Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

1 Preliminaries and Introduction

Welcome to the world of statistic inference and its rich connections to information theory. As we will see, it is intellectually rich, with a wealth of applications.

In these notes, we assume fluency with basic probabilistic system analysis. Accordingly, we need only to establish some notation to get started.

1.1 Notation

First, we denote random variables using san-serif fonts, e.g, x. By contrast, sample values of such variables, and other deterministic quantities, are denoted using regular serifed fonts, e.g., x. At times, we also need to distinguish between deterministic and randomized functions. We use our notation in the same way. In particular, $f(\cdot)$, for example, denotes a deterministic function, while $\mathsf{f}(\cdot)$ denotes a randomized function. With such notation, $\mathsf{f}(x)$ is a random variable, as is $f(\mathsf{x})$, as well as the doubly-random $\mathsf{f}(\mathsf{x})$.

Likewise, we use caligraphic letters to denote sets and events. An example of a set, when x is numeric, would be $\mathcal{E} = \{x \in \mathcal{X} : x > 0\}$. We will typically denote the alphabet of values that x can take on using $\mathcal{X}$. As additional notation, we use $|\cdot|$ to denote the cardinality of its set argument, i.e., the number of elements in the set, so, for example, $|\mathcal{X}|$ denotes the number of possible values x can take on. Similarly,

$$\mathcal{X}^N = \underbrace{\mathcal{X} \times \cdots \times \mathcal{X}}_{N\text{-fold}}$$

denotes the alphabet of N-tuples, each of whose elements are drawn from the alphabet $\mathcal{X}$, i.e., $(x_1, x_2, \ldots, x_N) \in \mathcal{X}^N$ is equivalent to $x_n \in \mathcal{X}$ for $n = 1, 2, \ldots, N$.

We denote the probability mass function for a discrete random variable x using $p_{\mathsf{x}}(\cdot)$, so with x denoting some fixed value in the alphabet $\mathcal{X}$, we have

$$\mathbb{P}(\mathsf{x} = x) = p_{\mathsf{x}}(x),$$

where $\mathbb{P}$ denotes the probability operator. The alphabet over which a discrete random variable is defined need not have any algebraic structure—it can be simply an arbitrary collection of symbols, e.g., $\mathcal{X} = \{\clubsuit, \heartsuit, \spadesuit, \diamondsuit\}$.

We likewise use $p_{\mathsf{x}}(\cdot)$ to denote the probability density function of a continuous random variable x. In addition, we use $\mathbb{E}[\cdot]$ as our notation for the expectation operator, and, when well-defined, use the notation $M_{\mathsf{x}}(jv) = \mathbb{E}[e^{jv\mathsf{x}}]$ to denote the characteristic function associated with the random variable x. Recall that the characteristic function is the Fourier transform of the probability mass or density function, and thus when it exists is an equivalent characterization of the random variable.

We will also frequently consider collections of random variables. For instance, a random variable pair (x, y) is characterized by the joint probability density $p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)$. However, it will often be convenient to assemble such collections into a vector and use more compact notation. For example, we can form the random vector $\mathbf{z}$ according to

$$\mathbf{z} = \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix},$$

and express the joint density for x and y, i.e., $p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)$, in the form $p_{\mathbf{z}}(\cdot)$, which is a scalar function of a vector argument. Of course, for discrete-valued quantities, the distinction between scalars and vectors is strictly unnecessary, though it can be useful at times in creating logical groupings of quantities of interest, such as for the purpose of computing conditional probabilities such as $p_{\mathsf{y}|\mathsf{x}}(\cdot|\cdot)$.

It will sometimes be useful to explicitly define classes of distributions. In particular, we use $\mathcal{P}^{\mathcal{X}}$ to denote the set of all possible probability mass (or density) functions defined over the alphabet $\mathcal{X}$. When the alphabet is clear from context, we sometimes omit the superscript. We analogously define the related notation $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ and $\mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ for joint and conditional probability distributions, respectively, etc.

Note that we are using bold face fonts to distinguish vector-valued quantities from scalar-valued ones. So z refers to a scalar, while $\mathbf{z}$ refers to a vector. Moreover, as demonstrated above, various combinations of our notation will be useful. For example, we distinguish a deterministic vector, e.g., $\mathbf{z}$, from a random vector, e.g., $\mathbf{z}$, using bold serifed font for the former and bold sans-serif font for the latter.

In general, we will reserve lowercase boldface letters for specifically column vectors, and uppercase boldface letters (e.g., $\mathbf{A}$) for matrices, i.e., when the row and column dimensions are each at least two. Row vectors can be denoted using transpose-operator notation (e.g., $\mathbf{z}^{\mathrm{T}}$). By contrast, for scalars, we attach no significance to whether the quantity is lowercase (e.g., z), or uppercase (e.g., Z).

When needed, we use bracket notation to identify elements of vector and matrix quantities. For example, $[\mathbf{A}]_{i,j}$ denotes the (i,j)th element of the matrix $\mathbf{A}$, and $[\mathbf{z}]_i$ denotes the ith element of the vector $\mathbf{z}$.

Finally, it will also be convenient at times to use script notation for sequences. In particular, for a sequence $x_1, x_2, \ldots$, we use

$$x_i^j = (x_i, x_{i+1}, \ldots, x_j)$$

when $j \geq i$ as subsequence notation. And, as a further shorthand, we often let $x^n = x_1^n$, for $n \geq 1$.

Of course, quantities such as x_i^j can also be represented in (column) vector notation (i.e., as $\mathbf{x}$), though the alternative subsequence notation makes explicit which elements constitute the vector, which will prove convenient. It is also worth emphasizing that vector and subsequence notation can and will be used at times in combination, where it serves to logically group quantities. An example would be $\mathbf{y}_i^j$ and $\mathbf{y}^n$, which would

refer to subsequences of the vector sequence $\mathbf{y}_1, \mathbf{y}_2, \ldots$. As always, such notation can be combined with serifed and sans-serif fonts to distinguish deterministic from random quantities, e.g., $\mathbf{y}^n$ vs. $\mathbf{y}^n$.

1.2 Special Functions

There are a variety of special functions that will be useful in our development. One example is the Kronecker function: for any event $\mathcal{A}$,

$$\mathbb{1}_{\mathcal{A}} \triangleq \begin{cases} 1 & \text{if } \mathcal{A} \text{ is true} \\ 0 & \text{otherwise} \end{cases}.$$

In addition, as a variant of this notation, for arbitrary variables x and y we will also use

$$\mathbb{1}_x(y) = \mathbb{1}_y(x) \triangleq \mathbb{1}_{x=y},$$

and, for a set $\mathcal{S}$,

$$\mathbb{1}_{\mathcal{S}}(x) \triangleq \mathbb{1}_{x \in \mathcal{S}},$$

and finally $\mathbb{1}_+(x) \triangleq \mathbb{1}_{x>0}$.

As matrix notation, in addition to transpose notation $^{\mathrm{T}}$ mentioned earlier, whereby $[\mathbf{A}^{\mathrm{T}}]_{i,j} = [\mathbf{A}]_{j,i}$, we use the superscript notation $^{-1}$ to denote matrix inversion, whereby for any nonsingular matrix $\mathbf{A}$ we have $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ if $\mathbf{y} = \mathbf{A}\mathbf{x}$.

In addition, if x and y are arbitrary random vectors, we denote their means via

$$\boldsymbol{\mu}_{\mathsf{x}} = \mathbb{E}[\mathsf{x}] \quad \text{and} \quad \boldsymbol{\mu}_{\mathsf{y}} = \mathbb{E}[\mathsf{y}],$$

respectively, and the covariance between them via

$$\mathrm{cov}(\mathsf{x}, \mathsf{y}) \triangleq \mathbb{E}\left[(\mathsf{x} - \boldsymbol{\mu}_{\mathsf{x}})(\mathsf{y} - \boldsymbol{\mu}_{\mathsf{y}})^{\mathrm{T}}\right].$$

1.3 Special Distributions

A few of basic distributions will arise regularly in our treatment, and thus warrant their own special notation. All are members of exponential families, the concept of which we will ultimately explore in more detail.

First is the Bernoulli distribution, denoted using $\mathtt{B}$. In particular, the notation $\mathsf{x} \sim \mathtt{B}(p)$ means that x is a binary random variable where one of the symbols has probability p (and the other with probability $1-p$).

Another is the uniform distribution, denoted using $\mathtt{U}$. In particular, the notation $\mathsf{x} \sim \mathtt{U}(\mathcal{X})$ means that x is uniformly distributed over the set $\mathcal{X}$.

Finally, there is the Gaussian (or "normal") distribution, denoted using $\mathtt{N}$. We use the notation $\mathsf{x} \sim \mathtt{N}(\mu, \sigma^2)$ to indicate that x is a scalar Gaussian random variable

with mean $\mathbb{E}[x] = \mu$ and variance $\mathbb{E}[(x-\mu)^2] = \sigma^2$. The tail probability under the unit Gaussian is denoted using $Q(\cdot)$-notation

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} \, \mathrm{d}t.$$

Although at times we also use $Q(\cdot)$ to denote a distribution, the risk of confusion will be minimal.

Moreover, $\mathbf{x} \sim \mathtt{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ will denote that $\mathbf{x}$ is a Gaussian random vector with mean vector $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and covariance matrix $\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right] = \boldsymbol{\Lambda}$. As we will later discuss, when $\boldsymbol{\Lambda}$ is nonsingular, such random vectors have a probability density function of the form[1]

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{|2\pi\boldsymbol{\Lambda}|^{1/2}},$$

the equiprobability contours of which are appropriately located, oriented and proportioned ellipses.

1.4 Inference: Introductory Remarks

Inference involves the problem of extracting information from a set of observations. When the observations are modeled as a collection of random variables, this extraction is referred to as *statistical inference.* This will be our focus, and our treatment will consider various approaches for modeling the information of interest—variables which may or may not be natural to view as random.

Problems of inference arise naturally in an extremely broad range of applications. A handful of examples include: finding the face of a person of interest from a database of images, determining and tracking position in a geolocation system, deciphering genomic data, decoding digital communication transmissions, diagnosing diseases from the results of medical tests, search query prediction, and even detecting extraterrestrial radio transmissions and playing games like RoShamBo (rock-paper-scissors).

Although the applications themselves will not be a focus, the underlying problems in such applications can be conveniently abstracted into a common form, to which a broad statistic inference framework we will develop can be applied.

Ultimately, the quality of inference that is possible in a given setting inherently depends on: 1) the quality of our inference algorithm; and 2) the quality of the data (observations) that are available. We will largely focus on the first of these: how to make (and evaluate) the best possible inferences from the available data.

At the same time, while largely beyond our scope, it is worth emphasizing that the second dependency above plays a major role and is often underappreciated in first exposures to the subject. Indeed, deciding what data (measurements) should be

[1]The operator $|\cdot|$ denotes the determinant of its matrix argument.

acquired and used in a given application is a central aspect of overall system (or experiment) design. When such decisions are not made carefully, even the best possible inference may not be able to achieve the performance required in the application of interest. (As the idiom goes, you can't make a silk purse out of a sow's ear.) In such cases, the inference can be viewed is "data starved."

Ultimately, though, to be able to make good data choices requires a strong understanding of the inference process, and thus we focus our development on such understanding, starting with the next installment of the notes.