for the convergence in the mean.

(b) Implement an adaptive filter of order $p = 4$ using the LMS algorithm. Set the initial weight vector equal to zero, and use a step size of $\mu = 0.1\mu_{max}$, where μ_{max} is the largest step size allowed for convergence in the mean. Let the adaptive filter adapt and record the final set of coefficients.

(c) Repeat part *(b)* using the normalized LMS algorithm with $\beta = 0.1$, and compare your results.

(d) Make a plot of the learning curve by repeating the experiment described in part *(b)* for 100 different realizations of $d[n]$, and plotting the average of the plots of $e^2[n]$ versus n. How many iterations are necessary for the mean-square error to fall to 10% of its peak value? Calculate the theoretical value for the excess mean-square error and compare it to what you observe in your plot of the learning curve.

Problem 6.6 (Missing Data)**.** We are given real-valued data with a single missing sample :

$$X_1, X_2, X_3, X_4, X_5, X_6, ?, X_8, X_9, \ldots \tag{6.60}$$

where we assume that the data is wide-sense stationary with autocorrelation function $R_X[k] = \alpha^{|k|}$, where $0 < \alpha < 1$. We would like to find a meaningful estimate for the missing sample X_7.

1. As a starting point, let us consider the estimate $\hat{X}_7 = wX_6$, where w is a real number. Find the value of w so as to minimize the mean-squared error $\mathbb{E}[(X_7 - \hat{X}_7)^2]$, and determine the incurred mean-squared error.

2. Now, consider the estimate $\hat{X}_7 = w_1X_6 + w_2X_8$. Again, find the values of w_1 and w_2 so as to minimize the mean-squared error $\mathbb{E}[(X_7 - \hat{X}_7)^2]$, and determine the incurred mean-squared error.

Problem 6.7 (Parameter Estimation and Fisher Information)**.** The Fisher information $J(\Theta)$ for the family $f_\theta(x), \theta \in \mathbf{R}$ is defined by

$$J(\theta) = \mathbb{E}_\theta \left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)} \right)^2 = \int \frac{(f_\theta')^2}{f_\theta}$$

Find the Fisher information for the following families:

(a) $f_\theta(x) = N(0, \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$

(b) $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$

(c) What is the Cramèr Rao lower bound on $\mathbb{E}_\theta(\hat{\theta}(X) - \theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

Problem 6.8 (Conditional Independence and MMSE)**.** For simplicity, throughout this problem, **all random variables are assumed to be zero-mean.**

(a) Show that if X and Y are conditionally independent given Z, then

$$\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] = 0. \tag{6.61}$$

(b) Now let X and Y be jointly Gaussian (zero-mean). It is well known that if $\mathbb{E}[XY] = 0$, then X and Y are independent. Establish this fact starting from the observation that for (zero-mean) Gaussian random variables X and Y, we may always write $Y = \alpha X + W$, for some constant α, where W is zero-mean Gaussian *independent of* X. *Note: This prepares you for Part (c).*

(c) Let X, Y, Z be jointly Gaussian (and zero-mean, as throughout this problem). Prove that if

$$\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] = 0, \tag{6.62}$$

then X and Y are conditionally independent given Z. *Hint:* Make sure to solve Part *(b)* first. Recall that for three jointly Gaussians X, Y, Z, we can always write $Y = \gamma X + \delta Z + V$, for some constants γ and δ, where V is Gaussian and independent of X and Z.

(d) Let X, Y, Z be jointly Gaussian (and zero-mean, as throughout this problem). Prove that X and Y are conditionally independent given Z if and only if

$$\mathbb{E}[XY]\mathbb{E}[Z^2] = \mathbb{E}[XZ]\mathbb{E}[YZ]. \tag{6.63}$$

(e) Continuing from Part *(d)*, let us simplify: $\mathbb{E}[X^2] = \mathbb{E}[Y^2] = \mathbb{E}[Z^2] = 1$, and use the notation $\rho = \mathbb{E}[XY]$. Define $a = \mathbb{E}[XZ]$ and $b = \mathbb{E}[YZ]$. Find

$$\arg\max_{a,b} \min_{f} \mathbb{E}[(Z - f(X,Y))^2], \tag{6.64}$$

where the inner minimum is over all measurable functions $f(x, y)$, and the maximum is over all choices a, b such that X and Y are conditionally independent given Z.

Problem 6.9 (Fisher Information and Divergence)**.** Suppose we are given a family of probability distributions $\{p(\,\cdot\,;\theta) : \theta \in \mathbb{R}\}$ on a set $\mathcal{X}$, parametrized by a real valued parameter θ. (Equivalently, a random variable X whose distribution depends on θ.) Assume that the parametrization is smooth, in the sense that

$$p'(x;\theta) := \frac{\partial}{\partial\theta} p(x;\theta) \quad \text{and} \quad p''(x;\theta) := \frac{\partial^2}{\partial\theta^2} p(x;\theta)$$

exist. (Note that the derivaties are with respect to the parameter θ, not with respect to x.) We will use the notation $\mathbb{E}_{\theta_0}[\cdot]$ to denote expectations when the parameter is equal to a particular value θ_0, i.e., $\mathbb{E}_\theta[g(X)] = \sum_x p(x;\theta) g(x)$.

Define the function $K(\theta, \theta') := D\big(p(\,\cdot\,;\theta) \| p(\,\cdot\,;\theta')\big)$.

(a) Show that for any θ_0, $\dfrac{\partial}{\partial\theta} K(\theta, \theta_0) = \displaystyle\sum_x p'(x;\theta) \log \frac{p(x;\theta)}{p(x;\theta_0)}$.

(b) Show that $\dfrac{\partial^2}{\partial\theta^2} K(\theta, \theta_0) = \displaystyle\sum_x p''(x;\theta_0) \log \frac{p(x;\theta)}{p(x;\theta_0)} + J(X;\theta)$ with

$$J(X;\theta) := \mathbb{E}_\theta\big[\big(p'(X;\theta)/p(X;\theta)\big)^2\big].$$

(c) Show that when θ is close to θ_0

$$K(\theta, \theta_0) = \tfrac{1}{2} J(X;\theta_0)(\theta - \theta_0)^2 + o((\theta - \theta_0)^2)$$

(d) Show that $J(X;\theta) = -\mathbb{E}_\theta\big[\frac{\partial^2}{\partial\theta^2} \log p(X;\theta)\big]$.

Chapter 7

Distribution Estimation, Property Testing and Property Estimation

Assume that we are given iid samples from an unknown distribution. How many samples do we need before we can estimate the distribution with an "acceptable" accuracy? And what if we are interested in only particular properties of the distribution, such as its support size, or perhaps it's entropy. These are the questions that we will discuss in this chapter.

This chapter closely follows the tutorial by Acharya, Orlitsky, and Suresh, see

https://people.ece.cornell.edu/acharya/papers/isit-tutorial-acharya-orlitsky-suresh.pdf

7.1 Distribution Estimation

7.1.1 Notation and Basic Task

Consider a random variable X taking values in the discrete set $\mathcal{X}$ and let $p(x), x \in \mathcal{X}$, describe the distribution of X. Of course we have $p : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $\sum_{x \in \mathcal{X}} p(x) = 1$. We assume that $\mathcal{X}$ has a finite support, $|\mathcal{X}| = k$. Without loss of generality we will assume that $\mathcal{X} = \{1, \cdots, k\}$. In this case we can think of $p(x)$ also as a vector of length k written as $p = (p_1, \cdots, p_k)$. Note that in this way p is an element of Δ_k, the simplex in $\mathbb{R}^k$.

In the sequel we will assume that we are given a sample of n elements from $\mathcal{X}$, call it $x^n = x_1, \cdots, x_n$, chosen iid according to $p(x)$, so that

$$p(x^n) = \prod_{i=1}^{n} p(x_i).$$

Given this sample x^n our task is to find a distribution $q = q(x^n)$, $q \in \Delta_k$, that is "close" to the distribution p.

7.1.2 Empirical Estimator

The perhaps most "natural" estimator is the *empirical* estimator q^{emp}. Given a sequence x^n let

$$t_i(x^n) = |\{j \in \{1, \cdots, n\} : x_j = i\}|.$$

In words, $t_i(x^n)$ counts how many times the symbol i appears in x^n. The empirical estimator is then defined as

$$q_i^{\text{emp}}(x^n) = t_i(x^n)/n.$$

Clearly,

$$\begin{aligned} q_i^{\text{emp}}(x^n) &\geq 0, \\ \sum_{i=1}^{k} q_i^{\text{emp}}(x^n) &= 1. \end{aligned}$$

In other words, $q^{\text{emp}}(x^n) \in \Delta_k$, so this estimator is well-defined.

Example 7.1. Let's assume that $n = 4$, $k = 3$, and $x^4 = 3112$. Then

$$q^{\text{emp}}(3112) = (q_1^{\text{emp}}(3112), q_2^{\text{emp}}(3112), q_3^{\text{emp}}(3112)) = (\frac{2}{4}, \frac{1}{4}, \frac{1}{4}).$$

The empirical estimator will play a prominent role in the following.

7.1.3 Loss Functions

Before we can analyse how a given estimator does we need to specify how we will measure the quality of the estimator. More precisely, given p and q, how do we measure their distance? The three most common choices are the ℓ_1 distance, the ℓ_2 distance, and the Kullback-Leibler divergence. Generically we call the loss $L(p, q)$. We will start by looking at ℓ_2^2 since it is mathematically the most convenient.

7.1.4 Min-Max Criterion

So assume that we have fixed the loss function L. We then still have various degrees of freedom. Let us go over these options.

- We are given a fixed distribution p and a fixed estimator q. It is then natural that we compute the expected loss, where the expectation is over the sample:

$$\mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

- We are given a fixed distribution p. For each estimator we can compute an expected loss as we discuss in the previous scenario. It is then natural to find that estimator that minimizes this expected loss:

$$q^* = \text{argmin}_q \mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

- We are given a fixed estimator q. For each fixed distribution we can compute an expected loss as we discussed in the first scenario. It is then natural that we find that distribution p^* that maximizes this expected loss:

$$p^* = \text{argmax}_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

What if we are neither given p nor q? We get a *robust* definition if we choose the estimator q in such a way that we *minimize* the expected risk for the *worst* distribution p. This is called the *min-max* criterion and in formulae it reads

$$r^L_{k,n} = \min_q \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

Let us emphasize: For each q (estimator) we look a that p (distribution) that gives the worst result. We then pick that q that minimizes the worst case.

We will be interested in finding what this min-max optimal estimator is and how it performs. Our strategy will be the following. We first compute the risk of the empirical estimator. By computing a lower bound on the risk, we will then see that a small variant of this natural estimator is min-max optimal.

7.1.5 Risk of Empirical Estimator in ℓ_2^2

We start by looking at the case where the difference between the true distribution and the estimated one is measured in ℓ_2^2 distance, i.e.,

$$\|p - q\|_2^2 = \sum_i (p_i - q_i)^2.$$

We want to compute

$$r^{q^{\text{emp}}}_{k,n} = \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p} \left[\sum_i (p_i - q_i^{\text{emp}}(X^n))^2 \right],$$

where $q_i^{\text{emp}}(x^n) = t_i(x^n)/n$.

Recall that the components of X^n are iid, chosen from $\mathcal{X}$ according the distribution p. This (the fact that they are iid) implies that for all $i \in \mathcal{X}$, $t_i(X^n)$ is a Binomial with parameters $\text{Binom}(p_i, n)$. Therefore,

$$\mathbb{E}_{X^n \sim p}[t_i(X^n)] = np_i,$$
$$\mathbb{E}_{X^n \sim p}[(t_i(X^n) - np_i)^2] = np_i(1 - p_i).$$

Hence,

$$\mathbb{E}_{X^n \sim p}[\sum_{i=1}^k (t_i(X^n)/n - p_i)^2] = \sum_{i=1}^k \frac{p_i(1-p_i)}{n} = \frac{1 - \sum_{i=1}^k p_i^2}{n} \overset{(a)}{\leq} \frac{1 - \frac{1}{k}}{n}.$$

In step we have used the Cauchy-Schwartz inequality,

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

To get our inequality, pick $x = (1, \cdots, 1)$ and $y = p$. Note that this inequality is achieved by the uniform distribution $p_i = 1/k$.

Note that the above bound is *universal* with respect to the underlying distribution p. This is good news. So we have shown that for ℓ_2^2 loss

$$\max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q^{\text{emp}}(X^n))] = \frac{1 - \frac{1}{k}}{n}.$$

7.1.6 Risk of "Add Constant" Estimator in ℓ_2^2

Is the empirical estimator min-max optimal? Not quite. Here is a slightly better estimator:

$$q_i^{+\sqrt{n}/k}(x^n) = \frac{t_i(x^n) + \frac{\sqrt{n}}{k}}{n + \sqrt{n}}.$$

This is an instance of an "add constant" estimator. We will soon see that this estimator is min-max optimal. Intuitively, adding a constant to our observations makes sense. If the number of samples is small, we cannot possibly have seen all elements, not because their probability is zero, but because it is small and there is randomness in sampling. What is the risk of this estimator? We claim that

$$\max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q^{+\sqrt{n}/k}(X^n))] = \frac{1 - \frac{1}{k}}{(\sqrt{n}+1)^2}$$

To see this claim note that

$$\mathbb{E}_{X^n \sim p}[q_i^{+\sqrt{n}/k}(x^n)] = \mathbb{E}_{X^n \sim p}[\frac{t_i(X^n) + \frac{\sqrt{n}}{k}}{n + \sqrt{n}}] = \frac{np_i + \frac{\sqrt{n}}{k}}{n + \sqrt{n}}.$$

Note that this is a *biased* estimator. We get

$$\begin{aligned}
\mathbb{E}_{X^n \sim p}[(q_i^{+\sqrt{n}/k}(X^n) - p_i)^2] &= \frac{\mathbb{E}_{X^n \sim p}[(t_i(X^n) - np_i - \frac{\sqrt{n}}{k}(kp_i - 1))^2]}{(n + \sqrt{n})^2} \\
&= \frac{\mathrm{Var}[t_i(X^n)] + \frac{n}{k^2}(kp_i - 1)^2}{(n + \sqrt{n})^2} \\
&= \frac{np_i(1 - p_i) + \frac{n}{k^2}(kp_i - 1)^2}{(n + \sqrt{n})^2} \\
&= \frac{np_i(1 - \frac{2}{k}) + \frac{n}{k^2}}{(n + \sqrt{n})^2}.
\end{aligned}$$

To compute the worst-case loss of the $q^{+\sqrt{n}/k}(x^n)$ estimator we need to sum this expression over all k components and then maximize with respect to the distribution p. This gives us

$$\max_{p \in \Delta_k} \sum_{i=1}^{k} \mathbb{E}_{X^n \sim p}[(q_i^{+\sqrt{n}/k}(X^n) - p_i)^2] = \frac{n(1 - \frac{1}{k})}{(n + \sqrt{n})^2} = \frac{(1 - \frac{1}{k})}{(\sqrt{n}+1)^2},$$

as claimed.

Note: This calculation shows that for this estimator the loss *does not* depend on the underlying distribution p. This will become important soon.

Note: If we had done this calculation with a general additive term β instead of the specific term $\beta^* = \frac{\sqrt{n}}{k}$ then it is easy to see that that the choice β^* is optimal.

7.1.7 Matching lower bound for ℓ_2^2

We will now derive a matching lower bound. We proceed as follows. Let π be a prior distribution on Δ_k. We then have

$$r_{k,n}^{\ell_2^2} = \min_q \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}\left[\sum_i (p_i - q_{X^n,i})^2\right] \geq \min_q \mathbb{E}_{P\sim\pi; X^n \sim P}\left[\sum_i (P_i - q_i(X^n))^2\right]$$
$$= \mathbb{E}_{P\sim\pi; X^n\sim P}\left[\sum_i (P_i - \mathbb{E}_{P\sim\pi;X^n\sim P}[P_i \mid X^n])^2\right].$$

where in the last step we have used the fact that the minimum-mean squared error estimator is given by the conditional expectation, a fact that is discussed in detail in Section 6.2.1.

Therefore, if we can guess a "good" prior then we will get a good bound. It turns out that a suitable Dirichlet prior gives us the matching lower bound. The Dirichlet distribution on Δ_k is characterized by a vector $\alpha \in (\mathbb{R}_+)^k$. Let $(x_1, \cdots, x_k) \in \Delta_k$. The associated density is given by

$$f(x_1, \cdots, x_k; \alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}.$$

Note that this is an exponential distribution, i.e., it can be written in a form

$$p_\Theta(x) = e^{\langle \Theta, \phi(x)\rangle - A(\Theta)},$$

where $x = (x_1, \cdots, x_k)$, $\phi(x) = (\ln(x_1), \cdots, \ln(x_k))$, and $\Theta = (\alpha_1 - 1, \cdots, \alpha_k - 1)$.

The important property of a Dirichlet distribution for our application is that it is the conjugate distribution of a multi-nomial distribution. So assume that the parameters $p_1, \cdots, p_k$ of a multi-nomial distribution are themselves random with prior $\mathrm{Dir}(\alpha)$. Assume that we sample from this Dirichlet distribution and then sample from the multi-nomial according to the chosen parameter. We get n samples and their counts are $T_1, \cdots, T_k$. Given this observation what is the posterior of the parameters? I.e., we want to determine

$$f(p_1, \cdots, p_k \mid T_1 = t_1, \cdots, T_k = t_k).$$

We claim that this is again a Dirichlet distribution, namely with parameters $\alpha + T$ (both vectors). Using Bayes' rule we get

$$\begin{aligned} f(p_1, \cdots, p_k \mid t_1, \cdots, t_k) &= \frac{f(t_1, \cdots, t_k \mid p_1, \cdots, p_k) f(p_1, \cdots, p_k)}{Z(t_1, \cdots, t_k)} \\ &= \frac{\binom{n}{t_1, \cdots, t_k} [\prod_{i=1}^k p_i^{t_i}] \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)} [\prod_{i=1}^k p_i^{\alpha_i - 1}]}{Z(t_1, \cdots, t_k)} \\ &= \frac{\prod_{i=1}^k p_i^{\alpha_i + t_i - 1}}{Z} \\ &= \mathrm{Dir}(p_1, \cdots, p_k; \alpha + t). \end{aligned}$$

The Dirichlet distribution has been well studied. In particular, its mean is known. If we assume that $\alpha_i = \alpha$ for all $i = 1, \cdots, k$, then we get

$$\mathbb{E}_{P\sim\pi; X^n\sim P}[P_i \mid X^n] = \frac{t_i(X^n) + \alpha}{n + k\alpha}.$$

Now note that if we pick $\alpha = \sqrt{n}/k$ then our previous calculations have shown that the loss does not in fact depend on the distribution p but is always equal to $\frac{1-1/k}{(\sqrt{n}+1)^2}$. This finishes our claim.

Although ℓ_2 is nice and easy from an analysis perspective it has also downsides. Perhaps the biggest one is that it is not a good measure if k is very large. Assume that p has $2/k$ in its first $k/2$ components and 0 in its remaining ones and assume that for q the roles of these two parts are exactly reversed. Their ℓ_2^2 distance is then equal to $4/k$ which quickly converges to 0 as k gets large. But it is hard to think of distributions that are more different! If instead we looked at their ℓ_1 distance we would get 2 and their KL "distance" is in fact infinity.

7.1.8 Risk in ℓ_1

This motivates us to look at ℓ_1, i.e., now we look at

$$\|p - q\|_1 = \sum_{i=1}^{k} |p_i - q_i|.$$

Note that this has a probabilistic interpretation. E.g., we can couple the two random variables so that up to a fraction of time $\|p - q\|_1$ they take the same value.

Lemma 7.1. *For k fixed and as n tends to infinity, the worst case min-max loss behaves like*

$$r_{k,n} \leq \sqrt{\frac{2(k-1)}{\pi n}} + O(n^{-\frac{3}{4}}).$$

Further, this is achieved by the empirical estimator.

The idea of the proof is similar to the technique we used before. We first compute the loss of the empirical estimator. We show that this loss is highest for the uniform distribution and this gives us an upper bound. Then we derive a lower bound that matches the dominant terms by computing again the Bayes loss with a proper Dirichlet prior. We skip the details.

If we are content with a sligthly looser upper bound we can proceed as follows. Consider the empirical estimator $q^{\text{emp}}(X^n)$. We have

$$\begin{aligned}
\mathbb{E}[\|p - q^{\text{emp}}(X^n)\|_1] &= \sum_{i=1}^{k} \mathbb{E}[|p_i - \frac{T_i(X^n)}{n}|] \\
&\overset{(a)}{\leq} \sum_{i=1}^{k} \sqrt{\mathbb{E}[|p_i - \frac{T_i(X^n)}{n}|^2]} \\
&\overset{(b)}{\leq} \sqrt{\frac{k-1}{n}}.
\end{aligned}$$

Step (a) follows by Jensen's inequality and in step (b) we have used our results for ℓ_2^2 and the Cauchy-Schwartz inequality. We see that we loose a factor $2/\pi$ compared to the previous result.

7.1.9 Risk in KL-Divergence

If we are using the KL divergence as our loss metric then we need to make sure that none of our estimated probabilities are 0 since otherwise our metric will be ∞. It is therefore natural to use an "add constant" estimator.

When the number of samples becomes large compared to the alphabet size one can show that the best "add constant" estimator is of the form $q_i^{+0.509}$ and this gives us an expected worst case loss of

$$\max_{p\in\Delta_k} \mathbb{E}[D(p\|q^{+0.509})] \sim 0.509\frac{k-1}{n}. \tag{7.1}$$

One can do slightly better ($\frac{1}{2}$ instead of 0.509) by using different constants depending on the observed frequency.

Note that by the Pinsker inequality we have

$$\sqrt{1.2\frac{k-1}{n}} \sim \max_{p\in\Delta_k} \sqrt{2\mathbb{E}[D(p\|q^{+0.509})]} \geq \max_{p\in\Delta_k} \mathbb{E}[\sqrt{2D(p\|q^{+0.509})}] \geq \max_{p\in\Delta_k} \mathbb{E}[\|p-q^{+0.509})\|_1].$$

The first step is (7.1). Note that this step is approximate (valid for large ratios n/k). In the second we have used Jensen's inequality. The final step is Pinsker's inequality $\|p-q\|_1 \leq \sqrt{2D(p\|q)}$. We see from this sequence of inequalities that these two results are related roughly as we would expect. (Compared to the result in Lemma 7.1 we loose only a factor 0.64 inside the square root.)

7.1.10 The problem with the min-max formulation

We have now surveyed how the min-max estimator behaves for various risks. One problem we encounter is that min-max is really quite pessimistic. Yes, the worst case is as good as we could hope for but the estimator could be quite bad for pretty much any case. In fact, we have seen that the whole trick of proving what the min-max estimator was for the ℓ_2 case was to come up with an estimator that was uniformly bad (so to speak). This brings us to a slightly different point of view.

7.1.11 Competitive distribution estimation

One possible view point is that we want an estimator q that is close to optimal for every distribution p. The key question is here: What is our comparison group? Let us make this more formal.

Look at the probability simplex Δ_k. Partition this space into groups $P_1, P_2, \cdots$ so that $\Delta_k = \cup_j P_j$. Call this partition $\mathcal{P}$. Let $L(\cdot,\cdot)$ be the loss as usual. Then we are interested in

$$r_{n,k}^{\mathcal{P}} = \min_q \max_j \left[\max_{p\in P_j} L(p,q) - \min_{q'} \max_{p'\in P_j} L(p',q')\right].$$

This is easy to interpret. Assume at first that we pick the partition so that every element of Δ_k forms a group on its own. Within each group we compare to $\min_{q'} \max_{p'\in P_j} L(p',q')$. This is the min-max estimator for that group. But since the group only exists of a single distribution we can use an estimator that *knows* that particular distribution. This measure

then collapses to our original min-max formulation and, as we have discussed, this is often simply to pessimistic.

On the other hand, if our partition consists only of a single group, i.e., the "genie" we compare ourselves to has no more knowledge than we have ourselves, then our loss is 0. This is not very useful either.

The key hence is to find for every case a suitable partition. And for each partition we compare ourselves in each group to a genie who knows that group a priori. E.g., the genie might know the entropy of the distribution a priori. Or perhaps the genie knows the *set* of probabilities but not which component has which of these probabilities.

This is a very rich setup and there are obviously many variations on the theme.

7.1.12 Multi-set genie estimator

Here is another way of avoiding the pessimistic nature of the min-max setup. Recall how we proceeded in the adversarial bandit setup. Rather than giving the genie in that setting only partial knowledge of the rewards table, we allowed it to see the whole table but we restricted the *choice* of the bandit. We can proceed in the same fashion in the present case.

7.1.13 Natural Genie and Good-Turing Estimator

Hence, assume that the genie is knows the distribution but is forced to give the same estimate to any group of symbols that appear the same number of times. E.g., $X^5 = 12213$. Then this genie will use the estimates

$$q_1 = q_2 = \frac{p_1 + p_2}{2},$$
$$q_3 = p_3.$$

Let $M(t)$ denote the total probability of all symbols that appear exactly t times. And let us assume that we are using the KL divergence as loss function. Further, let $\phi(t)$ denote the number of symbols that appeared t times.

The so-called Good-Turing estimator is then

$$q_i^{\mathrm{GT}} = \frac{T_i + 1}{n} \frac{\phi(T_i + 1)}{\phi(T_i)}.$$

This estimator has a fabled history and was supposedly one of the tools used in breaking the Enigma code. Good published it in 1953 based on an unpublished note by Turing.

Let us do an example. Assume that $X^9 = 121234555$. We then have $\phi(1) = 2$ since 3 and 4 appeared once, $\phi(2) = 2$ since 1 and 2 appeared twice, and $\phi(3) = 1$ since 5 appeared three times. We then have

$$q_3 = \frac{T_3 + 1}{9} \frac{\phi(2)}{\phi(1)} = \frac{2}{9} \frac{2}{2} = \frac{2}{9}.$$

What is the intuition for this estimator? The intuition comes by looking at what the natural genie will do. Recall, it will give the probability $M(t)/\phi(t)$ to each of the $\phi(t)$ symbols that appear t times. We are trying to compete against this genie. So it makes sense that we use an expression motivated by this estimate. Of course, we do not know what $M(t)$ is since we do not know the probabilities. We claim that

$$\mathbb{E}[M(t)] = \frac{t+1}{n} \mathbb{E}[\phi(t+1)]. \tag{7.2}$$

To be slightly more precise. We claim that we have this indentity if we us *Poisson* sampling. You will explore this more in the homework. It is a standard trick to get rid of the dependency that you get between the various coefficients when you sample a fixed number of samples.

Assume that we have given a distribution p on $\mathcal{X} = \{1, \cdots, k\}$. Let X^n denote a sequence of n iid samples. Let $T_i = T_i(X^n)$ be the number of times symbol i appears in X^n. Then

$$\mathbb{P}\{T_i = t_i\} = \binom{n}{t_i} p_i^{t_i} (1 - p_i)^{n - t_i}.$$

Note that the random variables T_i are *dependent*, since $\sum_i T_i = n$. This dependence can cause difficulties if we are using this distribution in a scheme and want to analyse its performance.

There is a convenient way of getting around this problem. This is called *Poisson* sampling. Let N be a random variable distributed according to a Poisson distribution with mean n. Let X^N be an iid sequence of N variables distributed according to p.

Then the following statements are true.

- $T_i(X^N)$ is distributed according to a Poisson random variable with mean $p_i n$.
- The $T_i(X^N)$ are independent.
- Conditioned on $N = n$, the induced distribution of the Poisson sampling scheme is equal to the distribution of the *original scheme*.

We will verify (7.2) in a moment. Equation (7.2) does not completely solve our problem since we do not know $\mathbb{E}[\phi(t+1)]$. But we do have its "instantaneous" value $\phi(t+1)$. Hence define $\hat{M}(t) = \frac{t+1}{n}\phi(t+1)$. If we now use $\hat{M}(t)/\phi(t)$ instead of $M(t)/\phi(t)$ then we get our Good-Turing estimator. Let us now verify (7.2). Note that

$$M(t) = \sum_{i=1}^{k} p_i \mathbb{1}_{\{T_i(X^N) = t\}},$$

where in the notation $M(t)$ we omit the fact that this quantity depends on X^n. We now have

$$\begin{aligned}
\mathbb{E}[\hat{M}(t)] &= \sum_i \frac{t+1}{n} \mathbb{E}[\mathbb{1}_{\{T_i(X^N) = t+1\}}] \\
&= \sum_i \frac{t+1}{n} e^{-(np_i)} \frac{(np_i)^{t+1}}{(t+1)!} \\
&= \sum_i \frac{np_i}{n} e^{-(np_i)} \frac{(np_i)^{t}}{(t)!} \\
&= \sum_i p_i e^{-(np_i)} \frac{(np_i)^{t}}{(t)!} \\
&= \sum_i p_i \mathbb{E}[\mathbb{1}_{\{T_i(X^N) = t\}}] \\
&= \mathbb{E}[M(t)].
\end{aligned}$$

We can now write down the competitive loss. We have.

$$r_{n,k}^{\text{nat}} = \min_{\hat{M}} \max_{p} \mathbb{E}[\sum_{t=0}^{n} M(t) \ln \frac{M(t)}{\hat{M}(t)}].$$

One can show that an estimator based on the Good-Turing estimator and the empirical estimator achieves

$$r_{n,k}^{\text{nat}} \sim \min\{n^{-\frac{1}{3}}, \frac{k}{n}\}.$$

7.2 Property Testing

We are given i.i.d. samples from an unknown distribution and we want to know if this distribution has a particular property or if it is at least an ϵ away from having this property. Here are a few examples.

1. We want to test if the distribution is uniform.
2. We want to test if the distribution is equal to a given distribution. This is called identity testing.
3. We want to test if a distribution over $\mathcal{X} \times \mathcal{X}$ is the product of two marginal distributions.
4. We want to test if the pdf is monotone.
5. We want to test if the pdf is log-concave.[1]

We can frame all these questions in the following manner. Let $\mathcal{P}$ and $\mathcal{Q}$ be two *families* of distributions with $\mathcal{P} \cap \mathcal{Q} = \emptyset$. Let $P \in \mathcal{P} \cup \mathcal{Q}$ and let X^n be n iid samples according P. We are given X^n but do not know P. We are asked to decide whether the samples where drawn according to a distribution in $\mathcal{P}$ or a distribution in $\mathcal{Q}$. More formally we are asked to design an estimator, $C(X^n) \to \{\mathcal{P}, \mathcal{Q}\}$ in such a way as to minimize the maximum probability of error,

$$p^n = \max_{P \in \mathcal{P} \cup \mathcal{Q}; X^n \sim P} \max\{\mathbb{P}\{C(X^n) = \mathcal{Q} \mid P \in \mathcal{P}\}, \mathbb{P}\{C(X^n) = \mathcal{P} \mid P \in \mathcal{Q}\}\}.$$

We could ask now how this probability of error behaves as n tends to infinity. This would bring us back to questions of large deviations. But in the current context we are more interested in how the quantity behaves for a small sample size; a sample size that is just big enough so that our error probability is bounded away from $\frac{1}{2}$.

7.2.1 General Idea

Assume that we can design a so-called *test statistics* $T(X^n) \to \mathbb{R}$ with the following properties: there exists a *threshold* τ so that

1. if $P \in \mathcal{P}$ then $\mathbb{P}\{T(X^n) > \tau\} < 0.1$,

[1] For discrete contiguous distributions we say that it is log-concave if $P_i^2 \geq P_{i-1} P_{i+1}$.

2. if $P \in \mathcal{Q}$ then $\mathbb{P}\{T(X^n) < \tau\} < 0.1$.

In this case, given the sample X^n we simply evaluate this test statistics and make our decision accordingly. I.e., we define

$$C(X^n) = \begin{cases} \mathcal{P}, & T(X^n) < \tau, \\ \mathcal{Q}, & T(X^n) > \tau. \end{cases}$$

In the above description we have assumed that the number of samples is fixed. As we have seen this for distribution estimation it is sometimes useful to allow this number to be itself a random variable distributed according to a Poisson distribution. We will then write X^N, $N \sim \mathrm{Poi}(n)$.

7.2.2 Testing Against a Uniform Distribution

Even though there are many questions that fall under the category of property testing we will consider only one – namely the question of testing whether samples come from a uniform distribution.

We will assume that the alphabet size k is known and we will ask wether the samples come from a uniform distribution with support on the *whole* alphabet size. It is important to note that, even though this is a meaningful question, and it is mathematically simpler, perhaps an even more meaningful question would be to allow distributions whose support is not all of $\mathcal{X}$.

Learning Approach

The first approach is obvious. Let us learn the distribution reasonably accurately and then compute the distance of this learned distribution to the uniform one. In the following let us assume that we measure the distance according to ℓ_1. Let U denote the uniform distribution on $\mathcal{X} = \{1, \cdots, k\}$. Then $\mathcal{P} = \{U\}$ and $\mathcal{Q}$ is the set of distributions that have ℓ_1 distance at least ϵ from U.

We then have the following algorithm.

1. Given X^n learn $\hat{P}$ so that $\|\hat{P} - P\|_1 \leq \epsilon/2$ with probability at least 0.9.
2. Output decision according to

$$C(X^n) = \begin{cases} \mathcal{P}, & \|\hat{P} - U\| < \epsilon/2, \\ \mathcal{Q}, & \text{otherwise.} \end{cases}$$

Let us quickly check that this scheme works as intended. If $P \in \mathcal{P}$, i.e., $P = U$ then by assumption $\|\hat{P} - U\|_1 \leq \epsilon/2$ with probability at least 0.9. So we make a mistake with probability at most 0.1. And if $P \in \mathcal{Q}$, then by assumption $\|P - U\|_1$ is at least ϵ. Since further by assumption $\|\hat{P} - P\|_1 < \epsilon/2$, it follows by the triangle inequality that $\|\hat{P} - U\|_1 > \epsilon/2$. So we see that indeed we have constructed an appropriate decision statistics and threshold for this case.

We have seen in Section 7.1.8 that in expectation the ℓ_1-distance is upper bounded by $\sqrt{\frac{k-1}{n}}$. This means that if we want the ℓ_1 risk to be bounded by $\epsilon/2$ with some fixed probability then $O(k/\epsilon^2)$ samples will suffice.

A Better Approach

We can do better than that. There is no reason we first have to learn the whole distribution if at the end we are only interested in this one bit of information. We will now see that $\sqrt{k}/\epsilon^2$ samples suffice. This might not seem a big deal if the alphabet size is small but for large alphabet sizes this is significant.

A Lower Bound

We claim that we need at least $\Omega(\sqrt{k})$ samples for any fixed ϵ. This bound does not give the correct scaling with respect to ϵ but it does tell us that we cannot hope to do better than $\sqrt{k}$ with respect to the alphabe size.

Recall that $\mathcal{P} = \{U\}$, where U is the uniform distribution on $\{1, \cdots, k\}$. Let X^N be iid samples according to U, where N is chosen according to a $\text{Poi}(n)$ distribution, and assume that $n \leq k$. Recall that in this setting $T_i(X^N)$ has distribution $\text{Poi}(n/k)$. The probability that symbol i is chosen 2 or more times is equal to

$$\sum_{j \geq 2} e^{-n/k} \frac{(n/k)^j}{j!}.$$

But since for $\lambda = n/k \leq 1$, $\sum_{j\geq 2} \frac{\lambda^j}{j!} \leq \lambda^2 \sum_{j \geq 2} \frac{1}{j!} \leq \lambda^2$, $\sum_{j \geq 2} e^{-n/k} \frac{(n/k)^j}{j!} \leq e^{-n/k}(n/k)^2$. Therefore, the expected number of symbols that appear more than once is upper bounded by $ke^{-n/k}(n/k)^2 = e^{-n/k}n^2/k$.

Assume that we pick $n < \sqrt{k}/10$. Then this expected value is upper bounded by $1/100$. So let us recap. We have a random variable, call it Z, that is integer-valued and non-negative and whose expected value is upper bounded by $1/100$. So

$$1/100 \geq \mathbb{E}[Z] = \sum_{i \geq 0} \mathbb{P}\{Z = i\} i \geq \sum_{i \geq 1} \mathbb{P}\{Z = i\} = \mathbb{P}\{Z \geq 1\}.$$

Thefore the probability that none of the symbols appear at least twice is upper bounded by $1/100$. This is called the first moment method.

Now consider a distribution, call it $\tilde{U}$ that is also uniform, but uniform on a subset of $\{1, \cdots, k\}$ of size $k/2$. By exactly the same argument, replacing k with $k/2$ everywhere, we have that with probability at most $1/50$ we see any of the symbols repeated more than once. We conclude that, under these conditions, we cannot hope to be able to distinguish between those two distributions. And clearly these two distributions are quite different. In fact, their ℓ_1 distance is 1!

We recognize that $n \sim \sqrt{k}$ is not an arbitrary threshold. This is the threshold that we know from the *birthday* paradox. This is not a coincidence. As we will discuss in more depth when we discuss property estimation, essentially the only information that is contained in the samples *are* the overlaps. So $n \sim \sqrt{k}$ is when we start getting useful information.

An Upper Bound

Now where we have the "right" (we don't know this yet, but soon ...) scaling in k let us look at an actual algorithm that gives us the desired result of $\sqrt{k}/\epsilon^2$ samples. Let us first relate ℓ_1 to ℓ_2.

Lemma 7.2. *Let $P, Q \in \Delta_k$. If $\|P - Q\|_1 \geq \epsilon$ then $\|P - Q\|_2^2 \geq \epsilon^2/k$.*

Proof. By Cauchy-Schwarz $|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle$, with $u^\top = (|P_1 - Q_1|, \cdots, |P_k - Q_k|)$, and $v^\top = (1, \cdots, 1)$ we have

$$\underbrace{\sum_i (P_i - Q_i)^2}_{\langle u,u\rangle} \underbrace{k}_{\langle v,v\rangle} \geq (\sum_i |P_i - Q_i|)^2 \geq \epsilon^2.$$

□

Lemma 7.2 might give you pause. Why do we go via an ℓ_2 route? Did we not claim a few pages ago that ℓ_2 is not a good metric when it comes to large alphabet sizes? Indeed this is the case. In the current context we assumed that "uniform" means that the support of U is *all* of the alphabet $\mathcal{X} = \{1, \cdots, k\}$. If we considered a slightly more general scenario where we allowed U to have a support that was strictly smaller than k, as long as all symbols with non-negative weight have equal weight, then we would have to proceed in a different manner.

Before we proceed let us quickly recall some facts about Poisson distributions.

Lemma 7.3. *Let X be a random variable with Poisson distribution Poi(λ). Then for $l \geq 1$*

$$\mathbb{E}[X(X-1)\cdots(X-l+1)] = \lambda^l.$$

Further, if μ is any real number then

$$\mathrm{Var}((X-\mu)^2 - X) = 2\lambda^2 + 4\lambda(\lambda - \mu)^2.$$

Proof. Consider the first statement. The generating function of the Poisson distribution is $e^{\lambda(x-1)}$. Taking the derivative with respect to x and then setting $x = 1$ gives us the mean since this corresponds to the weighted sum with weight i. We get $\lambda e^{\lambda(x-1)}|_{x=1} = \lambda$. More generally, taking the l-th derivative of $e^{\lambda(x-1)}$ and then setting $x = 1$ gives us λ^l and this corresponds to the stated expression.

Now consider the second statement. Expanding $\mathbb{E}[(X-\mu)^2 - X]$ and using the previous result we see that $\mathbb{E}[(X-\mu)^2 - X] = (\mu - \lambda)^2$. To compute the variance write down the corresponding expected value and expand in terms of $X(X-1)\cdots(X-l+1)$ for $l = 0$ up to $l = 3$. Use the previous trick to get the result. □

Recall that we need a test statistics. We claim that

$$T(X^n) = \sum_i (T_i(X^n) - \frac{n}{k})^2 - T_i(X^n)$$

is a good candidate.

NOTE: This is perhaps not the best of notation. $T_i(X^n)$ refers to the count of the symbol i in the sample X^n, whereas $T(X^n)$ refers to the test statistics.

Lemma 7.4. *Let $P, Q \in \Delta_k$. Let N be chosen according to Poi(n) and let X^N be N iid samples according to P. Then*

$$\mathbb{E}[\sum_i (T_i(X^N) - nQ_i)^2 - T_i(X^N)] = n^2 \sum_i (P_i - Q_i)^2.$$

Proof. Recall that $T_i(X^N)$ has distribution $\mathrm{Poi}(nP_i)$. Therefore,

$$\begin{aligned}\mathbb{E}[\sum_i (T_i(X^N) - nQ_i)^2 - T_i(X^N)] &= \mathbb{E}[\sum_i T_i(X^N)(T_i(X^N)-1) - 2nT_i(X^N)Q_i + n^2Q_i^2] \\ &\overset{\text{Lemma 7.3}}{=} \sum_i [n^2P_i^2 - 2n^2P_iQ_i + n^2Q_i^2] \\ &= \sum_i n^2(P_i - Q_i)^2.\end{aligned}$$

□

Assume now that $P \in \mathcal{P} = \{U\}$, i.e., $P = U$. Then Lemma 7.4 tells us that

$$\mathbb{E}[T(X^N)] = 0.$$

But if P is such that $\|P - U\|_1 \geq \epsilon$ then

$$\mathbb{E}[T(X^N)] \overset{\text{Lemma 7.4}}{=} n^2 \sum_i (P_i - \frac{1}{k})^2 \overset{\text{Lemma 7.2}}{\geq} \frac{n^2\epsilon^2}{k}.$$

We are now ready to state the algorithm that has the claimed performance.

1. Obtain $N \sim \mathrm{Poi}(n)$ iid samples X^N from P, where P is unknown.
2. Ouput decision according to

$$\begin{cases} \mathcal{P}, & T(X^N) < \tau = \frac{n^2\epsilon^2}{2k}, \\ \mathcal{Q}, & T(X^N) > \tau. \end{cases}$$

Lemma 7.5. *Consider the previous algorithm and assume that $n > \sqrt{80k}/\epsilon^2$. Then*

1. *if $P \in \mathcal{P}$ then $\mathbb{P}\{T(X^N) > \tau\} < 0.1$,*
2. *if $P \in \mathcal{Q}$ then $\mathbb{P}\{T(X^N) < \tau\} < 0.1$.*

Proof. Let us look at the two cases separately. If $P = U$ then we know that $\mathbb{E}[T(X^N)] = 0$. From Lemma 7.3 with $\mu = \lambda = n/k$, and taking into account that the $T_i(X^N)$ are independent we get

$$\mathrm{Var}(T(X^N)) = k2(n/k)^2 = 2\frac{n^2}{k}.$$

By the Chebyshev inequality

$$\mathbb{P}\{T(X^N) > \tau\} \leq \frac{\mathrm{Var}(T(X^N))}{\tau^2} = \frac{2n^2}{k}\frac{4k^2}{n^4\epsilon^4} = \frac{8}{n^2\epsilon^4}$$

If we want the right-hand-side to no more than say 0.1 then solving $\frac{8k}{n^2\epsilon^4} = 0.1$ for n shows that $n \geq \sqrt{80k}\epsilon^2$, as promised. The second case follows in a similar manner and we skip the details. □

7.3 Property Estimation

We now get the last topic. We have seen how to estimate distributions and how to test properties of distributions. Let us now look how we can estimate properties of distributions.

There are plenty of properties that one might be interested in: entropy, support size, or perhaps the mutual information between two densities.

The simplest approach is to use so-called *plug-in* estimators. This means, estimate the distribution and then plug in these estimates into the functional that computes the desired quantity. But the question is if it is really necessary (and optimal) first to learn the whole distribution if all that we are interested in is one number.

7.3.1 Entropy Estimation

The set-up is very similar than what we used for in the distribution estimation scenario. We have an alphabet $\mathcal{X} = \{1, \cdots, k\}$. We get iid samples $X^n = X_1, \cdots, X_n$ that are drawn according to a fixed but unknown distribution p.

We are given a functional $f(p)$,

$$f(p) = \sum_{i=1}^{k} f(p_i).$$

E.g., if we are interested in the entropy then we want to compute $f(p) = \sum_i p_i \log_2 \frac{1}{p_i}$.

Our aim therefore is to design a functional $\hat{f} : \mathcal{X}^n \to R$ that is best in our usual min-max sense,

$$\min_{\hat{f}} \max_{p \in \Delta_k} \mathbb{E}[(f(p) - \hat{f}(X^n))^2]$$

As always one of the most natural such estimators is to use the *empirical one*

$$\hat{f}^{\text{emp}}(X^n) = \sum_i f(\frac{T_i(X^n)}{n}).$$

7.4 Problems

Problem 7.1 (ℓ_1 versus Total Variation)**.** In class we defined the ℓ_1 distance as

$$\|p - q\|_1 = \sum_{i=1}^{k} |p_i - q_i|.$$

Another important distance is the total variation distance $d_{\text{TV}}(p, q)$. It is defined as

$$d_{\text{TV}}(p, q) = \max_{S \subseteq \{1, \cdots, k\}} |\sum_{i \in S} (p_i - q_i)|.$$

Show that if p, q are two probability mass vectors (i.e. elements of the simplex) we have that $d_{\text{TV}}(p, q) = \frac{1}{2}\|p - q\|_1$.

Problem 7.2 (Poisson Sampling)**.** Assume that we have given a distribution p on $\mathcal{X} = \{1, \cdots, k\}$. Let X^n denote a sequence of n iid samples. Let $T_i = T_i(X^n)$ be the number of times symbol i appears in X^n. Then

$$\mathbb{P}\{T_i = t_i\} = \binom{n}{t_i} p_i^{t_i} (1 - p_i)^{n - t_i}.$$

Note that the random variables T_i are *dependent*, since $\sum_i T_i = n$. This dependence can sometimes be inconvenient.

There is a convenient way of getting around this problem. Thit is called *Poisson* sampling. Let N be a random variable distributed according to a Poisson distribution with mean n. Let X^N be then an iid sequence of N variables distributed according to p.

Conditioned on $N = n$, what is the induced distribution of the Poisson sampling scheme? Show that

1. $T_i(X^N)$ is distributed according to a Poisson random variable with mean $p_i n$.
2. The $T_i(X^N)$ are independent.

Problem 7.3 (Add-β Estimator)**.** The add-β estimator $q_{+\beta}$ over $[k]$, assigns to symbol i a probability proportional to its number of occurrences plus β, namely,

$$q_i \stackrel{\text{def}}{=} q_i(X^n) \stackrel{\text{def}}{=} q_{+\beta,i}(X^n) \stackrel{\text{def}}{=} \frac{T_i + \beta}{n + k\beta}$$

where $T_i \stackrel{\text{def}}{=} T_i(X^n) \stackrel{\text{def}}{=} \sum_{j=1}^n \mathbf{1}(X_j = i)$. Prove that for all $k \geq 2$ and $n \geq 1$,

$$\min_{\beta \geq 0} r_{k,n}^{l_2^2}(q_{+\beta}) = r_{k,n}^{l_2^2}(q_{+\sqrt{n}/k}) = \frac{1 - \frac{1}{k}}{(\sqrt{n} + 1)^2}$$

Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every distribution $p \in \Delta_k$.

Problem 7.4 (Uniformity Testing)**.** Let us reconsider the problem of testing against uniformity. In the lecture we saw a particular *test statistics* that required only $O(\sqrt{k}/\epsilon^2)$ samples where ϵ was the ℓ_1 distance.

Let us now derive a test from scratch. To make things simple let us consider the ℓ_2^2 distance. Recall that the alphabet is $\mathcal{X} = \{1, \cdots, k\}$, where k is known. Let U be the uniform distribution on $\mathcal{X}$, i.e., $u_i = 1/k$. Let P be a given distribution with components p_i. Let X^n be a set of n iid samples. A pair of samples (X_i, X_j), $i \neq j$, is said to *collide* if $X_i = X_j$, if they take on the same value.

1. Show that the expected number of collisions is equal to $\binom{n}{2}\|p\|_2^2$.
2. Show that the uniform distribution minimizes this quantity and compute this minimum.
3. Show that $\|p - u\|_2^2 = \|p\|_2^2 - \frac{1}{k}$.

   NOTE: In words, if we want to distinguish between the uniform distribution and distributions P that have an ℓ_2^2 distance from U of at least ϵ, then this implies that for those distributions $\|p\|_2^2 \geq 1/k + \epsilon$. Together with the first point this suggests the following test: compute the number of collisions in a sample and compare it to $\binom{n}{2}(1/k + \epsilon/2)$. If it is below this threshold decide on the uniform one. What remains is to compute the variance of the collision number as a function of the sample size. This will tell us how many samples we need in order for the test to be reliable.

4. Let $a = \sum_i p_i^2$ and $b = \sum_i p_i^3$. Show that the variance of the collision number is equal to

$$\begin{aligned}&\binom{n}{2}a + \binom{n}{2}\left[\binom{n}{2} - \left(1 + \binom{n-2}{2}\right)\right]b + \binom{n}{2}\binom{n-2}{2}a^2 - \binom{n}{2}^2 a^2\\ =&\binom{n}{2}\left[2b(n-2) + a(1 + a(3-2n))\right]\end{aligned}$$

   by giving an interpretation of each of the terms in the above sum.

   NOTE: If you don't have sufficient time, skip this step and go to the last point.

   For the uniform distribution this is equal to

$$\binom{n}{2}\frac{(k-1)(2n-3)}{k^2} \leq \frac{n^2}{2k}.$$

   NOTE: You don't have to derive this from the previous result. Just assume it.

5. Recall that we are considering the ℓ_2^2 distance which becomes generically small when k is large. Therefore, the proper scale to consider is $\epsilon = \kappa/k$. Use the Chebyshev inequality and conclude that if we have $\Theta(\sqrt{k}/\kappa)$ samples then with high probability the empirical number of collisions will be less than $\binom{n}{2}(1/k + \kappa/(2k))$ assuming that we get samples from a uniform distribution.

NOTE: The second part, namely verifying that the number of collisions is with high probability smaller than $\binom{n}{2}(1/k+\kappa/(2k))$ when we get $\Theta(\sqrt{k}/\kappa)$ samples from a distribution with ℓ_2^2 distance at least κ/k away from a uniform distribution follows in a similar way.

HINT: Note that if p represents a vector with components p_i then $\|p\|_1 = \sum_i |p_i|$ and $\|p\|_2^2 = \sum_i p_i^2$.

Problem 7.5 (Estimating Support Size)**.** You are attending Balelec. You want to estimate how many people are attending. Let this number be m. Here is a very simple algorithm. You walk around randomly. Every 5 minutes you take a picture of the person who is right next to at this moment. Assume that 5 minutes is sufficiently long so that in this manner you sample participants at Balelec with uniform probability. Assume further that during the whole time you do your experiment no person joins or leaves Balelec.

You do this N times, where N is a Poisson random variable with mean $n = 100$. Once you are done you look at the photos. Assume that in total you have encountered $K = 102$ distinct people. Out of those 102, 100 you have seen only once, one you saw twice, and one you saw three times. Give an estimate of the number of people attending Balelec (the support size of the distribution). Call this number $\hat{m}$. We do not expect a number as answer since the estiate might involve an optimization step which might not be trivial to do by hand. Simplify as far as you can and then write down how you would get final answer.

Hint: Follow your own path or answer the question according to the following steps.

1. Assume that there are m people attending Balelec. Take a specific person at Balelec. Call this person "1". Given the procedure outlined above, what is the probability that this person appears c_1, $c_1 \geq 0$, times on your photos?

2. Now take two specific people. Call them "1" and "2". What is the probability that they appear $\{c_i\}_{i=1}^2$ times on your photos?

3. Now consider all people all Balelec together. Assume as before that each has a specific identity. What is the probability that the m people appear $\{c_i\}_{i=1}^m$ times on your photos?

4. Assume again that m people attend Balelec and also as before that we have the counts $\{c_i\}_{i=1}^m$. But this time we do not know who has what count, i.e., we do not know the identites of the people. All we know is the counts themselves. What is the probability of getting the counts $\{c_i\}_{i=1}^m$? [Note: What we see are the non-zero counts, but since we also assume that we know m, we know in fact all counts.]

5. How can you use the last expression to derive an estimate?

Problem 7.6 (Estimating Entropy)**.** You are given n iid samples of a Bernoulli random variable with parameter μ. The parameter is known to be in the range $[\kappa, 1-\kappa]$, where $0 < \kappa \leq \frac{1}{2}$. Let the samples be denoted by $S = \{X_1, X_2, \cdots, X_n\}$, $X_i \in \{0, 1\}$, $i = 1, \cdots n$.

Your task is to estimate the entropy of the underlying distribution accurately. Let h denote the true entropy of the distribution and $\hat{h} = \hat{h}(S)$ be your estimate.

(i) Design a scheme to accurately estimate h. Give an explicit epression for $\hat{h}$ as a function of the samples S.

(ii) Since the samples S are random your estimate $\hat{h}(S)$ is a random variable. Let $\delta, \epsilon > 0$. Derive a bound of the form

$$\mathbb{P}\{|\hat{h}(S) - h| \geq \epsilon\} \leq \delta.$$

(iii) [5pts] In the expression of (ii) assume that you set δ to some fixed constant. How does the gap ϵ behave as a function of n?

Hint: Simple does it.

Problem 7.7 (l_2 Estimation)**.** Assume that we have two distributions p and q on $\{1, \cdots K\}$. Let $n \in \mathbb{N}$. Let $N_1, N_2 \sim \text{Poi}(n)$ be independent random variables. We are given N iid samples from each, call them $\{X_j\}_{j=1}^{N_1}$ and $\{Y_j\}_{j=1}^{N_2}$, respectively. Let $t_k(X^{N_1})$, $k = 1, \cdots, K$, respectively, $t_k(Y^{N_2})$, denote the empirical counts. E.g.,

$$t_k(x^n) = |\{j \in \{1, \ldots n\} : x_j = k\}|.$$

We want to estimate $\|p - q\|_2^2$.

Define $Z = \sum_{k=1}^K (t_k(X^{N_1}) - t_k(Y^{N_2}))^2 - t_k(X^{N_1}) - t_k(Y^{N_2})$. We claim that Z/n^2 is a good estimator for $\|p - q\|_2^2$.

(a) Show that Z is an unbiased estimator of $n^2\|p - q\|_2^2$.

Hint: The expression for Z should look somewhat familiar. The notes are your best friend.

(b) Assuming that $\|p\|_2^2 \leq b$ and $\|q\|_2^2 \leq b$ show that the variance of Z can be upper bounded in the following way:

$$\begin{aligned}\mathrm{Var}(Z) &\overset{(i)}{=} \sum_{k=1}^{K} 4n^3(p_k - q_k)^2(p_k + q_k) + 2n^2(p_k + q_k)^2 \\ &\overset{(ii)}{\leq} \sum_{k=1}^{K} 8n^3(p_k - q_k)^2 + 2n^2(p_k^2 + q_k^2 + 2p_k q_k) \\ &\overset{(iii)}{\leq} 8n^3\|p - q\|_2^2 + 8n^2 b.\end{aligned}$$

Justify each of the three steps.

Hint: Define $R = (U - V)^2 - U - V$, where $U \sim \mathrm{Poi}(\lambda)$ and $V \sim \mathrm{Poi}(\mu)$. A straighforward but tedious calculation shows that $\mathrm{Var}(R) = 4(\lambda-\mu)^2(\lambda+\mu)+2(\lambda+\mu)^2$.

(c) Show that $\mathbb{P}\{|Z/n^2 - \|p - q\|_2^2| \geq \epsilon\} \leq \frac{8n\|p-q\|_2^2+8b}{n^2\epsilon^2}$.

Chapter 8

Exponential Families and Maximum Entropy Distributions

Exponential families are a class of parametrized distributions. They are important for several reasons. First, many "standard" distributions we are well acquainted with (like the Gaussian distribution) are members of this family. Therefore, they appear frequently in applications. Second, all members of this family have nice theoretical properties. Hence, rather than discussing these properties for each member of this family, it is convenient to discuss them for the whole family at once.

To give some "practical motivations," in machine learning exponential families are used in the context of classification, giving rise to so-called *generalized linear models*. Further, these are the distributions that maximize the *entropy* given constraints on the moments. E.g., we will see that the Gaussian has the maximum entropy of any family with a given second moment constraint. It is therefore natural to consider such distributions as prior distributions since they make "the least assumptions" if all we know are constraints on moments.

We will be relatively terse. If you want to dig deeper, we recommend the lecture notes by John Duchi [10], Chapter 6 and 7, or the extensive monograph by Wainwright and Jordan [11].

8.1 Definition

Definition 8.1. Let $\mathcal{X}$ be a given alphabet and let $\phi : \mathcal{X} \to \mathbb{R}^d$, $d \in \mathbb{N}$. The *exponential family* associated with ϕ is the set of distributions parametrized by $\theta \in \mathbb{R}^d$ with densities given by

$$p_\theta(x) = h(x)e^{\langle \theta, \phi(x) \rangle - A(\theta)}.$$

Note that $A(\theta)$ is a *normalizing* constant. As such it might not seem to play an important role. But, as we will discuss soon, it in fact encodes (in its derivatives) crucial information. The function $A(\theta)$ is some-times called the *log-partition function* (the partition function is a term used in statistical physics for the normalization constant and A is the log of this). In statistics it is known as the *cumulant* function.

In our definition of an exponential family we included the term $h(x)$. In principle this term can be absorbed by the underlying measure $\nu(x)$ and is in this sense redundant. But it might sometimes be more "natural" to represent a distribution in this way. For all our

subsequent computations and proofs of properties it does not really matter which point of view we take, i.e., if we explicitly write out the term $h(x)$ think of it as being included in the underlying measure.

8.2 Examples

Example 8.1 (Gaussian)**.** Let $\mathcal{X} = \mathbb{R}$ and let ν be the Lebesgue measure on $\mathbb{R}$. Then the density of the normal distribution with mean μ and variance σ^2 can be written as

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= e^{x\frac{\mu}{\sigma^2} - x^2\frac{1}{2\sigma^2} - [\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2)]} \\ &= e^{\langle\theta,\phi(x)\rangle - A(\theta)}, \end{aligned}$$

where $h(x) = 1$, $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^\top$, $\phi(x) = (x, x^2)^\top$, and

$$A = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\ln(-\theta_2/\pi).$$

Note that $\phi(x)$ is a vector of dimension 2, reflecting the fact that the Gaussian has two degrees of freedom. Further, we have the following bijective relationships

$$\begin{aligned} \theta = (\theta_1, \theta_2)^\top &= (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^\top, \\ (\mu, \sigma^2)^\top &= (-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2})^\top. \end{aligned}$$

Example 8.2 (Poisson)**.** Let $\mathcal{X} = \mathbb{N}$ and let ν be the counting measure on $\mathcal{X}$. We can represent the Poisson distribution with parameter λ in the form

$$\begin{aligned} P(X = x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{1}{x!} e^{x\ln(\lambda) - \lambda} \\ &= \frac{1}{x!} e^{\theta x - e^\theta} \\ &= h(x) e^{\langle\theta,\phi(x)\rangle - A(\theta)}, \end{aligned}$$

where $h(x) = 1/x!$, $\theta = \ln(\lambda)$, $\phi(x) = x$, and $A(\theta) = e^\theta$.

Example 8.3 (Bernoulli)**.** Let $\mathcal{X} = \{0, 1\}$ and let ν be the couting measure on $\mathcal{X}$. We can represent the Bernoulli distribution with $P(X = 1) = p$ in the form

$$\begin{aligned} P(X = x) &= p^x (1-p)^{1-x} \\ &= e^{x\ln p + (1-x)\ln(1-p)} \\ &= e^{x\ln\frac{p}{1-p} + \ln(1-p)} \\ &= h(x) e^{\langle\theta,\phi(x)\rangle - A(\theta)}, \end{aligned}$$

where $h(x) = 1$, $\theta = \ln\frac{p}{1-p}$, so that $p = \frac{e^\theta}{1+e^\theta}$, $\phi(x) = x$, and $A(\theta) = -\ln(1-p) = -\ln(1 - \frac{e^\theta}{1+e^\theta}) = \ln(1 + e^\theta)$.

Example 8.4 (Multinomial)**.** A generalization of the Bernoulli measure is the multinomial. Let $\mathcal{X} = \{0, \cdots, n\}^d$ and let ν be the counting measure on $\mathcal{X}$. Then the multinomial distribution with parameter $\alpha = (\alpha_1, \cdots, \alpha_d)$ can be expressed as

$$\begin{aligned} P(X_1 = x_1, \cdots, X_d = x_d) &= \binom{n}{x_1, \cdots, x_d} \prod_{i=1}^{d} \alpha_i^{x_i} \\ &= \binom{n}{x_1, \cdots, x_d} e^{\sum_{i=1}^{d} \ln(\alpha_i) x_i} \\ &= h(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \end{aligned}$$

where

$$h(x) = \binom{n}{x_1, \cdots, x_d},$$

$\theta = (\ln(\alpha_1), \cdots, \ln(\alpha_d))$, $\phi(x) = x = (x_1, \cdots, x_d)$, and $A(\theta) = 0$.

Example 8.5 (Dirichlet)**.** The Dirichlet distribution of order $d \geq 2$ with parameter $\alpha = (\alpha_1, \cdots, \alpha_d)$, $\alpha_i > 0$, has a density with respect to the Lebesgue measure on $\mathbb{R}^{d-1}$ of the form

$$\begin{aligned} p_\theta(x) &= \frac{1}{B(\alpha)} \prod_{i=1}^{d} x_i^{\alpha_i - 1} \\ &= h(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)} \end{aligned}$$

where x belongs to the $(d-1)$-dimensional simplex, i.e., $\sum_{i=1}^{d} x_i = 1$, $x_i \geq 0$, and where

$$B(\alpha) = \frac{\prod_{i=1}^{d} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{d} \alpha_i)}.$$

Further, $h(x) = 1$, $\theta = (\alpha_1 - 1, \cdots, \alpha_d - 1)$, $\phi(x) = (\ln(x_1), \cdots, \ln(x_d))$, and $A(\theta) = \ln(B(\alpha))$.

If $d = 2$ then the Dirichlet distribution is called the Beta distribution.

8.3 Convexity of $A(\theta)$

Theorem 8.1. *Let $\Theta = \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$. The log-partition function $A(\theta)$ is convex in θ on Θ.*

Proof. Let $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2$, $\theta_1, \theta_2 \in \Theta$. Let $p = \frac{1}{\lambda}$ and $q = \frac{1}{1-\lambda}$ so that $p\lambda = q(1-\lambda) = 1$. Note that $1/p + 1/q = \lambda + (1 - \lambda) = 1$ and that $p, q \in [1, \infty]$. Hölder's inequality states that $\|fg\|_1 \leq \|f\|_p \|g\|_q$. In more detail,

$$\left(\int |f(x)g(x)| d\nu(x) \right) \leq \left(\int |f(x)|^p d\nu(x) \right)^{\frac{1}{p}} \left(\int |g(x)|^q d\nu(x) \right)^{\frac{1}{q}}.$$

Recall that $p_\theta(x) = h(x)e^{\langle\theta,\phi(x)\rangle - A(\theta)}$ so that $1 = \int p_\theta(x)d\nu(x) = \left[\int h(x)e^{\langle\theta,\phi(x)\rangle}d\nu(x)\right]e^{-A(\theta)}$, or, $A(\theta) = \ln\left[\int h(x)e^{\langle\theta,\phi(x)\rangle}d\nu(x)\right]$. We have

$$\begin{aligned}
A(\theta_\lambda) &= \ln\left[\int h(x)e^{\langle\theta_\lambda,\phi(x)\rangle}d\nu(x)\right] \\
&= \ln\left[\int \underbrace{\left(h(x)e^{\langle\theta_1,\phi(x)\rangle}\right)^{\lambda}}_{f(x)}\underbrace{\left(h(x)e^{\langle\theta_2,\phi(x)\rangle}\right)^{(1-\lambda)}}_{g(x)}d\nu(x)\right] \\
&\overset{\text{Hölder}}{\leq} \ln\left[\left(\int\left(h(x)e^{\langle\theta_1,\phi(x)\rangle}\right)^{p\lambda}d\nu(x)\right)^{\frac{1}{p}}\left(\int\left(h(x)e^{\langle\theta_2,\phi(x)\rangle}\right)^{q(1-\lambda)}d\nu(x)\right)^{\frac{1}{q}}\right] \\
&= \ln\left[\left(\int\left(h(x)e^{\langle\theta_1,\phi(x)\rangle}\right)^{p\lambda}d\nu(x)\right)^{\frac{1}{p}}\right] + \ln\left[\left(\int\left(h(x)e^{\langle\theta_2,\phi(x)\rangle}\right)^{q(1-\lambda)}d\nu(x)\right)^{\frac{1}{q}}\right] \\
&\overset{p\lambda=q(1-\lambda)=1}{=} \frac{1}{p}\ln\left[\left(\int h(x)e^{\langle\theta_1,\phi(x)\rangle}d\nu(x)\right)\right] + \frac{1}{q}\ln\left[\int\left(h(x)e^{\langle\theta_2,\phi(x)\rangle}d\nu(x)\right)\right] \\
&= \lambda A(\theta_1) + (1-\lambda)A(\theta_2).
\end{aligned}$$

□

8.4 Derivatives of $A(\theta)$

Without proof we state that $A(\theta)$ is infinitely often differentiable on Θ. In particular the first two derivatives are of interest to us.

Let us compute the first derivative (gradient). We have

$$\begin{aligned}
\nabla A(\theta) &= \nabla \ln \int h(x)e^{\langle\theta,\phi(x)\rangle}d\nu(x) \\
&= \frac{\int \nabla h(x)e^{\langle\theta,\phi(x)\rangle}d\nu(x)}{\int h(x)e^{\langle\theta,\phi(x)\rangle}d\nu(x)} \\
&= \frac{\int h(x)e^{\langle\theta,\phi(x)\rangle}\phi(x)d\nu(x)}{e^{A(\theta)}} \\
&= \int h(x)e^{\langle\theta,\phi(x)\rangle - A(\theta)}\phi(x)d\nu(x) \\
&= \mathbb{E}[\phi(x)].
\end{aligned}$$

For future reference, let us record that

$$\nabla A(\theta) = \mathbb{E}[\phi(x)]. \tag{8.1}$$

In a similar manner we have

$$\nabla^2 A(\theta) = \mathbb{E}[\phi(x)\phi(x)^\top] - \mathbb{E}[\phi(x)]\mathbb{E}[\phi(x)^\top].$$

Note that this gives us a second proof that $A(\theta)$ is convex since we see that the Hessian of $A(\theta)$ is a convariance matrix and hence positive semidefinite.

Example 8.6 (Bernoulli)**.** For the Bernoulli distribution we have seen that $A(\theta) = \ln(1+e^\theta)$ and $\theta = \ln \frac{p}{1-p}$. Therefore,

$$\frac{dA(\theta)}{d\theta} = \frac{d\ln(1+e^\theta)}{d\theta} = \frac{e^\theta}{1+e^\theta} = \sigma(\theta) = p,$$
$$\frac{d^2A(\theta)}{d\theta^2} = \frac{d\sigma(\theta)}{d\theta} = \sigma(\theta)(1-\sigma(\theta)) = p(1-p).$$

8.5 Application to Parameter Estimation and Machine Learning

The convexity of $A(\theta)$ is one of the main reason why this family of distributions is so convenient to work with.

Assume that we have a set of samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$ and that we assume that they are iid according to an exponential family with an unknown parameter θ. We want to estimate this parameter.

We can then write down the likelihood as

$$p_\theta(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \prod_{n=1}^{N} h(\mathbf{x}_n) e^{\langle \theta, \phi(\mathbf{x}_n)\rangle - A(\theta)}. \tag{8.2}$$

Instead of maximizing this likelihood we can equivalently take the log of this expression, multiply by minus one, and minimize instead. This gives us

$$-\ln p_\theta(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \sum_{n=1}^{N} [-\ln(h(\mathbf{x}_n)) - \langle \theta, \phi(\mathbf{x}_n)\rangle + A(\theta)] \tag{8.3}$$

Now note that the function on the right is convex – it is the sum of the constant (with respect to θ) $-\sum_{n=1}^{N} \ln(h(\mathbf{x}_i))$, the linear function $-\sum_{n=1}^{N} \langle \theta, \phi(\mathbf{x}_n)\rangle$ and the convex function $NA(\theta)$. Greedy, local algorithms are therefore expected to work well in locating the optimal parameter θ.

If we take the gradient of the above expression and set it to zero then we get the equation

$$\mathbb{E}_\theta[\phi(x)] = \frac{1}{N} \sum_{n=1}^{N} \phi(\mathbf{x}_n). \tag{8.4}$$

In words, we should choose the parameter θ in such a way that the expected value of $\phi(x)$ equals its empirical value. From this we see why $\phi(\cdot)$ is called the *sufficient statistics*. We only need this quantity for the parameter estimation.

A word of caution is in order here. Just because a function is convex, it does not mean that it is easy to minimize. We need in addition that the function itself (and perhaps its derivative) is easy to compute. If you look ahead at the *Ising model* described in Example 8.13, then you will see that even though the function $A(\theta)$ to be minimized is convex, there is no low-complexity algorithm known to accomplish this minimization since the computation of $A(\theta)$ requires in general exponential effort.

8.6 Conjugate Priors

In the previous section we considered an application in ML where we estimated the underlying parameter θ, given some iid samples from the distribution by maximizing the

likelihood. We have seen that for exponential distributions the underlying maximization problem is "simple" since the underlying function is convex. The justification for maximizing the likelihood is that under some technical conditions this leads to a consistent estimator (see Section 8.10.2).

Alternatively, in the Baysian setting, we assume that there is a prior on the set of parameters and we will then maximize the posterior instead. In this case the question is what prior we should pick. One part of the question is what priors are "meaningful" or "appropriate." Leaving out this question for the moment, there is still the question what priors lead to "manageable" computational tasks, e.g., convex functions to be minimized. Here is where conjugate priors enter. If we start with a likelihood that is a member of an exponential family and use as a conjugate prior then we end up again with an element for the exponential family. Rather than discussing this in the abstract, let us look at some important examples.

Example 8.7 (Bernoulli)**.** Consider a Bernoulli distribution with parameter p,

$$P_p(X = x) = p^x(1-p)^{1-x},$$

where we recall $x \in \{0, 1\}$. Assume that the parameter $p \in [0, 1]$ is unknown and follows a a Beta distribution $q(p)$, i.e.,

$$q(p) = K(\alpha_1, \alpha_2)p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1},$$

where $\alpha, \alpha_2 > 0$ so that the density can be normalized and where $K(\alpha_1, \alpha_2)$ is the normalization constant, $K(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}$.

Let us now quickly discuss, why this prior is convenient. Assume that we have a set of iid samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$. As in Section 8.5 we assume that they are iid according to a Bernoulli distribution with an unknown parameter p. In addition we assume that the parameter itself is distributed according to $q(p)$ with the parameters α_1 and α_2 fixed. Let us then write down the posterior distribution for the parameter p given the samples. We have

$$p(p \mid \mathbf{x}_1, \cdots, \mathbf{x}_N) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \prod_{n=1}^{N} p^{\mathbf{x}_n}(1-p)^{1-\mathbf{x}_n} \tag{8.5}$$

$$\propto p^{\alpha_1 - 1 + \sum_{n=1}^{N} \mathbf{x}_n}(1-p)^{\alpha_2 - 1 + N - \sum_{n=1}^{N} \mathbf{x}_n}. \tag{8.6}$$

The key is to notice that this is again a beta distribution but now with parameters $\sum_{n=1}^{N} \mathbf{x}_n + \alpha_1$ and $N - \sum_{n=1}^{N} \mathbf{x}_n + \alpha_2$. In particular, this is again an exponential distribution and so the optimization of this expression is again "simple."

Example 8.8 (Multinomial)**.** The conjugate prior of a Multinonomial is a Dirichlet distribution.

Example 8.9 (Gaussian)**.** The conjuate prior of a Gaussian is a Gaussian.

8.7 Maximum Entropy Distributions

Assume that we are given a function $\phi : \mathcal{X} \to \mathbb{R}^d$ and a vector $\alpha \in \mathbb{R}^d$. What distribution on $\mathcal{X}$ maximizes the entropy subject to the condition that the expected value of $\phi(X)$ is equal to α? Mathematically, we are looking for

$$P^* = \text{argmax}_{P : \mathbb{E}_P[\phi(X)] = \alpha} H(P)$$

Why are we interested in this problem? If all that we know is a constraint on the mean it makes sense to look at the "most random" distribution that fulfills this constraint. This is the distribution that makes the least "assumptions" if we use it as a prior.

When looking at maximum entropy distributions we will drop the factor $h(x)$ from exponential families. As we have mentioned earlier, any specific factor $h(x)$ can be absorbed into the underlying measure $\nu(x)$ and this is indeed the natural view point for our current purpose.

Theorem 8.2. *For $\theta \in \mathbb{R}^d$, let P_θ have density*

$$p_\theta(x) = \exp\{\langle \theta, \phi(x)\rangle - A(\theta)\}$$

with respect to the measure ν. If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then P_θ maximizes $H(P)$ over $\{P : \mathbb{E}_{P_\theta}[\phi(X)] = \alpha\}$ and it is the unique distribution with this property.

Proof. Let θ be a parameter so that $\mathbb{E}_{p_\theta}[\phi(X)] = \alpha$ and let P be any other distribution so that $\mathbb{E}_p[\phi(X)] = \alpha$. Then

$$\begin{aligned}
H(P) &= -\int p(x)\log p(x) d\nu(x) \\
&= -\int p(x)\log p_\theta(x) d\nu(x) + \int p(x)\log p_\theta(x) d\nu(x) - \int p(x)\log p(x) d\nu(x) \\
&= -\int p(x)\log p_\theta(x) d\nu(x) - \int p(x)\log \frac{p(x)}{p_\theta(x)} d\nu(x) \\
&= -\int p(x)[\langle \theta, \phi(x)\rangle - A(\theta)] d\nu(x) - D(p(x)\|p_\theta(x)) \\
&= -\int p_\theta(x)[\langle \theta, \phi(x)\rangle - A(\theta)] d\nu(x) - D(p(x)\|p_\theta(x)) \\
&= H(P_\theta) - \underbrace{D(p(x)\|p_\theta(x))}_{\geq 0} \\
&\leq H(P_\theta).
\end{aligned}$$

□

In all three of the following examples we pick $\phi(x) = x^2$ and $\alpha = 1$, i.e., we are constraining the distribution P by asking that the second moment is equal to 1, $\mathbb{E}_P[X^2] = 1$. The general form of the density that maximizes the entropy is then

$$p_\theta(x) = \frac{\exp\{x^2\theta\}}{Z}, \tag{8.7}$$

where Z is the normalizing constant.

Example 8.10. If the measure ν is the counting measure on $\{-1, 1\}$ then the distribution P is of the form

$$(P(x = -1) = \frac{e^\theta}{Z}, P(X = 1) = \frac{e^\theta}{Z}),$$

where we have the condition $\mathbb{E}_P[X^2] = 2\frac{e^\theta}{Z} = 1$. Hence the maximum entropy distribution $P(x)$ is $(P(x = -1) = \frac{1}{2}, P(X = 1) = \frac{1}{2})$, the uniform distribution.

Example 8.11. If the measure ν is the counting measure on $\mathbb{Z}$ then the distribution P is of the form

$$p_\theta(x) = \frac{e^{-\theta x^2}}{\sum_i e^{-\theta i^2}}, x \in \mathbb{Z},$$

where θ is chosen so that

$$\sum_{x\in\mathbb{Z}} x^2 \frac{e^{-\theta x^2}}{\sum_i e^{-\theta i^2}} = 1.$$

Example 8.12. If the measure ν is the Lebesque measure on $\mathbb{R}$ then we recognize from the basic form of the density given in (8.7) that the density that maximizes the entropy is the Gaussian distribution with mean 0 and variance 1,

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

In the proof of Theorem 8.2 we have seen that *if* an exponential distribution exists that yields the right moment then it is the maximum entropy distribution with this moment. Assume for a moment that we did not already *know* the form of this distribution. It is then perhaps insightful to "derive" the form of the distribution from first principles. Consider the following Lagrangian:

$$L = \int p(x) \log p(x) d\nu(x) + \theta^\top (\mu - \int p(x)\phi(x) d\nu(x)) + \kappa(1 - \int p(x) d\nu(x)).$$

Our aim is to minimize this Lagrangian. The first term is equal to $-H(P)$. Indeed, we want to maximize $H(P)$, i.e., equivalenty we want to minimize $-H(P)$. The second term corresponds to all the constraints on the moments. Here, θ is a vector of length d. And the third term corresponds to the normalization constraint on the density. Note that we have not included any term to ensure that the "density" is non-negative. We will see in a second that even without adding this constraint the solution will automatically fulfill this constraint, hence there is no need to add this constraint explicitly.

If we now take the "derivative" with respect to $p(x)$ and set it to 0 we get

$$0 = 1 + \log(p(x)) - \langle \theta, \phi(x) \rangle - \kappa.$$

Solving for $p(x)$,

$$p(x) = e^{\langle \theta, \phi(x) \rangle + \kappa - 1}.$$

This is of course an exponential distribution as expected. Note that due to the special structure of the solution $p(x) \geq 0$ is automatically fulfilled.

8.8 Application To Physics

Let us re-derive one of the basic laws of physics – the Maxwell-Boltzmann distribution.

Assume that we have particles in $\mathbb{R}^3$. They each have a position and a velocity vector associated to them. We will not be interested in the position but we are asking how the

velocity vectors are distributed. Let $\mathbf{v} = (v_1, v_2, v_2)$ be the velocity vector associated to a particular particle.

We associate an average "kinetic energy" E (per particle) to the distribution

$$\int p(\mathbf{v}) \frac{1}{2} m(\mathbf{v}_1^2 + \mathbf{v}_2^2 + \mathbf{v}_3^2) d\mathbf{v} = E, \tag{8.8}$$

where m is the mass of a particle (all are assumed to have equal mass).

Let $s = \sqrt{\mathbf{v}_1^2 + \mathbf{v}_2^2 + \mathbf{v}_3^2}$, the speed. What is the maximum entropy distribution $p(s)$? Note that in this case $\phi(\mathbf{v}) = \mathbf{v}_1^2 + \mathbf{v}_2^2 + \mathbf{v}_3^2$. Therefore, the form of the maximizing distribution is

$$p(\mathbf{v}) = e^{\theta(\mathbf{v}_1^2 + \mathbf{v}_2^2 + \mathbf{v}_3^2) - A(\theta)}.$$

We recognize this to be a three-dimensional zero-mean Gaussian distribution with indepenen and identically distributed components. We conclude that each component is distributed according to

$$p(v) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{v^2}{2\sigma^2}},$$

for some value of σ^2. Going back to our orignal constraint (8.8) we see that $\sigma^2 = \frac{2E}{3m}$. Summarizing, the velocity distribution of each component has the form

$$p(v) = \sqrt{\frac{3m}{4\pi E}} e^{-\frac{3mv^2}{4E}}.$$

What is the induced distribution of the overall speed s? Recall that the surface of a sphere (in 3 D) of radius s has area $4\pi s^2$. Hence,

$$\mathbb{P}\{s \leq S \leq s + ds\} = \left(\frac{3m}{4\pi E}\right)^{\frac{3}{2}} e^{-\frac{3ms^2}{4E}} 4\pi s^2 ds,$$

so that

$$p(s) = \sqrt{\frac{27m^3}{4\pi E^3}} s^2 e^{-\frac{3ms^2}{4E}}. \tag{8.9}$$

Appealing to thermodynamics, we write E as $E = 3/2kT$, where k is the Boltzmann constant and T is the temperature. The factor 3 accounts for the three degrees of freedom and $\frac{1}{2}kT$ is the kinetic energy per degree of freedom. Then we get the usual form

$$p(s) = \sqrt{\frac{2m^3}{\pi k^3 T^3}} s^2 e^{-\frac{ms^2}{2kT}}.$$

As an alternative derivation, we could have found $p(s)$ by directly finding the maximum entropy distribution on $\mathcal{X} = \mathbb{R}^+$ with measure $d\nu(s) = 4\pi s^2 ds$, for $s \geq 0$. In this case we know that

$$p(s) = e^{-\theta s^2 - A(\theta)} \mathbb{1}_{\{s \geq 0\}}.$$

The normalizing condition reads

$$\int_{s\geq 0} p(s)d\nu(s) = \int_{s\geq 0} e^{-\theta s^2 - A(\theta)} 4\pi s^2 ds = \frac{\pi^{3/2}}{\theta^{3/2}} e^{-A(\theta)} = 1.$$

This tells us that $A(\theta) = \frac{3}{2}\ln\frac{\pi}{\theta}$. The second moment requires that

$$\mathbb{E}_{p(s)}[s^2] = \int_{s\geq 0} \left(\frac{\theta}{\pi}\right)^{3/2} e^{-\theta s^2} s^2 d\nu(s) = \frac{3}{2\theta} = \frac{2E}{m},$$

so that $\theta = \frac{3m}{4E}$. It follows that the distribution that maximizes this entropy when written with respect to the Lebesque measure on $\mathbb{R}^+$ is equal to

$$p(s) = \left(\frac{\theta}{\pi}\right)^{3/2} 4\pi s^2 e^{-\theta s^2} \mid_{\theta=\frac{3m}{4E}},$$

which is equal to what we got in (8.9).

8.9 I-Projections

In previous lectures we have discussed at length Sanov's theorem. Recall that if we have given a family of distributions, call the family Π, and a fixed distribution P, then the chance that we will mistake samples from P for samples from one of the elements of Π, is exponentially small and the exponent is asymptotically equal to $\operatorname{argmin}_{Q\in\Pi} D(Q\|P)$. The operation of finding the "closest" element of Π is called an I-projection. In general it is difficult to compute this projection. But if the family is linear then the projection is again easy to compute as we will see now.

Theorem 8.3. *Let P be a fixed distribution with density $p(x)$ and let Π be the set of all distributions so that $\mathbb{E}_q[\phi(x)] = \mu$ for $q \in \Pi$. If P_θ has density*

$$p_\theta = p(x)e^{\langle\theta,\phi(x)\rangle - A(\theta)}$$

and $\mathbb{E}_{P_\theta}[\phi(x)] = \mu$ then

$$P_\theta = \operatorname{argmin}_{Q\in\Pi} D(Q\|P).$$

In words, P_θ is the I-projection of P onto Π.

Proof. The proof uses the same idea as we used to show that exponential distributions solve the maximum entropy problem. We have

$$\begin{aligned}
D(Q\|P) &= \int q(x)\log\frac{q(x)}{p(x)}d\nu(x) \\
&= \int q(x)\log\frac{p_\theta(x)}{p(x)}d\nu(x) + \int q(x)\log\frac{q(x)}{p_\theta(x)}d\nu(x) \\
&= \int q(x)[\langle\theta,\phi(x)\rangle - A(\theta)]d\nu(x) + D(Q\|P_\theta) \\
&= \int p_\theta(x)[\langle\theta,\phi(x)\rangle - A(\theta)]d\nu(x) + D(Q\|P_\theta) \\
&= \int p_\theta(x)\log\frac{p_\theta(x)}{p(x)}d\nu(x) + D(Q\|P_\theta) \\
&= D(P_\theta\|P) + D(Q\|P_\theta) \\
&\geq D(P_\theta\|P).
\end{aligned}$$

□

8.10 Relationship between θ and $\mathbb{E}[\phi(x)]$

8.10.1 The forward map $\nabla A(\theta)$

Assume that we fix $h(x)$ and $\phi(x)$. Then for every $\theta \in \Theta$ there is a distribution $p_\theta(x)$ and an associated "mean" $\mathbb{E}_{P_\theta}[\phi(x)]$. This mapping $\theta \mapsto \mu = \mathbb{E}_{P_\theta}[\phi(x)]$ is called the "forward" map. We have seen in (8.1) that it is given by $\nabla_\theta A(\theta)$.

Clearly this mapping is important. For simple distributions as for Bernoulli, Poisson, or Gaussian this map is simple to state and simple to compute. But there are important classes of distributions in the exponential family where this map is computationally difficult. Let us give one such example.

Example 8.13 (Ising Model)**.** The *Ising* model is a classical example from statistical physics which was initially introduced in order to study magnetism. The associated exponential distribution has the form

$$p_\theta(x) = e^{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - A(\theta)}.$$

Here, the x_s take values in $\{\pm 1\}$ and they are called *spins*. The set of spins is V and there is an underlying undirected graph with vertex set V and edge set E. The strength of the "interactions" between two spins that are connected by an edge (s,t) is θ_{st}. There are also the "local fields" θ_s for every spin $s \in V$.

Note that we are here in a much "higher dimensional" setting – $\phi(x)$ has dimension $|V| + |E|$. Given the local fields and the strength of the interactions we are typically interested in the marginals and the pairwise correlations, i.e., we are exactly interested in $\mu = \mathbb{E}_{P_\theta}[\phi(x)]$. In particular we are interested if, e.g., some marginals become strongly "biased" or pairs become strongly correlated. If we stay with the physical interpretation of this model then such an "emergent" bias would represent the emergence of a global magnetic field given the local interaction rules. "Emergent" here means that we envision that we change the parameters of the model (the θ_s and θ_{st}) and that for some such parameters even a small extra change might suddenly lead to biased marginals.

In summary, we are interested in the foward map $\theta \mapsto \mu$. But this map is in general exponentially complex to compute. E.g., if you look at the expression of $A(\theta)$, it has the form

$$A(\theta) = \ln \sum_{x\in\{0,1\}^{|V|}} e^{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t},$$

requiring a sum over an exponential number of terms. And the same is true if you look at the gradient of this expression since $A(\theta)$ is part of this gradient computation. So even though the map is well-defined and mathematically simple to describe, it might be difficult to compute.

Definition 8.2 (Set of Feasible Means)**.** For a fixed $\phi(x)$ let

$$\mathcal{M} = \{\mu \in \mathbb{R}^d : \exists P \text{ so that } \mathbb{E}_P[\phi(x)] = \mu\}.$$

In words, for a fixed sufficient statistics, $\mathcal{M}$ is the set of all means that can be achieved by *some* distribution. It is important here that *P is not assumed to be an element of the exponential family.*

Definition 8.3 (Regular Families)**.** Let $\phi(x)$ be given. We say that the associated exponential family is *regular* if Θ is *open*.

Definition 8.4 (Minimal Families)**.** Let $\phi(x)$ be given. We say that the associated exponential family is *minimal* if there *does not exist* a vector η so that

$$\eta^\top \phi(x) = \text{const},$$

$\nu(x)$-almost everywhere.

Theorem 8.4. *For a regular family the gradient $\nabla_\theta A(\theta) : \Theta \to \mathcal{M}$ is one-to-one if and only if the exponential representation is minimal.*

Proof. Assume at first that the family is not minimal, i.e., there does exist an η so that $\eta^\top \phi(x) = c$, a constant, $\nu(x)$-almost everywhere. Pick a $\theta_1 \in \Theta$. Then for a sufficiently small ϵ, $\theta_2 = \theta_1 + \epsilon\eta \in \Theta$ since we assumed that Θ was open (the family is regular). Note that $A(\theta_2) = A(\theta_1) + \epsilon c$. Therefore

$$\nabla_\theta A(\theta_1) = \nabla_\theta A(\theta_2).$$

Conversely, assume that the family is minimal. We claim that in this case $A(\theta)$ is strictly convex. This implies that

$$\begin{aligned} A(\theta_2) &> A(\theta_1) + \langle \nabla_\theta A(\theta_1), \theta_2 - \theta_1 \rangle, \\ A(\theta_1) &> A(\theta_2) + \langle \nabla_\theta A(\theta_2), \theta_1 - \theta_2 \rangle. \end{aligned}$$

We therefore have

$$\langle \nabla_\theta A(\theta_1), \theta_1 - \theta_2 \rangle > A(\theta_1) - A(\theta_2) > \langle \nabla_\theta A(\theta_2), \theta_1 - \theta_2 \rangle.$$

This implies that

$$\langle \nabla_\theta A(\theta_1) - \nabla_\theta A(\theta_2), \theta_1 - \theta_2 \rangle > 0.$$

It remains to explain why $A(\theta)$ is strictly convex for a minimal family. Recall that $\nabla^2_\theta A(\theta)$ is the convariance matrix of $\phi(x)$. So $A(\theta)$ is always convex. If the family is minimal then for no η is $\langle \eta, \phi(x) \rangle$ a constant. We conclude that the covariance of $\langle \eta, \phi(x) \rangle$ is strictly positive. But this covariance is equal to $\eta^\top \nabla^2_\theta A(\theta) \eta$, and so this quantity is strictly positive for any $\eta \in \Theta$. □

8.10.2 The backward map

When we discussed the maximum entropy problem we had to assume that for a given mean vector μ there exists a parameter θ so that $\mathbb{E}_{P_\theta}[\phi(x)] = \mu$. Only then could we conclude that the maximum entropy solution is an element of the exponential distribution. As we will see now, this is not really much of a restriction as long as there is *some* distribution that has this mean.

Theorem 8.5. *In a minimal exponential family, the gradient map $\nabla_\theta A(\theta) : \Theta \to \mathcal{M}$ is onto the interior of $\mathcal{M}$.*

We will not provide a proof here but refer the reader to [11].

We can therefore define a backward map from the interior of $\mathcal{M}$ onto Θ. This has the pleasing consequence that if we are looking for a maximum entropy distribution then as long as we pick a mean vector from the interior of $\mathcal{M}$ then the solution will be an element of the exponential family.

This has another important consequence. Let us go back to the parameter estimation problem discussed in Section 8.5. Assume that the samples do come from a minimal exponential family with sufficient statistic $\phi(x)$ and that the parameter θ_0 is such that $\nabla A(\theta_0) = \mu$ is in the interior of $\mathcal{M}$. Assume that we compute the empirical mean

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} \phi(x).$$

We know that $\hat{\mu} \overset{N\to\infty}{\to} \mu$ almost surely. Therefore, if we estimate the parameter by applying the inverse map to $\hat{\mu}$ then this estimate will converge almost surely to the true parameter θ_0. In other words, this estimator is *consistent*. This gives us an well-founded justifcation for using the ML estimator in the first place.

8.11 Problems

Problem 8.1. Find the parametric form of the maximum entropy density f satisfying the Laplace transform condition

$$\int f(x)e^{-x}dx = \alpha,$$

and give the constraints on the parameter.

Problem 8.2 (Exponential Families and Maximum Entropy 2)**.** Find the maximum entropy density f, defined for $x \geq 0$, satisfying $\mathbb{E}[X] = \alpha_1$, $\mathbb{E}[\ln X] = \alpha_2$. That is, maximize $-\int f \ln f$ subject to $\int xf(x)dx = \alpha_1$, $\int (\ln x)f(x)dx = \alpha_2$, where the integral is over $0 \leq x < \infty$. What family of densities is this?

Problem 8.3. What is the maximum entropy distribution $p(x, y)$ that has the following marginals?

x \ y	1	2	3	
1	p_{11}	p_{12}	p_{12}	$\frac{1}{2}$
2	p_{21}	p_{22}	p_{23}	$\frac{1}{4}$
3	p_{31}	p_{32}	p_{33}	$\frac{1}{4}$
	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	

Problem 8.4. *(a)* What is the parametric-form maximum entropy density $f(x)$ satisfying the two conditions

$$\mathbb{E}[X^8] = a \qquad \mathbb{E}[X^{16}] = b$$

(b) What is the maximum entropy density satisfying the condition

$$\mathbb{E}[X^8 + X^{16}] = a + b$$

(c) Which entropy is higher?

Problem 8.5. What is the maximum entropy distribution, call it $p(x, i)$, on $[0, \infty] \times \mathbb{N}$, both of whose marginals have mean $\mu > 0$. (I.e., in one axis the distribution is over the positive reals, whereas in the other one it is over the natural numbers.)

Problem 8.6 (Exponential Families and Maximum Entropy: I-projections)**.** Let P denote the zero-mean and unit-variance Gaussian distribution. Assume that you are given N iid samples distributed according to P and let $\hat{P}_N$ be the empirical distribution.

Let Π denote the set of distributions with second moment $\mathbb{E}[X^2] = 2$. We are interested in

$$\lim_{N\to\infty} \frac{1}{N} \log \Pr\{\hat{P_N} \in \Pi\} = - \inf_{Q\in\Pi} D(Q\|P).$$

(a) Determine $-\text{arginf}_{Q\in\Pi} D(Q\|P)$, i.e., determine the element Q for which the infinum is taken on.

(b) Determine $-\inf_{Q\in\Pi} D(Q\|P)$.

Problem 8.7. We learned in the course that as long as the set of feasible means is open then every such mean can be realized by an element of the exponential family. In the following verify this explicitly (by not referring to the above statement for the following scenario).

(i) Let $\phi(x) = (x^2)$.

(ii) Let $\phi(x)$ consist of all elements $x_i x_j$, where i and j go from 1 to K.

Problem 8.8 (Exponential Families and Maximum Entropy)**.** Let $Y = X_1 + X_2$. Find the maximum entropy of Y under the constraint $\mathbb{E}[X_1^2] = P_1$, $\mathbb{E}[X_2^2] = P_2$:

(a) If X_1 and X_2 are independent.

(b) If X_1 and X_2 are allowed to be dependent.

Problem 8.9 (Exponential Families and Maximum Entropy)**.** For $t > 0$, consider a family of distributions supported on $[t, +\infty]$ such that $\mathbb{E}[\ln X] = \frac{1}{\alpha} + \ln t$, $\alpha > 0$.

1. What is the parametric form of a maximum entropy distribution satisfying the constraint on the support and the mean?

2. Find the exact form of the distribution.

Chapter 9

Signal Representations

It is generally assumed that the student has basic familiarity with the early topics in this chapter, such as bases, projections, and so on. Canonical references include [12, 13, 14].

Introduction

"Signal representation" refers to the act of representing a "signal" x as a linear combination of elements in a "dictionary" composed of elements ϕ_ℓ, where ℓ runs over integers :

$$x = \sum_{\ell \in \mathbb{Z}} X_\ell \phi_\ell. \tag{9.1}$$

In (9.1) the signal is represented exactly but sometimes it is also useful to consider approximations. The coefficients X_ℓ (9.1) are real or complex numbers. The basic idea is that once we have such a representation, instead of working with the signal x, we may work with its representation. Often this turns out to be convenient.

The primary object of study is to find good and "suitable" dictionaries $\{\phi_\ell\}_{\ell \in \mathbb{Z}}$. In this module, we discuss the main methods and arguments relating to this quest. What is meant by "suitable" above will vary on the application. But common criteria are that this leads to *sparse* representations or to efficient processing.

Why do we discuss such representations in a course about Data Science? Before the latest ML revolution, scientists and engineers spent a considerable effort into how to represent signals efficiently. Much of this effort was seemingly swept away by the latest developments in neural networks and the attitude right now is to let the system itself learn a suitable representation. This works quite well. But it also has drawbacks. Neural networks can represent a very large class of signals. The downside is that for a particular application you might in fact not need this level of generality and you pay a price by potentially needing more samples or having a considerable larger computational cost. Therefore, depending on the application, it might be of considerable advantage to start with a representation that is taylored to the given use case. And this means being aware of some of the underlying trade-offs.

There is a second important connection. The representations we discuss can be considered as one form of "signal compression". We want to use as few dimensions as possible to represent a signal. We will discuss several other forms of compression, namely the Johnson-Lindenstrauss dimensionality reduction scheme discussed in Section 10.2.2, as well as the information-theoretic notion of compression. All three of these schemes try in some sense

to accomplish a similar aim, but there are important distinctions between them. This does not only make for good exam questions but it is important to clarify their differences when deciding in an application how to proceed.

9.1 Fourier Representations

It is assumed that you have come across Fourier representations at least three times in your previous education (specifically, in your classes *Analysis III, Circuits & Systems II,* and *Signal Processing for Communications*). We here present a very brief overview, emphasizing some of the more advanced aspects.

Among all signal representations, Fourier representations are arguably the most important ones. This is due to several important reasons. First of all, they represent eigenvectors of LTI systems. Further reasons include important connections to wide-sense stationary signals and observations that many naturally occurring signal classes (audio, images, etc) have specific characteristics in the frequency domain. Moreover, Fourier representations can be calculated efficiently and have many desirable properties.

9.1.1 DFT and FFT

For the *discrete Fourier transform* (DFT), we follow the notation used in your prerequisite class, see [15, Section 4.2]. In line with this, let

$$W_N \;=\; e^{-j\frac{2\pi}{N}}. \tag{9.2}$$

The Fourier matrix W is the matrix whose entry in row k, column n, is given by

$$\{W\}_{kn} \;=\; W_N^{(k-1)(n-1)}, \text{ for } n,k \in \{1,2,\ldots,N\}. \tag{9.3}$$

and the DFT of the vector $\mathbf{x}$ is the vector $\mathbf{X}$ defined as

$$\mathbf{X} \;=\; W\mathbf{x}. \tag{9.4}$$

With this, the inverse transform is

$$\mathbf{x} \;=\; \frac{1}{N}W^H\mathbf{X}. \tag{9.5}$$

Explicitly, we can express the $(k+1)$ entry of the vector $\mathbf{X}$ (for $k = 0,1,\ldots,N-1$) as

$$X[k] \;=\; \langle \mathbf{x}, \mathbf{w}_k \rangle = \mathbf{w}_k^H\mathbf{x} = \sum_{n=0}^{N-1} x[n]e^{-j2\pi\frac{kn}{N}}, \tag{9.6}$$

where the very last expression illustrates a slight notational anomaly, namely, we denote the elements of the signal vector $\mathbf{x}$ by $x[0], x[1], \cdots, x[N-1]$, that is, we number them from 0 to $N-1$. With this, we choose to follow the terminology used in [15, Section 4.2]. Also, we have used the notation $\mathbf{w}_k = (1, W_N^k, W_N^{2k}, \cdots, W_N^{(N-1)k})^H$.

Properties of the DFT

One of the most important reasons for the importance of the DFT is the wealth of useful properties it has. You have encountered these in detail in [15, Chapter 4].

Cyclic shifts. Consider the signal vector $\mathbf{x}$ of length N and with entries denoted $x[0], \ldots, x[N-1]$. Let $\mathbf{y}$ be the signal vector $\mathbf{x}$, cyclically shifted to the right by n_0 positions. We have the following DFT pair:

$$y[n] = x[(n - n_0) \mod N] \quad \circ\!\!-\!\!\bullet \quad Y[k] = W_N^{kn_0} X[k], \tag{9.7}$$

where $X[0], \cdots, X[N-1]$ denote the entries of the DFT vector $\mathbf{X} = W\mathbf{x}$.

To establish this property, it is more convenient to use the sum representation than the matrix-vector representation. Namely,

$$Y[k] = \sum_{n=0}^{N-1} y[n] e^{-j\frac{2\pi}{N}kn} = \sum_{n=0}^{N-1} x[(n - n_0) \mod N] e^{-j\frac{2\pi}{N}kn} \tag{9.8}$$

and change summation variables by defining $m = n - n_0$, which yields

$$Y[k] = \sum_{m=-n_0}^{N-1-n_0} x[m \mod N] e^{-j\frac{2\pi}{N}k(m+n_0)}. \tag{9.9}$$

We can rewrite the exponent as $e^{-j\frac{2\pi}{N}k(m+n_0)} = e^{-j\frac{2\pi}{N}k(m \mod N)} e^{-j\frac{2\pi}{N}kn_0}$ and introduce $\ell = m \mod N$ to obtain $Y[k] = e^{-j\frac{2\pi}{N}kn_0} \sum_{\ell=0}^{N-1} x[\ell] e^{-j\frac{2\pi}{N}k\ell}$, which completes the proof.

Modulation property. Consider the signal vector $\mathbf{x}$ of length N and with entries denoted $x[0], \ldots, x[N-1]$. Let $\mathbf{y}$ be the signal vector with entries

$$y[n] = W_N^{-k_0 n} x[n] \quad \circ\!\!-\!\!\bullet \quad Y[k] = X[(k - k_0) \mod N], \tag{9.10}$$

and the proof can be done following exactly the same steps as for the cyclic shift property.

Duality. A key observation is that these two properties are essentially one and the same. This is a reflection of the fact that DFT and inverse DFT are essentially the same (up to a complex-conjugate), and thus, the time and frequency variables can be exchanged.

9.1.2 The Other Fourier Representations

In your prerequisite classes, you have encountered several Fourier representations. For the theoretical understanding of the underpinnings and underlying ideas, the most important is the Fourier transform,

$$X(j\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt, \tag{9.11}$$

whose inverse is given by

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega) e^{j\omega t} d\omega. \tag{9.12}$$

As you recall, there are several subtle aspects as to whether this inversion formula will indeed give back the original signal. Those will not be of particular interest to our class.

9.2 The Hilbert Space Framework for Signal Representation

Perhaps the most powerful framework to understand signal representation and approximation is that of *Hilbert space* which you have briefly encountered in the class *Signal Processing for Communications* [15, Chapter 3].[1]

A (real or complex) *vector space* is a set of vectors $\mathbf{x} \in E$ with an addition for vectors, denoted by $+$, and a scalar multiplication (i.e., multiplication of a vector $\mathbf{x}$ by a real- or complex-valued scalar α) such that for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in E$ and all scalars α, β :

- Commutativity : $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
- Associativity : $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$, and $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$.
- Distributive laws : $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$, and $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$.
- There exists a vector $\mathbf{0} \in E$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x} \in E$.
- For all $\mathbf{x} \in E$, there exists an element $-\mathbf{x} \in E$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
- For all $\mathbf{x} \in E$, $1 \cdot \mathbf{x} = \mathbf{x}$.

A (real or complex) *inner product space* is a (real or complex) vector space together with an inner product $\langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{R}$ or $\mathbb{C}$ satisfying, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in E$ and scalars α,

- $\langle \mathbf{x} + \mathbf{z}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle$.
- $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$.
- $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle^*$.
- $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$.

The induced norm of the inner product space is defined as $\|\mathbf{x}\| \stackrel{\text{def}}{=} \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. This definition directly implies the following important and useful facts:

- Cauchy-Schwarz inequality: $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \ \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in E$, with equality if and only if $\mathbf{x} = \alpha\mathbf{y}$ for some scalar α.
- Triangle inequality: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, with equality if and only if $\mathbf{x} = \alpha\mathbf{y}$ for some real-valued non-negative scalar α.
- Paralellogram identity: $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$.

For example, to establish the Cauchy-Schwarz inequality, we may start by observing that $\|\mathbf{x} - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}\mathbf{y}\|^2 \geq 0$, which holds by the definition of the norm. Writing this norm in terms of inner products and repeatedly applying the properties of the inner product leads to the Cauchy-Schwarz inequality.

The final key ingredient pertains to the *convergence* of sequences of vectors $\mathbf{x}_n \in E$. Quite naturally, we say that such a sequence converges to $\mathbf{x} \in E$ if $\lim_{n\to\infty} \|\mathbf{x}_n - \mathbf{x}\| = 0$. A sequence $\mathbf{x}_n \in E$ is called a *Cauchy sequence* if $\lim_{m,n\to\infty} \|\mathbf{x}_m - \mathbf{x}_n\| = 0$. Then, a *Hilbert space* is an inner product space with the additional property that *every* Cauchy sequence converges to a vector $\mathbf{x} \in E$. (This can be thought of as a technical condition which for the purpose of our class will not matter too much since it is satisfied for all examples of interest to us.)

Example 9.1 (n-dimensional complex vector space)**.** This is the usual vector space with inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i^*$. The induced norm is $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}$.

[1] Our treatment here closely follows the development in the excellent textbook by Pierre Brémaud [13]. Alternatively, you may consult [14, Chapter 2] and/or follow the class COM-514 *Mathematical Foundations of Signal Processing*.

Example 9.2 (Square-integrable functions (often denoted as $L^2(\mathbb{R})$ or $L_2(\mathbb{R})$))**.** The set of all functions $f(t)$ satisfying $\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty$, with inner product $\langle f, g\rangle = \int_{-\infty}^{\infty} f(t)g^*(t)dt$, is a Hilbert space. The induced norm is $\|f\| = \sqrt{\int_{-\infty}^{\infty} |f(t)|^2 dt}$.

Projection Theorem

For signal representation problems, the main reason why the Hilbert space framework is powerful is the projection theorem. This theorem tackles the following question : Given a Hilbert space H and a subspace G of H such that G is also a Hilbert space (meaning that G is also closed). Then, a very common task is that of representing any element $\mathbf{x} \in H$ only using elements from the subspace G "in the best possible way." Of course, since G is smaller than H, this leads to a more compact (approximate) representation of x, and is thus of obvious interest for many applications. More precisely, for any $\mathbf{x} \in H$, we are looking for an approximation $\hat{\mathbf{x}} \in G$ such that $\|\mathbf{x} - \hat{\mathbf{x}}\|$ is as small as possible. The projection theorem guarantees existence and uniqueness of this miminizer, and it establishes that the minimizer has the very useful property that it is *orthogonal* to the approximation error, i.e., $\langle \hat{\mathbf{x}}, \mathbf{x} - \hat{\mathbf{x}}\rangle = 0$. This last relationship is often called the *orthogonality principle* and considerably simplifies the problem of finding the best approximation $\hat{\mathbf{x}}$. For any Hilbert subspace G, let us define $G^{\perp} = \{\mathbf{z} \in H : \langle \mathbf{z}, \mathbf{x}\rangle = 0, \forall \mathbf{x} \in G\}$. Then, we have the following statement:

Theorem 9.1. *Let $\mathbf{x} \in H$. There exists a unique element $\mathbf{y} \in G$ such that $\mathbf{x} - \mathbf{y} \in G^{\perp}$. Moreover, $\|\mathbf{y} - \mathbf{x}\| = \inf_{\mathbf{u}\in G} \|\mathbf{u} - \mathbf{x}\|$.*

A proof can be found e.g. in [13, Sec. C1] or in [14, Ch.2]. We will also explore it to some extent in the Homework.

Orthonormal Basis

A collection of vectors $\{\mathbf{e}_n\}_{n\geq 0}$ in a Hilbert space H is called an *orthonormal system* if $\langle \mathbf{e}_n, \mathbf{e}_k\rangle = 0$ for all $n \neq k$, and $\|\mathbf{e}_n\| = 1$, for all $n \geq 0$.

Theorem 9.2 (Hilbert Basis Theorem)**.** *$\{\mathbf{e}_n\}_{n\geq 0}$ is an orthonormal system in H. Then, the following statements are equivalent:*

- *$\{\mathbf{e}_n\}_{n\geq 0}$ generates the Hilbert space H.*
- *For all $\mathbf{x} \in H$, we have $\|\mathbf{x}\|^2 = \sum_n |\langle \mathbf{x}, \mathbf{e}_n\rangle|^2$.*
- *For all $\mathbf{x} \in H$, we have $\mathbf{x} = \sum_n \langle \mathbf{x}, \mathbf{e}_n\rangle \mathbf{e}_n$*

Theorem 9.3 (Projection theorem, revisited)**.** *Suppose G is spanned by the orthonormal basis $\{\mathbf{g}_n\}_{n\geq 0}$. Then, the element $\mathbf{y} \in G$ that attains $\min_{\mathbf{u}\in G} \|\mathbf{u} - \mathbf{x}\|$ is given by $y = \sum_n \langle \mathbf{x}, \mathbf{g}_n\rangle \mathbf{g}_n$.*

9.3 General Bases, Frames, and Time-Frequency Analysis

9.3.1 The General Transform

Definition

A useful general way of thinking of transforms is in the shape of inner products with a set of "basis" functions:

$$T_x(\gamma) = \langle x(t), \phi_\gamma(t) \rangle \tag{9.13}$$

$$= \int_{-\infty}^{\infty} x(t)\phi_\gamma^*(t)dt, \tag{9.14}$$

where * denotes the complex conjugate.

The idea here is that 'T' denotes what kind of "basis" functions are being used and γ is the index of a basis function. The basis functions are $\phi_\gamma(t)$ for all values of γ.

A good way of thinking about this is that for a fixed γ, the transform coefficient $T_x(\gamma)$ is the result of projecting the original signal $x(t)$ onto the "basis" element $\phi_\gamma(t)$.

An example is the Fourier transform, where instead of the letter γ, we more often use the letter Ω, and where $\phi_\Omega(t) = e^{j\Omega t}$. Hence, in line with the above general notation, we could write

$$FT_x(\Omega) = \langle x(t), \phi_\Omega(t) \rangle \tag{9.15}$$

$$= \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt, \tag{9.16}$$

Of course, we more often simply write $X(\Omega)$ (or $X(j\Omega)$) in place of $FT_x(\Omega)$.

Alternative Formulation

For our next step, we need the (general) Parseval/Plancherel formula, which asserts that

$$\int_{-\infty}^{\infty} f(t)g^*(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\Omega)G^*(j\Omega)d\Omega. \tag{9.17}$$

Using this, we can rewrite the general transform as

$$T_x(\gamma) = \langle x(t), \phi_\gamma(t) \rangle \tag{9.18}$$

$$= \int_{-\infty}^{\infty} x(t)\phi_\gamma^*(t)dt \tag{9.19}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\Omega)\Phi_\gamma^*(j\Omega)d\Omega \tag{9.20}$$

$$= \langle X(j\Omega), \frac{1}{2\pi}\Phi_\gamma(j\Omega) \rangle \tag{9.21}$$

Hence, we now have *two* good ways of thinking about transforms: For a fixed γ, the transform coefficient $T_x(\gamma)$ is the result of projecting the original signal $x(t)$ onto the "basis" element $\phi_\gamma(t)$, and *equivalently*, of projecting the original spectrum $X(j\Omega)$ onto the spectrum of the "basis" element $\phi_\gamma(t)$, which is $\frac{1}{2\pi}\Phi_\gamma(j\Omega)$.

Consider Figure 9.1: Merely as a thought experiment, let us think of a "basis" element $\phi_\gamma(t)$ that lives[2] only inside the box illustrated in Figure 9.1. Then, a great way of thinking about the transform coefficient $T_x(\gamma)$ is that it tells us "how much" of the original signal $x(t)$ sits inside that box.

[2] In the next section, we will make precise what "lives" means.

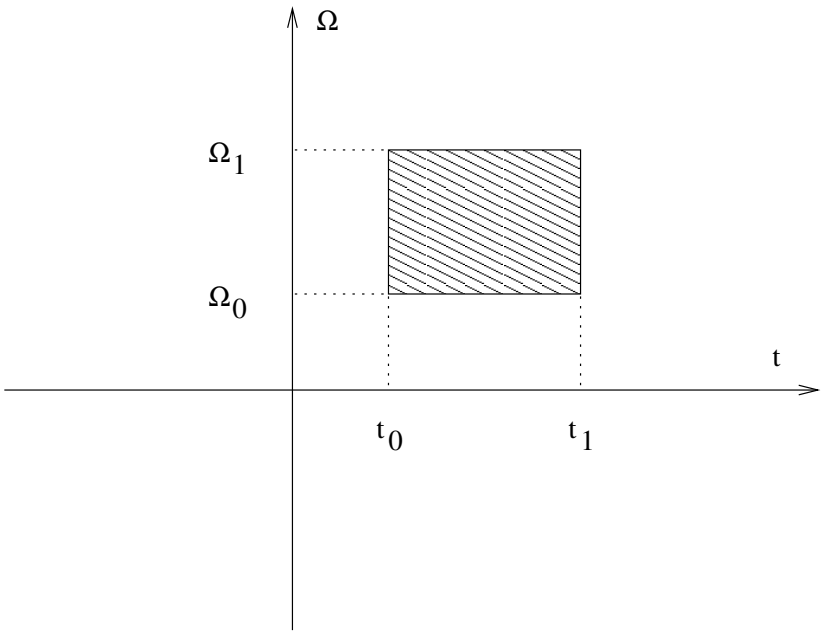


Figure 9.1: A conceptual picture: We imagine that the basis element $\phi_\gamma(t)$ only lives in the shaded box, i.e., that the signal is very small outside the interval $t_0 \leq t \leq t_1$, and that its spectrum $\Phi_\gamma(j\Omega)$ is very small outside of the interval $\Omega_0 \leq \Omega \leq \Omega_1$.

In line with this intuition, for the Fourier transform, the transform coefficient $T_x(\Omega)$ tells us "how much" of the original signal $x(t)$ sits at frequency Ω, and the "box" shown in Figure 9.1 is infinitesimally thin in frequency and infinitely long in time.

9.3.2 The Heisenberg Box Of A Signal

Reconsider the conceptual picture given in Figure 9.1. Now, we want to make this precise. In order to do so, consider any signal $\phi(t)$. For simplicity (and without loss of generality), we assume that the signal is "normalized" such that

$$\int_{-\infty}^{\infty} |\phi(t)|^2 dt = 1. \tag{9.22}$$

Note that by Parseval, this also means that $\frac{1}{2\pi}\int_{-\infty}^{\infty} |\Phi(j\Omega)|^2 d\Omega = 1$.

We define the following quantities. The "middle" of the signal $\phi(t)$ is given by

$$m_t = \int_{-\infty}^{\infty} t|\phi(t)|^2 dt. \tag{9.23}$$

If you have taken a class in probability, you will recognize this to be the mean value of the distribution $|\phi(t)|^2$.

Similarly, we define the "middle" of the spectrum $\Phi(j\Omega)$ to be

$$m_\Omega = \int_{-\infty}^{\infty} \Omega \frac{1}{2\pi} |\Phi(j\Omega)|^2 d\Omega, \tag{9.24}$$

with a similar probability interpretation.

Moreover, we define:

$$\sigma_t^2 = \int_{-\infty}^{\infty} (t - m_t)^2 |\phi(t)|^2 dt, \tag{9.25}$$

$$\sigma_\Omega^2 = \int_{-\infty}^{\infty} (\Omega - m_\Omega)^2 \frac{1}{2\pi} |\Phi(j\Omega)|^2 d\Omega. \tag{9.26}$$

Again, these can be understood as the respective *variances* of the two "probability distributions."

With these definitions, we can now draw a more precise picture of the time-frequency box of the signal $\phi(t)$, as given in Figure 9.2.

We should also point out that for the Fourier transform, the basis functions are of the form $\phi(t) = e^{j\Omega_0 t}$, and for those, the above integrals do not all converge, so special care is required mathematically. However, the right intuition is to say that the Heisenberg box (the term appears in [12], and perhaps earlier) of the function $\phi(t) = e^{j\Omega_0 t}$ is a horizontal line at frequency Ω_0.

9.3.3 The Uncertainty Relation

So, what are the possible Heisenberg boxes?

Theorem 9.4 (uncertainty relation)**.** *For any function $\phi(t)$, the Heisenberg box must satisfy*

$$\sigma_t \sigma_\Omega \geq \frac{1}{2}. \tag{9.27}$$

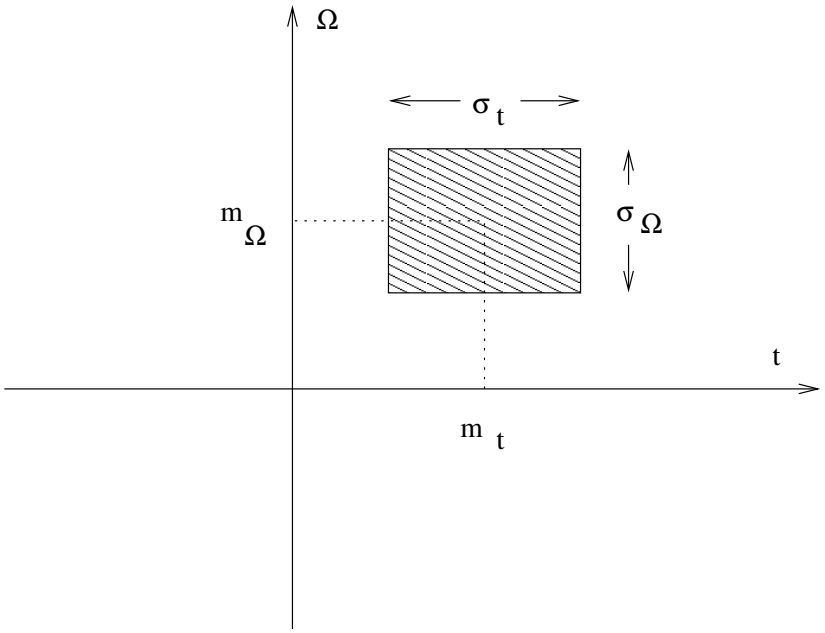


Figure 9.2: The Heisenberg box of the function $\phi(t)$ (i.e., the place in time and frequency where the function $\phi(t)$ is really alive).

That is, Heisenberg boxes cannot be too small. Or: transforms cannot have a very high time resolution *and* a very high frequency resolution at the same time.

Proof. Without loss of generality assume that $m_t = 0$ and $\|\phi(t)\| = 1$, i.e., the signal has mean zero and norm 1. Further, let us assume that the signal $\phi(t)$ is real-valued to simplify the proof somewhat. This in particular implies that $m_f = 0$ as well. We then have

$$\begin{aligned}\left|\int_{-\infty}^{\infty} t\phi(t)\phi'(t)dt\right|^2 &\leq \left(\int_{-\infty}^{\infty} |t\phi(t)|^2 dt\right)\left(\int_{-\infty}^{\infty} |\phi'(t)|^2 dt\right)\\ &\leq \left(\int_{-\infty}^{\infty} |t\phi(t)|^2 dt\right)\frac{1}{2\pi}\left(\int_{-\infty}^{\infty} |j\Omega\Phi(j\Omega)|^2 d\Omega\right)\\ &\leq \sigma_t^2\sigma_\Omega^2.\end{aligned}$$

To finish the proof note that

$$\begin{aligned}\int_{-\infty}^{\infty} t\phi(t)\phi'(t)dt &= \frac{1}{2}\int_{-\infty}^{\infty} t\frac{d\phi(t)^2}{dt}dt\\ &= \underbrace{\frac{1}{2}t\phi(t)^2\,|_{-\infty}^{\infty}}_{=0} - \frac{1}{2}\underbrace{\int_{-\infty}^{\infty}\phi(t)^2 dt}_{=1}\\ &= -\frac{1}{2}.\end{aligned}$$

□

9.3.4 The Short-time Fourier Transform

It has long been recognized that one of the most significant drawbacks of the Fourier transform is its lack of *time localization:* An event that is localized in time (such as a signal discontinuity) affects all of the frequencies (remember the Gibbs phenomenon). This feature is clearly undesirable for many engineering tasks, including compression and classification.

To regain some of the time localization, one could do a "short-time" Fourier transform, essentially chopping up the signal into "short" pieces and taking Fourier transforms separately for each piece. Kind of trivially, this gives back some time localization.

More generally, the following form can be given:

$$STFT_x(\tau,\Omega) = \int_{-\infty}^{\infty} x(t)g^*(t-\tau)e^{-j\Omega t}dt, \tag{9.28}$$

where the function $g(t)$ is an appropriate "window" function that cuts out a piece of the signal $x(t)$. With the parameter τ, we can place the window wherever we want.

With regard to the general transform, here, instead of the letter γ, we use the pair (τ,Ω), and

$$\phi_{\tau,\Omega}(t) = g(t-\tau)e^{j\Omega t}. \tag{9.29}$$

Many different window functions $g(t)$ are being used, but one of the easiest to understand is the Gaussian window:

$$g(t) = \frac{1}{\sqrt[4]{\pi\sigma^2}}e^{-\frac{t^2}{2\sigma^2}}. \tag{9.30}$$

Note that strictly speaking, this window is never zero, so it does not really "cut" the signal. However, if $|t|$ is large, $g(t)$ is tiny, so this is "almost the same as zero," but much easier to analyze. With this window, we find the "basis" elements to be

$$\phi_{\tau_0,\Omega_0}(t) = \frac{1}{\sqrt[4]{\pi\sigma^2}}e^{-\frac{(t-\tau_0)^2}{2\sigma^2}}e^{j\Omega_0 t}. \tag{9.31}$$

Now, we want to find explicitly the Heisenberg box of this "basis" function. To this end, we need the Fourier transform of the Gaussian window, which is known to be

$$G(j\Omega) = \sqrt[4]{4\pi\sigma^2}e^{-\frac{\Omega^2\sigma^2}{2}}, \tag{9.32}$$

and thus, using the standard time- and frequency-shift properties of the Fourier transform,

$$\Phi_{\tau_0,\Omega_0}(j\Omega) = \sqrt[4]{4\pi\sigma^2}e^{-\frac{(\Omega-\Omega_0)^2\sigma^2}{2}}e^{-j\Omega\tau_0}. \tag{9.33}$$

Now, we can find the corresponding parameters of the Heisenberg box as:

$$m_t = \tau_0, \tag{9.34}$$

$$m_\Omega = \Omega_0 \tag{9.35}$$

$$\sigma_t^2 = \frac{\sigma^2}{2} \tag{9.36}$$

$$\sigma_\Omega^2 = \frac{1}{2\sigma^2}, \tag{9.37}$$

and so, we can draw the corresponding Figure 9.2. It is also interesting to note that for the Gaussian window, the Heisenberg uncertainty relation (Theorem 9.4) is satisfied with equality. It can be shown that the Gaussian window is (essentially) the only function that satisfies the uncertainty relation with equality, see e.g. [12, p.31].

9.4 Problems

Problem 9.1 (The Fourier matrix diagonalizes all circulant matrices.)**.** The discrete Fourier transform (DFT) $\mathbf{X}$ of the vector $\mathbf{x}$ is given by

$$\mathbf{X} = W\mathbf{x} \quad \text{and} \quad \mathbf{x} = \frac{1}{N} W^H \mathbf{X}. \tag{9.38}$$

In this homework problem, you will prove that the Fourier matrix diagonalizes all *circulant* matrices.

(a) To cut the derivation into two simpler steps, we introduce an auxiliary matrix M, defined as

$$M = WA = W \underbrace{\begin{pmatrix} b_0 & b_{N-1} & b_{N-2} & b_{N-3} & \dots & b_1 \\ b_1 & b_0 & b_{N-1} & b_{N-2} & \dots & b_2 \\ b_2 & b_1 & b_0 & b_{N-1} & \dots & b_3 \\ b_3 & b_2 & b_1 & b_0 & \dots & b_4 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{N-1} & b_{N-2} & b_{N-3} & b_{N-4} & \dots & b_0 \end{pmatrix}}_{\text{This is a circulant matrix}}. \tag{9.39}$$

Let us denote the unitary DFT of the sequence $\{b_0, b_1, \dots, b_{N-1}\}$ by $\{B_0, B_1, \dots, B_{N-1}\}$. Write out the matrix M in terms of $\{B_0, B_1, \dots, B_{N-1}\}$. *Hint:* The first column of the matrix M is simply given by

$$W \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{N-1} \end{pmatrix} = \begin{pmatrix} B_0 \\ B_1 \\ B_2 \\ B_3 \\ \vdots \\ B_{N-1} \end{pmatrix} \tag{9.40}$$

To find the second column, you will need to use some Fourier properties.

(b) Using the matrix M from above, compute the full matrix product

$$WAW^H = MW^H. \tag{9.41}$$

Hint: Handle every *row* of the matrix M separately. Define the vector $\mathbf{m}$ such that $\mathbf{m}^H$ is simply the first row of the matrix M. But the product $\mathbf{m}^H W^H$ is easily computed, recalling that $\mathbf{m}^H W^H = (W\mathbf{m})^H$.

Problem 9.2 (Inner Products)**.** Consider the standard n-dimensional vector space $\mathbb{R}^n$.

1. Characterize the set of matrices W for which $\mathbf{y}^T W \mathbf{x}$ is a valid inner product for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

2. Prove that *every* inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ on $\mathbb{R}^n$ can be expressed as $\mathbf{y}^T W \mathbf{x}$ for an appropriately chosen matrix W.

3. For a subspace of dimension $k < n$, spanned by the basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k \in \mathbb{R}^n$, express the orthogonal projection operator (matrix) with respect to the general inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T W \mathbf{x}$. *Hint:* For any vector $\mathbf{x} \in \mathbb{R}^n$, express its projection as $\widehat{\mathbf{x}} = \sum_{j=1}^{k} \alpha_j \mathbf{b}_j$.

Problem 9.3 (A Hilbert space of matrices)**.** In this problem, we consider the set of matrices $A \in \mathbb{R}^{m\times n}$ with standard matrix addition and multiplication by scalar.

(a) Briefly argue that this is indeed a vector space, using the definition given in class.

(b) Show that $\langle A, B\rangle = \text{trace}(B^H A)$ is a valid inner product.

(c) Explicitly state the norm induced by this inner product. Is this a norm that you have encountered before?

(d) Consider as a further inner product candidate the form $\langle A, B\rangle = \text{trace}(B^H W A)$, where W is a square $(m \times m)$ matrix. Give conditions on W such that this is a valid inner product. Explicit and detailed arguments are required for full credit.

Problem 9.4 (Canonical Correlation Analysis)**.** Let $\mathbf{X}$ and $\mathbf{Y}$ be zero-mean real-valued random vectors with covariance matrices $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$, respectively. Moreover, let $R_{\mathbf{XY}} = \mathbb{E}[\mathbf{X}\mathbf{Y}^T]$. Our goal is to find vectors $\mathbf{u}$ and $\mathbf{v}$ such as to maximize the correlation between $\mathbf{u}^T\mathbf{X}$ and $\mathbf{v}^T\mathbf{Y}$, that is,

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbb{E}[\mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}]}{\sqrt{\mathbb{E}[|\mathbf{u}^T\mathbf{X}|^2]}\sqrt{\mathbb{E}[|\mathbf{v}^T\mathbf{Y}|^2]}}. \tag{9.42}$$

Show how we can find the optimizing choices of the vectors $\mathbf{u}$ and $\mathbf{v}$ from the problem parameters $R_{\mathbf{X}}, R_{\mathbf{Y}}$, and $R_{\mathbf{XY}}$.

Hint: Recall for the singular value decomposition that

$$\max_{\mathbf{v}} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \sigma_1(A), \tag{9.43}$$

where $\sigma_1(A)$ denotes the maximum singular value of the matrix A. The corresponding maximizer is the right singular vector $\mathbf{v}_1$ (i.e., eigenvector of $A^T A$) corresponding to $\sigma_1(A)$.

Chapter 10

Compression and Dimensionality Reduction

In this chapter we will investigate two kinds of "compressions." First, we will go back to classical data compression and explore the connection between entropy (entropy rate) and the amount of data we need in order to losslessly describe a source. Second, and closer to the SVD example, we explore how data in high dimensions can often be represented in lower dimensions without essential loss in accuracy.

10.1 Data compression

Notation. *Given a set A we denote by A^* the set of all finite sequences $\{(a_1, \ldots, a_n) : n \geq 0, a_i \in A\}$ (including the null sequence λ of length 0). In particular $\{0,1\}^* = \{\lambda, 0, 1, 00, 01, 10, 11, 000, \ldots\}$.*

Consider the problem of assigning binary sequences (also called binary strings) to elements of a finite set $\mathcal{U}$. Such an assignment $c : \mathcal{U} \to \{0,1\}^*$ is called a *binary code* for the set $\mathcal{U}$. The binary string $c(u)$ is called the *codeword* for u. The collection $\{c(u) : u \in \mathcal{U}\}$ is thus the set of codewords.

Definition 10.1. A code c is called *injective* if for all $u \neq v$ we have $c(u) \neq c(v)$.

Definition 10.2. A code c is called *prefix-free* if $c(u)$ is not a prefix of $c(v)$ for all $u \neq v$. In particular, if c is prefix-free then c is injective. (To be clear: a string $a_1 \ldots a_m$ is a prefix of a string $b_1 \ldots b_n$ if $m \leq n$ and $a_i = b_i$ for $i = 1, \ldots, m$. Thus, the null string is a prefix of any string, and each string is a prefix of itself.)

Lemma 10.1. *Suppose $c : \mathcal{U} \to \{0,1\}^*$ is injective. Then, $\sum_u 2^{-length(c(u))} \leq \log_2(1 + |\mathcal{U}|)$.*

Proof. Without loss of generality, we can assume that whenever $k = \text{length}(c(u))$ for some u, then for every binary string b of length $i < k$ there is a v with $b = c(v)$. (Otherwise, there is a b with $\text{length}(b) < k$ which is not a codeword, and replacing $c(u)$ with b will preserve the injectiveness of c and increase the left hand side of the inequality.)

For such a code c, with k denoting the length of the longest codeword, the set of codewords is the union of $\bigcup_{i=0}^{k-1}\{0,1\}^i$ with a non-empty subset of $\{0,1\}^k$. With $1 \leq r \leq 2^k$ denoting the cardinality of this last subset, we have $|\mathcal{U}| = 2^k - 1 + r$ and $\sum_u 2^{-\text{length}(c(u))} = k + r2^{-k}$. As $\log_2(1 + |\mathcal{U}|) = k + \log_2(1 + r2^{-k})$ and $0 < r2^{-k} \leq 1$, all we need to show

is $x \leq \log_2(1+x)$ for $0 < x \leq 1$. As equality obtains for $x = 0$ and $x = 1$, the inequality follows from the concavity of log. □

Lemma 10.2. *Suppose $c : \mathcal{U} \to \{0,1\}^*$ is prefix-free. Then, $\sum_u 2^{-\mathit{length}(c(u))} \leq 1$. Conversely, if $\ell : \mathcal{U} \to \{0,1,2,\dots\}$ with $\sum_u 2^{-\ell(u)} \leq 1$, then there exists a prefix-free code $c : \mathcal{U} \to \{0,1\}^*$ with $\mathit{length}(c(u)) = \ell(u)$.*

Proof. Given a binary sequence $a = a_1 \dots a_m$, let $p(a) = \sum_{i=1}^m a_i 2^{-i}$ denote the rational number whose binary expansion is $0.a_1 \dots a_m$. With this notation, a binary sequence $a = a_1 \dots a_m$ is a prefix of a binary sequence $b = b_1 \dots b_n$ if and only the $p(b)$ lies in the interval $I(a) = [p(a), p(a) + 2^{-m})$.

For the first claim, observe that c being prefix-free thus implies that the intervals $I(c(u))$ are disjoint. As $I(c(u))$ is of size $2^{-\mathrm{length}(c(u))}$ and all of the intervals are included in $[0,1)$, the inequality follows.

For the second claim, order the elements of $\mathcal{U}$ as $u_1, \dots, u_K$ such that $\ell_1 := \ell(u_1) \leq \dots \leq \ell_K := \ell(u_K)$. Let $p_k = \sum_{i<k} 2^{-\ell_i}$ and set $I_k = [p_k, p_k + 2^{-\ell_k})$. Observe that the intervals $I_1, \dots, I_K$ are disjoint, and $I_k \subset [0,1)$. Furthermore, for each k, $2^{\ell_k} p_k$ is an integer, thus p_k can be expressed in binary as $0.b^{(k)}$ with $b^{(k)}$ a binary string of length ℓ_k. The code $c(u_k) = b^{(k)}$ now has the required properties — it being prefix free a consequence of the disjointness of the collection intervals I_k. □

Lemma 10.3. *Suppose $P \in \Pi(\mathcal{U})$ is a probability distribution on $\mathcal{U}$ and U is random variable with distribution P. Then, with $H(U) = -\sum_u P(u) \log_2 P(u)$ denoting the entropy of U,*

(i) for any prefix-free $c : \mathcal{U} \to \{0,1\}^*$, $\mathbb{E}[\mathit{length}(c(U))] \geq H(U)$;

(ii) there exists a prefix-free $c : \mathcal{U} \to \{0,1\}^*$ with $\mathbb{E}[\mathit{length}(c(U))] \leq H(U) + 1$;

(iii) for any injective $c : \mathcal{U} \to \{0,1\}^*$, $\mathbb{E}[\mathit{length}(c(U))] \geq H(U) - \log_2 \log_2(1 + |\mathcal{U}|)$,

(iv) there exists an injective $c : \mathcal{U} \to \{0,1\}^*$ with $\mathbb{E}[\mathit{length}(c(U))] \leq H(U)$.

Proof. For (i) and (iii) let $Q(u) = 2^{-\mathrm{length}(c(u))}$ and observe that

$$H(U) - \mathbb{E}[\mathrm{length}(c(U))] = \sum_u P(u) \log_2 \frac{Q(u)}{P(u)} \leq \log_2 \sum_u Q(u),$$

where the inequality is because log is concave. When c is prefix-free $\sum_u Q(u) \leq 1$ by Lemma 10.2, and when c is injective $\sum_u Q(u) \leq \log_2(1 + |\mathcal{U}|)$ by Lemma 10.1. The inequalities (i) and (iii) thus follow.

For (ii) set $\ell(u) = \lceil -\log_2 P(u) \rceil$. As $2^{-\ell(u)} \leq P(u)$, we see that $\sum_u 2^{-\ell(u)} \leq 1$ and by Lemma 10.2 there exists a prefix-free code c with $\mathrm{length}(c(u)) = \ell(u)$. As $\ell(u) < -\log_2 P(u) + 1$, (ii) follows.

For (iv) order the elements of $\mathcal{U}$ as $u_1, \dots, u_K$ with $P(u_1) \geq \dots \geq P(u_K)$. Let $c(u_k) = b_k$ where b_k is the kth element of the sequence $\lambda, 0, 1, 00, 01, 10, 11, 000, 001, \dots$, (e.g., $b_1 = \lambda$, $b_2 = 0$, $b_3 = 1$, $b_4 = 00$, $\dots$, $b_9 = 001$, $\dots$). Observe that $\mathrm{length}(b_k) = \lfloor \log_2 k \rfloor \leq \log_2 k$. Also note that $1 \geq \sum_{i=1}^k P(u_i) \geq k P(u_k)$, and thus $\log_2 k \leq -\log_2 P(u_k)$. Consequently, for this c, $\mathbb{E}[\mathrm{length}(c(U))] \leq -\sum_k P(u_k) \log_2 P(u_k) = H(U)$. □

Corollary 10.4. *Suppose $U_1, U_2, \ldots$ is a stochastic process. Then for any sequence $c_n : \mathcal{U}^n \to \{0,1\}^*$ of injective codes*

$$\liminf_n \frac{1}{n}\mathbb{E}[\mathit{length}(c_n(U^n))] \geq \liminf_n \frac{1}{n} H(U^n),$$

and there exists a sequence c_n of prefix-free codes for which

$$\limsup_n \frac{1}{n}\mathbb{E}[\mathit{length}(c_n(U^n))] \leq \limsup_n \frac{1}{n} H(U^n).$$

In particular, if $r = \lim_n \frac{1}{n} H(U^n)$ exists, all faithful representations of the process $U_1, U_2, \ldots$ with bits will asymptotically require at least r bits per letter, and there is a representation that asymptotically requires exactly as much.

Proof. The first inequality follows from noting that

$$\mathbb{E}[\mathrm{length}(c_n(U^n))] \geq H(U^n) - \log_2 \log_2(1 + |\mathcal{U}|^n)$$

and observing that $\lim_n \frac{1}{n} \log_2 \log_2(1+|\mathcal{U}|^n) = 0$. The second inequality follows from noting that there exist prefix-free c_n with

$$\mathbb{E}[\mathrm{length}(c_n(U^n))] \leq H(U^n) + 1$$

and that $\lim_n 1/n = 0$. □

Remark. Lemma 10.2 gives evidence of a strong connection between prefix-free codes and probability distributions. On the one hand, given a prefix-free code c, one can construct a probability distribution Q that assigns the letter u the probability $Q(u) = 2^{-\mathrm{length}(c(u))}$. By the lemma, $\sum_u Q(u) \leq 1$; if equality holds Q is indeed a probability distribution, otherwise, we can assign $1 - \sum_u Q(u)$ as the probability $Q(u_0)$ of a fictitious symbol $u_0 \notin \mathcal{U}$. If U is a random variable with distribution P, we then have (by assigning $P(u_0) = 0$ if necessary),

$$\mathbb{E}[\mathrm{length}(c(U))] - H(U) = \sum_u P(u)[\mathrm{length}(c(U)) + \log P(u)] = D(P\|Q).$$

On the other hand, given a distribution $Q \in \Pi(\mathcal{U})$, by Lemma 10.2 we can construct a prefix-free code $c : \mathcal{U} \to \{0,1\}^*$ with $\mathrm{length}(c(u)) = \lceil -\log_2 Q(u) \rceil$. As $-\log_2 Q(u) \leq \mathrm{length}(c(u)) < -\log_2 Q(u) + 1$, we see that

$$\mathbb{E}[\mathrm{length}(c(U))] - H(U) = \sum_u P(u)[\mathrm{length}(c(u)) + \log_2 P(u)]$$

is bounded from below by $D(P\|Q)$, and from above $D(P\|Q) + 1$.

These observations give the divergence $D(P\|Q)$ an interpretation as the expected number of "excess" bits (beyond the minimum possible $H(U)$) a code based on Q requires when describing a random variable with distribution P.

Consequently, if we are given $S \subset \Pi$ and told that the distribution P of a random variable U belongs to S, a reasonable strategy to design a code c is to look for a distribution $Q \in \Pi$ such that

$$\sup_{P \in S} D(P\|Q)$$

is small (e.g., by finding the Q that minimizes this quantity) and construct a code c based on Q as above.

Example 10.1. To illustrate the remark above, suppose we are told that $U_1, U_2, \ldots$ are binary and i.i.d. random variables. The distribution of U^n can be parametrized by $\theta = \Pr(U_1 = 1)$, and is given by

$$\Pr(U^n = u^n) = P_\theta^n(u^n) = (1-\theta)^{n_0(u^n)}\theta^{n_1(u^n)}$$

where $n_0(u^n)$ and $n_1(u^n)$ are the number of zeros and ones in the sequence $u_1 \ldots u_n$. With this notation, $S_n = \{P_\theta^n : 0 \le \theta \le 1\}$ is the class of distributions that we are told the distribution of U^n belongs to.

Consider now a sequence of conditional distributions

$$Q_{U_{k+1}|U^k}(u|u^k) = \frac{n_u(u^k)+1}{k+2}$$

where $n_u(u^k)$ is as above, denoting the number of u's in $u_1 \ldots u_k$. Note that $Q_{U_1}(0) = Q_{U_1}(1) = 1/2$. Define

$$Q_n(u^n) = \prod_{i=1}^n Q_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

One can prove by induction on n, that for any $n \ge 1$ and any $u^n \in \{0,1\}^n$,

$$Q_n(u^n) \ge \frac{1}{n+1}\left(\frac{n_0(u^n)}{n}\right)^{n_0(u^n)}\left(\frac{n_1(u^n)}{n}\right)^{n_1(u^n)}.$$

If $U_1, \ldots, U_n$ are i.i.d. with common distribution P_θ,

$$\begin{aligned}
D(P_\theta^n \| Q_n) &= \mathbb{E}\Big[\log \frac{P_\theta^n(U^n)}{Q_n(U^n)}\Big] \\
&\le \log(n+1) + \mathbb{E}\Big[\log \frac{P_\theta^n(U^n)}{(n_0(U^n)/n)^{n_0(U^n)}(n_1(U^n)/n)^{n_1(U^n)}}\Big] \\
&= \log(n+1) + \mathbb{E}\Big[n_0(U^n)\log\frac{n(1-\theta)}{n_0(U^n)} + n_1(U^n)\log\frac{n\theta}{n_1(U^n)}\Big] \\
&\le \log(n+1) + n(1-\theta)\log\frac{n(1-\theta)}{n(1-\theta)} + n\theta\log\frac{n\theta}{n\theta} = \log(n+1),
\end{aligned}$$

where the inequality in the last line is because $x \mapsto x\log[1/x]$ is concave and $\mathbb{E}[n_0(U^n)] = n(1-\theta)$, and $\mathbb{E}[n_1(U^n)] = n\theta$.

Consequently, we see that $\sup_{P^n \in S_n} D(P^n\|Q_n) \le \log(n+1)$. If Q_n were used to construct a prefix-free code $c_n : \{0,1\}^n \to \{0,1\}^*$, by the remark above, c_n will satisfy

$$\frac{1}{n}\mathbb{E}[\text{length}(c_n(U^n))] - H(P) \le \frac{1}{n}[\log(n+1)+1]$$

whenever U^n is i.i.d. with distribution P. As the right hand side vanishes as n gets large, it would be appropriate to call the sequence of codes c_n "asymptotically universal for the class of binary i.i.d. data". In the exercises we will see another choice of Q_n which improves the upper bound on $D(P^n\|Q_n)$ to $\frac{1}{2}\log n$.

Note that, had we chosen Q_n to be a member of S_n, say $Q_n = P_{\theta_0}^n$ for some θ_0, then $D(P_\theta^n\|Q_n)$ would have grown linearly in n for any $\theta \ne \theta_0$. Thus, even if we know that the true distribution P is in S, choosing Q outside of S (as we have done above) may lead to a better code construction.

Remark. The example above also illustrates a connection between compression and prediction. (One can also use the term 'learning' instead of prediction.) Suppose we have a family S_n of distributions on $\mathcal{U}^n$, and we are given a prefix-free code $c_n : \mathcal{U}^n \to \{0,1\}^*$ performs well, in the sense that

$$\sup_{P \in S_n} \frac{1}{n} \mathbb{E}_P[\text{length}(c_n(U^n))] - \frac{1}{n} H(U^n)$$

is small. Construct the distribution Q associated with the code c, i.e., $Q(u^n) = 2^{-\text{length}(c_n(u^n))}$ and factorize it as $Q(u^n) = \prod_{i=1}^n Q(u_i|u^{i-1})$. As the code c performs well, $\frac{1}{n} D(P\|Q)$ is small for all $P \in S_n$. But

$$\begin{aligned}
\frac{1}{n} D(P\|Q) &= \frac{1}{n} \sum_{u^n} P(u^n) \log \frac{P(u^n)}{Q(u^n)} \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{u^n} P(u^n) \log \frac{P(u_i|u^{i-1})}{Q(u_i|u^{i-1})} \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{u^i} P(u^i) \log \frac{P(u_i|u^{i-1})}{Q(u_i|u^{i-1})} \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{u^{i-1}} P(u^{i-1}) \sum_{u_i} P(u_i|u^{i-1}) \log \frac{P(u_i|u^{i-1})}{Q(u_i|u^{i-1})} \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{u^{i-1}} P(u^{i-1}) D(P(\cdot|u^{i-1})\|Q(\cdot|u^{i-1})),
\end{aligned}$$

so we conclude that for a large fraction of i's in $1, \ldots, n$, and for a set of u^{i-1}'s with large P probability, the quantity $D(P(\cdot|u^{i-1})\|Q(\cdot|u^{i-1}))$ is small.[1] Which is to say, no matter what P from S_n is the true distribution of the data, if after observing u^{i-1} we predicted the distribution of the next symbol u_i to be $Q(\cdot|u^{i-1})$, our prediction will be close to the true distribution $P(\cdot|u^{i-1})$ for most i's and for a high probability set of u^{i-1}'s.

10.2 Dimensionality Reduction

Assume that we are given n data points in $\mathbb{R}^d$, call them $u_1, \cdots u_n$. If d is very large then it might be challenging to store and process this "raw" data set. We are asking if we can represent this data in lower dimension, let us say $\mathbb{R}^k$, while maintaining some of its basic properties. In the sequel, we discuss two versions of this.

10.2.1 PCA

In this section, we discuss Principal Components Analysis (PCA), sometimes also referred to as the Karhunen-Loève Transform (KLT). This goes back to [16, 17, 18]. We start by fixing a certain $k < d$. The goal is to find a good k-dimensional basis such that most of the data points can be represented quite accurately in terms of this basis. It is not initially

[1] To be concrete, if $\frac{1}{n} D(P\|Q)$ is less than ϵ, then, except for a $\epsilon^{1/3}$ fraction of the i's, we have $\sum_{u^{i-1}} P(u^{i-1}) D(P(\cdot|u^{i-1})\|Q(\cdot|u^{i-1})) < \epsilon^{2/3}$, and except for a set of P probability $\epsilon^{1/3}$ of u^{i-1}'s, we have $D(P(\cdot|u^{i-1})\|Q(\cdot|u^{i-1})) < \epsilon^{1/3}$.

clear what "most" and "quite accurately" should mean. One intuitively pleasing metric is to select the basis (and corresponding coefficients for each data sample) so as to minimize the overall mean-squared error, that is:

$$\sum_{j=1}^{n} \|\mathbf{x}^{(j)} - \widehat{\mathbf{x}^{(j)}}\|^2, \tag{10.1}$$

where $\widehat{\mathbf{x}^{(j)}}$ represents the best approximation to $\mathbf{x}^{(j)}$ within the chosen basis. In spite of appearance, this problem actually has a clean solution : This is precisely the Eckart-Young theorem.

To see this, let us denote the (yet unknown) basis vectors by $\phi_1, \phi_2, \cdots, \phi_k \in \mathbb{R}^d$ and collect them (as column vectors) into the $d \times k$ matrix

$$\Phi \;=\; \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_k \end{pmatrix}. \tag{10.2}$$

Then, we can express

$$\widehat{\mathbf{x}^{(j)}} \;=\; \Phi \mathbf{f}^{(j)}, \tag{10.3}$$

where $\mathbf{f}^{(j)} \in \mathbb{R}^k$ is the feature vector corresponding to data sample $\mathbf{x}^{(j)}$. Hence, we are looking for

$$\min_{\text{feature vectors } \{\mathbf{f}^{(j)}\}_{j=1}^n \in \mathbb{R}^k} \left\{ \min_{\text{basis vectors } \{\phi_i\}_{i=1}^k \in \mathbb{R}^d} \sum_{j=1}^{n} \|\mathbf{x}^{(j)} - \Phi \mathbf{f}^{(j)}\|^2 \right\}. \tag{10.4}$$

To see how to proceed, we can rewrite

$$\begin{aligned} \sum_{j=1}^{n} \|\mathbf{x}^{(j)} - \Phi \mathbf{f}^{(j)}\|^2 &= \sum_{j=1}^{n} \operatorname{trace}\left((\mathbf{x}^{(j)} - \Phi \mathbf{f}^{(j)})(\mathbf{x}^{(j)} - \Phi \mathbf{f}^{(j)})^H \right) \\ &= \operatorname{trace}\left(\sum_{j=1}^{n} (\mathbf{x}^{(j)} - \Phi \mathbf{f}^{(j)})(\mathbf{x}^{(j)} - \Phi \mathbf{f}^{(j)})^H \right) \\ &= \operatorname{trace}\left((X - \Phi F)(X - \Phi F)^H \right) \\ &= \|X - \Phi F\|_F^2, \end{aligned} \tag{10.5}$$

where we have collected all the data samples into the $d \times n$ matrix X and all the feature vectors into the $k \times n$ matrix F.

This problem is precisely addressed by the Eckart–Young theorem that we have discussed earlier. The answer is simply to determine the SVD of the matrix X, and retain only the p largest singular values along with their corresponding singular vectors. Explicitly:

$$X \;=\; U \Sigma V^H = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \tag{10.6}$$

where r is the rank of the matrix X and where, as always, we assume that the singular values are ordered in decreasing order. Then, from the Eckart–Young theorem, we know that our error criterion is minimized if

$$\Phi F \;=\; \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^H. \tag{10.7}$$

In other words, we may select our basis vectors of length d to be the left singular vectors of X (that is, the eigenvectors of XX^H),

$$\phi_1 = \mathbf{u}_1, \;\; \phi_2 = \mathbf{u}_2, \;\; \cdots, \;\; \phi_k = \mathbf{u}_k, \tag{10.8}$$

in which case the matrix of feature vectors (of dimension $k \times n$) is given by

$$F \;\; = \;\; \begin{pmatrix} \sigma_1 \mathbf{v}_1^H \\ \sigma_2 \mathbf{v}_2^H \\ \vdots \\ \sigma_k \mathbf{v}_k^H \end{pmatrix}. \tag{10.9}$$

For example, the first column of this matrix is the feature vector corresponding to the first data sample, $\mathbf{x}^{(1)}$. Of course, this feature vector (of length k) can also be found by projecting the data sample successively into the k basis elements $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k$. It is left as an exercise to the reader to show that this indeed leads to the same answer.

Let us also remark that this is, quite obviously, not the unique basis — we can always rotate the basis to find an alternative basis (spanning exactly the same space). This will change the feature vectors, but it will not change the approximation quality.

We should note that many textbooks present PCA without proof simply as follows: We first find the covariance matrix of the data samples (using the above notation, this is the matrix XX^H), and then find its eigendecomposition. Clearly, the eigenvectors of the matrix XX^H are precisely the vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d$ above, and the PCA stipulates to use the eigenvectors corresponding to the k largest eigenvalues of XX^H as the approximate basis — exactly the same solution as the one we found above.

10.2.2 Johnson-Lindenstrauss

In this section, we present an insight found by Johnson and Lindenstrauss in [19]. We start from n data points in $\mathbb{R}^d$, call them $u_1, \cdots u_n$. We ask if we can find a map $F : \mathbb{R}^d \mapsto \mathbb{R}^k$, for $1 \leq k \leq d$, so that for a given $\delta \in (0, 1)$ and all $1 \leq i, j \leq n$

$$(1 - \delta)\|u_i - u_j\|_2^2 \leq \|F(u_i) - F(u_j)\|_2^2 \leq (1 + \delta)\|u_i - u_j\|_2^2.$$

We then talk of an *embedding* of the data in $\mathbb{R}^n$ that approximately preserves Euclidean distances. We then interested in the following questions.

1. For what dimensions $1 \leq k \leq d$ and parameters $0 < \delta < 1$ does such a mapping exist?
2. Can this mapping be found efficiently?

It turns out that a random linear mapping will do the trick as long as k is of order $\log(n)$. Note that in this statement the required dimension depends on the number of points we are embedding but not on the dimension that the original points came from. Our plan is simple. We define a random linear map so that in expectation the squared norm of vectors stays preserved. Since the map is linear this implies that the squared distance between two points also stays preserved in expectation under such a map. We will then use a suitable tail to show that with high probability the squared distance is not too far away from this expected value and, finally, use the union bound to show that by a suitable choice of parameters we can guarantee that for all $\binom{n}{2}$ pairs the squared distances stay approximately preserved.

Lemma 10.5 (Johnson-Lindenstrauss)**.** *Let $S = \{u_1, \cdots, u_n\}$ be a set of n points in $\mathbb{R}^d$, $n, d \in \mathbb{N}$. Let $0 < \delta < 1$ and let $k \in \mathbb{N}$,*

$$k > \frac{16}{\delta^2} \log(n).$$

Then there exists a function $F : \mathbb{R}^d \mapsto \mathbb{R}^k$ so that for all $u_i, u_j \in S$

$$(1-\delta)\|u_i - u_j\|^2 \leq \|F(u_i) - F(u_j)\|^2 \leq (1+\delta)\|u_i - u_j\|^2.$$

In words, we can embed the set S of points in $\mathbb{R}^d$ into the k-dimensional space $\mathbb{R}^k$ with only a small distortion of the pairwise Euclidean distances.

Proof. Let X be a $k \times d$ real-valued random matrix with i.i.d. entries that are zero-mean, unit-variance Gaussians. Let F be given by

$$F(u) = \frac{1}{\sqrt{k}} Xu.$$

We claim that $\mathbb{E}[\|F(u)\|_2^2] = \|u\|_2^2$. Let $X^{(i)}$ denote the i-th row of X. Fix $u \in \mathbb{R}^d$. Note that $\frac{X^{(i)}u}{\|u\|_2}$ is a Gaussian random variable with mean zero and variance 1, call it Z_i (the randomness here comes from the map). Then $F(u) = \frac{1}{\sqrt{k}}(Z_1, \cdots, Z_n)^T \|u\|_2$ and $\|F(u)\|_2^2 = \frac{1}{k}(\sum_{i=1}^k Z_i^2)\|u\|_2^2$. The claim follows by taking the expectation over this equality.

Consider the statement

$$(1-\delta)\|u\|^2 \leq \|F(u)\|^2 \leq (1+\delta)\|u\|^2.$$

This statement is equivalent to the statement that

$$1 - \delta \leq \frac{1}{k} \sum_{i=1}^k Z_i^2 \leq 1 + \delta.$$

In other words we are looking for tail bounds for the sum of squares of Gaussians. Note that if Z is a Gaussian then Z^2 follows a χ^2 distribution (with one degree of freedom). Note that a χ^2 distribution is *not* subgaussian, but it is subexponential. Therefore, from Lemma 2.7, we can obtain the following tail bound

$$\mathbb{P}\{|\frac{1}{k} \sum_{i=1}^k Z_i^2 - 1| > \delta\} \leq 2e^{-\delta^2 k/8}.$$

Now note that the map F is linear. If $u = u_i - u_j$, for $u_i, u_j \in S$, this means that $F(u_i) - F(u_j) = F(u_i - u_j) = F(u)$. Therefore, the above condition is equivalent to the condition that the random linear map preserves the distance of a fixed pair.

To finish the proof note that there are $\binom{n}{2}$ pairs of points and that we want to approximately preserve all these distances. Using the union bound we see that the probability that not all of these distances are preserved is upper bounded by

$$\binom{n}{2} 2e^{-\delta^2 k/8}.$$

Let us say that we want this probability to be upper bounded by ϵ, $0 < \epsilon < 1$. Solving for k we then get the condition

$$k \geq \frac{8}{\delta^2} \log(n^2/\epsilon).$$

If all we want is to show the existence of a suitable map then we can let ϵ tend to 1 and make the inequality strict. This would apply if we are allowed to generate such a random map, check if it works, and repeat the procedure until we have found a suitable map. If we cannot check (e.g., perhaps we do not even have the set S at the point in time when we need to decide) we likely want to pick a sufficiently small ϵ and choose the dimension k according to this ϵ. □

10.3 Problems

Problem 10.1 (Elias coding)**.** Let 0^n denote a sequence of n zeros. Consider the code (the subscript U a mnemonic for 'Unary'), $\mathcal{C}_U : \{1, 2, \dots\} \to \{0,1\}^*$ for the positive integers defined as $\mathcal{C}_U(n) = 0^{n-1}$.

(a) Is $\mathcal{C}_U$ injective? Is it prefix-free?

Consider the code (the subscript B a mnenonic for 'Binary'), $\mathcal{C}_B : \{1, 2, \dots\} \to \{0,1\}^*$ where $\mathcal{C}_B(n)$ is the binary expansion of n. I.e., $\mathcal{C}_B(1) = 1$, $\mathcal{C}_B(2) = 10$, $\mathcal{C}_B(3) = 11$, $\mathcal{C}_B(4) = 100$, ... Note that

$$\text{length}\mathcal{C}_B(n) = \lceil \log_2(n+1) \rceil = 1 + \lfloor \log_2 n \rfloor.$$

(b) Is $\mathcal{C}_B$ injective? Is it prefix-free?

With $k(n) = \text{length}\mathcal{C}_B(n)$, define $\mathcal{C}_0(n) = \mathcal{C}_U(k(n))\mathcal{C}_B(n)$.

(c) Show that $\mathcal{C}_0$ is a prefix-free code for the positive integers. To do so, you may find it easier to describe how you would recover $n_1, n_2, \dots$ from the concatenation of their codewords $\mathcal{C}_0(n_1)\mathcal{C}_0(n_2)\dots$.

(d) What is length$(\mathcal{C}_0(n))$?

Now consider $\mathcal{C}_1(n) = \mathcal{C}_0(k(n))\mathcal{C}_B(n)$.

(e) Show that $\mathcal{C}_1$ is a prefix-free code for the positive integers, and show that length$(\mathcal{C}_1(n)) = 2 + 2\lfloor \log(1 + \lfloor \log n \rfloor) \rfloor + \lfloor \log n \rfloor \leq 2 + 2\log(1 + \log n) + \log n$.

Suppose U is a random variable taking values in the positive integers with $\Pr(U = 1) \geq \Pr(U = 2) \geq \dots$.

(f) Show that $\mathbb{E}[\log U] \leq H(U)$, [Hint: first show $i \Pr(U = i) \leq 1$], and conclude that

$$E[\text{length}\mathcal{C}_1(U)] \leq H(U) + 2\log(1 + H(U)) + 2.$$

Problem 10.2 (Code Extension)**.** Suppose $|\mathcal{U}| \geq 2$. For $n \geq 1$ and a code $c : \mathcal{U} \to \{0,1\}^*$ we define its *n-extension* $c^n : \mathcal{U}^n \to \{0,1\}*$ via $c^n(u^n) = c(u_1) \dots c(u_n)$. In other words $c^n(u^n)$ is the concatenation of the binary strings $c(u_1)$, ..., $c(u_n)$. A code c is said to be *uniquely decodeable* if for any u^k and $\tilde{u}^m$ with $u^k \neq \tilde{u}^m$, $c^k(u^k) \neq c^m(\tilde{u}^m)$.

(a) Show that if c is uniquely decodable, then for all $n \geq 1$, c^n is injective.

(b) Show that if c is not uniquely decodable, there are u^k and $\tilde{u}^m$ with $u_1 \neq \tilde{u}_1$ and $c^k(u^k) = c^m(\tilde{u}^m)$.

(c) Show that if c is not uniquely decodable, then there is an n for which c^n is not injective. [Hint: try $n = k + m$.]

Problem 10.3 (Prediction and coding)**.** After observing a binary sequence $u_1, \ldots, u_i$, that contains $n_0(u^i)$ zeros and $n_1(u^i)$ ones, we are asked to estimate the probability that the next observation, u_{i+1} will be 0. One class of estimators are of the form

$$\hat{P}_{U_{i+1}|U^i}(0|u^i) = \frac{n_0(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha} \quad \hat{P}_{U_{i+1}|U^i}(1|u^i) = \frac{n_1(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha}.$$

We will consider the case $\alpha = 1/2$, this is known as the Krichevsky–Trofimov estimator. Note that for $i = 0$ we get $\hat{P}_{U_1}(0) = \hat{P}_{U_1}(1) = 1/2$.

Consider now the joint distribution $\hat{P}(u^n)$ on $\{0,1\}^n$ induced by this estimator,

$$\hat{P}(u^n) = \prod_{i=1}^{n} \hat{P}_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

(a) Show, by induction on n that, for any n and any $u^n \in \{0,1\}^n$,

$$\hat{P}(u_1, \ldots, u_n) \geq \frac{1}{2\sqrt{n}} \Big(\frac{n_0}{n}\Big)^{n_0} \Big(\frac{n_1}{n}\Big)^{n_1},$$

where $n_0 = n_0(u^n)$ and $n_1 = n_1(u^n)$.

[Hint: if $0 \leq m \leq n$, then $(1 + 1/n)^{n+1/2} \geq \frac{m+1}{m+1/2}(1 + 1/m)^m$]

(b) Conclude that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that

$$\text{length}_{\mathcal{C}}(u_1, \ldots, u_n) \leq n h_2\left(\frac{n_0(u^n)}{n}\right) + \frac{1}{2}\log n + 2,$$

with $h_2(x) = -x \log x - (1 - x)\log(1 - x)$.

(c) Show that if $U_1, \ldots, U_n$ are i.i.d. Bernoulli, then

$$\frac{1}{n}\mathbb{E}[\text{length}_{\mathcal{C}}(U_1, \ldots, U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

Problem 10.4 (Lempel Ziv 78)**.** Suppose $\ldots, U_{-1}, U_0, U_1, \ldots$ is a stationary process, i.e., for any $k = 1, 2, \ldots$, any $u_0, \ldots, u_{k-1}$, and any $n = \ldots, -1, 0, 1, \ldots$

$$\mathbb{P}(U_n \ldots U_{n+k-1} = u_0 \ldots u_{k-1}) = \mathbb{P}(U_0 \ldots U_{k-1} = u_0 \ldots u_{k-1}).$$

Suppose also that U is a recurrent process, i.e., any letter u_0 with $\mathbb{P}(U_0 = u_0) > 0$, the event $A = \{\text{there exists } i \geq 0 \text{ and } j > 0 \text{ such that } U_i = U_{-j} = u_0\}$ has $\mathbb{P}(A) = 1$. (That is, a positive probability letter u_0 will occur infinitely often.)

Fix u_0 with $\mathbb{P}(U_0 = u_0) > 0$. For $i \geq 0$ and $j < 0$, let

$$A_{ij} = \{U_i = u_0\} \cap \{U_{-j} = u_0\} \cap \bigcap_{k=-j+1}^{i-1} \{U_k \neq u_0\}$$

denote the event that j is the last time before time 0 that u_0 was seen and i was the first time after time 0 that u_0 is seen.

(a) Show that $\sum_{i\geq 0, j>0} \mathbb{P}(A_{ij}) = \mathbb{P}(A) = 1$.

(b) Show that $\mathbb{P}(A_{ij}) = f(i+j)$, where

$$f(k) = \mathbb{P}(U_{-k} = u_0, U_{-l} \neq u_0 \text{ for } l = 1, \ldots, k-1, U_0 = u_0).$$

(c) Using (a) and (b), show that

$$1 = \sum_{k\geq 1} k f(k) = 1.$$

(d) Let $K = \inf\{k > 0 : U_{-k} = u_0\}$ (i.e., the negative index of the most recent time before time 0 u_0 was seen). Observe that the event $\{K = k, U_0 = u_0\}$ is the event whose probability is $f(k)$. Using (c) show that

$$E[K|U_0 = u_0] = 1/\mathbb{P}(U_0 = u_0)$$

and that $\mathbb{E}[\log K] \leq H(U_0)$.

Suppose we have a stationary and ergodic source $\ldots, X_{-1}, X_0, X_1, \ldots$. This means, in particular, that for any $n > 0$, the process $\{U_i\}$ defined by $U_i = (X_i, X_{i+1}, \ldots, X_{i+n-1})$ is stationary and recurrent.

Fix a sequence $x_0, \ldots, x_{n-1}$ with $\mathbb{P}\big((X_0 \ldots X_{n-1}) = (x_0 \ldots x_{n-1})\big) > 0$. Let

$$K = \inf\{k > 0 : (X_{-k} \ldots X_{-k+n-1}) = (x_0 \ldots x_{n-1})\}.$$

(e) Show that $\mathbb{E}[\log K] \leq H(X_0 \ldots X_{n-1})$.

(f) Consider the following data compression method. Assuming that the encoder has already described the infinite past $\ldots, X_{-2}, X_{-1}$ to the decoder, he describes $X_0, \ldots, X_{n-1}$ by (i) finding the most recent occurrence $X_0 \ldots X_{n-1}$ in the past, (ii) describing the index K of this occurrence by the method of problem 2(f). Now that the decoder knows $\ldots, X_{n-1}$, the encoder describes $X_n \ldots X_{2n-1}$ is the same way, etc. Show that this method uses fewer than

$$\frac{1}{n} H(X_0 \ldots X_{n-1}) + \frac{2}{n} \log(1 + H(X_0 \ldots X_{n-1})) + \frac{2}{n}$$

bits per letter on the average.

Problem 10.5 (Lower bound on Expected Length)**.** Suppose U is a random variable taking values in $\{1, 2, \ldots\}$. Set $L = \lfloor \log_2 U \rfloor$. (I.e., $L = j$ if and only if $2^j \leq U < 2^{j+1}$; $j = 0, 1, 2, \ldots$.

(a) Show that $H(U|L = j) \leq j$, $j = 0, 1, \ldots$.

(b) Show that $H(U|L) \leq \mathbb{E}[L]$.

(c) Show that $H(U) \leq \mathbb{E}[L] + H(L)$.

(d) Suppose that $\Pr(U = 1) \geq \Pr(U = 2) \geq \ldots$. Show that $1 \geq i \Pr(U = i)$.

(e) With U as in (d), and using the result of (d), show that $\mathbb{E}[\log_2 U] \leq H(U)$ and conclude that $\mathbb{E}[L] \leq H(U)$.

(f) Suppose that N is a random variable taking values in $\{0, 1, \dots\}$ with distribution p_N and $\mathbb{E}[N] = \mu$. Let G be a geometric random variable with mean μ, i.e., $p_G(n) = \mu^n/(1+\mu)^{1+n}$, $n \geq 0$.

Show that $H(G) - H(N) = D(p_N \| p_G)$, and conclude that $H(N) \leq g(\mu)$ with $g(x) = (1+x)\log_2(1+x) - x\log_2 x$.

[Hint: Let $f(n, \mu) = -\log_2 p_G(n) = (n+1)\log_2(1+\mu) - n\log_2(\mu)$. First show that $\mathbb{E}[f(G, \mu)] = \mathbb{E}[f(N, \mu)]$, and consequently $H(G) = \sum_n p_N(n)\log_2(1/p_G(n))$.]

(g) Show that for U as in (d) and $g(x)$ as in (f),

$$E[L] \geq H(U) - g(H(U)).$$

[Hint: combine (f), (e), (c).]

(h) Now suppose U is a random variable taking values on an alphabet $\mathcal{U}$, and $c : \mathcal{U} \to \{0,1\}^*$ is an injective code. Show that

$$E[\text{lengthc}(U)] \geq H(U) - g(H(U)).$$

[Hint: the best injective code will label $\mathcal{U} = \{a_1, a_2, a_3, \dots\}$ so that $\Pr(U = a_1) \geq \Pr(U = a_2) \geq \dots$, and assign the binary sequences $\lambda, 0, 1, 00, 01, 10, 11, ...$ to the letters $a_1, a_2, \dots$ in that order. Now observe that the i'th binary sequence in the list $\lambda, 0, 1, 00, 01, \dots$ is of length $\lfloor \log_2 i \rfloor$.]

Problem 10.6 (Nonsingular and Uniquely Decodable Codes)**.** Recall that for a code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ we define $\mathcal{C}^n : \mathcal{U}^n \to \{0,1\}^*$ as $\mathcal{C}^n(u_1, \dots, u_n) = \mathcal{C}(u_1)\dots\mathcal{C}(u_n)$.

If a code $\mathcal{C}$ is uniquely decodable, it is clear that for each n, $\mathcal{C}^n$ is non-singular (indeed $\mathcal{C}^n$ is uniquely decodable).

1. Suppose $\mathcal{C}$ is *not* uniquely decodable. Show that there is an $n \geq 1$ such that $\mathcal{C}^n$ is singular.

2. Suppose $\mathcal{K} : \{0, 1, 2, \dots\} \to \{0,1\}^*$ is a *prefix-free* code for non-negative integers. Show that for any *non-singular* code $\mathcal{C}$ for any alphabet $\mathcal{U}$, the code $\mathcal{C}' : \mathcal{U} \to \{0,1\}^*$ with

$$\mathcal{C}'(u) = \mathcal{K}(\text{length}(\mathcal{C}(u)))\mathcal{C}(u)$$

is prefix free.

Recall from Homework 4, Problem 1 that there is a prefix-free $\mathcal{C}_1 : \{1, 2, \dots\} \to \{0,1\}^*$ for positive-integers for which $\text{length}(\mathcal{C}_1(n)) \leq 2 + 2\log(1 + \log n) + \log n$. Let $\mathcal{K} : \{0, 1, \dots\} \to \{0,1\}^*$ be defined as $\mathcal{K}(n) = \mathcal{C}_1(n+1)$.

3. Show that for any non-singular code $\mathcal{C}$ for $\mathcal{U}$ with $\mathbb{E}[\text{length}(\mathcal{C}(U))] = L$, there is a prefix-free code $\mathcal{C}'$ for $\mathcal{U}$ with

$$\mathbb{E}[\text{length}(\mathcal{C}'(U))] \leq L + 2 + 2\log(1 + \log(1 + L)) + \log(1 + L).$$

Problem 10.7 (Quantization with two criteria)**.** Suppose U^n has i.i.d. components with distribution P. We want to describe U^n at rate R, i.e., we want to design a function $f : \mathcal{U}^n \to \{1, \dots, 2^{nR}\}$.

We are given two distortion measures $d_1 : \mathcal{U} \times \mathcal{V}_1 \to \mathbb{R}$ and $d_2 : \mathcal{U} \times \mathcal{V}_2 \to \mathbb{R}$, and we wish to ensure that from $i = f(U^n)$ we can reconstruct $V_1^n = g_1(i) \in \mathcal{V}_1^n$ and $V_2^n = g_2(i) \in \mathcal{V}_2^n$ so that

$$\mathbb{E}[d_1(U^n, V_1^n)] \le D_1 \quad \text{and} \quad \mathbb{E}[d_2(U^n, V_2^n)] \le D_2$$

with given distortion criteria D_1 and D_2. (As in class $d(U^n, V^n) = \frac{1}{n}\sum_{i=1}^n d(U_i, V_i)$.)

(a) What is the rate distortion function $R(D_1, D_2)$?

(b) Suppose $R_1(D_1)$ is the rate distortion function with the first distortion criterion alone, and $R_2(D_2)$ is the rate distortion function with the second criterion alone. What relationship exists between $R(D_1, D_2)$ and $R_1(D_1) + R_2(D_2)$?

Problem 10.8 (Choose the Shortest Description)**.** Suppose $\mathcal{C}_0 : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{C}_1 : \mathcal{U} \to \{0,1\}^*$ are two prefix-free codes for the alphabet $\mathcal{U}$. Consider the code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ defined by

$$\mathcal{C}(u) = \begin{cases} [0, \mathcal{C}_0(u)] & \text{if length}\mathcal{C}_0(u) \le \text{length}\mathcal{C}_1(u) \\ [1, \mathcal{C}_1(u)] & \text{else.} \end{cases}$$

Observe that $\text{length}(\mathcal{C}(u)) = 1 + \min\{\text{length}(\mathcal{C}_0(u)), \text{length}(\mathcal{C}_1(u))\}$.

(a) Is $\mathcal{C}$ a prefix-free code? Explain.

(b) Suppose $\mathcal{C}_0, \dots, \mathcal{C}_{K-1}$ are K prefix-free codes for the alphabet $\mathcal{U}$. Show that there is a prefix-free code $\mathcal{C}$ with

$$\text{length}(\mathcal{C}(u)) = \lceil \log_2 K \rceil + \min_{0 \le k < K-1} \text{length}(\mathcal{C}_k(u)).$$

(c) Suppose we are told that U is a random variable taking values in $\mathcal{U}$, and we are also told that the distribution p of U is one of K distributions $p_0, \dots, p_{K-1}$, but we do not know which. Using (b) describe how to construct a prefix-free code $\mathcal{C}$ such that

$$\mathbb{E}[\text{length}(\mathcal{C}(U))] \le \lceil \log_2 K \rceil + 1 + H(U).$$

[Hint: From class we know that for each k there is a prefix-free code $\mathcal{C}_k$ that descibes each letter u with at most $\lceil -\log_2 p_k(u) \rceil$ bits.]

Problem 10.9 (Universal codes)**.** Suppose we have an alphabet $\mathcal{U}$, and let Π denote the set of distributions on $\mathcal{U}$. Suppose we are given a family of S of distributions on $\mathcal{U}$, i.e., $S \subset \Pi$. For now, assume that S is finite.

Define the distribution $Q_S \in \Pi$

$$Q_S(u) = Z^{-1} \max_{P \in S} P(u)$$

where the normalizing constant $Z = Z(S) = \sum_u \max_{P \in S} P(u)$ ensures that Q_S is a distribution.

(a) Show that $D(P\|Q) \le \log Z \le \log |S|$ for every $P \in S$.

(b) For any S, show that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that for any random variable U with distribution $P \in S$,

$$E[\text{length}\mathcal{C}(U)] \leq H(U) + \log Z + 1.$$

(Note that $\mathcal{C}$ is designed on the knowledge of S alone, it cannot change on the basis of the choice of P.) [Hint: consider $L(u) = -\log_2 Q_S(u)$ as an 'almost' length function.]

(c) Now suppose that S is not necessarily finite, but there is a finite $S_0 \subset \Pi$ such that for each $u \in \mathcal{U}$, $\sup_{P \in S} P(u) \leq \max_{P \in S_0} P(u)$. Show that $Z(S) \leq |S_0|$.

Now suppose $\mathcal{U} = \{0,1\}^m$. For $\theta \in [0,1]$ and $(x_1, \ldots, x_m) \in \mathcal{U}$, let

$$P_\theta(x_1, \ldots, x_n) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i}.$$

(This is a fancy way to say that the random variable $U = (X_1, \ldots, X_n)$ has i.i.d. Bernoulli θ components). Let $S = \{P_\theta : \theta \in [0,1]\}$.

(d) Show that for $u = (x_1, \ldots, x_m) \in \{0,1\}^m$

$$\max_\theta P_\theta(x_1, \ldots, x_m) = P_{k/m}(x_1, \ldots, x_m)$$

where $k = \sum_i x_i$.

(e) Show that there is a prefix-free code $\mathcal{C} : \{0,1\}^m \to \{0,1\}^*$ such that whenever $X_1, \ldots, X_n$ are i.i.d. Bernoulli,

$$\frac{1}{m}\mathbb{E}[\text{length}\mathcal{C}(X_1, \ldots, X_m)] \leq H(X_1) + \frac{1 + \log_2(1+m)}{m}.$$

Problem 10.10 (Universality via Typicality)**.** Given an alphabet $\mathcal{U}$, and a rate $0 \leq R \leq \log|\mathcal{U}|$, consider the sequence of sets

$$A_n = \bigcup_{Q \in \Pi_n : H(Q) < R} T^n(Q), \quad n = 1, 2, \ldots$$

(i.e., A_n is the union of the typical sets of all empirical probability distributions with entropy at most R.)

(a) Find $\lim_{n\to\infty} \frac{1}{n} \log|A_n|$.

Hint: For a lower bound, fix Q with $H(Q) < R$, and a sequence of types $Q_1, Q_2, \ldots$ with $\lim_{n\to\infty} Q_n = Q$. Now observe that for large n, A_n includes $T^n(Q_n)$.

Suppose $P \in \Pi$ with $H(P) < R$ (i.e., P is probability distribution on $\mathcal{U}$ with entropy strictly less than R.)

(b) With A_n^c denoting the complement of A^n, find $\lim_{n\to\infty} P^n(A_n^c)$.

(c) Show that there is a injective code $\mathcal{C}_n : \mathcal{U}^n \to \{0,1\}^*$ such that

$$\text{length}(\mathcal{C}_n(u^n)) = \begin{cases} 1 + \lceil \log|A_n| \rceil & u^n \in A^n \\ 1 + \lceil n \log|\mathcal{U}| \rceil & \text{else} \end{cases}$$

(d) Show that there is a sequence of injective codes $\mathcal{C}_n : \mathcal{U}^n \to \{0,1\}^*$ such that for any $P \in \Pi$ with $H(P) < R$ and any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\text{length}(\mathcal{C}_n(U^n)) > n(R+\epsilon)\big) = 0.$$

Problem 10.11 (Fibonacci Coding)**.** Consider the following binary encoding of a positive integer n: $\mathcal{C}_F(n) = I_1 \dots I_r 1$, where $n = \sum_{i=1}^r I_i F_{i+1}$ and F_i is i-th Fibonacci number, $F_0 = 0$, $F_1 = 1$, $F_2 = F_0 + F_1 = 1$, $\cdots$, $F_i = F_{i-1} + F_{i-2}$, $i \geq 2$, and $I_i \in \{0,1\}$. E.g., 1011 denotes the integer $1 \times 1 + 0 \times 2 + 1 \times 3 = 4$.

For every positive integer n such a representation exists. In order to make it unique, given an integer, find the largest Fibonacci number that it contains. Note it and remove its value from the integer. Proceed recursively to find the unique representation. E.g., for $n = 4$, $F_4 = 3$ is the largest Fibonacci number that is contained in 4 and $F_2 = 1$ is the largest Fibonanni number that is contained in $n - F_4 = 1$. This gives us the representation 1011.

Recall that besides a recursive description of the Fibonacci numbers there exists an explicit formula $F_i = \lfloor \frac{\phi^i}{\sqrt{5}} + \frac{1}{2} \rfloor$, where $\phi = \frac{1+\sqrt{5}}{2} \sim 1.618$ is the golden ratio.

(i) What is the length of $\mathcal{C}_F(n)$?

(ii) Show that the code is prefix-free. *Hint*: Use the property of Fibonacci numbers.

(iii) Show that $\log_\phi(\sqrt{5}i) \leq 3 + 2\log_2 i$.

(iv) Consider a random variable U that takes values on the positive integers s.t. $P(U = i)$ is decreasing. Show that $\mathbb{E}[\text{length}(\mathcal{C}_F(U))] \leq 3 + 2H(U)$. *Hint*: First show that $iP(U = i) \leq 1$.

Problem 10.12 (Projections)**.** Assume that we get m samples in $\mathbb{R}^d$, call them $u_1, \dots, u_m$. The dimension d is very large. Therefore we would like to compress the data. We fix $n < d$ and we would like to produce n-dimensional representations $\hat{u}_1, \dots, \hat{u}_m$ that are close to the original ones. Assume that we collect our data samples into a $d \times m$ matrix U and the desired representations into a $n \times m$ matrix $\hat{U}$.

In the course we learned that two possible compression techniques for this scenario are the PCA and random projections.

Recall that random projections are linear maps $f(u) : \mathbb{R}^d \to \mathbb{R}^n$, defined as $f(u) = \frac{1}{\sqrt{n}} Xu$, where X is a real-valued matrix with iid zero-mean unit-variance entries.

(i) Assume that your "goodness" criterion is the spectral norm $||U^T U - \hat{U}^T \hat{U}||_2$. What guarantees to you get for both methods? You can assume that the smallest eigenvalue of $X^T X$ is 0.

(ii) Assume your "goodness" criterion is $\max_{i,j} |||u_i - u_j||^2 - ||\hat{u}_i - \hat{u}_j||^2|$. What guarantees do you get for both methods? No need for complicated computations.

Problem 10.13. *(Johnson-Lindenstrauss for subgaussians)*

(a) In preparation for this problem, establish the following facts:

- If U is a subexponential random variable with parameters (ν, b), then αU (where we assume $\alpha > 0$) is a subexponential random variable with parameters $(\alpha\nu, \alpha b)$.
- If U and V are independent subexponential random variables with parameters (ν_u, b_u) and (ν_v, b_v), respectively, then $U+V$ is a subexponential random variable with parameters $(\sqrt{\nu_u^2 + \nu_v^2}, \max(b_u, b_v))$.

In this problem, we reconsider the Johnson-Lindenstrauss Lemma (Lemma 10.5 in the lecture notes). The only change is that inside the real-valued $k \times d$ matrix X in the proof of the Lemma, we no longer assume that the entries are independent Gaussians. We still assume the entries X_{ij} to be independent. We also still assume that they each have mean zero and variance 1. But beyond this, we only assume that they are subgaussian with variance proxy σ^2.

To proceed, exactly as in the Johnson-Lindenstrauss Lemma, consider an arbitrary real-valued vector u of length d. As in the proof of the Johnson-Lindenstrauss Lemma, we define, for $i = 1, 2, \ldots, k$,

$$Z_i = \frac{1}{\|u\|_2} \sum_{j=1}^{d} u_j X_{ij}.$$

(b) Show the following facts (short justifications are sufficient, and you may refer freely to the lecture notes)

- The random variables Z_i are independent of each other.
- Each Z_i is subgaussian. Find the corresponding variance proxy.
- We have $\mathbb{E}[Z_i^2] = 1$.

To continue, we will need the following theorem:

Theorem. If Y is subgaussian with variance proxy σ^2, then Y^2 with mean $\mathbb{E}[Y^2]$ is subexponential with parameters $(c\sigma^2, d\sigma^2)$ for some absolute constants c and d.

(c) Exactly as in the proof of the Johnson-Lindenstrauss Lemma, we next need to analyze $S = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$. Leveraging the theorem, show that S is subexponential with mean 1 and find the corresponding parameters.

(d) Give a concentration bound, that is, an upper bound of the form

$$\mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k} Z_i^2 - 1\right| > \delta\right\} \leq \ldots$$

(e) Discuss the differences of the resulting lemma with respect to what is proved in the lecture notes.

Chapter 11

Information Measures and Generalization Error

In this chapter, we consider the *generalization error* of a learning algorithm. Roughly speaking, it tries to capture the idea of model "overfitting".

11.1 Setup and Problem Statement

The standard setup of statistical learning theory is as follows: we have an instance space $\mathcal{X}$, a hypothesis space $\mathcal{H}$, and a loss function $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+$.

We observe $D = (X_1, X_2, \ldots, X_n)$ i.i.d samples from some *unknown* distribution P_X. Using these training samples, the learning algorithm picks a hypothesis in $\mathcal{H}$. We can think of the hypotheses as models we use to explain our data, and we use the loss function to evaluate the performance of our chosen model.

Example 11.1. $\mathcal{X} = \mathbb{R}\times\mathbb{R}$, $\mathcal{H} = \{\text{affine functions from } \mathbb{R} \text{ to } \mathbb{R}\}$, and $\ell\left(h, (x, y)\right) = (y - h(x))^2$. That is, we observe pairs of values $\{(x_i, y_i\}_{i=1}^n$, and we want to find the best linear approximation of y_i in terms of x_i. This setup with the choice of the squared error loss is referred to as linear regression.

Definition 11.1. Given $h \in \mathcal{H}$, the *population risk* of h is defined as:

$$L_{P_X}(h) := \mathbb{E}_X[\ell(h, X)].$$

The population risk indicates how well the hypothesis h models the data. We would like this risk to be small, but we cannot evaluate directly since P_X is unknown. On the other hand, given a data set D, we can evaluate the empirical risk (or *training error*).

Definition 11.2. Given a data set $D = (X_1, X_2, \ldots, X_n)$ and a hypothesis h, the *empirical risk* is defined as:

$$L_D(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, X_i).$$

The learning algorithm that chooses a hypothesis $h \in \mathcal{H}$ based on D does not need to be deterministic. Therefore, we model the *learning algorithm* as a conditional distribution $P_{H|D}$. We observe that this induces a joint distribution over D and H, which we will denote as P_{DH}. The key definition for the present section is the (expected) generalization error:

Definition 11.3. Given a learning algorithm $P_{H|D}$, the *generalization error* is defined as

$$\operatorname{gen}(P_X, P_{H|D}) = \left|\mathbb{E}_{P_{DH}}[L_D(H) - L_{P_X}(H)]\right|,$$

that is, the difference between the loss on the training data and the loss on fresh test data.

Let us now give some intuition about the generalization error. If the chosen hypothesis "depends too much" on the given data, then the generalization error can be large, i.e., we are "overfitting". We can bound the error by controlling the degree of dependence, and we will use again information measures to do so. Note that, in this setting, if H is chosen independently from the data, then the generalization error will be zero but the population risk will be large, which we ultimately want to be small. So we have a tension where on the one hand, if H is independent from D so the learning algorithm is not "learning" anything; and on the other hand, if H depends too much on the data, it will be overfitting.

11.2 Bounds on the Generalization Error

In order to analyze the generalization error, it is instructive to rewrite Definition 11.1 slightly differently as

$$L_{P_X}(h) = \mathbb{E}_{\tilde{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, \tilde{X}_i)\right], \tag{11.1}$$

where $\tilde{X}_i$ are i.i.d. samples from the distribution P_X, independent of the training set D, and we use the notation $\tilde{D} = (\tilde{X}_1, \ldots, \tilde{X}_n)$. This shows the close relationship to the empirical risk defined in Definition 11.2.

With this, we can rewrite the generalization error as follows:

$$\begin{aligned}\operatorname{gen}(P_X, P_{H|D}) &= \left|\mathbb{E}_{H,D}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(H, X_i) - \mathbb{E}_{\tilde{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(H, \tilde{X}_i)\right]\right]\right| &(11.2)\\ &= \left|\mathbb{E}_{H,D}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(H, X_i)\right] - \mathbb{E}_{H,\tilde{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(H, \tilde{X}_i)\right]\right|. &(11.3)\end{aligned}$$

This is the difference of two expected values of *the same* function, but under two different probability distributions: In the first expectation, since H was indeed chosen (via the learning procedure) as a function of the training data D, we have that H and D are dependent on each other. In the second expectation, H is independent of $\tilde{D} = (\tilde{X}_1, \ldots, \tilde{X}_n)$. To drive home this point, let us call $Z = (H, X_1, \ldots, X_n)$. Then, our goal is to understand the following quantity:

$$|\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)]|, \tag{11.4}$$

where P and Q are two different distributions. In the special case of the generalization error, P is the joint distribution of the training set and the hypothesis chosen by the learning procedure, and Q is the product of the corresponding marginals. In the sequel, we derive bounds on the quantity from Equation (11.4). These bounds can then be applied to the generalization error, but are also of independent interest.

11.2.1 L_1-Distance Bound

Our first bound for the quantity from Equation (11.4) starts from the following well-known relationship:

Lemma 11.1. *Let P and Q be two probability mass functions on a finite set $\mathcal{Z}$. Then,*

$$\|P - Q\|_1 = 2 \max_{S \subseteq \mathcal{Z}} P(S) - Q(S).$$

Proof. Let $A = \{z \in \mathcal{Z} : P(z) \geq Q(z)\}$, and A^c the complement of A in $\mathcal{Z}$. Then,

$$\begin{aligned}\|P - Q\|_1 &= \sum_{z \in A} P(z) - Q(z) + \sum_{z \in A^c} Q(z) - P(z) = P(A) - Q(A) + Q(A^c) - P(A^c) \\ &= P(A) - Q(A) + 1 - Q(A) - 1 + P(A) = 2(P(A) - Q(A)).\end{aligned}$$

To complete the proof, start by considering any other set S by adding elements to A to observe that for such S, we have $P(S) - Q(S) \leq P(A) - Q(A)$. Then, consider any other set S by removing elements from A, and again observe that for such S, we have $P(S) - Q(S) \leq P(A) - Q(A)$. □

Note that for any subset S, $P(S)$ can also be seen as $\mathbb{E}_P[f_S(Z)]$ where $f_S(z) = \begin{cases} 1, & z \in S, \\ 0, & z \notin S \end{cases}$. Moreover, the proof of Lemma 11.1 can be simply modified to show the following:

Lemma 11.2.

$$\|P - Q\|_1 = 2 \max_{f : \mathcal{Z} \to [0,1]} \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)].$$

Proof: Let $A = \{z \in \mathcal{Z} : P(z) \geq Q(z)\}$. Then,

$$\begin{aligned}\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] &= \sum_{z \in A} f(z)\,(P(z) - Q(z)) + \sum_{z \notin A} f(z)\,(P(z) - Q(z)) \\ &\leq \sum_{z \in A} (P(z) - Q(z)) \\ &= \frac{\|P - Q\|_1}{2}.\end{aligned}$$

Equality can be achieved if we choose $f(z) = \begin{cases} 1, & z \in A, \\ 0, & z \notin A \end{cases}$. ■

Remark. The form of the equality in Lemma 11.2 is called the *variational representation* of L_1-distance. More generally, we can represent any convex function (such as the L_1-distance) as the supremum of affine functions.

We now have a bound on (11.4) for bounded f:

Corollary 11.3. *For any distributions P and Q of a finite set $\mathcal{Z}$, and any function $f : \mathcal{Z} \to [0, 1]$, we have*

$$|\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)]| \leq \frac{\|P - Q\|_1}{2}.$$

Exercise 11.1. *The statement of Lemma 11.2 does not include the absolute value. Verify that the corollary follows by applying Lemma 11.2 twice: once using f, and another using $g = 1 - f$.*

As noted earlier, our initial setup corresponds to choosing P to be some joint distribution P_{XY}, and Q to be the product of the marginals $P_X P_Y$. Then, the closer P_{XY} is to $P_X P_Y$, the closer they are to independence (i.e., the less Y depends on X), which makes the exploration bias smaller, as captured in the corollary.

One disadvantage of the above bound is that it is restricted to bounded functions. And as noted in the remark, the main property that allowed us to derive such bound is the convexity of the L_1-distance. Hence, we can derive similar bounds using other convex dependence measures. In particular, we will turn to the KL divergence.

11.2.2 Mutual Information Bound

Our second bound for the quantity from Equation (11.4) can be expressed in the following compact form:

Lemma 11.4. *Suppose that $f(Z) - \mathbb{E}_Q[f(Z)]$ is subgaussian with variance proxy σ^2 when Z is distributed according to Q. Then,*

$$|\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)]| \leq \sigma\sqrt{2D(P||Q)}. \tag{11.5}$$

Proof. This lemma is a clever application of the Donsker-Varadhan variational representation, see Lemma 4.12. Specifically,

$$D(P||Q) = \sup_g \mathbb{E}_P[g(Z)] - \log \mathbb{E}_Q[e^{g(Z)}] \tag{11.6}$$

$$= \sup_g \mathbb{E}_P[g(Z)] - \mathbb{E}_Q[g(Z)] - \log \mathbb{E}_Q[e^{g(Z) - \mathbb{E}_Q[g(\tilde{Z})]}] \tag{11.7}$$

Hence, in particular, selecting $g(z) = \lambda f(z)$, we have

$$D(P||Q) \geq \lambda\left(\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)]\right) - \log \underbrace{\mathbb{E}_Q[e^{\lambda(f(Z) - \mathbb{E}_Q[f(\tilde{Z})])}]}_{M(\lambda)}, \tag{11.8}$$

where $M(\lambda)$ is the moment generating function of the random variable $f(Z) - \mathbb{E}_Q[f(Z)]$ when Z is distributed according to Q. The assumption given in the lemma statement is precisely the requirement that $M(\lambda) \leq e^{\lambda^2\sigma^2/2}$, see Lemma 2.1. Hence,

$$D(P||Q) \geq \lambda\left(\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)]\right) - \lambda^2\sigma^2/2. \tag{11.9}$$

This bound holds for all λ. Maximizing over λ gives the claimed bound. □

From Lemma 11.4, we obtain the following theorem.

Theorem 11.5. *If for all $h \in \mathcal{H}$, $\ell(h, X) - \mathbb{E}[\ell(h, X)]$ is subgaussian with variance proxy σ^2, then*

$$\text{gen}(P_X, P_{H|D}) \leq \sqrt{\frac{2\sigma^2}{n} I(D; H)}. \tag{11.10}$$

Proof. First of all, let us recall (see for example in the proof of Hoeffding's bound, Lemma 2.6) that if $\ell(h, X) - \mathbb{E}[\ell(h, X)]$ is subgaussian with variance proxy σ^2, then the function

$$f(h, D) = \frac{1}{n}\sum_{i=1}^{n}(\ell(h, X_i) - \mathbb{E}[\ell(h, X_i)]) \tag{11.11}$$

is subgaussian with variance proxy σ^2/n.

We apply Lemma 11.4 separately for each instance h. Specifically, in Lemma 11.4, we select $P = P_{D|H=h}$ and $Q = P_D$ (the marginal distribution of the data). Then, Lemma 11.4 establishes the bound

$$|\mathbb{E}_P[f(h, D)] - \mathbb{E}_Q[f(h, \tilde{D})]| \leq \frac{\sigma}{\sqrt{n}}\sqrt{2D(P_{D|H=h}||Q_D)}. \tag{11.12}$$

The theorem follows by taking expectations over H on both sides. Specifically, for the left hand side, we observe

$$\begin{aligned} \text{gen}(P_X, P_{H|D}) &= |\mathbb{E}_{P_{DH}}[L_D(H) - L_{P_X}(H)]| && (11.13)\\ &\leq \mathbb{E}_{P_H}\left|\mathbb{E}_{P_{D|H}}[L_D(H) - L_{P_X}(H)]\right|, && (11.14)\end{aligned}$$

and the last expression is precisely the left hand side of Equation (11.12) with expectation over H. For the right hand side, we observe that by Jensen (square root is concave)

$$\mathbb{E}_{P_H}\left[\frac{\sigma}{\sqrt{n}}\sqrt{2D(P_{D|H=h}||Q_D)}\right] \leq \frac{\sigma}{\sqrt{n}}\sqrt{2\mathbb{E}_{P_H}\left[D(P_{D|H=h}||Q_D)\right]}, \tag{11.15}$$

and the last expectation is by definition simply the mutual information $I(D; H)$. □

11.3 Exploration Bias

In this section, we study a seemingly unrelated question that arises frequently in Data Science: A given data set is first used to test a certain hypothesis. Unfortunately, the statistician ends up concluding that the answer is not statistically significant. She then dreams up a new hypothesis to test against the same data set, and so on. If we proceed in this way, selecting the hypothesis to be tested *after* seeing the data, we introduce what is referred to as an *exploration bias*. As we will see, the methods and bounds derived in the previous section will be useful here, too.

Let $\mathcal{X}$ denote the sample space, and $D \in \mathcal{X}^n$ denote the data set. Assume that we have m hypotheses, $i \in \{1, \cdots, m\}$. Let $\phi_i(D)$, $i \in \{1, 2, \ldots, m\}$, denote the test statistics for hypothesis i. It might be useful to think of the m hypotheses as the m functions we have to choose from in a learning task and the $\phi_i(D)$ might be the empirical risk for hypothesis i based on the sample i in case we use an empirical risk minimizer as learning algorithm.

Since D is random, so is $\phi_i(D)$. The true mean associated to ϕ_i is $\mu_i = \mathbb{E}[\phi_i(D)]$, where the expectation is over the randomness of the dataset. On a particular dataset D, if the learning algorithm decides on $T(D) = i$ the output of data exploration is the value $\phi_{T(D)}(D)$. The reported value is thus $\mathbb{E}[\phi_{T(D)}(D)]$, resulting in a bias of $\mathbb{E}[\phi_{T(D)}(D)] - \mu_{T(D)}$.

We would like to bound

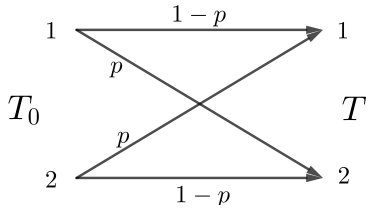
$$\left|\mathbb{E}[\phi_{T(D)}(D) - \mu_{T(D)}]\right|. \tag{11.16}$$

Note that T is not necessarily a deterministic function of D, rather there exists a conditional distribution $P_{T|D}$. In the remainder, we will assume that T is chosen based on the measurements $\phi = (\phi_1, \phi_2, \ldots, \phi_m)$ and suppress D in the notation. That is, we can rewrite (11.16) as

$$\left|\mathbb{E}[\phi_T - \mu_T]\right|. \tag{11.17}$$

The next observation is that this quantity is exactly of the type of Equation (11.4). To see this, in Equation (11.4), we select $\mathcal{Z} = \mathbb{R}^m \times \{1, 2, \ldots, m\}$ (where $\mathbb{R}^m$ and $\{1, 2, \ldots, m\}$ represent the sets in which ϕ and T live, respectively), $f(\phi, t) = \phi_t$, $P = P_{\phi T}$ (i.e., the joint distribution of T and ϕ), and $Q = P_\phi P_T$ (i.e., the product of the marginals of T and ϕ).

Example 11.2. Let ϕ_1 and $\phi_2 \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Let $T_0 = \operatorname{argmax}_{i\in\{1,2\}} \phi_i$ and generate T as follows: $T = \begin{cases} T_0, & \text{with probability } 1-p \\ 3 - T_0, & \text{with probability } p \end{cases}$ for some $p \in [0, 1]$. Now to compute the



exploration bias, note that $\mathbb{E}[\mu_T] = \mu$. On the other hand,

$$\begin{aligned}\mathbb{E}[\phi_T] &= \mathbb{P}(T = T_0)\mathbb{E}[\phi_{T_0}] + \mathbb{P}(T = 3 - T_0)\mathbb{E}[\phi_{(3-T_0)}] \\ &= (1-p)\mathbb{E}[\max\{\phi_1, \phi_2\}] + p\mathbb{E}[\min\{\phi_1, \phi_2\}].\end{aligned}$$

Now let $S = \phi_1 + \phi_2$ and $\Delta = \phi_1 - \phi_2$. It is straightforward to check that $S \sim \mathcal{N}(2\mu, 2\sigma^2)$, $\Delta \sim \mathcal{N}(0, 2\sigma^2)$, $\max\{\phi_1, \phi_2\} = \frac{S+|\Delta|}{2}$, and $\min\{\phi_1, \phi_2\} = \frac{S-|\Delta|}{2}$. Then,

$$\begin{aligned}\mathbb{E}[\phi_T] &= \frac{1}{2}\Big((1-p)\mathbb{E}[S + |\Delta|] + p\mathbb{E}[S - \Delta]\Big) \\ &= \frac{1}{2}\Big(\mathbb{E}[S] + (1-2p)\mathbb{E}[|\Delta|]\Big) \\ &= \frac{1}{2}\left(2\mu + (1-2p)\sqrt{\frac{4\sigma^2}{\pi}}\right) \\ &= \mu + (1-2p)\sigma\sqrt{\frac{1}{\pi}}.\end{aligned}$$

Hence, the exploration bias is given by

$$\left|\mathbb{E}[\phi_T] - \mathbb{E}[\mu_T]\right| = |1-2p|\sigma\sqrt{\frac{1}{\pi}}.$$

Note that for $p = \frac{1}{2}$, the bias is zero. Indeed, for $p = 1/2$, T is independent of (ϕ_1, ϕ_2); hence the index does not depend on the data, so we are not introducing any bias. As we decrease p from $\frac{1}{2}$ to 0, we "increase the dependence" between T_0 and (ϕ_1, ϕ_2), and the exploration bias increases accordingly.

As we saw in the above example, the exploration bias depends on the degree to which T depends on ϕ. Hence, we will use dependence measures to find good bounds on the bias.

11.3.1 Mutual Information Bound

We are now ready to prove the main bound for this section on the exploration bias $|\mathbb{E}[\phi_T - \mu_T]|$, where $\phi = (\phi_1, \phi_2, \ldots, \phi_m)$ and $T \in \{1, 2, \ldots, m\}$.

Theorem 11.6. *Suppose for each $i \in \{1, 2, \ldots, m\}$, $\phi_i - \mu_i$ is σ^2-subgaussian. Then,*

$$|\mathbb{E}[\phi_T - \mu_T]| \leq \sigma\sqrt{2I(T;\phi)}.$$

Remark. As expected, if T is independent of ϕ, then the exploration bias is zero. If T does not depend "too much" on ϕ, as captured by mutual information, then we can guarantee a small bias.

Proof. We apply Lemma 11.4 separately for each instance $T = i$, letting $P = P_{\phi_i|T=i}$ and $Q = P_{\phi_i}$. This gives the bound

$$\left|\mathbb{E}_{P_{\phi_i|T=i}}[\phi_i] - \mu_i\right| \leq \sigma\sqrt{2D(P_{\phi_i|T=i}||P_{\phi_i})}. \tag{11.18}$$

The theorem then follows by averaging both sides of T. Specifically,

$$\begin{aligned}
|\mathbb{E}[\phi_T - \mu_T]| &= \left|\sum_{i=1}^m \mathbb{P}(T=i)\mathbb{E}_{P_{\phi_i|T=i}}[\phi_i] - \mu_i\right| \\
&\overset{(a)}{\leq} \sum_{i=1}^m \mathbb{P}(T=i)\left|\mathbb{E}_{P_{\phi_i|T=i}}[\phi_i] - \mu_i\right| \\
&\leq \sum_{i=1}^m \mathbb{P}(T=i)\sigma\sqrt{2D(P_{\phi_i|T=i}||P_{\phi_i})} \\
&\overset{(b)}{\leq} \sum_{i=1}^m \mathbb{P}(T=i)\sigma\sqrt{2D(P_{\phi_1,\phi_2,\ldots,\phi_m|T=i}||P_{\phi_1,\phi_2,\ldots,\phi_m})} \\
&\overset{(c)}{\leq} \sigma\sqrt{2\sum_{i=1}^m \mathbb{P}(T=i)D(P_{\phi|T=i}||P_\phi)} \\
&= \sigma\sqrt{2D(P_{\phi T}||P_\phi P_T)} = \sigma\sqrt{2I(T;\phi)},
\end{aligned}$$

where (a) and (c) follow from Jensen's inequality, and (b) follows from the data processing inequality. □

Exercise 11.2. *Show that, if $\phi_i - \mu_i$ is σ_i^2-subgaussian for each $i \in \{1, 2, \ldots, m\}$, then*

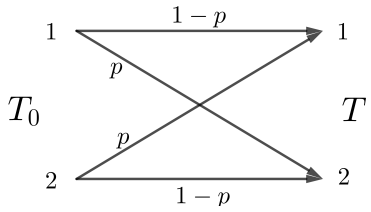
$$|\mathbb{E}[\phi_T - \mu_T]| \leq \sqrt{\mathbb{E}[\sigma_T^2]}\sqrt{2I(T;\phi)}.$$

Let's revisit the initial example:

Example 11.3. Let ϕ_1 and $\phi_2 \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Let $T_0 = \text{argmax}_{i\in\{1,2\}}\phi_i$ and generate T as follows: $T = \begin{cases} T_0, & \text{with probability } 1-p \\ 3 - T_0, & \text{with probability } p \end{cases}$ for some $p \in [0, 1]$.

Since $\phi_i - \mu_i \sim \mathcal{N}(0, \sigma^2)$, it is σ^2-subgaussian, thus satisfying the assumption of Theorem 11.6. To compute $I(T;\phi)$:

$$I(T;\phi) = H(T) - H(T|\phi).$$



Since ϕ_1 and ϕ_2 are i.i.d, then $\mathbb{P}(T_0 = 1) = \mathbb{P}(T_0 = 2) = \frac{1}{2}$. Hence, $H(T) = \log 2$. Since both $\phi - T_0 - T$ and $T_0 - \phi - T$ are Markov chains, we get $H(T|\phi, T_0) = H(T|\phi)$ and $H(T|\phi, T_0) = H(T|T_0)$. Hence,

$$I(T;\phi) = H(T) - H(T|\phi) = \log 2 - H(T|T_0) = \log 2 - H(p).$$

Hence, by the above theorem,

$$|\mathbb{E}[\phi_T - \mu_T]| \leq \sigma\sqrt{2\left(\log 2 - H(p)\right)}.$$

Example 11.4. Suppose $\phi_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for $i \in \{1, 2, \ldots, m\}$, and $T = \text{argmax}_i \phi_i$. Then,

$$I(T;\phi) = H(T) = \log m,$$

and

$$\mathbb{E}\left[\max\{\phi_1, \phi_2, \ldots, \phi_m\}\right] \leq \sigma\sqrt{2\log m}.$$

11.4 Problems

Problem 11.1 (Cumulant Generating Function)**.** Given a real random variable X taking values on a finite set $\mathcal{X} \subset \mathbb{R}$, define $\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right]$. Show that

(a) $\psi'(\lambda) = \mathbb{E}[X_\lambda]$ where $\mathbb{E}[X_\lambda]$ is a random variable taking values on $\mathcal{X}$, with distribution $p_\lambda(x) = p(x)\exp(\lambda x)\exp(-\psi(\lambda))$. Hence $\psi'(0) = \mathbb{E}[X]$.

(b) $\psi''(\lambda) = \text{Var}(X_\lambda)$. Conclude that ψ is convex.

Problem 11.2 (Exploration Bias)**.** (a) Let $X_1, X_2, \ldots, X_n \sim$ i.i.d. $\mathcal{N}(0,1)$. Let $Y = \text{argmax}_i X_i$ and $T \in \{1, 2, \ldots, n\}$ is such that

$$P_{T|Y}(t|y) = \begin{cases} p, & t = y \\ \frac{1-p}{n-1}, & t \neq y \end{cases} \quad \text{for some } p \in [0,1]. \tag{11.19}$$

1. Compute $I(X;T)$ where $X = (X_1, X_2, \ldots, X_n)$. (Hint: write $I(X;T) = H(T) - H(T|X)$. What is the marginal distribution of T?)

(b) Let $X_1, \ldots, X_4 \sim$ i.i.d. $\mathcal{N}(0,1)$ and $X_5 \sim \mathcal{N}(0,4)$. Let Y and T be as in part (a) with $p = 0.3$.

1. Show that $\mathbb{P}(Y = 5) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{8\pi}}(1 - Q(x))^4 e^{-x^2/8} dx$, where we are using $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$, and find a corresponding numerical approximation (using Mathematica, for example).

2. Using the previous numerical approximation, find the marginal distributions P_Y and P_T.

Problem 11.3 (Gibbs Algorithm)**.** Let $\mathcal{X}$ be the sample space, $\mathcal{W}$ the hypothesis space, and let $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_+$ be a corresponding loss function. On a dataset $D = (X_1, X_2, \ldots, X_n)$, the empirical risk for a hypothesis w is given by $L_D(w) = \frac{1}{n}\sum_{i=1}^n \ell(w, X_i)$. We saw in class that $I(D;W)$ can be used to bound the generalization error. Hence, we can use it as a *regularizer* in empirical risk minimization.

(a) First, show that given any joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$ and marginal distribution Q on $\mathcal{Y}$, $D(P_{XY}||P_X P_Y) \leq D(P_{XY}||P_X Q)$.

Since we cannot directly compute $D(P_{DW}||P_D P_W)$, we will use $D(P_{DW}||P_D Q)$ as a proxy, where Q is a distribution on $\mathcal{W}$.

(b) Let

$$P^\star_{W|D} = \arg\min_{P_{W|D}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right).$$

1. Show that

$$\min_{P_{W|D}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right) = \mathbb{E}_D \left[\min_{P_{W|D=d}} \left(\mathbb{E}[L_d(W)] + \frac{1}{\beta} D(P_{W|D=d}||Q) \right) \right].$$

2. Show that the minimizer on the right-hand side $P^\star_{W|D=d}$ is given by

$$P^\star_{W|D=d} = \frac{e^{-\beta L_d(w)} Q(w)}{\mathbb{E}_Q\left[e^{-\beta L_d(W)}\right]}.$$

This is known in the literature as the Gibbs algorithm. (Hint: Write $\mathbb{E}[\beta L_d(W)] = \mathbb{E}[\log e^{\beta L_d(W)}]$, combine with the KL divergence term and use non-negativity of KL divergence.)

3. Show that $P^\star_{W|D=d}$ is $2\beta/n$-differential private if $\ell \in [0,1]$.

Problem 11.4 (Dependence and large error events)**.** In the lecture notes we have seen how to bound the expected generalization error using information measures. With this exercise we will work on large error events and provide bounds on the probabilities of such events. The setting is the same: we observe n iid samples $D = (X_1, \ldots, X_n)$ (according to some unknown distribution P) and based on this observation we will choose a hypothesis $w \in W$. We also consider the usual definition of empirical and population risk, *i.e.* given a loss function ℓ, some hypothesis w, $L_D(w) = \frac{1}{n}\sum_{i=1}^n \ell(w, X_i)$, and $L_P(w) = \mathbb{E}_P[\ell(w, X)]$. We are interested in controlling the following quantity:

$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right). \tag{11.20}$$

(a) Suppose that the loss is such that $\ell(w, x) \in \{0, 1\}$ for every $w \in W$ and $x \in \mathcal{X}$. Suppose also that $|\mathcal{W}| < \infty$, *i.e.*, the number of hypotheses is finite.

1. Show that for every **fixed** $w \in W$ $\Pr\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$;

2. Show that
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq |W| \cdot 2\exp(-2n\epsilon^2); \tag{11.21}$$
Hint: denote with $\mathbb{E} = \{(d, w) : |L_P(w) - L_d(w)| > \epsilon\}$.
You have that $\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) = \Pr(E) = \sum_{(w,d)\in E} P(w, d)$.
(be careful: $\Pr\left(|L_P(W) - L_D(W)| > \epsilon | W = w\right)$ is not necessarily $\leq 2\exp(-2n\epsilon^2).Why?$)

(b) Now consider the following information measure, given two discrete random variables X, Y:
$$\mathcal{L}(X \to Y) = \log \sum_y \max_{x:P_X(x)>0} P_{Y|X}(y|x). \tag{11.22}$$
This quantity is known in the literature as Maximal Leakage and quantifies the leakage of information between X and Y.

1. Show that if the alphabet of Y (denoted with $\mathcal{Y}$) is finite then
$$\mathcal{L}(X \to Y) \leq \log |\mathcal{Y}|,$$
which distributions achieve the bound with equality?
2. It is possible to show that
$$\mathcal{L}(X \to Y) \geq 0,$$
which distributions achieve the bound with equality?
3. Let X be a binary random variable and let Y be an observation of X after passing through a Binary Symmetric Channel with parameter δ. More precisely we have $P_{Y|X=x}(x) = 1 - \delta$, for $x \in \{0, 1\}$.
What is the maximal leakage $\mathcal{L}(X \to Y)$?
Which values of δ allow you to achieve the bounds in $(1), (2)$ with equality?
4. Suppose further that the space of samples $\mathcal{D}$ is finite. Denote with $E_w = \{d : (d, w) \in E\}$, for $w \in \mathcal{W}$; Show that:
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq \exp(\mathcal{L}(D \to W)) \max_{w\in\mathcal{W}} \Pr(E_w);$$
5. Conclude that
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq 2\exp(\mathcal{L}(D \to W) - 2n\epsilon^2);$$
6. Compare the two bound retrieved in *(a2)* and *(b4)*, what do you notice? Is one of the two better than the other? When are they equal? What conclusions can you draw?

Bibliography

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 2nd ed., 2006.

[2] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. New York: Cambridge, 2020.

[3] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[4] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[5] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*. New York: Wiley, 2nd ed., 2011.

[6] D. A. Berry and B. Fristedt, *Bandit Problems : Sequential Allocation of Experiments*. New York: Springer, 1985.

[7] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *IRE WESON Convention Record*, vol. 4, (Los Angeles, CA), pp. 96–104, 1960.

[8] M. H. Hayes, *Statistical Digital Signal Processing and Modelling*. Wiley, 1996.

[9] S. O. Haykin, *Adaptive Filter Theory*. Prentice Hall, 5th ed., 2013.

[10] J. Duchi, *Lecture Notes for Statistics 311/Electrial Engieering 377*. Stanford, 2016.

[11] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families and Variational Inference*. NOW Publishers, 2008.

[12] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[13] P. Brémaud, *Mathematical Principles of Signal Processing*. New York: Springer-Verlag, 2002.

[14] M. Vetterli, J. Kovacevic, and V. Goyal, *Foundations of Signal Processing*. Cambridge University Press, 2014.

[15] P. Prandoni and M. Vetterli, *Signal Processing for Communications*. EPFL Press, 2008. Available online at http://www.sp4comm.org.

[16] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[17] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441 and 498–520, 1933.

[18] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[19] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, 1984.

Index

adaptive filters, 65
  LMS, 67
algorithm
  Exp3, 54
  UCB, 47

bandits, 43
  adversarial, 54
  contextual, 56
  Exp3 algorithm, 54
  information theoretic lower bound, 51
  UCB algorithm, 47

distribution
  Bernoulli, 11, 96, 100
  Dirichlet, 12, 79, 97
  Gaussian, 11, 96, 100
  Multinomial, 12, 97, 100
  Poisson, 11, 96
divergence
  KL, 29
    Donsker-Varadhan variational characterization, 33, 140
    Variational Representation, 33

entropy, 31
  relative, 29, 33
estimation
  linear MMSE, 63, 64
  MMSE, 61, 79
  parameter, 67
Exp3 algorithm, 54
expectation
  conditional, 16, 61

Fisher information, 67, 69

Gaussian distribution
  MMSE, 62
generalization error, 137

Hilbert space, 63, 112
  definition, 112
Hoeffding's bound, 14
hypothesis testing
  binary, 59
  repeated independent observations, 61

inequality
  Jensen, 12
  Markov, 13
  Pinsker, 31
information
  Fisher, 67, 69
  mutual, 34, 140

Jensen's inequality, 12

lemma
  Hoeffding, 14
  Johnson-Lindenstrauss, 127
LMS, 67

Markov inequality, 13
moment generating function, 13, 140

PCA, 125
principal components analysis, 125

subexponential, 15, 128
subgaussian, 13

theorem
  Eckart-Young, 23, 126
  projection, 113
Total Variation Distance, 33

UCB algorithm, 47

Variational Representation
  KL, 33
  Total Variation, 33

Wiener filter, 63, 64