Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2017

Recitation 4: Bayesian Estimation Theory

Date: Friday, March 3, 2017 **TA:** Anuran Makur

1 Introduction

We have studied several approaches to statistical inference so far, each suited to a particular set of assumptions on the problem model. Let us briefly understand the common abstract setting of all these frameworks. There is always an underlying hidden variable $x \in \mathcal{X}$ (which could be deterministic or random) that defines an observation model (or likelihood) for a random variable $\mathsf{y} \in \mathcal{Y}$. We typically have access to this likelihood model ($\{p_{\mathsf{y}|\mathsf{x}}(\cdot|x) : x \in \mathcal{X}\}$ or $\{p_{\mathsf{y}}(\cdot; x) : x \in \mathcal{X}\}$). Upon observing a realization $\mathsf{y} = y$ generated by this model, our goal is to infer the hidden variable x.

In the early 1900's, the radar community was interested in models where $|\mathcal{X}|$ was finite (and small), and often $|\mathcal{X}| = 2$. For instance, radar engineers would observe some measurement and have to detect if there was a signal in the measurement, or if the measurement was just random noise. This problem could be set up as a binary hypothesis testing problem where $\mathcal{X} = \{H_0 = \text{no signal}, H_1 = \text{signal}\}$ (and $|\mathcal{X}| = 2$). Today, such problems are classified under the category of *detection theory.* In contrast, the branch of statistics that deals with inference questions where $\mathcal{X}$ is a discrete or "continuous" set with $|\mathcal{X}| = +\infty$ (or has very large cardinality) is known as *estimation theory.* In the radar story, after detecting an analog signal, engineers would have to approximate its value from noisy measurements, which would correspond to a parameter estimation problem.

In the statistics community, there was another divide among inference problems. Bayesian statisticians believed that the underlying parameter x was random and had a prior distribution p_{x} which represented their *belief* about x. So, the "right" way to proceed after observing $\mathsf{y} = y$ was to compute the posterior distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$ in order to update their belief (or use the joint distribution $p_{\mathsf{x},\mathsf{y}}$ to infer x). In contrast, non-Bayesian (or *frequentist*) statisticians did not impose such a prior over x and assumed it was a deterministic parameter. The matrix below classifies the various approaches we have developed in lectures according to these categories.

	Bayesian	**Non-Bayesian**
Detection	Bayesian Hypothesis Testing	Neyman-Pearson, Minimax
Estimation	BLS, LLS	MVU, Efficient

This recitation is concerned with Bayesian estimation theory.

Note: Sections 2.1 and 2.4 are edited, revised, and extended versions of material from previous TAs of the course.

2 Bayesian Parameter Estimation

2.1 Problem Setup and Basic Results

As discussed above, in the Bayesian framework, we assume that a prior distribution $p_{\mathsf{x}}(\cdot)$, and a likelihood model $p_{\mathsf{y}|\mathsf{x}}(\cdot|\cdot)$ are available to us. The goal is to find a good estimator $\hat{x}(y)$ that provides an estimate of x given any observation $\mathsf{y} = y$. Although we work with scalar random variables x and y here, the analysis remains unchanged for random vectors $\mathbf{y}$, and can be easily generalized for random vectors $\mathbf{x}$ (as shown in the lecture notes).

In order to find a good estimator $\hat{x}(\mathsf{y})$, we minimize the expected cost:

$$\hat{x}(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}_{p_{\mathsf{x},\mathsf{y}}}\left[C(\mathsf{x}, f(\mathsf{y}))\right]$$

and different choices of the cost function $C(\cdot,\cdot)$ lead to different optimal estimators. In the lecture notes, we considered the following cost functions:

1. (**Minimum Absolute Error**) $C(a,\hat{a}) = |a - \hat{a}|$
   The optimal estimator $\hat{x}_{MAE}(y)$ is the **median** of the posterior distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$.

2. (**Minimum Uniform Cost**) $C(a,\hat{a}) = \begin{cases} 0\,, & \text{if } |a-\hat{a}| \le \epsilon \\ 1\,, & \text{otherwise} \end{cases}$
   As $\epsilon \to 0$, the optimal estimator $\hat{x}_{MAP}(y)$ is the **mode** of the posterior distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$.

3. (**Minimum Mean-Square Error**) $C(a,\hat{a}) = (a - \hat{a})^2$
   The optimal estimator $\hat{x}_{BLS}(y)$ is the **mean** of the posterior distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$.

The derivations of these results can be found in the lecture notes. Moreover, we note that in order to measure the performance of estimators under the mean-square error criterion, the following quantities were introduced in the lecture notes:

$$\begin{aligned} e(\mathsf{x},\mathsf{y}) &\triangleq \hat{x}(\mathsf{y}) - \mathsf{x} \text{ (\textbf{error})} \\ b &\triangleq \mathbb{E}_{p_{\mathsf{x},\mathsf{y}}}\left[\hat{x}(\mathsf{y}) - \mathsf{x}\right] = \mathbb{E}_{p_{\mathsf{x},\mathsf{y}}}\left[e(\mathsf{x},\mathsf{y})\right] \text{ (\textbf{bias})} \\ \lambda_e &\triangleq \mathbb{E}_{p_{\mathsf{x},\mathsf{y}}}\left[(e(\mathsf{x},\mathsf{y}) - b)^2\right] \text{ (\textbf{error variance})} \\ \text{MSE} &\triangleq \mathbb{E}_{p_{\mathsf{x},\mathsf{y}}}\left[e(\mathsf{x},\mathsf{y})^2\right] = \lambda_e + b^2 \text{ (\textbf{mean-square error})} \end{aligned}$$

2.2 Hilbert Spaces and the Geometry of MMSE Estimation

The most popular cost criterion in Bayesian estimation is the *minimum mean-square error (MMSE)* criterion. When MMSE optimization is performed without any constraints, it outputs the *Bayes' least-squares (BLS)* estimator. On the other hand, when it is performed over linear estimators, it outputs the *linear least-squares (LLS)* estimator. We now develop a unified framework for constrained MMSE estimation that characterizes both BLS and LLS estimators in a single shot.

To this end, let $\mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}}) \triangleq \{f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \,|\, \mathbb{E}\left[f(\mathsf{x},\mathsf{y})^2\right] < +\infty\}$ denote the *Hilbert space* of real-valued functions $f(x, y)$ that have finite second moment (over the field $\mathbb{R}$). The finite second moment constraint is an analytical condition that ensures that the space is complete (you may ignore this for the purposes of this course if you are unfamiliar with it). By Hilbert space, we mean that such functions form a (complete) vector space and is endowed with an inner product. Informally, the vector space structure refers to the fact that linear combinations of functions in the space also belong to the space (this is easy to check, but checking that the finite second moment constraint holds is harder), and the inner product permits us to measure distances and angles between vectors (or functions) in the space. Note that the zero vector in this Hilbert space is the everywhere zero function $f_0(x, y) = 0$.

We endow $\mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$ with the following inner product:

$$\langle f, g\rangle \triangleq \mathbb{E}\left[f(\mathsf{x},\mathsf{y})g(\mathsf{x},\mathsf{y})\right]$$

for any two functions $f, g \in \mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$ (where the expectation is taken with respect to the joint distribution $p_{\mathsf{x},\mathsf{y}}$). It is straightforward to verify that this is indeed an inner product by checking the inner product axioms:

1. (positive definiteness) $\langle f, f\rangle = \mathbb{E}\left[f(\mathsf{x},\mathsf{y})^2\right] \geq 0$ with equality iff $f = f_0$
2. (symmetry) $\langle f, g\rangle = \mathbb{E}\left[f(\mathsf{x},\mathsf{y})g(\mathsf{x},\mathsf{y})\right] = \mathbb{E}\left[g(\mathsf{x},\mathsf{y})f(\mathsf{x},\mathsf{y})\right] = \langle g, f\rangle$
3. (linearity) $\langle af + bg, h\rangle = \mathbb{E}\left[(af(\mathsf{x},\mathsf{y}) + bg(\mathsf{x},\mathsf{y}))h(\mathsf{x},\mathsf{y})\right] = a\,\mathbb{E}\left[f(\mathsf{x},\mathsf{y})h(\mathsf{x},\mathsf{y})\right] + b\,\mathbb{E}\left[g(\mathsf{x},\mathsf{y})h(\mathsf{x},\mathsf{y})\right] = a\,\langle f, h\rangle + b\,\langle g, h\rangle$

where $f, g, h \in \mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$ and $a, b \in \mathbb{R}$. Furthermore, this inner product induces the norm:

$$\|f\| \triangleq \langle f, f\rangle^{\frac{1}{2}} = \mathbb{E}\left[f(\mathsf{x},\mathsf{y})^2\right]^{\frac{1}{2}}$$

for every function $f \in \mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$.

It is worth noting that various well-known inequalities for Hilbert spaces are carried over to this probabilistic setting. For instance, for any two functions $f, g \in \mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$, we have the well-known *Cauchy-Schwarz-Bunyakovsky inequality*:

$$|\langle f, g\rangle|^2 = \mathbb{E}\left[f(\mathsf{x},\mathsf{y})g(\mathsf{x},\mathsf{y})\right]^2 \leq \mathbb{E}\left[f(\mathsf{x},\mathsf{y})^2\right]\mathbb{E}\left[g(\mathsf{x},\mathsf{y})^2\right] = \|f\|^2\,\|g\|^2 \ ,$$

as well as the *triangle inequality*:

$$\|f + g\| = \mathbb{E}\left[(f(\mathsf{x},\mathsf{y}) + g(\mathsf{x},\mathsf{y}))^2\right]^{\frac{1}{2}} \leq \mathbb{E}\left[f(\mathsf{x},\mathsf{y})^2\right]^{\frac{1}{2}} + \mathbb{E}\left[g(\mathsf{x},\mathsf{y})^2\right]^{\frac{1}{2}} = \|f\| + \|g\| \ .$$

Let $\mathcal{S}$ be a linear subspace of $\mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$ (for rigorous mathematicians, we really mean a closed subspace, or a sub-Hilbert space). This means $\mathcal{S}$ is a non-empty subset of $\mathcal{L}^2(\mathcal{X} \times \mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$ that is itself a Hilbert space with the same inner product. Then, we have the following **orthogonality principle** (which can be shown to follow from the *Hilbert projection theorem* in convex analysis).

Theorem 1 (Orthogonality Principle). *Given $g \in \mathcal{L}^2(\mathcal{X} \times \mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$, we have:*

$$h = \arg\min_{f \in \mathcal{S}} \|g - f\|^2 = \arg\min_{f \in \mathcal{S}} \mathbb{E}\left[(g(\mathsf{x}, \mathsf{y}) - f(\mathsf{x}, \mathsf{y}))^2\right]$$

if and only if for every $f \in \mathcal{S}$:

$$\langle h - g, f \rangle = \mathbb{E}\left[(h(\mathsf{x}, \mathsf{y}) - g(\mathsf{x}, \mathsf{y}))f(\mathsf{x}, \mathsf{y})\right] = 0 \, .$$

Proof. Since you are not required to know real analysis to take this course, we omit analytical details that guarantee the existence and uniqueness of h as the solution to the extremization $\min_{f \in \mathcal{S}} \|g - f\|^2$. However, rest assured that the statement of the theorem is rigorous.

To prove the forward direction, consider the function $h - \epsilon f \in \mathcal{S}$ for any fixed $f \in \mathcal{S}$ and $\epsilon \neq 0$, and observe using $h = \arg\min_{f \in \mathcal{S}} \|g - f\|^2$ that:

$$\|g - h\|^2 \leq \|g - h + \epsilon f\|^2 = \|g - h\|^2 + 2\epsilon \langle g - h, f \rangle + \epsilon^2 \|f\|^2$$

which implies that:

$$2\epsilon \langle g - h, f \rangle + \epsilon^2 \|f\|^2 \geq 0 \, .$$

If $\langle g - h, f \rangle > 0$, then taking ϵ to be small (in magnitude) and negative contradicts the non-negativity above. Likewise, if $\langle g - h, f \rangle < 0$, then taking ϵ to be small and positive contradicts the non-negativity above. Hence, we must have $\langle g - h, f \rangle = 0$ for every $f \in \mathcal{S}$.

To prove the converse direction, note that for every $f \in \mathcal{S}$:

$$\begin{aligned}
\|g - f\|^2 &= \|g - h + h - f\|^2 \\
&= \|g - h\|^2 + 2 \langle g - h, h - f \rangle + \|h - f\|^2 \\
&= \|g - h\|^2 + \|h - f\|^2 \\
&\geq \|g - h\|^2
\end{aligned}$$

where the third equality follows from $\langle g - h, h - f \rangle = 0$ since $h - f \in \mathcal{S}$. This completes the proof. □

Geometrically, this principle states that given a function $g \in \mathcal{L}^2(\mathcal{X} \times \mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$, the closest function to g in a subspace $\mathcal{S}$ (in the norm sense) is a function $h \in \mathcal{S}$ such that the error $h - g$ is orthogonal to the subspace $\mathcal{S}$. Suppose $g(\mathsf{x}, \mathsf{y}) = \mathsf{x}$ and $\mathcal{S}$ is some set of possible estimators $\hat{x}(\mathsf{y})$ (which only includes functions that depend on y). Then, the minimization in Theorem 1 corresponds to the constrained MMSE problem:

$$\min_{f \in \mathcal{S}} \mathbb{E}\left[(\mathsf{x} - f(\mathsf{y}))^2\right]$$

where each $f(\cdot)$ is a function of y only, and the set $\mathcal{S}$ may also impose further constraints on the estimators. The solution to this constrained MMSE problem is the estimator $\hat{x}_{\mathcal{S}}(\cdot) = \arg\min_{f \in \mathcal{S}} \mathbb{E}\left[(x - f(y))^2\right]$ that satisfies the orthogonality principle:

$$\mathbb{E}\left[(\hat{x}_{\mathcal{S}}(y) - x) f(y)\right] = 0$$

for all functions $f \in \mathcal{S}$. Intuitively, this means that the error $e(x, y) = \hat{x}_{\mathcal{S}}(y) - x$ of this optimal estimator is orthogonal to (or uncorrelated with) any function of y. Equivalently, $\hat{x}_{\mathcal{S}}(y)$ is the *projection* of x onto the subspace $\mathcal{S}$. We will use this idea to establish orthogonality characterizations of BLS and LLS estimators as corollaries of Theorem 1 in the ensuing subsections.

2.3 Bayes' Least-Squares Estimator

Recall that the BLS estimator $\hat{x}_{BLS}(y) = \mathbb{E}\left[x | y = y\right]$ is the solution to the optimization problem:

$$\hat{x}_{BLS}(\cdot) = \underset{f(\cdot)}{\arg\min}\, \mathbb{E}\left[(x - f(y))^2\right]$$

where we minimize over all functions with domain $\mathcal{Y}$. It has the desirable property that $\mathbb{E}\left[\hat{x}_{BLS}(y)\right] = \mathbb{E}\left[x\right]$ (which follows from the tower property of expectation), i.e. the BLS estimator $\hat{x}_{BLS}(y)$ is unbiased. Now consider the sub-Hilbert space $\mathcal{S} = \{f : \mathcal{Y} \to \mathbb{R} \,|\, \mathbb{E}\left[f(y)^2\right] < +\infty\}$ of $\mathcal{L}^2(\mathcal{X} \times \mathcal{Y}, p_{x,y})$ that contains all real-valued functions $f(y)$ (only depending on y) with finite second moment. Informally verifying that $\mathcal{S}$ is indeed a subspace is straightforward, as linear combinations of functions of y are functions of y. Letting $g(x, y) = x$ in Theorem 1 provides the following *orthogonality characterization* of BLS estimators:

$$\hat{x}_{BLS}(\cdot) = \underset{f \in \mathcal{S}}{\arg\min}\, \mathbb{E}\left[(x - f(y))^2\right]$$

if and only if for every function $f : \mathcal{Y} \to \mathbb{R}$ (with finite second moment):

$$\mathbb{E}\left[(\hat{x}_{BLS}(y) - x) f(y)\right] = 0\,.$$

Therefore, the BLS estimator is defined by the property that the *error* $e(x, y) \triangleq \hat{x}_{BLS}(y) - x$ is orthogonal to every function of the data y (or equivalently, the subspace of all functions of y). In other words, the BLS estimator is the projection of x onto the subspace $\mathcal{S}$ of estimators (which are functions of y).

In fact, we can derive the explicit form of the BLS estimator from its orthogonality characterization. If for every $f : \mathcal{Y} \to \mathbb{R}$ with finite second moment, we have:

$$\mathbb{E}\left[\hat{x}_{BLS}(y) f(y)\right] = \mathbb{E}\left[x f(y)\right] = \mathbb{E}\left[\mathbb{E}\left[x|y\right] f(y)\right] \;\Rightarrow\; \mathbb{E}\left[(\hat{x}_{BLS}(y) - \mathbb{E}\left[x|y\right]) f(y)\right] = 0$$

then $\hat{x}_{BLS}(y) = \mathbb{E}\left[x|y\right]$, since we can take $f(y) = \hat{x}_{BLS}(y) - \mathbb{E}\left[x|y\right]$ and use the fact that the second moment of a random variable vanishes iff the random variable is zero with probability one.

2.4 Linear Least-Squares Estimator

Although BLS estimators are nice, they require complete knowledge of the joint distribution $p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)$. Such information may not be available in some situations. Even when it is available, finding the BLS estimator may be computationally challenging, especially when performing real-time calculations. Under these circumstances, it makes sense for us to trade optimality for the sake of speed and simplicity.

In particular, we often restrict our attention to the class of linear estimators. The LLS estimator is defined as follows:

$$\hat{x}_{LLS}(\cdot) = \arg\min_{f(\cdot)\in\mathcal{S}} \mathbb{E}\left[(\mathsf{x} - f(\mathsf{y}))^2\right]$$

where $\mathcal{S} = \{f : \mathcal{Y} \to \mathbb{R} \,|\, f(y) = ay + d \text{ for some } a, d \in \mathbb{R}\}$. If $\mathbb{E}\left[\mathsf{y}^2\right] < +\infty$, one can verify that $\mathcal{S}$ is a sub-Hilbert space of $\mathcal{L}^2(\mathcal{X}\times\mathcal{Y}, p_{\mathsf{x},\mathsf{y}})$ that contains all real-valued linear functions $f(y)$ (only depending on y). Note that we are essentially performing MMSE estimation over a smaller subspace (than that in BLS estimation) here. As before, letting $g(\mathsf{x},\mathsf{y}) = \mathsf{x}$ in Theorem 1 provides the following *orthogonality characterization* of LLS estimators:

$$\hat{x}_{LLS}(\cdot) = \arg\min_{f\in\mathcal{S}} \mathbb{E}\left[(\mathsf{x} - f(\mathsf{y}))^2\right]$$

if and only if for every linear function $f(y) = ay + d$:

$$\mathbb{E}\left[(\hat{x}_{LLS}(\mathsf{y}) - \mathsf{x})f(\mathsf{y})\right] = 0\,.$$

Therefore, the LLS estimator is defined by the property that the *error* $\mathsf{e}(\mathsf{x},\mathsf{y}) \triangleq \hat{x}_{LLS}(\mathsf{y}) - \mathsf{x}$ is orthogonal to every *linear* function of the data y. In the lecture notes, this characterization is used to prove that the LLS estimator has the form:

$$\hat{x}_{LLS}(y) = \frac{\mathrm{cov}(\mathsf{x},\mathsf{y})}{\mathrm{var}(\mathsf{y})}(y - \mathbb{E}\left[\mathsf{y}\right]) + \mathbb{E}\left[\mathsf{x}\right]$$

which is clearly unbiased, and has error variance:

$$\lambda_{LLS} = \mathbb{E}\left[\mathsf{e}(\mathsf{x},\mathsf{y})^2\right] = \mathrm{var}(\mathsf{x}) - \frac{\mathrm{cov}(\mathsf{x},\mathsf{y})^2}{\mathrm{var}(\mathsf{y})}\,.$$

Finally, we mention a few more remarks. Firstly, the LLS estimator has the advantage that it can be computed using only first and second order moments of the joint distribution $p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)$; these quantities are much easier to obtain in practice than $p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)$ itself. Secondly, it is proved in the lectures notes that if only the first and second order moments of $p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)$ are available, the LLS estimator is actually the minimax optimal MSE estimator in a precise sense. Lastly, when (x,y) are jointly Gaussian random variables, the BLS and LLS estimators coincide.

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2017

Recitation 5: Exponential Families and Sufficient Statistics

Date: Friday, March 10, 2017 **TA:** Anuran Makur

1 Exponential Families

Exponential families form an important class of distributions in statistics. Unfortunately, it is difficult to motivate their importance on a first sighting. This is because the utility of exponential families is often only obvious once their basic properties have been developed. Nevertheless, we collect some salient features of exponential families below without going into any detailed explanations (since many of these features will be covered later in the course).

1. Exponential families admit certain conjugate families of distributions (which are themselves exponential families). In the Bayesian estimation setting, such *conjugate priors* make posterior belief updates particularly efficient. We will see this later in the course.

2. In the non-Bayesian estimation setting, if we are performing i.i.d. sampling from some likelihood model, then exponential families are the only models for which there are *sufficient statistics whose dimensions remain bounded* as the the sample size grows. This is the content of the well-known *Pitman-Koopman-Darmois theorem* (which is proved using various regularity conditions). Although we will touch upon this later, a thorough treatment is beyond the scope of this course.

3. *Efficient estimators* exist for a likelihood model if and only if the model is described by an exponential family with certain additional constraints. A detailed exposition of this result can be found in the lecture notes.

4. Exponential families are *maximum entropy distributions* subject to linear (expectation) constraints such as moment constraints. For this reason, they are sometimes used as priors for different models in order to capture the maximum amount of uncertainty about the latent variable. We will see this later in the course.

5. Exponential *tilting* is a very useful tool in probability theory. As we will see later in the course, it will be indispensable in proving and deriving intuition about results from *large deviations theory* such as the Cramér-Chernoff theorem.

6. Exponential families are *analytically tractable* models that allow us to prove things rigorously about them. For instance, in a canonical exponential family, the *cumulants* of y can be easily calculated from the log-partition function.

Note: Section 1.2 is an edited and revised version of material from previous TAs of the course.

1.1 Definition

Formally, we define a (one-parameter) **exponential family** as the parametrized set of distributions $\{p_Y(\cdot;x); x \in \mathcal{X}\}$ over the alphabet $\mathcal{Y}$ that have the form:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ p_Y(y;x) = \exp\left(\lambda(x)t(y) - \alpha(x) + \beta(y)\right) \tag{1}$$

for some choice of *natural parameter* $\lambda : \mathcal{X} \to \mathbb{R}$, *natural statistic* $t : \mathcal{Y} \to \mathbb{R}$ (which is also a sufficient statistic), and *log-base function* $\beta : \mathcal{Y} \to \mathbb{R}$. The *log-partition function* $\alpha : \mathcal{X} \to \mathbb{R}$ (where the name comes from statistical mechanics) is the logarithm of the normalization constant:

$$\forall x \in \mathcal{X},\ \ \exp(\alpha(x)) = \sum_{y \in \mathcal{Y}} \exp\left(\lambda(x)t(y) + \beta(y)\right)$$

where we assume that $\mathcal{Y}$ is discrete; the continuous setting is analogous with the sum replaced by an integral. Typically, we require that such an exponential family is *regular*, i.e. $\mathcal{Y}$ does not depend on the parameter x. Furthermore, the parameter space $\mathcal{X}$ is chosen such that the normalization constants are finite, and is typically some interval so that we can freely differentiate $\alpha(\cdot)$ on $\mathcal{X}^\circ$ (the interior of $\mathcal{X}$). Finally, we often consider models where the base function $\exp(\beta(\cdot))$ is actually a valid *base distribution*. In this context, the exponential family can be interpreted as a (general) tilting of the base distribution. Since several properties of exponential families are explained at great length in the lecture notes, we omit a discussion of them here.

1.2 A Distribution that is not an Exponential Family

In the lecture notes, a considerably wide range of distributions are shown to be exponential families. However, not all distributions can be parametrized as exponential families with a finite number of natural parameters. In this subsection, we present an example of a family of heavy-tailed distributions that cannot be written as a one-parameter exponential family. Consider the parametrized family of pdfs $\{p_Y(\cdot;\mu) : \mu > 0\}$ with support $\mathbb{R}^+$:

$$\forall \mu > 0, \forall y \geq 0,\ \ p_Y(y;\mu) = \frac{\mu}{(\mu + y)^2} \tag{2}$$

which are a specialization of the so called *Burr or Singh-Maddala distributions* (and are one-sided analogs of the better known Cauchy distributions). We claim that this model does not belong to a one-parameter exponential family.

We prove this by contradiction. Suppose that the model $p_Y(\cdot;\mu)$ belongs to a one-parameter exponential family:

$$\forall \mu > 0, \forall y \geq 0,\ \ \ln p_Y(y;\mu) = \lambda(\mu)t(y) - \alpha(\mu) + \beta(y)\,.$$

On the one hand, taking the second partial derivative $\frac{\partial^2}{\partial\mu\partial y}$ of this leads to the following factorized form:

$$\frac{\partial^2}{\partial\mu\partial y}\ln p_{\mathsf{y}}(y;\mu) = \lambda'(\mu)t'(y)\,.$$

So, if we define $g(\mu, y) = \frac{\partial^2}{\partial\mu\partial y}\ln p_{\mathsf{y}}(y;\mu)$, then the formula above implies that:

$$\frac{g(\mu_1, y)}{g(\mu_2, y)} = \frac{\lambda'(\mu_1)t'(y)}{\lambda'(\mu_2)t'(y)} = \frac{\lambda'(\mu_1)}{\lambda'(\mu_2)} \tag{3}$$

which is not a function of y. On the other hand, using the form of the distribution in (2), we know that:

$$g(\mu, y) = \frac{\partial^2}{\partial\mu\partial y}\ln p_{\mathsf{y}}(y;\mu) = \frac{2}{(\mu+y)^2}$$

which in turn implies that:

$$\frac{g(\mu_1, y)}{g(\mu_2, y)} = \frac{(\mu_2+y)^2}{(\mu_1+y)^2}\,.$$

Since $g(\mu_1, y)/g(\mu_2, y)$ depends on y here, it contradicts the form presented in (3) (which is derived from the one-parameter exponential family assumption). Therefore, the model $p_{\mathsf{y}}(\cdot;\mu)$ is not a one-parameter exponential family. This completes the proof. In closing this section, we remark that other notable distributions that are not exponential families include the Cauchy distributions and their generalizations, the Student's t-distributions.

2 Sufficient Statistics

In this section, we will consider sufficient statistics from a non-Bayesian standpoint as this stays truer to the historical development of the subject. The lecture notes offer Bayesian analogs of some of the topics we will cover. Recall that the setup of non-Bayesian parameter estimation involves a deterministic parameter $x \in \mathcal{X}$ that determines the likelihood $p_{\mathsf{y}}(\cdot;x)$ of a random variable $\mathsf{y} \in \mathcal{Y}$. Let us assume throughout this discussion that $\mathcal{X}$ and $\mathcal{Y}$ are fixed intervals in $\mathbb{R}$ (and $\{p_{\mathsf{y}}(\cdot;x)|x \in \mathcal{X}\}$ are pdfs). In general, a *statistic* for such a model refers to a deterministic function $t : \mathcal{Y} \to \mathbb{R}$, whose purpose is often to capture some useful information about the underlying parameter x. In the sequel, we will let $\mathsf{t} = t(\mathsf{y})$ be the random variable corresponding to this statistic, and $\mathcal{T} \subseteq \mathbb{R}$ be the image of t. The remaining subsections introduce various notions of statistics that are standard in the statistical inference literature. Some of this ensuing exposition is inspired by and structured like the wonderful online resource http://www.math.uah.edu/stat/point/Sufficient.html.

2.1 Sufficiency

In the early 1900's, Fisher introduced two important ideas (actually, he introduced many more, but we have only seen two so far) that revolutionized the study of estimation. One was the notion of *intrinsic accuracy* (which is known as *Fisher information* today), and the other was the related notion of *sufficient statistics*. We have already seen Fisher information in this course, so we now turn to the other idea. (The relationship between these two notions is that the *data processing inequality* for Fisher information is met with equality iff we have a sufficient statistic. This is partly hinted at in the problem set, and will become more evident when we study data processing inequalities for mutual information and Kullback-Leibler divergence later in the course.) Intuitively, a sufficient statistic captures all the information about x that is relevant for inference. The next definition rigorizes this intuition.

Definition 1 (Sufficient Statistic)**.** *A statistic $t : \mathcal{Y} \rightarrow \mathbb{R}$ of y is said to be sufficient for x (with respect to the likelihood $p_{\mathsf{y}}(\cdot; x)$) if the conditional distribution $p_{\mathsf{y}|\mathsf{t}}(\cdot|\cdot; x)$ is not a function of x for all $x \in \mathcal{X}$.*

While this defines sufficient statistics, constructing explicit sufficient statistics is often difficult. The next theorem characterizes a factorization structure that sufficient statistics impart on the likelihood model. This structure can sometimes be used to identify sufficient statistics as shown in the lecture notes.

Theorem 1 (Fisher-Neyman Factorization)**.** *A statistic $\mathsf{t} = t(\mathsf{y})$ of y is sufficient for x if and only if there exist functions $a : \mathcal{T} \times \mathcal{X} \rightarrow \mathbb{R}$ and $b : \mathcal{Y} \rightarrow \mathbb{R}$ such that:*

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; p_{\mathsf{y}}(y; x) = a(t(y), x) b(y) \, .$$

We omit a proof of this result as it can be found in the lecture notes.

2.2 Sufficiency and Non-Bayesian Estimation

We next look at the relationship between sufficient statistics and non-Bayesian parameter estimation. Our first result here establishes that maximum likelihood (ML) estimators (when they exist) can be taken to be functions of the sufficient statistic $t(y)$ rather than the variable y. This agrees with the intuition that a sufficient statistic contains all the information necessary for inference about x.

Theorem 2 (Sufficiency and ML Estimation)**.** *If $\mathsf{t} = t(\mathsf{y})$ is a sufficient statistic of y for x, and an ML estimator for x exists, then there exists an ML estimator $\hat{x}_{ML}(\mathsf{t})$ for x that is a function of t.*

Proof. Since an ML estimator exists, it can be found by maximizing the log-likelihood $x \mapsto \ln p_{\mathsf{y}}(y; x)$. Using the Fisher-Neyman factorization theorem, this corresponds to maximizing $x \mapsto \ln a(t(y), x) + \ln b(y)$. The maximizing argument to this latter function depends only on $t(y)$ since $\ln b(y)$ is a constant as a function of x. So, there exists an ML estimator $\hat{x}_{ML}(\mathsf{t})$ for x that is a function of t. □

Our second result is a well-known theorem of Rao and Blackwell that offers a systematic way to improve unbiased estimators in a mean-square error sense.

Theorem 3 (Rao-Blackwell Theorem)**.** *If $\mathsf{t} = t(\mathsf{y})$ is a sufficient statistic of y for x, and $\hat{x}(\mathsf{y})$ is an unbiased estimator for x, then $\hat{x}_{RB}(\mathsf{t}) \triangleq \mathbb{E}\left[\hat{x}(\mathsf{y})|\mathsf{t}\right]$ is an unbiased estimator for x that is uniformly better in a mean-square error sense:*

$$\forall x \in \mathcal{X},\ \ \mathbb{E}\left[\left(\hat{x}_{RB}(t(\mathsf{y})) - x\right)^2\right] \leq \mathbb{E}\left[\left(\hat{x}(\mathsf{y}) - x\right)^2\right] .$$

Proof. First notice that $\hat{x}_{RB}(\mathsf{t}) = \mathbb{E}\left[\hat{x}(\mathsf{y})|\mathsf{t}\right]$ is a valid estimator (that does not depend on x) because the conditional distribution $p_{\mathsf{y}|\mathsf{t}}(\cdot|\cdot; x)$ does not depend on x since t is a sufficient statistic. Moreover, it is unbiased because $\mathbb{E}\left[\hat{x}_{RB}(\mathsf{t})\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{x}(\mathsf{y})|\mathsf{t}\right]\right] = \mathbb{E}\left[\hat{x}(\mathsf{y})\right] = x$ (using the tower property of expectation). To complete the proof, observe that for any $x \in \mathcal{X}$:

$$\begin{aligned}\mathbb{E}\left[\left(\hat{x}_{RB}(t(\mathsf{y})) - x\right)^2\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[\hat{x}(\mathsf{y})|\mathsf{t}\right] - x\right)^2\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\left(\hat{x}(\mathsf{y}) - x\right)^2 |\mathsf{t}\right]\right] \\ &= \mathbb{E}\left[\left(\hat{x}(\mathsf{y}) - x\right)^2\right]\end{aligned}$$

where the inequality follows from conditional Jensen's inequality. □

The process of constructing the estimator $\hat{x}_{RB}(\mathsf{t})$ from an unbiased estimator $\hat{x}(\mathsf{y})$ is known as *Rao-Blackwellization* in statistics. The Rao-Blackwell estimator $\hat{x}_{RB}(\mathsf{t})$ averages the values of the unbiased estimator $\hat{x}(\mathsf{y})$ over all values of y that produce $\mathsf{t} = t(\mathsf{y})$. This can intuitively be perceived as a form of "filtering" that reduces the sensitivity of the estimator to particular choices of y that correspond to the same value of t. This is useful because all the information for inference about x can be found in the sufficient statistic t anyway. So, extra variations in y given t only add to the overall variance without helping in the estimation of x. From the perspective of constructing good unbiased estimators, this tells us that we can restrict our attention to estimators that are functions of t since Rao-Blackwellization will leave such estimators unchanged. Finally, we note that although we only proved that $\hat{x}_{RB}(\mathsf{t})$ has uniformly lower expected mean-square error, it is straightforward to see from the proof that $\hat{x}_{RB}(\mathsf{t})$ actually achieves uniformly lower expected cost for any convex cost function (to which we can apply conditional Jensen's inequality).

2.3 Minimality

Since a particular model can have many different sufficient statistics, it is worthwhile to find sufficient statistics that are intuitively the "most compact"; indeed, it is conceivable that such statistics are useful in applications. We dub such sufficient statistics as *minimal*, and formally define them as follows.

Definition 2 (Minimal Sufficient Statistic)**.** *A sufficient statistic $\mathsf{t}^\star = t^\star(\mathsf{y})$ of y for x is minimal if for any other sufficient statistic $\mathsf{t} = t(\mathsf{y})$ of y for x, there exists a function $g : \mathcal{T} \to \mathbb{R}$ such that $\mathsf{t}^\star = g(\mathsf{t})$.*

Much like the definition of sufficiency, it is difficult to identify minimal sufficient statistics from this definition. So, there are several sufficient conditions in the literature to determine minimality of sufficient statistics. For example, we can use Theorem 2 to deduce that if a model has a unique ML estimator $\hat{x}_{ML}(\mathsf{y})$ for x that is also a sufficient statistic, then $\hat{x}_{ML}(\mathsf{y})$ must be a minimal sufficient statistic since it is a function of every sufficient statistic. There are other sufficient conditions for minimality that depend on exponential family structure; see the problem set. We remark that minimal sufficient statistics usually always exist, but there are pathological instances when they do not exist. On the other hand, when they exist, they are not unique as we can apply invertible maps to generate other minimal statistics. In the next subsection, we turn to the concept of *completeness*, which provides yet another sufficient condition for minimality.

2.4 Completeness

As shown next, complete sufficient statistics have a rather analytical definition. The definition is justified by the various results in mathematical statistics that depend on it, rather than any explicitly tangible intuition. Sometimes, a finer notion of *boundedly complete* statistics is required for a rigorous treatment of the ensuing results, but we will omit this aspect from the discussion.

Definition 3 (Complete Sufficient Statistic)**.** *A sufficient statistic $\mathsf{t} = t(\mathsf{y})$ of y for x is said to be complete if every function $f : \mathcal{T} \to \mathbb{R}$, we have:*

$$\forall x \in \mathcal{X},\, \mathbb{E}\left[f(\mathsf{t})\right] = 0 \quad \Rightarrow \quad \forall x \in \mathcal{X},\, \mathbb{P}\left(f(\mathsf{t}) = 0\right) = 1$$

where the expectation $\mathbb{E}\left[\cdot\right]$ and the probability measure $\mathbb{P}\left(\cdot\right)$ are determined by $p_{\mathsf{y}}(\cdot; x)$.

Note that the condition $\forall x \in \mathcal{X},\, \mathbb{P}\left(f(\mathsf{t}) = 0\right) = 1$ can be interpreted as $f(\cdot) \equiv 0$ (as shown in the lecture notes) because $f(\cdot)$ behaves like the everywhere zero function with respect to the likelihood model under consideration. Completeness can be interpreted in the following functional analytic sense that explains why we call such statistics "complete." We can perceive the set of distributions $\{p_{\mathsf{t}}(\cdot; x) : x \in \mathcal{X}\}$ as vectors in a functional space. Such a set of vectors is called "complete" when it spans the entire space, or equivalently, for any function $f(\cdot)$:

$$\forall x \in \mathcal{X},\;\; \mathbb{E}\left[f(\mathsf{t})\right] = \langle f(\cdot), p_{\mathsf{t}}(\cdot; x)\rangle = \int_{\mathcal{T}} f(t) p_{\mathsf{t}}(t; x)\, dt = 0 \quad \Rightarrow \quad f(\cdot) \equiv 0$$

i.e. if f is orthogonal to the spanning set of vectors, then f must be the zero vector. Hence, a sufficient statistic $\mathsf{t} = t(\mathsf{y})$ is complete if its corresponding likelihoods are complete in the functional analytic sense. We remark that complete sufficient statistics need not exist for every model, and when they exist, are usually not unique.

From the perspective of estimation, we can interpret the definition of completeness as follows. Think of $f(\mathsf{t})$ as a statistic of t that is constructed to be an unbiased estimator for 0 (where 0 is perceived as a function of x). Then, completeness of t implies that the statistic that is 0 with probability one is the only such unbiased estimator. As we will see, the main use of completeness in various proofs is to argue that different estimators (that are functions of t) are equal by showing that their difference is zero. The next result illustrates that completeness is a sufficient condition for minimality.

Theorem 4 (Bahadur's Theorem)**.** *If $\mathsf{t} = t(\mathsf{y})$ is a complete sufficient statistic of y for x, and a minimal sufficient statistic exists, then t is also a minimal sufficient statistic.*

A proof of this result can be found in the lecture notes. The utility of this theorem arises from the fact that completeness can sometimes be an easy condition to check (often due to the invertibility of *Laplace and $\mathcal{Z}$-transforms*); see the lecture notes and exercises for examples. Note however that completeness is not a necessary condition for minimality.

We next introduce a celebrated classical result due to Lehmann and Scheffé which shows that complete sufficient statistics can be used to generate minimum-variance unbiased (MVU) estimators via Rao-Blackwellization.

Theorem 5 (Lehmann-Scheffé Theorem)**.** *If $\mathsf{t} = t(\mathsf{y})$ is a complete sufficient statistic of y for x, and $\hat{x}(\mathsf{t})$ is an unbiased estimator for x, then $\hat{x}(\mathsf{t})$ is an MVU estimator for x.*

Proof. Fix any unbiased estimator $\tilde{x}(\mathsf{y})$ for x. Then, $\hat{x}_{RB}(\mathsf{t}) = \mathbb{E}\left[\tilde{x}(\mathsf{y})|\mathsf{t}\right]$ is an unbiased estimator for x with uniformly lower mean-square error by the Rao-Blackwell theorem since t is a sufficient statistic. Furthermore, $\hat{x}_{RB}(\mathsf{t}) - \hat{x}(\mathsf{t})$ is a function of t, and $\mathbb{E}\left[\hat{x}_{RB}(\mathsf{t}) - \hat{x}(\mathsf{t})\right] = x - x = 0$ for every $x \in \mathcal{X}$ (using the tower property of expectation). Since t is complete, we have $\hat{x}_{RB}(\mathsf{t}) = \hat{x}(\mathsf{t})$ with probability one for every $x \in \mathcal{X}$. Hence, $\hat{x}(\mathsf{t})$ has uniformly lower mean-square error than every unbiased estimator $\tilde{x}(\mathsf{y})$ for x. This means that it is the MVU estimator for x. □

Some remarks are in order. Firstly, the Lehmann-Scheffé theorem can be easily generalized for convex cost functions (since the Rao-Blackwell theorem admits such a generalization). Secondly, $\hat{x}(\mathsf{t})$ is the *unique* function of t that is an unbiased estimator of x. This can be easily argued using completeness as shown in the proof. Thirdly, the Lehmann-Scheffé theorem offers a systematic way to find MVU estimators: find a complete sufficient statistic t and an unbiased estimator $\tilde{x}(\mathsf{y})$, and then Rao-Blackwellize $\tilde{x}(\mathsf{y})$ to get the MVU estimator $\hat{x}_{MVU}(\mathsf{y}) = \hat{x}_{RB}(\mathsf{t})$. While this provides an alternative approach to constructing MVU estimators (in contrast to verifying whether an efficient estimator exists that meets the Cramér-Rao bound with equality for all $x \in \mathcal{X}$), it can sometimes be difficult to find a complete sufficient statistic and/or evaluate the conditional expectation required for Rao-Blackwellization.

2.5 Ancillarity

Finally, we introduce the concept of *ancillary* statistics, which is an idea due to Fisher.

Definition 4 (Ancillary Statistic)**.** *If $\mathsf{t} = t(\mathsf{y})$ is a statistic whose distribution $p_{\mathsf{t}}(\cdot; x)$ does not depend on x, then it is known as an ancillary statistic for x.*

Since an ancillary statistic has no information about the parameter x, we intuitively expect it to be independent of a sufficient statistic that contains all of the relevant information for inference about x "in a compact fashion." It turns out that completeness (rather than minimality) is the right way to define compactness here. The next theorem formally states this result.

Theorem 6 (Basu's Theorem)**.** *If $\mathsf{t} = t(\mathsf{y})$ is a complete sufficient statistic of y for x, and $\mathsf{s} = s(\mathsf{y})$ is an ancillary statistic of y, then t and s are independent.*

The proof of this result can be found in the problem set, which also offers examples of ancillary statistics. In closing, we note that together, Basu's theorem, Bahadur's theorem, and the Lehmann-Scheffé theorem demonstrate the utility of completeness in mathematical statistics.

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2017

Recitation 8: Differential Entropy

Date: Friday, April 14, 2017 **TA:** Anuran Makur

1 Differential Entropy

In our study of continuous information measures, we came across a natural analog of discrete Shannon entropy for probability density functions (PDFs) known as *differential entropy.*

Definition 1 (Differential Entropy)**.** *Given a continuous random vector $\mathsf{x} \in \mathbb{R}^n$ with PDF $p_{\mathsf{x}}(\cdot)$ that has support $\mathcal{X} \subseteq \mathbb{R}^n$, we define the differential entropy of x as:*

$$h(\mathsf{x}) \triangleq \mathbb{E}\left[\log\left(\frac{1}{p_{\mathsf{x}}(\mathsf{x})}\right)\right] = \int_{\mathcal{X}} p_{\mathsf{x}}(\mathbf{x}) \log\left(\frac{1}{p_{\mathsf{x}}(\mathbf{x})}\right) d\mathbf{x}$$

when the expectation is well-defined. All logarithms will be assumed to be natural.

Unlike KL divergence and mutual information, which remain non-negative and invariant to coordinate transformations in the continuous setting, it was shown in the lecture notes that differential entropy is neither (necessarily) non-negative nor (necessarily) invariant to coordinate transformations. As a result, one should be careful when using differential entropy arguments. The next subsection illustrates using examples the various values that differential entropy can take on (in the scalar setting).

1.1 Possible Values of Differential Entropy

Example 1 ($h(\mathsf{x}) \in \mathbb{R}$)**.** Suppose x has a uniform PDF on $[0, \Delta]$ with $\Delta > 0$. Then, the differential entropy of x is:

$$h(\mathsf{x}) = \int_0^\Delta \frac{1}{\Delta} \log(\Delta)\, dx = \log(\Delta)$$

which is positive for $\Delta > 1$, zero for $\Delta = 1$, and negative for $\Delta < 1$. This shows that differential entropy can be any real number.

Example 2 ($h(\mathsf{x}) = +\infty$ [1])**.** Suppose x has PDF:

$$p_{\mathsf{x}}(x) = \begin{cases} \frac{1}{x \log(x)^2} & , \quad x \geq e \\ 0 & , \quad x < e \end{cases}$$

which is non-negative and satisfies:

$$\int_{-\infty}^{+\infty} p_{\mathsf{x}}(x)\,dx = \int_{e}^{+\infty} \frac{1}{x\log(x)^2}\,dx = \int_{1}^{+\infty} \frac{1}{u^2}\,du = 1$$

using the substitution $u = \log(x)$. Then, the differential entropy of x is $+\infty$ because:

$$\begin{aligned} h(\mathsf{x}) &= \int_{e}^{+\infty} \frac{\log(x\log(x)^2)}{x\log(x)^2}\,dx = \int_{e}^{+\infty} \frac{\log(x)}{x\log(x)^2} + \underbrace{\frac{2\log(\log(x))}{x\log(x)^2}}_{\geq 0}\,dx \\ &\geq \int_{e}^{+\infty} \frac{1}{x\log(x)}\,dx = \int_{1}^{+\infty} \frac{1}{u}\,du = +\infty \end{aligned}$$

where we again use the substitution $u = \log(x)$. We note that if $\mathrm{var}(\mathsf{x}) < +\infty$, then $h(\mathsf{x}) < +\infty$.

Example 3 ($h(\mathsf{x}) = -\infty$ [1])**.** Suppose x has PDF whose support is the union of disjoint intervals $\{I_k : k = 2, 3, 4, \ldots\}$ in $\mathbb{R}$ such that the length of I_k is $1/(k\log(k))^2$. Define the constant:

$$C = \sum_{k=2}^{\infty} \frac{1}{k\log(k)^2} < +\infty$$

whose finiteness follows from the integral test: $\displaystyle\int_{2}^{+\infty} \frac{1}{x\log(x)^2}\,dx < +\infty$. The PDF of x is:

$$p_{\mathsf{x}}(x) = \begin{cases} \frac{k}{C} & , \quad x \in I_k \text{ for } k = 2, 3, 4, \ldots \\ 0 & , \quad \text{otherwise} \end{cases}$$

which is non-negative and satisfies:

$$\int_{-\infty}^{+\infty} p_{\mathsf{x}}(x)\,dx = \sum_{k=2}^{\infty} \int_{I_k} \frac{k}{C}\,dx = \sum_{k=2}^{\infty} \frac{k}{C(k\log(k))^2} = \frac{1}{C}\sum_{k=2}^{\infty} \frac{1}{k\log(k)^2} = 1\,.$$

The differential entropy of x is:

$$\begin{aligned} h(\mathsf{x}) &= \sum_{k=2}^{\infty} \int_{I_k} \frac{k}{C}\log\left(\frac{C}{k}\right)\,dx = \frac{1}{C}\sum_{k=2}^{\infty} \frac{\log(C) - \log(k)}{k\log(k)^2} \\ &= \log(C) - \frac{1}{C}\sum_{k=2}^{\infty} \frac{1}{k\log(k)} = -\infty \end{aligned}$$

where the final equality follows from the integral test: $\displaystyle\int_{2}^{+\infty} \frac{1}{x\log(x)}\,dx = +\infty$. We note that if the PDF of x is bounded, then $h(\mathsf{x}) > -\infty$.

Example 4 ($h(\mathsf{x})$ undefined [2])**.** Suppose x has PDF whose support is the union of disjoint intervals $\{I_k : k = 1, 2, 3, \dots\}$ in $\mathbb{R}$ such that the length of I_k is $\frac{C}{k^2}\exp\big(-(-1)^k \cdot k\big)$, where $C = 6/\pi^2$. In particular, the PDF of x is:

$$p_{\mathsf{x}}(x) = \begin{cases} \exp\big((-1)^k k\big) & , \ x \in I_k \text{ for } k = 1, 2, 3, \dots \\ 0 & , \ \text{otherwise} \end{cases}$$

which is non-negative and satisfies:

$$\int_{-\infty}^{+\infty} p_{\mathsf{x}}(x)\,dx = \sum_{k=1}^{\infty} \int_{I_k} \exp\big((-1)^k k\big)\ dx = C \sum_{k=1}^{\infty} \frac{1}{k^2} = 1\,.$$

The differential entropy of x is:

$$\begin{aligned} h(\mathsf{x}) &= -\sum_{k=1}^{\infty} \int_{I_k} \exp\big((-1)^k k\big) \log\big(\exp\big((-1)^k k\big)\big)\ dx \\ &= -\sum_{k=1}^{\infty} \frac{C}{k^2} \exp\big(-(-1)^k k\big) \exp\big((-1)^k k\big)\ (-1)^k k \\ &= C \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \end{aligned}$$

which is a conditionally convergent series. By the Riemann series (rearrangement) theorem, the terms in this series can be permuted to diverge or converge to any chosen value. So, the differential entropy is undefined. This shows that the expectation in Definition 1 indeed may not be well-defined.

1.2 Properties of Differential Entropy

We next present some basic properties of differential entropy (cf. [2]). Note that conditional and joint differential entropies can be defined from Definition 1 analogously to the discrete case.

Theorem 1 (Properties of Differential Entropy)**.** *Suppose* $\mathbf{x} = (\mathsf{x}_1, \dots, \mathsf{x}_n) \in \mathbb{R}^n$ *is a continuous random vector such that all joint, conditional, and marginal PDFs exist, and* $\mathbf{y}$ *is another arbitrary random vector such that the conditional PDFs* $p_{\mathbf{x}|\mathbf{y}}(\cdot|\cdot)$ *exist. Assuming all the differential entropy terms below exist and are finite, we have the following results:*

1. *(Uniform maximizes entropy) If the support of the PDF* $p_{\mathbf{x}}(\cdot)$ *is a bounded set* $\mathcal{X} \subseteq \mathbb{R}^n$ *with* $\mathrm{vol}(\mathcal{X}) > 0$*, then:*

$$h(\mathbf{x}) \leq \log(\mathrm{vol}(\mathcal{X}))$$

with equality if and only if $\mathbf{x}$ *is uniform on* $\mathcal{X}$*, where* $\mathrm{vol}(\cdot)$ *denotes the volume (or more precisely, Lebesgue measure) in* $\mathbb{R}^n$*.*

2. *(Chain rule)*

$$h(\mathbf{x}^n) = \sum_{k=1}^{n} h\big(x_k|\mathbf{x}^{k-1}\big)$$

where $\mathbf{x}^0$ *is a placeholder representing no conditioning, and* $\mathbf{x}^k = (x_1, \dots, x_k)$ *for* $k = 1, \dots, n$ *(so that* $\mathbf{x} = \mathbf{x}^n$*).*

3. *(Conditioning reduces entropy)*

$$h(\mathbf{x}|\mathbf{y}) \leq h(\mathbf{x})$$

with equality if and only if $\mathbf{x}$ *and* $\mathbf{y}$ *are independent.*

4. *(Tensorization)*

$$h(\mathbf{x}) \leq \sum_{k=1}^{n} h(x_k)$$

with equality if and only if $x_1, \dots, x_n$ *are mutually independent.*

Proof.

1. Let $q(\cdot)$ denote the uniform PDF on $\mathcal{X}$, which is well-defined as $\mathcal{X}$ is a bounded set with $\mathrm{vol}(\mathcal{X}) > 0$. Then, we have via Gibbs' inequality:

$$0 \leq D(p_{\mathbf{x}}||q) = \int_{\mathcal{X}} p_{\mathbf{x}}(\mathbf{x}) \log\left(\frac{p_{\mathbf{x}}(\mathbf{x})}{1/\mathrm{vol}(\mathcal{X})}\right) d\mathbf{x} = \log(\mathrm{vol}(\mathcal{X})) - h(\mathbf{x})$$

which proves the result. Note that we have equality in Gibbs' inequality if and only if the input distributions are the same.

2. This easily follows from telescoping as shown below:

$$p_{\mathbf{x}^n}(\mathbf{x}^n) = \prod_{k=1}^{n} p_{x_k|\mathbf{x}^{k-1}}\big(x_k|\mathbf{x}^{k-1}\big)$$

$$-\log(p_{\mathbf{x}^n}(\mathbf{x}^n)) = -\sum_{k=1}^{n} \log\big(p_{x_k|\mathbf{x}^{k-1}}\big(x_k|\mathbf{x}^{k-1}\big)\big)$$

$$h(\mathbf{x}^n) = -\mathbb{E}\left[\log(p_{\mathbf{x}^n}(\mathbf{x}^n))\right] = -\sum_{k=1}^{n} \mathbb{E}\left[\log\big(p_{x_k|\mathbf{x}^{k-1}}\big(x_k|\mathbf{x}^{k-1}\big)\big)\right] = \sum_{k=1}^{n} h\big(x_k|\mathbf{x}^{k-1}\big)\,.$$

3. This can be verified using Gibbs' inequality as mentioned in the lecture notes.

4. This follows from parts 2 and 3. □

2 Applications of Differential Entropy

Differential entropy turns out to have applications in various areas of mathematics. The next two subsections present two seemingly unrelated results from linear algebra and geometry that admit elegant proofs using the aforementioned properties of differential entropy.

2.1 Hadamard's Inequality

In matrix theory, *Hadamard's inequality* is an upper bound on the absolute value of the determinant of a matrix in terms of the Euclidean ℓ^2-norms of its columns. The inequality turns out to be a direct consequence of the tensorization property of differential entropy.

Theorem 2 (Hadamard's Inequality). *For every $n \times n$ real matrix $A \in \mathbb{R}^{n\times n}$ with columns $\{a_k \in \mathbb{R}^n : k = 1, \dots, n\}$, we have:*

$$|\det(A)| \leq \prod_{k=1}^{n} \|a_k\|$$

where $\|\cdot\|$ denotes the Euclidean ℓ^2-norm. Furthermore, if A is full rank, then equality is achieved if and only if the columns $\{a_k \in \mathbb{R}^n : k = 1, \dots, n\}$ are orthogonal.

Proof. If A is not full rank, then $\det(A) = 0$ and the inequality trivially holds. So, we assume without loss of generality that A is full rank. Suppose $\mathsf{x} \sim \mathcal{N}(0, I_n)$ is an n-length i.i.d. standard Gaussian random vector with mean 0 and covariance I_n (the $n \times n$ identity matrix). Define the jointly Gaussian random vector $\mathsf{y} = A^T\mathsf{x} \sim \mathcal{N}(0, A^T A)$, where the mean and covariance can be calculated as follows:

$$\begin{aligned} \mathbb{E}\left[\mathsf{y}\right] &= A^T\,\mathbb{E}\left[\mathsf{x}\right] = 0\,, \\ \mathbb{E}\left[\mathsf{y}\mathsf{y}^T\right] &= A^T\,\mathbb{E}\left[\mathsf{x}\mathsf{x}^T\right] A = A^T A\,. \end{aligned}$$

Note that the random variables $\mathsf{y}_k \sim \mathcal{N}\big(0, \|a_k\|^2\big)$ for $k = 1, \dots, n$ in y are also Gaussian. Using the tensorization property of differential entropy, and the formulae for differential entropies of Gaussian random variables and vectors derived in the lecture notes, we have:

$$\frac{1}{2}\log\big((2\pi e)^n \det\big(A^T A\big)\big) = h(\mathsf{y}) \leq \sum_{k=1}^{n} h(\mathsf{y}_k) = \sum_{k=1}^{n} \frac{1}{2}\log\big(2\pi e\,\|a_k\|^2\big)$$

where upon exponentiating and simplifying, we get the desired inequality:

$$\det\big(A^T A\big) = |\det(A)|^2 \leq \prod_{k=1}^{n} \|a_k\|^2\,.$$

Equality is achieved in this inequality if and only if $\mathsf{y}_1, \dots, \mathsf{y}_n$ are mutually independent. This happens if and only if the off-diagonal entries of $A^T A$ (the covariance matrix of y) are all zero, or equivalently, when the columns of A are all orthogonal. □

Geometrically, this inequality states that the volume of a hyper-parallelepiped (which is given by the absolute value of the determinant) is bounded by the product of all its lengths. Equivalently, it says that the volume of a hyper-parallelepiped with given lengths is maximized when it is actually a hyper-rectangle (or n-orthotope). In information theory, this result turns out to be useful in the analysis of *water-filling* solutions for channels with colored Gaussian noise. A discussion of water-filling is beyond the scope of this course, so we refer readers to [3] for further details.

2.2 Bollobás-Thomason Box Theorem

The *Bollobás-Thomason box theorem* from the geometry and isoperimetry literature is yet another result that follows in a straightforward manner from basic properties of differential entropy. It illustrates that a hyper-rectangle simultaneously minimizes the volumes of all its projections. We follow the exposition in [2].

Theorem 3 (Bollobás-Thomason Box Theorem)**.** *Suppose $K \subseteq \mathbb{R}^n$ is a closed and bounded set. For $S \subseteq [n] \triangleq \{1, \dots, n\}$, let K_S be the projection of K onto the subset S of coordinate axes. Then, there exists a hyper-rectangle $R \subseteq \mathbb{R}^n$ such that $\text{vol}(R) = \text{vol}(K)$, and for every non-empty $S \subseteq [n]$:*

$$\text{vol}(R_S) \leq \text{vol}(K_S)$$

where $\text{vol}(\cdot)$ denotes the volume (or more precisely, Lebesgue measure) in the appropriate $|S|$-dimensional Euclidean space.

Proof. If $\text{vol}(K) = 0$, then we can take a single point as R. So, we assume without loss of generality that $\text{vol}(K) > 0$. Suppose $\mathsf{x} = (\mathsf{x}_1, \dots, \mathsf{x}_n)$ is uniformly distributed on K such that $h(\mathsf{x}) = \log(\text{vol}(K))$. Define the constants $\{r_k > 0 : k = 1, \dots, n\}$ such that:

$$\forall k \in \{1, \dots, n\}, \;\; \log(r_k) = h\big(\mathsf{x}_k|\mathsf{x}^{k-1}\big) \; .$$

Using these constants, we define a hyper-rectangle R with lengths $r_1, \dots, r_n$ such that $\text{vol}(R) = \prod_{k=1}^n r_k$. The chain rule establishes that $\text{vol}(R) = \text{vol}(K)$:

$$\log(\text{vol}(K)) = h(\mathsf{x}) = \sum_{k=1}^{n} h\big(\mathsf{x}_k|\mathsf{x}^{k-1}\big) = \log\left(\prod_{k=1}^{n} r_k\right) = \log(\text{vol}(R)) \, .$$

For a set $A \in [n]$, let x_A denote the random vector $\{\mathsf{x}_k : k \in A\}$ (where conditioning on $\mathsf{x}_\varnothing$ corresponds to no conditioning at all). Fix any non-empty set $S \subseteq [n]$. Then, we have the following sequence of inequalities:

$$\begin{aligned} \log(\text{vol}(K_S)) &\geq h(\mathsf{x}_S) \\ &= \sum_{k=1}^{n} \mathbb{1}_S(k) h\big(\mathsf{x}_k|\mathsf{x}_{[k-1]\cap S}\big) \\ &\geq \sum_{k\in S} h\big(\mathsf{x}_k|\mathsf{x}^{k-1}\big) = \log\left(\prod_{k\in S} r_k\right) \\ &= \log(\text{vol}(R_S)) \end{aligned}$$

where the first line holds because the uniform distribution maximizes differential entropy, the second line follows from the chain rule, and the third line holds because conditioning reduces entropy. Hence, $\mathrm{vol}(R_S) \leq \mathrm{vol}(K_S)$ for every non-empty set $S \subseteq [n]$. This completes the proof. □

3 Relation to Fisher Information

Finally, we present an intriguing relation between differential entropy and Fisher information (in the scalar setting) based on [4]. Before we present this relation, we have to define the notion of Fisher information for a single PDF. Given a random variable x with PDF $p_{\mathsf{x}}(\cdot)$ that has support $\mathbb{R}$, we can define a parametrized family of PDFs $\{p_{\mathsf{x}}(\cdot;\phi) : \phi \in \mathbb{R}\}$ such that $p_{\mathsf{x}}(x;\phi) = p_{\mathsf{x}}(x - \phi)$. The Fisher information of x is given by the Fisher information of this translation family:

$$J(\mathsf{x}) \triangleq J_{\mathsf{x}}(\phi) = \mathrm{var}\left(\frac{p'_{\mathsf{x}}(\mathsf{x})}{p_{\mathsf{x}}(\mathsf{x})}\right) = \mathbb{E}\left[\frac{p'_{\mathsf{x}}(\mathsf{x})^2}{p_{\mathsf{x}}(\mathsf{x})^2}\right] \tag{1}$$

where $p'_{\mathsf{x}}(x) = \frac{\partial}{\partial x} p_{\mathsf{x}}(x)$, and we assume sufficient regularity conditions (as in the lecture notes) so that (1) is well-defined. The next subsection derives the well-known *de Bruijn's identity.*

3.1 De Bruijn's Identity

Theorem 4 (De Bruijn's Identity)**.** *Given independent random variables x and z such that $J(\mathsf{x})$ exists and* $\mathrm{var}(\mathsf{z}) < +\infty$*, we have:*

$$\left.\frac{d}{dt} h(\mathsf{x} + \sqrt{t}\mathsf{z})\right|_{t=0} = \frac{1}{2}\,\mathrm{var}(\mathsf{z}) J(\mathsf{x})\,.$$

Proof. We follow the proof in [4], but neglect all issues of rigor (to rigorously prove this result, one needs to apply the dominated convergence theorem at appropriate places). Letting $\theta = \sqrt{t}$ and $\mathsf{y} = \mathsf{x} + \theta\mathsf{z}$, since the left-hand side satisfies:

$$\left.\frac{d}{dt} h(\mathsf{x} + \sqrt{t}\mathsf{z})\right|_{t=0} = \lim_{\theta \to 0} \frac{h(\mathsf{x} + \theta\mathsf{z}) - h(\mathsf{x})}{\theta^2}$$

by definition of derivative, it suffices to prove that:

$$I(\mathsf{y};\mathsf{z}) = h(\mathsf{x} + \theta\mathsf{z}) - h(\mathsf{x}) = \frac{1}{2}\theta^2\,\mathrm{var}(\mathsf{z}) J(\mathsf{x}) + o\big(\theta^2\big) \tag{2}$$

where $I(\mathsf{y};\mathsf{z}) = h(\mathsf{x} + \theta\mathsf{z}) - h(\mathsf{x} + \theta\mathsf{z}|\mathsf{z}) = h(\mathsf{x} + \theta\mathsf{z}) - h(\mathsf{x})$ by the translation invariance of differential entropy (check this!), and $o(\theta^2)$ denotes a function satisfying $\lim_{\theta\to 0} o(\theta^2)/\theta^2 = 0$. Now observe that:

$$I(\mathsf{y};\mathsf{z}) = \mathbb{E}_{p_{\mathsf{z}}}\big[D(p_{\mathsf{y}|\mathsf{z}}||p_{\mathsf{y}})\big]\ .$$

So, for every fixed $\mathsf{z} = z$, we must compute the KL divergence $D(p_{\mathsf{y}|\mathsf{z}=z}||p_{\mathsf{y}})$. We construct a parametric family of PDFs $\{p_{\mathsf{u}}(\cdot;\theta) : \theta \in \mathbb{R}\}$ such that $D(p_{\mathsf{y}|\mathsf{z}=z}||p_{\mathsf{y}}) = D(p_{\mathsf{u}}(\cdot;0)||p_{\mathsf{u}}(\cdot;\theta))$. Suppose $p_{\mathsf{u}}(u;\theta) = p_{\mathsf{y}}(y) = p_{\mathsf{x}+\theta\mathsf{z}}(y)$ and $p_{\mathsf{u}}(u;0) = p_{\mathsf{y}|\mathsf{z}}(y|z) = p_{\mathsf{x}}(y-\theta z)$, then we must also let $u = y - \theta z$ for consistency (note that z is fixed, and u and y vary). Hence, we define:

$$p_{\mathsf{u}}(u;\theta) \triangleq p_{\mathsf{x}+\theta\mathsf{z}}(u+\theta z) = p_{\mathsf{y}}(y)$$

and plugging in $\theta = 0$ produces $p_{\mathsf{u}}(u;0) = p_{\mathsf{y}|\mathsf{z}}(y|z) = p_{\mathsf{x}}(u)$. Since we have merely translated the PDFs $p_{\mathsf{y}|\mathsf{z}}(\cdot|z)$ and $p_{\mathsf{y}}(\cdot)$ to produce $p_{\mathsf{u}}(\cdot;0)$ and $p_{\mathsf{u}}(\cdot;\theta)$ respectively, we have $D(p_{\mathsf{y}|\mathsf{z}=z}||p_{\mathsf{y}}) = D(p_{\mathsf{u}}(\cdot;0)||p_{\mathsf{u}}(\cdot;\theta))$. This means that:

$$I(\mathsf{y};\mathsf{z}) = \mathbb{E}_{p_{\mathsf{z}}}[D(p_{\mathsf{u}}(\cdot;0)||p_{\mathsf{u}}(\cdot;\theta))] \ .$$

Next, recall from the problem sets that (for any fixed $\mathsf{z} = z$):

$$D(p_{\mathsf{u}}(\cdot;0)||p_{\mathsf{u}}(\cdot;\theta)) = \frac{1}{2}\theta^2 J_{\mathsf{u}}(0) + o\big(\theta^2\big)$$

where $J_{\mathsf{u}}(0)$ is the Fisher information that u carries about θ when $\theta = 0$. To compute this quantity, notice that for fixed $\mathsf{y} = y$:

$$\begin{aligned}
p_{\mathsf{u}}(u;0) &= p_{\mathsf{y}|\mathsf{z}}(y|z) = p_{\mathsf{x}}(y-\theta z) = p_{\mathsf{x}}(y) - \theta z p_{\mathsf{x}}'(y) + o(\theta) \\
\Rightarrow \quad p_{\mathsf{u}}(u;\theta) &= p_{\mathsf{y}}(y) = \mathbb{E}_{p_{\mathsf{z}}}\left[p_{\mathsf{y}|\mathsf{z}}(y|\mathsf{z})\right] = p_{\mathsf{x}}(y) - \theta\,\mathbb{E}\left[\mathsf{z}\right] p_{\mathsf{x}}'(y) + o(\theta) \\
\Rightarrow \quad p_{\mathsf{u}}(u;\theta) &= p_{\mathsf{u}}(u;0) + \theta(z - \mathbb{E}\left[\mathsf{z}\right])p_{\mathsf{x}}'(u+\theta z) + o(\theta) \\
\Rightarrow \quad \frac{d}{d\theta}p_{\mathsf{u}}(u;\theta)\bigg|_{\theta=0} &= \lim_{\theta\to 0}\frac{p_{\mathsf{u}}(u;\theta) - p_{\mathsf{u}}(u;0)}{\theta} = (z - \mathbb{E}\left[\mathsf{z}\right])p_{\mathsf{x}}'(u)
\end{aligned}$$

where the first equation follows from Taylor's theorem, the second equation follows from taking expectations with respect to z, and the third equation is obtained by subtracting the first equation from the second. Thus, we get for any fixed $\mathsf{z} = z$:

$$J_{\mathsf{u}}(0) = \mathbb{E}_{p_{\mathsf{u}}(\cdot;0)}\!\left[\frac{(z - \mathbb{E}\left[\mathsf{z}\right])^2 p_{\mathsf{x}}'(\mathsf{u})^2}{p_{\mathsf{u}}(\mathsf{u};0)^2}\right] = (z - \mathbb{E}\left[\mathsf{z}\right])^2\,\mathbb{E}_{p_{\mathsf{x}}}\!\left[\left(\frac{p_{\mathsf{x}}'(\mathsf{u})}{p_{\mathsf{x}}(\mathsf{u})}\right)^2\right] = (z - \mathbb{E}\left[\mathsf{z}\right])^2 J(\mathsf{x})$$

which implies that:

$$\begin{aligned}
D(p_{\mathsf{u}}(\cdot;0)||p_{\mathsf{u}}(\cdot;\theta)) &= \frac{1}{2}\theta^2(z - \mathbb{E}\left[\mathsf{z}\right])^2 J(\mathsf{x}) + o\big(\theta^2\big) \\
\Rightarrow \quad I(\mathsf{y};\mathsf{z}) &= \frac{1}{2}\theta^2\operatorname{var}(\mathsf{z})J(\mathsf{x}) + o\big(\theta^2\big)
\end{aligned}$$

where the second equation follows from taking expectations with respect to z. This completes the proof. □

Intuitively, de Bruijn's identity portrays that the "sensitivity" of a random variable to some independent and additive noise is given by the Fisher information of the random variable. To further elaborate on this, we interpret x as a sender's signal, z as channel noise with $\text{var}(z) = 1$, θ as the noise standard deviation (or amplification), and y as the received signal. Then, de Bruijn's identity or its equivalent version in (2) states that the mutual information between the received signal and the noise is locally quadratic as a function of θ. The *curvature* of this quadratic function is given by the Fisher information $J(x)$. This means that larger values of $J(x)$ make the received signal more dependent on the noise for a given value of θ. Therefore, a higher Fisher information implies greater sensitivity to noise. This offers another interpretation for Fisher information.

One popular application of de Bruijn's identity is to prove the so called *entropy power inequality.* It is also used extensively in the information theory literature on (Bayesian and non-Bayesian) Cramér-Rao bounds and *uncertainty principles.* In closing, we remark that the result from the problem set relating mutual information (between the input and output of an additive Gaussian noise channel) and *minimum mean-square error* as functions of *signal-to-noise ratio* (originally derived in [5]) is closely related to de Bruijn's identity as shown in [4].

References

[1] R. B. Ash, *Information Theory*, ser. Interscience Tracts in Pure and Applied Mathematics. New York, NY, USA: John Wiley & Sons, Inc., 1965, no. 19.

[2] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," May 2016, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, Lecture Notes 6.441.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[4] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 33–55, January 2011.

[5] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2017

Recitation 10: Mixing of Markov Chains

Date: Friday, April 28, 2017 **TA:** Anuran Makur

1 Introduction

Markov Chain Monte Carlo (MCMC) methods address the problem of sampling from a given distribution by first constructing a Markov chain whose stationary distribution is the given distribution, and then sampling from this Markov chain. Since there are broad classes of Markov chains for which the distribution over states converges to the stationary distribution, MCMC methods such as the Metropolis-Hastings algorithm eventually produce samples from the desired distribution (or some distribution "close" to it). A natural question that arises here is:

How long do we have to run a Markov chain before its distribution over states is "close" to the stationary distribution?

In the Markov chain literature, the length of time a Markov chain must run until its distribution is close to the stationary distribution is known as its *mixing time*. As mentioned in the lecture notes, mixing times determine the *burn-in period* of MCMC algorithms (i.e. they determine how many samples must be discarded before useful samples are produced). In the sequel, we will try to answer this question.

1.1 Basics of Markov Chains

We begin by recalling some basic definitions of Markov chains.

Definition 1 (Markov Chain)**.** *A* Markov chain *is a discrete-time stochastic process* $\{\mathsf{x}_n : n \geq 0\}$ *with each random variable taking values in a countable* state space $\mathcal{X}$, *that satisfies the* Markov property*:*

$$\mathbb{P}\left(\mathsf{x}_n = x_n | \mathsf{x}_{n-1} = x_{n-1}, \dots, \mathsf{x}_0 = x_0\right) = \mathbb{P}\left(\mathsf{x}_n = x_n | \mathsf{x}_{n-1} = x_{n-1}\right)$$

for every $n \geq 1$ *and* $x_0, \dots, x_n \in \mathcal{X}$ *such that* $\mathbb{P}\left(\mathsf{x}_0 = x_0, \dots, \mathsf{x}_{n-1} = x_{n-1}\right) > 0$. *We say* $\{\mathsf{x}_n : n \geq 0\}$ *is a* finite state Markov chain *if* $\mathcal{X}$ *is a finite set, and we say it is* time-homogeneous *if for every* $n \geq 1$ *and every* $x, y \in \mathcal{X}$*:*

$$\mathbb{P}\left(\mathsf{x}_n = y | \mathsf{x}_{n-1} = x\right) = \mathbb{P}\left(\mathsf{x}_1 = y | \mathsf{x}_0 = x\right) \, .$$

We will only consider time-homogeneous finite state Markov chains in our discussion. So, we will refer to a time-homogeneous finite state Markov chain as an "MC" from hereon. Without loss of generality, let $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$ with $|\mathcal{X}| \geq 2$, and let $\mathcal{P}$ denote the simplex of all probability distributions on $\mathcal{X}$. We will assume that all distributions in $\mathcal{P}$ are row vectors, i.e. each $\mu \in \mathcal{P}$ can be represented as:

$$\mu = [\mu(1) \;\; \mu(2) \; \cdots \; \mu(|\mathcal{X}|)] \,. \tag{1}$$

Observe that given the initial state, the distribution of an MC can be succinctly described by its one-step transition probabilities:

$$\forall x, y \in \mathcal{X}, \;\; W(x,y) \triangleq \mathbb{P}\left(\mathsf{x}_1 = y | \mathsf{x}_0 = x\right) \tag{2}$$

which we usually stack into an $|\mathcal{X}| \times |\mathcal{X}|$ *stochastic matrix* W whose (x, y)th element is $W(x, y)$ for all $x, y \in \mathcal{X}$ (as shown in the lecture notes). W is an entry-wise non-negative matrix whose rows sum to 1. In particular, we will denote the xth row of W as $W(x, \cdot) \in \mathcal{P}$ (which is the conditional distribution of the next state given the current state is $x \in \mathcal{X}$).[1]

It is straightforward to verify that for every $n \geq 1$ and every $x, y \in \mathcal{X}$:

$$W^n(x,y) = \mathbb{P}\left(\mathsf{x}_n = y | \mathsf{x}_0 = x\right) \tag{3}$$

which shows that $W^n(x, \cdot) \in \mathcal{P}$ is the conditional distribution of the nth state given the initial state is $x \in \mathcal{X}$ (Chapman-Kolmogorov equation). Moreover, if the initial distribution of the MC is $p_{\mathsf{x}_0} \in \mathcal{P}$, then the distribution of x_n for every $n \geq 1$ can be obtained by:

$$p_{\mathsf{x}_n} = p_{\mathsf{x}_0} W^n \,. \tag{4}$$

Typically, we study properties of MCs that only depend on the transition probabilities. As a result, we usually do not specify an initial distribution, and represent an MC with its stochastic transition probability matrix W.

We next present some more definitions relevant to our discussion.

Definition 2 (Irreducibility and Aperiodicity)**.** *An MC with stochastic transition probability matrix W is called* irreducible *if for every pair of states $x, y \in \mathcal{X}$, there exists some $n \geq 0$ such that $W^n(x, y) > 0$. If W is irreducible, then we say it is* aperiodic *if every state $x \in \mathcal{X}$ has* period $d_x \triangleq \gcd\{n \geq 1 : W^n(x, x) > 0\} = 1$.

These definitions turn out to be the precise conditions under which we observe the behavior of MCs converging to their stationary distributions over time. Intuitively, an MC is irreducible if it is possible to get to any state from any other state after a finite sequence of transitions. This condition allows probabilities to "flow" to different states even if all the probability is initially concentrated at a particular state. However,

[1] Note that we use slightly different notation from the lecture notes here for ease of exposition. For example, we do not use boldface for matrices and vectors.

irreducibility does not preclude the MC with $\mathcal{X} = \{0, \ldots, |\mathcal{X}| - 1\}$ where transitions happen from state $x \in \mathcal{X}$ to state $x + 1 \,(\text{mod}\ |\mathcal{X}|)$ with probability 1. This MC is irreducible, and each of its states has period $|\mathcal{X}|$. If we start this chain at state 0 (i.e. $\mathbb{P}\left(\mathsf{x}_0 = 0\right) = 1$), we cannot hope for this probability to "diffuse" to all states over time, because the probability mass will periodically cycle over all the states. The condition of aperiodicity is needed to preclude such chains. As we will see, irreducibility and aperiodicity together, allow us to prove convergence to stationary distributions over time. We remark that although Definition 2 requires us to check that $d_x = 1$ for every $x \in \mathcal{X}$ to deduce aperiodicity, it suffices to only check this for one state. Indeed, it is a simple exercise to prove that $d_x = d_y$ for any two states $x, y \in \mathcal{X}$ of an irreducible MC (try it!).

We now state some well-known results about MCs. The proofs are omitted since many readers are probably familiar with these results.

Theorem 1 (Properties of Markov Chains)**.** *Let W be the stochastic transition probability matrix of an MC. Then, the following are true:*

1. *There exists a* stationary distribution $\pi \in \mathcal{P}$ *such that* $\pi W = \pi$.
2. *If W is irreducible, then the stationary distribution π is unique and entry-wise strictly positive.*
3. *If W is irreducible and aperiodic, then there exists some $n \geq 0$ such that $W^n(x, y) > 0$ for all $x, y \in \mathcal{X}$ (i.e. W^n is entry-wise strictly positive).*

We remark that the first result is actually an immediate consequence of *Brouwer's fixed-point theorem.* It also admits a short proof using linear programming duality, but a probabilistic proof requires some work. For those familiar with matrix theory, we also remark that the third result simply says that irreducible and aperiodic stochastic matrices are *primitive* matrices.

2 Total Variation Distance

In order to prove convergence to stationary distributions, we require a notion of distance between distributions. The classical choice for this is the so called total variation distance (which you were introduced to in the problem sets).

Definition 3 (Total Variation Distance)**.** *Given two distributions $\mu, \nu \in \mathcal{P}$, we define the* total variation distance *between them as:*

$$\|\mu - \nu\|_{\mathsf{TV}} \triangleq \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|$$

where $\mu(A) = \sum_{x \in A} \mu(x)$ for any event $A \subseteq \mathcal{X}$.

This definition perceives distributions (or probability measures) as maps from the set of all events to $[0, 1]$, and measures the maximum deviation between the two distributions over all events. We will prove a few more equivalent characterizations of total variation distance. To this end, we first introduce the notion of couplings.

2.1 Coupling

Coupling is a powerful proof technique in probability theory, and we will see some uses of it in the ensuing sections. For now, we formally define it below.

Definition 4 (Coupling)**.** *Given two probability distributions $\mu, \nu \in \mathcal{P}$, a coupling between them corresponds to a pair of random variables (x, y) (defined on the same probability space) with joint distribution $p_{\mathsf{x},\mathsf{y}}$ on $\mathcal{X} \times \mathcal{X}$ whose marginal distributions satisfy $p_{\mathsf{x}} = \mu$ and $p_{\mathsf{y}} = \nu$.*

There are several possible couplings between any two $\mu, \nu \in \mathcal{P}$. For example, we can always define the independent coupling where x and y are independent random variables with $p_{\mathsf{x},\mathsf{y}}(x, y) = \mu(x)\nu(y)$ for every $x, y \in \mathcal{X}$. This is typically not a very useful coupling. If $\mu = \nu$, then we can also define the "identical" coupling with $\mathsf{x} = \mathsf{y}$ and $p_{\mathsf{x}} = \mu$. As we will see next, a particular mixture of these two couplings is closely related to the total variation distance.

2.2 Equivalent Characterizations of Total Variation Distance

The next result presents some equivalent characterizations of total variation distance.

Theorem 2 (Characterizations of Total Variation Distance)**.** *For any two probability distributions $\mu, \nu \in \mathcal{P}$, we have:*

$$\begin{aligned} \|\mu - \nu\|_{\mathsf{TV}} &= \sum_{x \in \mathcal{X}:\, \mu(x) \geq \nu(x)} \mu(x) - \nu(x) \\ &= \frac{1}{2} \|\mu - \nu\|_1 \\ &= \min \{\mathbb{P}(\mathsf{x} \neq \mathsf{y}) : (\mathsf{x}, \mathsf{y}) \text{ coupling of } \mu \text{ and } \nu\} \end{aligned}$$

where the first equality illustrates that the event $S \triangleq \{x \in \mathcal{X} : \mu(x) \geq \nu(x)\}$ achieves the maximum in the definition of total variation distance, the second equality is the ℓ^1-norm characterization (recall that ℓ^1-norm is defined as $\|x\|_1 \triangleq \sum_{i=1}^n |x_i|$ for any $x \in \mathbb{R}^n$), and the final equality is the optimal coupling representation.

Proof. To prove the first characterization, let $S \triangleq \{x \in \mathcal{X} : \mu(x) \geq \nu(x)\}$. Then, for any event $A \subseteq \mathcal{X}$, we have:

$$\mu(A) - \nu(A) \leq \mu(S) - \nu(S)$$

because $x \in A \backslash S \Rightarrow \mu(x) - \nu(x) < 0$ and $x \in S \backslash A \Rightarrow \mu(x) - \nu(x) \geq 0$. Likewise, we also get $\nu(A) - \mu(A) \leq \nu(S^c) - \mu(S^c) = \mu(S) - \nu(S)$, which implies that:

$$|\mu(A) - \nu(A)| \leq \mu(S) - \nu(S) \, .$$

We can maximize over all $A \subseteq \mathcal{X}$ on the left hand side and obtain $\|\mu - \nu\|_{\mathsf{TV}} = \mu(S) - \nu(S)$ (where equality is achieved by $A = S$). Since we have:

$$\mu(S) - \nu(S) = \sum_{x \in \mathcal{X} :\, \mu(x) \geq \nu(x)} \mu(x) - \nu(x)$$

this proves the first characterization.

To prove the ℓ^1-norm characterization, notice that $\|\mu - \nu\|_{\mathsf{TV}} = \nu(S^c) - \mu(S^c) = \mu(S) - \nu(S)$ also gives us:

$$\|\mu - \nu\|_{\mathsf{TV}} = \frac{1}{2}\left(\mu(S) - \nu(S) + \nu(S^c) - \mu(S^c)\right) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)| = \frac{1}{2} \|\mu - \nu\|_1 \, .$$

To prove the optimal coupling representation, observe that for any coupling (x, y) of μ and ν where $\mu = p_{\mathsf{x}}$ and $\nu = p_{\mathsf{y}}$, and any event $A \subseteq \mathcal{X}$, we have:

$$\begin{aligned}
\mu(A) - \nu(A) &= \mathbb{P}(\mathsf{x} \in A) - \mathbb{P}(\mathsf{y} \in A) \\
&= \mathbb{P}(\mathsf{x} \in A, \mathsf{y} \in A) + \mathbb{P}(\mathsf{x} \in A, \mathsf{y} \notin A) - \mathbb{P}(\mathsf{y} \in A) \\
&\leq \mathbb{P}(\mathsf{x} \in A, \mathsf{y} \in A) + \mathbb{P}(\mathsf{x} \in A, \mathsf{y} \notin A) - \mathbb{P}(\mathsf{x} \in A, \mathsf{y} \in A) \\
&= \mathbb{P}(\mathsf{x} \in A, \mathsf{y} \notin A) \\
&\leq \mathbb{P}(\mathsf{x} \neq \mathsf{y}) \, .
\end{aligned}$$

Likewise, $\nu(A) - \mu(A) \leq \mathbb{P}(\mathsf{x} \neq \mathsf{y})$, and hence, $|\mu(A) - \nu(A)| \leq \mathbb{P}(\mathsf{x} \neq \mathsf{y})$, which implies that:

$$\|\mu - \nu\|_{\mathsf{TV}} \leq \min\left\{\mathbb{P}(\mathsf{x} \neq \mathsf{y}) : (\mathsf{x}, \mathsf{y}) \text{ coupling of } \mu \text{ and } \nu\right\} \, .$$

So, it suffices to construct a particular joint distribution $p_{\mathsf{x},\mathsf{y}}$ that achieves equality here by making x equal to y as much as possible. Let $a \wedge b \triangleq \min\{a, b\}$ for any $a, b \in \mathbb{R}$. Notice that:

$$\|\mu - \nu\|_{\mathsf{TV}} + \sum_{x \in \mathcal{X}} \mu(x) \wedge \nu(x) = \sum_{x \in S} \mu(x) - \nu(x) \;\; + \;\; \sum_{x \in \mathcal{X}} \mu(x) \wedge \nu(x) = \sum_{x \in \mathcal{X}} \mu(x) = 1$$

which means we can define $\delta \triangleq 1 - \|\mu - \nu\|_{\mathsf{TV}} = \sum_{x \in \mathcal{X}} \mu(x) \wedge \nu(x)$. Furthermore, define the probability distributions (check this!):

$$\begin{aligned}
\forall x \in \mathcal{X}, \;\; p_1(x) &= \frac{\mu(x) \wedge \nu(x)}{\delta} \, , \\
\forall x \in \mathcal{X}, \;\; p_2(x) &= \frac{\mu(x) - \nu(x)}{1 - \delta} \, \mathbb{1}_S(x) \, , \\
\forall x \in \mathcal{X}, \;\; p_3(x) &= \frac{\nu(x) - \mu(x)}{1 - \delta} \, \mathbb{1}_{S^c}(x) \, .
\end{aligned}$$

Let Z be a Bernoulli random variable with $\mathbb{P}(Z=1) = 1 - \mathbb{P}(Z=0) = \delta$. Given $Z = 1$, let (x, y) be coupled identically with conditional distribution:

$$\forall x, y \in \mathcal{X}, \;\; p_{\mathsf{x},\mathsf{y}|\mathsf{z}}(x, y|1) = p_1(x)\mathbb{1}_{x=y}$$

so that $\mathsf{x} = \mathsf{y}$. Given $Z = 0$, let (x, y) be coupled independently with conditional distribution:

$$\forall x, y \in \mathcal{X}, \;\; p_{\mathsf{x},\mathsf{y}|\mathsf{z}}(x, y|0) = p_2(x)p_3(y)$$

where $\mathsf{x} \neq \mathsf{y}$ as p_2 and p_3 have disjoint supports. It is straightforward to verify that $p_{\mathsf{x}} = \delta p_1 + (1-\delta)p_2 = \mu$ and $p_{\mathsf{y}} = \delta p_1 + (1-\delta)p_3 = \nu$, which means $p_{\mathsf{x},\mathsf{y}}$ is a valid coupling with $\mathbb{P}(\mathsf{x} \neq \mathsf{y}) = 1 - \delta = \|\mu - \nu\|_{\mathsf{TV}}$. This completes the proof. □

In the first characterization, the event S has a useful interpretation in terms of binary hypothesis testing as explored in the problem sets. Indeed, if μ and ν are likelihoods corresponding to two equiprobable hypotheses, then S is precisely the decision region where the maximum likelihood (ML) decision rule chooses μ, and the total probability of error P_{e} of the ML decision rule is given by:

$$P_{\mathsf{e}} = \frac{1}{2}\mu(S^c) + \frac{1}{2}\nu(S) = \frac{1}{2}(1 - (\mu(S) - \nu(S))) = \frac{1}{2}(1 - \|\mu - \nu\|_{\mathsf{TV}}) \,. \tag{5}$$

The ℓ^1-norm characterization of total variation distance illustrates that it is a valid distance (or metric) between distributions that is symmetric and satisfies the triangle inequality. To interpret the third characterization, recall from subsection 2.1 that if $\mu = \nu$, then we can define a coupling with $\mathsf{x} = \mathsf{y}$. The optimal coupling representation portrays that the closest a coupling can get to having x identical to y is the coupling corresponding to total variation distance. Furthermore, couplings that achieve total variation distance and maximize $\mathbb{P}(\mathsf{x} = \mathsf{y})$ are known as *maximal couplings*. Finally, for those familiar with the Monge-Kantorovich problem from transportation theory, we remark that the optimal coupling representation of total variation distance shows that it is a *Wasserstein distance* of order 1 with respect to the Hamming metric.

3 Convergence and Ergodic Theorems

In this section, we present two fundamental results from the basic theory of MCs. The first is an analog of the *strong law of large numbers* (SLLN) for irreducible MCs.

Theorem 3 (Ergodic Theorem). *Given a function $f : \mathcal{X} \to \mathbb{R}$ and an irreducible MC $\{\mathsf{x}_n : n \geq 0\}$ with stationary distribution $\pi \in \mathcal{P}$, for any initial distribution $p_{\mathsf{x}_0} \in \mathcal{P}$, we have:*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(\mathsf{x}_k) = \mathbb{E}_\pi[f(\mathsf{x})]\right) = 1$$

where $\mathbb{E}_\pi[f(\mathsf{x})] = \sum_{x\in\mathcal{X}} \pi(x)f(x)$.

One way to prove this result is to segment the MC into blocks using carefully chosen *stopping times* and then employing the SLLN. We omit this proof since we do not assume a thorough understanding of such topics. However, it is worth mentioning certain special cases of this result. If the irreducible MC is actually an i.i.d. process (which means the stochastic transition probability matrix W has all rows equal to π), then Theorem 3 reduces to the SLLN for the sequence of i.i.d. random variables $\{f(\mathsf{x}_n) : n \geq 1\}$. If the function $f(y) = \mathbb{1}_{y=x}$ for some $x \in \mathcal{X}$, then Theorem 3 states that:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{\mathsf{x}_k = x} = \pi(x)\right) = 1\,. \tag{6}$$

This illustrates that an irreducible MC asymptotically spends roughly $\pi(x)$ fraction of its time in state $x \in \mathcal{X}$.

Even when (6) holds, we may not have convergence to the stationary distribution due to periodicity of the MC under consideration. So, we will present a second result that guarantees convergence to stationary distributions for irreducible and aperiodic MCs. Before stating the result, we formalize the "distance from stationarity" using total variation distance. For an irreducible MC with stochastic transition probability matrix W and stationary distribution $\pi \in \mathcal{P}$, we define:

$$\forall n \geq 0,\;\; d(n) \triangleq \max_{x\in\mathcal{X}} \left\|W^n(x,\cdot) - \pi\right\|_{\mathsf{TV}} = \max_{\mu\in\mathcal{P}} \left\|\mu W^n - \pi\right\|_{\mathsf{TV}} \tag{7}$$

which represents the "distance from stationarity" at time n. The maximum in the rightmost extremal problem in (7) can indeed be achieved due to the extreme value theorem. To prove the second equality in (7), notice that $\max_{x\in\mathcal{X}} \|W^n(x,\cdot) - \pi\|_{\mathsf{TV}} \leq \max_{\mu\in\mathcal{P}} \|\mu W^n - \pi\|_{\mathsf{TV}}$ is clearly true, and:

$$\max_{\mu\in\mathcal{P}} \left\|\mu W^n - \pi\right\|_{\mathsf{TV}} \leq \max_{\mu\in\mathcal{P}} \sum_{x\in\mathcal{X}} \mu(x) \left\|W^n(x,\cdot) - \pi\right\|_{\mathsf{TV}} \leq \max_{x\in\mathcal{X}} \left\|W^n(x,\cdot) - \pi\right\|_{\mathsf{TV}}$$

where the first inequality follows from the triangle inequality. We also note that $d(n)$ is non-increasing in n (you can try to prove this by establishing a *data processing inequality* for total variation distance). The next result presents the convergence theorem for irreducible and aperiodic MCs.

Theorem 4 (Convergence Theorem)**.** *Given an irreducible and aperiodic MC with stochastic transition probability matrix W and stationary distribution $\pi \in \mathcal{P}$, there exist constants $\lambda \in (0,1)$ and $C > 0$ such that:*

$$\forall n \geq 0,\;\; d(n) \leq C\lambda^n\,.$$

Proof. First observe from part 3 of Theorem 1 that since W is irreducible and aperiodic, there exists some $m \geq 0$ such that $P = W^m$ is entry-wise strictly positive. This means that there exists some $\delta \in (0,1)$ such that P satisfies the *Doeblin minorization* condition:

$$\forall x, y \in \mathcal{X}, \;\; P(x,y) \geq (1-\delta)\pi(y) \, .$$

Let $\mathbf{1}$ denote the $|\mathcal{X}| \times 1$ column vector with all entries equal to 1. Then, we can decompose the MC P into a mixture of independent sampling from π and another MC Q:

$$P = (1-\delta)\mathbf{1}\pi + \delta Q \tag{8}$$

where $\mathbf{1}\pi$ is a unit rank stochastic matrix with all rows equal to π, and $Q \triangleq \frac{1}{\delta}(P - (1-\delta)\mathbf{1}\pi)$ is a valid stochastic matrix due to the Doeblin minorization condition. We claim that in fact:

$$\forall n \geq 1, \;\; P^n = (1-\delta^n)\mathbf{1}\pi + \delta^n Q^n \, . \tag{9}$$

For $n = 1$, this is simply (8). Suppose (9) holds for some $n = k \geq 1$: $P^k = (1-\delta^k)\mathbf{1}\pi + \delta^k Q^k$. Then, using this and (8) we have:

$$\begin{aligned}
P^{k+1} &= P^k P = ((1-\delta^k)\mathbf{1}\pi + \delta^k Q^k)((1-\delta)\mathbf{1}\pi + \delta Q) \\
&= (1-\delta)(1-\delta^k)\mathbf{1}\pi\mathbf{1}\pi + \delta(1-\delta^k)\mathbf{1}\pi Q + (1-\delta)\delta^k Q^k \mathbf{1}\pi + \delta^{k+1}Q^{k+1} \\
&= ((1-\delta)(1-\delta^k) + \delta(1-\delta^k) + (1-\delta)\delta^k)\mathbf{1}\pi + \delta^{k+1}Q^{k+1} \\
&= (1-\delta^{k+1})\mathbf{1}\pi + \delta^{k+1}Q^{k+1}
\end{aligned}$$

where the third line holds because $\pi\mathbf{1} = 1$ (π is a distribution that sums to 1), $\pi Q = \frac{1}{\delta}(\pi P - (1-\delta)\pi\mathbf{1}\pi) = \pi$, and $Q^k\mathbf{1} = \mathbf{1}$ (rows of a stochastic matrix sum to 1). By induction, this implies that (9) is true for all $n \geq 1$.

Now observe from (9) that:

$$\begin{aligned}
\forall n \geq 0, \;\; P^n - \mathbf{1}\pi &= \delta^n(Q^n - \mathbf{1}\pi) \\
\forall n \geq 0, 0 \leq r < m, \;\; W^{mn+r} - \mathbf{1}\pi &= \delta^n(Q^n W^r - \mathbf{1}\pi) \\
\forall n \geq 0, 0 \leq r < m, \;\; \max_{x \in \mathcal{X}} \left\| W^{mn+r}(x,\cdot) - \pi \right\|_{\mathsf{TV}} &= \delta^n \max_{x \in \mathcal{X}} \left\| Q^n W^r(x,\cdot) - \pi \right\|_{\mathsf{TV}} \\
\forall k \geq 0, \;\; \max_{x \in \mathcal{X}} \left\| W^k(x,\cdot) - \pi \right\|_{\mathsf{TV}} &\leq \delta^{\lfloor k/m \rfloor} \\
\forall k \geq 0, \;\; d(k) &\leq \frac{1}{\delta}(\delta^{1/m})^k
\end{aligned}$$

where the second line follows from substituting $P = W^m$ and then multiplying both sides by W^r, the third line follows from equating the ℓ^1-norms of the rows on both sides and using the ℓ^1-norm characterization of total variation distance in Theorem 2, and the fourth line holds because total variation distance is always bounded by 1 and $mn + r$ (for fixed m and varying $n \geq 0$, $0 \leq r < m$) runs over all $k \geq 0$. This completes the proof. □

We remark that another well-known method of proving this result uses coupling ideas, but we omit this alternative proof for brevity. An immediate corollary of this result is that for an irreducible and aperiodic MC W:

$$\forall x, y \in \mathcal{X}, \ \lim_{n \to \infty} W^n(x, y) = \pi(y) \tag{10}$$

which conveys that the MC converges to its stationary distribution regardless of its initial distribution. Moreover, the convergence theorem shows that irreducible and aperiodic MCs converge exponentially fast to their stationary distributions. Unfortunately, it does not provide explicit estimates of the constants C and λ. So, it does not directly address the question of how many time steps we need to wait in order to guarantee we are close to the stationary distribution in total variation distance (i.e. it does not give us explicit bounds on mixing times). It is worth mentioning that for *reversible* MCs, the asymptotic rate of convergence to stationarity can be easily shown (via the Perron-Frobenius theorem) to be the *second largest eigenvalue modulus* (SLEM) of W. However, since C is still unknown and could be very large, such SLEM estimates are still not very useful as we only run MCs for finitely many time steps in practice. In the next section, we present a brief introduction to the use of couplings to find explicit upper bounds on mixing times of MCs.

4 Upper Bounds on Mixing Times

We first define the notion of a mixing time, which formally captures the minimum amount of time needed for the distance $d(n)$ to be less than some prescribed constant.

Definition 5 (Mixing Time)**.** *Given an irreducible MC with stochastic transition probability matrix W and stationary distribution $\pi \in \mathcal{P}$, we define the* ϵ-mixing time *of this MC for any $\epsilon \in (0, 1)$ as:*

$$t_{\mathsf{mix}}(\epsilon) \triangleq \min \{n \geq 0 : d(n) \leq \epsilon\} \ .$$

Furthermore, we refer to $t_{\mathsf{mix}} \triangleq t_{\mathsf{mix}}(1/4)$ as the mixing time *of the MC.*

The ensuing subsections illustrate a simple technique to upper bound the mixing times of irreducible and aperiodic MCs.

4.1 Markovian Coupling

The upper bounding technique relies on the idea of couplings. Recall that a coupling of two distributions $\mu, \nu \in \mathcal{P}$ is a pair of jointly distributed random variables (x, y) with joint distribution $p_{\mathsf{x},\mathsf{y}}$ on $\mathcal{X} \times \mathcal{X}$ such that $p_{\mathsf{x}} = \mu$ and $p_{\mathsf{y}} = \nu$. We now define Markovian couplings, which are couplings between MCs.

Definition 6 (Markovian Coupling)**.** *Suppose $\{\mathsf{x}_n : n \geq 0\}$ and $\{\mathsf{y}_n : n \geq 0\}$ are two MCs on the state space $\mathcal{X}$ with stochastic transition probability matrices W_1 and W_2,* *respectively. Then, a* Markovian coupling *of these MCs is the MC* $\{\mathsf{z}_n = (\mathsf{x}_n, \mathsf{y}_n) : n \geq 0\}$ *on the state space* $\mathcal{X} \times \mathcal{X}$ *with stochastic transition probability matrix* P *that satisfies:*

$$\forall x, x', y \in \mathcal{X}, \;\; \sum_{y' \in \mathcal{X}} P((x,y),(x',y')) = W_1(x,x')\,,$$
$$\forall x, y, y' \in \mathcal{X}, \;\; \sum_{x' \in \mathcal{X}} P((x,y),(x',y')) = W_2(y,y')\,.$$

Therefore, the Markovian coupling $\{\mathsf{z}_n = (\mathsf{x}_n, \mathsf{y}_n) : n \geq 0\}$ is a "joint" MC whose "marginals" are themselves the original MCs $\{\mathsf{x}_n : n \geq 0\}$ and $\{\mathsf{y}_n : n \geq 0\}$. Note that as before, there are two trivial Markovian coupling examples. If $W_1 = W_2$ and $p_{\mathsf{x}_0} = p_{\mathsf{y}_0}$, then we can simply let $\mathsf{x}_n = \mathsf{y}_n$ for all $n \geq 0$ to obtain the "identical" Markovian coupling $\{z_n : n \geq 0\}$ with stochastic transition probability matrix:

$$\forall x, x', y' \in \mathcal{X}, \;\; P((x,x),(x',y')) = W_1(x,x')\mathbb{1}_{x'=y'} \tag{11}$$

and initial distribution: $\forall x, y \in \mathcal{X}, \;\; p_{\mathsf{z}_0}(x,y) = p_{\mathsf{x}_0}(x)\mathbb{1}_{x=y}$. Alternatively, we can run the MCs $\{\mathsf{x}_n : n \geq 0\}$ and $\{\mathsf{y}_n : n \geq 0\}$ independently and obtain the Markovian coupling $\{z_n : n \geq 0\}$ with stochastic transition probability matrix:

$$\forall x, x', y, y' \in \mathcal{X}, \;\; P((x,y),(x',y')) = W_1(x,x')W_2(y,y') \tag{12}$$

and initial distribution: $\forall x, y \in \mathcal{X}, \;\; p_{\mathsf{z}_0}(x,y) = p_{\mathsf{x}_0}(x)p_{\mathsf{y}_0}(y)$.

When $W = W_1 = W_2$ (but p_{x_0} is not necessarily equal to p_{y_0}), a particularly useful fact is that any Markovian coupling can be modified so that the two "marginal" MCs run together after the first time they meet. Formally, given a Markovian coupling with initial distribution p_{z_0} and stochastic transition probability matrix P, this modified Markovian coupling has the same initial distribution and stochastic transition probability matrix Q given by:

$$Q((x,y),(x',y')) = \begin{cases} P((x,y),(x',y')) & , \;\; x \neq y \\ W(x,x') & , \;\; x = y \text{ and } x' = y' \\ 0 & , \;\; x = y \text{ and } x' \neq y' \end{cases} \tag{13}$$

for every $x, x', y, y' \in \mathcal{X}$. We use couplings of this kind in the next two subsections.

4.2 Upper Bounds via Coupling

We now prove an upper bound on the "distance from stationarity" $d(n)$ using Markovian couplings.

Theorem 5 (Coupling Upper Bound)**.** *Let* $\{\mathsf{x}_n : n \geq 0\}$ *be an irreducible MC with stochastic transition probability matrix* W *and stationary distribution* $\pi \in \mathcal{P}$*. For* *each pair of states $x, y \in \mathcal{X}$, suppose $\{(\mathsf{x}_n, \mathsf{y}_n) : n \geq 0\}$ is a Markovian coupling of $\{\mathsf{x}_n : n \geq 0\}$ with itself, that has been modified to satisfy* (13)*, and starts at the state $(\mathsf{x}_0, \mathsf{y}_0) = (x, y) \in \mathcal{X} \times \mathcal{X}$ (i.e. has initial distribution $p_{\mathsf{x}_0,\mathsf{y}_0}(x, y) = 1$). Let $\mathbb{P}_{x,y}(\cdot)$ be the probability distribution of $\{(\mathsf{x}_n, \mathsf{y}_n) : n \geq 0\}$ with $(\mathsf{x}_0, \mathsf{y}_0) = (x, y)$, and define $\mathsf{t}_{\mathsf{coup}} \triangleq \min\{n \geq 0 : \mathsf{x}_n = \mathsf{y}_n\}$ to be the first time the "marginal" MCs meet for this Markovian coupling. Then, we have:*

$$\forall n \geq 1, \;\; d(n) \leq \max_{x,y \in \mathcal{X}} \mathbb{P}_{x,y}(\mathsf{t}_{\mathsf{coup}} > n) \, .$$

Proof. First fix any two states $x, y \in \mathcal{X}$, and consider the Markovian coupling $\{(\mathsf{x}_n, \mathsf{y}_n) : n \geq 0\}$ that starts at $(\mathsf{x}_0, \mathsf{y}_0) = (x, y)$ and runs the two "marginal" MCs together for all $n \geq \mathsf{t}_{\mathsf{coup}}$. Since $W^n(x, x') = \mathbb{P}_{x,y}(\mathsf{x}_n = x')$ and $W^n(y, y') = \mathbb{P}_{x,y}(\mathsf{y}_n = y')$ for every $x', y' \in \mathcal{X}$ and any fixed $n \geq 1$, we see that $(\mathsf{x}_n, \mathsf{y}_n)$ is a coupling of the distributions $W^n(x, \cdot) \in \mathcal{P}$ and $W^n(y, \cdot) \in \mathcal{P}$. The optimal coupling characterization of total variation distance in Theorem 2 allows us to upper bound the total variation distance between $W^n(x, \cdot)$ and $W^n(y, \cdot)$:

$$\left\| W^n(x, \cdot) - W^n(y, \cdot) \right\|_{\mathsf{TV}} \leq \mathbb{P}_{x,y}(\mathsf{x}_n \neq \mathsf{y}_n) = \mathbb{P}_{x,y}(\mathsf{t}_{\mathsf{coup}} > n)$$

where the equality holds because our Markovian coupling runs the two "marginal" MCs together after they meet. This implies that:

$$\forall n \geq 1, \;\; \max_{x,y \in \mathcal{X}} \left\| W^n(x, \cdot) - W^n(y, \cdot) \right\|_{\mathsf{TV}} \leq \max_{x,y \in \mathcal{X}} \mathbb{P}_{x,y}(\mathsf{t}_{\mathsf{coup}} > n) \, .$$

So, it suffices to prove that:

$$\forall n \geq 0, \;\; d(n) = \max_{x \in \mathcal{X}} \left\| W^n(x, \cdot) - \pi \right\|_{\mathsf{TV}} \leq \max_{x,y \in \mathcal{X}} \left\| W^n(x, \cdot) - W^n(y, \cdot) \right\|_{\mathsf{TV}} \, .$$

This holds due to the following sequence of inequalities:

$$\begin{aligned}
\left\| W^n(x, \cdot) - \pi \right\|_{\mathsf{TV}} &\triangleq \max_{A \subseteq \mathcal{X}} \left| W^n(x, A) - \pi(A) \right| \\
&= \max_{A \subseteq \mathcal{X}} \left| \sum_{y \in \mathcal{X}} \pi(y) (W^n(x, A) - W^n(y, A)) \right| \\
&\leq \max_{A \subseteq \mathcal{X}} \sum_{y \in \mathcal{X}} \pi(y) \left| W^n(x, A) - W^n(y, A) \right| \\
&\leq \sum_{y \in \mathcal{X}} \pi(y) \max_{A \subseteq \mathcal{X}} \left| W^n(x, A) - W^n(y, A) \right| \\
&= \sum_{y \in \mathcal{X}} \pi(y) \left\| W^n(x, \cdot) - W^n(y, \cdot) \right\|_{\mathsf{TV}} \\
&\leq \max_{y \in \mathcal{X}} \left\| W^n(x, \cdot) - W^n(y, \cdot) \right\|_{\mathsf{TV}}
\end{aligned}$$

where the second line holds because $\pi(A) = \sum_{y \in \mathcal{X}} \pi(y) W^n(y, A)$ (since π is the stationary distribution), the third line uses the triangle inequality, the fourth line holds because the maximum of a sum is always upper bounded by the sum over the maximum, and the final line holds because a weighted average is always upper bounded by the maximum element. Taking the maximum over all $x \in \mathcal{X}$ in the final inequality completes the proof. □

This result can be used to find explicit upper bounds on mixing times of irreducible and aperiodic MCs as the next example illustrates. We note that there are many other techniques to upper bound mixing times, as well as to lower bound mixing times, but a discussion of these techniques is beyond our scope.

4.3 Example: Lazy Random Walk on the k-Cycle

As an example, consider the *random walk on the k-cycle* for some $k \geq 2$. This is an MC with state space $\mathcal{X} = \mathbb{Z}_k = \{0, \ldots, k-1\}$ (which is the finite additive cyclic group of integers modulo k) and stochastic transition probability matrix given by:

$$\forall i, j \in \mathcal{X}, \ \ W(i,j) = \begin{cases} \frac{1}{2} & , \ \ j = i+1 \pmod{k} \text{ or } j = i-1 \pmod{k} \\ 0 & , \ \ \text{otherwise} \end{cases} . \tag{14}$$

Equivalently, we can think of the states as vertices of an undirected *cycle graph C_k*, where at each time step, the MC randomly and uniformly chooses an adjacent vertex and moves to it (i.e. at each step, it moves clockwise or anti-clockwise with probability $\frac{1}{2}$ each). The random walk on the k-cycle is an irreducible MC that is aperiodic if k is odd, and periodic with all states having period 2 if k is even (check this!).

One way to make this MC aperiodic for all $k \geq 2$ is to construct the *lazy random walk on the k-cycle.* This MC has stochastic transition probability matrix given by:

$$\forall i, j \in \mathcal{X}, \ \ W_{\mathsf{lazy}}(i,j) = \begin{cases} \frac{1}{2} & , \ \ j = i \pmod{k} \\ \frac{1}{4} & , \ \ j = i+1 \pmod{k} \text{ or } j = i-1 \pmod{k} \\ 0 & , \ \ \text{otherwise} \end{cases} \tag{15}$$

and is irreducible and aperiodic for all $k \geq 2$ (check this!). As before, we can perceive this lazy random walk as an MC on the cycle graph C_k where at each time step, the walk moves clockwise or anti-clockwise with probability $\frac{1}{4}$ each, and does not move at all with probability $\frac{1}{2}$.

We will upper bound the mixing time of the lazy random walk on the k-cycle. Let $\{\mathsf{x}_n : n \geq 0\}$ denote the MC corresponding to the lazy random walk. For any two states $x, y \in \mathcal{X}$, let $\{(\mathsf{x}_n, \mathsf{y}_n) : n \geq 0\}$ denote a Markovian coupling of $\{\mathsf{x}_n : n \geq 0\}$ with itself that starts at $(\mathsf{x}_0, \mathsf{y}_0) = (x, y)$. This Markovian coupling is governed by the following dynamics at each time step $0 \leq n < \mathsf{t}_{\mathsf{coup}}$ (before the "marginal" MCs meet):

- flip an unbiased coin (independent of all other coin tosses),

- if we get heads, then let $\mathsf{y}_{n+1} = \mathsf{y}_n$ and generate x_{n+1} from x_n according to W (i.e. move clockwise or anti-clockwise with probability $\frac{1}{2}$ each),
- if we get tails, then let $\mathsf{x}_{n+1} = \mathsf{x}_n$ and generate y_{n+1} from y_n according to W (i.e. move clockwise or anti-clockwise with probability $\frac{1}{2}$ each).

Furthermore, at each time step $n \geq \mathsf{t}_{\mathsf{coup}}$, $\mathsf{x}_n = \mathsf{y}_n$ and transitions occur according to W_{lazy} (i.e. after meeting, the "marginal" MCs run together). It is straightforward to verify that this describes a valid Markovian coupling of $\{\mathsf{x}_n : n \geq 0\}$ with itself starting at $(\mathsf{x}_0, \mathsf{y}_0) = (x, y)$ for every $x, y \in \mathcal{X}$. Using Theorem 5, we get that the "distance from stationarity" of the lazy random walk on the k-cycle is upper bounded by:

$$\forall n \geq 1, \;\; d(n) \leq \max_{x,y \in \mathcal{X}} \mathbb{P}_{x,y}(\mathsf{t}_{\mathsf{coup}} > n) \leq \frac{\max_{x,y \in \mathcal{X}} \mathbb{E}_{x,y}\left[\mathsf{t}_{\mathsf{coup}}\right]}{n} \tag{16}$$

where the second inequality follows from Markov's inequality, and $\mathbb{E}_{x,y}\left[\cdot\right]$ denotes the expectation with respect to $\mathbb{P}_{x,y}(\cdot)$.

It is a classical exercise in probability theory when analyzing the *gambler's ruin* model to establish that $\mathbb{E}_{x,y}\left[\mathsf{t}_{\mathsf{coup}}\right] = b(x,y)(k - b(x,y))$, where $b(x,y)$ denotes the "clockwise distance" between x and y. We omit a proof of this result for brevity, but using it, we obtain the bound:

$$\forall n \geq 1, \;\; d(n) \leq \frac{\max_{x,y \in \mathcal{X}} b(x,y)(k - b(x,y))}{n} \leq \frac{k^2}{4n} \tag{17}$$

where the second inequality holds because $r(1-r) \leq \frac{1}{4}$ for all $r \in [0,1]$. For any $\epsilon \in (0,1)$, we can let $\frac{k^2}{4n} \leq \epsilon$, and see that $d(n) \leq \epsilon$ if $n \geq \frac{k^2}{4\epsilon}$. This produces the following upper bound on the ϵ-mixing time:

$$\mathsf{t}_{\mathsf{mix}}(\epsilon) \leq \frac{k^2}{4\epsilon} \tag{18}$$

which we can specialize (by setting $\epsilon = \frac{1}{4}$) to get the following upper bound on the mixing time:

$$\mathsf{t}_{\mathsf{mix}} \leq k^2 \,. \tag{19}$$

We note that this result is tight in the sense that $\mathsf{t}_{\mathsf{mix}} \geq Ck^2$ for some constant $C > 0$, which can also be proved using fairly simple techniques.

The upper bound in (18) guarantees that after $\frac{k^2}{4\epsilon}$ time steps, the distribution over the states of the lazy random walk on the k-cycle is ϵ-close to its stationary distribution in total variation distance, regardless of the choice of initial distribution. Hence, such upper bounds on ϵ-mixing times indeed address our motivating question from the introduction. Moreover, (18) conveys (as we would expect) that using smaller ϵ or larger k (state space size) increases the ϵ-mixing time.

References

The material presented here is largely based on the text:

D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, 1st ed. Providence, RI, USA: American Mathematical Society, 2009.