Probability Primer

Juho Lee

The graduate school of AI, KAIST

Disclaimer

This lecture note depends heavily on the following materials:

- *https://ben-br.github.io/stat-547c-fall-2019/assets/notes/lecture-notes.pdf*.
- Çinlar, E. Probability and Stochastics. Springer New York, 2011

Measure-theoretic probability

Why measure theory?

- Recommended thread: *https://math.stackexchange.com/questions/393712/why-measure-theory-for-probability*.
- In short: more principled (and natural) way of dealing with
    - Mixture of continuous and discrete random variables. (e.g., $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathrm{Ber}(0.5)$, then $Z = (X, Y)$?).
    - Infinite dimensional random variables (stochastic processes, random probability measures, ...).
    - Non-trivial objects cannot be defined with Lebesgue measure.

Definition 1.1 (σ-algebra)

A collection $\mathcal{E}$ of subsets of a set E is a σ-algebra on E if it is closed under countable unions and complements.

1. $A \in \mathcal{E} \to E \setminus A \in \mathcal{E}$.
2. $A_1, A_2, \cdots \in \mathcal{E} \to \bigcup_{n \geq 1} A_n \in \mathcal{E}$.

Corollary

A σ-algebra $\mathcal{E}$ on E is closed under countable intersections.

$$\bigcap_{n \geq 1} A_n = E \setminus \bigcup_{n \geq 1} (E \setminus A_n). \tag{1}$$

Definition 1.2 (Topological space)

A topology τ_E of a set E is a colleciton of subsets such that

1. $\varnothing, E \in \tau_E$.
2. τ_E is closed under finite intersections.
3. τ_E is closed under any (finite, countable, uncountable) unions.

A nonemtpy set with its topology is called a topological space (E, τ_E). Subsets in τ_E are called open sets.

Think of a set of open intervals (a, b) on $\mathbb{R}$.

Definition 1.3 (Generated σ-algebra)

Let $\mathcal{E}$ be a collection of subsets of E. A σ-algebra generated from $\mathcal{E}$, denoted as $\sigma(\mathcal{E})$, is the interaction of all σ-algebras containing $\mathcal{E}$.

Corollary

$\sigma(\mathcal{E})$ is the smallest σ-algebra containing $\mathcal{E}$.

Definition 1.4 (Borel σ-algebra)

Let (E, τ_E) be a topological space. Then, the σ-algebra generated from τ_E is called the Borel σ-algebra (i.e., the smallest σ-algebra containing all open sets in E), and denoted as $\mathcal{B}(E)$. The sets in $\mathcal{B}(E)$ are called the Borel sets.

Borel σ-algebra

Take an example: $\mathcal{B}(\mathbb{R})$.

- Definition: the smallest σ-algebra containing all open sets in $\mathbb{R}$.
- Does it contain all open intervals (a, b) - yes, by definition.
- Does it contain all semi-open intervals $(a, b]$? - yes, $(a, b] = \bigcap_{n \geq 1} (a, b + 1/n)$.
- Does it contain all singleton sets $\{a\}$? - yes, $\{a\} = \bigcap_{n \geq 1} (a - 1/n, a + 1/n)$.
- Does it contain all closed intervals $[a, b]$? - yes, $[a, b] = (a, b] \cup \{b\}$.

Borel σ-algebra

- We want to have the most generic subsets as our σ-algebra.
- Why not consider $2^{\mathbb{R}}$ - the powerset of $\mathbb{R}$? - some subsets are not measurable (w.r.t. Lebesgue measure)!
- $\mathcal{B}(\mathbb{R})$ is the most generic collections of subsets that we can comfortably work with Lebesgue measure.

Measurable space, measures, and measurable mappings.

Definition 1.5 (Measurable space)

Let E be a set and $\mathcal{E}$ be its σ-algebra on E. A pair $(E, \mathcal{E})$ is called measurable space, and the elements in $\mathcal{E}$ are called measurable sets.

Definition 1.6 (Measures and measure spaces)

Let $(E, \mathcal{E})$ be a measurable space. A measure on $(E, \mathcal{E})$ is a mapping $\mu : \mathcal{E} \to \mathbb{R}^+$ such that

1. $\mu(\varnothing) = 0$.
2. $\mu(\bigcup_{n \geq 1} A_n) = \sum_{n \geq 1} \mu(A_n)$ for every disjoint $(A_n)_{n \geq 1}$.

The triplet $(E, \mathcal{E}, \mu)$ is called measure space.

Definition 1.7 (Measurable mappings)
Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. A mapping $f : E \to F$ is measurable if for any inverse image of $B \in \mathcal{F}$ is measurable.

$$f^{-1}(B) := \{x \in E | f(x) \in B\} \in \mathcal{E}. \tag{2}$$

In such case, we write f is $\mathcal{E}/\mathcal{F}$-measurable. If $(F, \mathcal{F})$ is obvious (e.g., $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$), we say that f is $\mathcal{E}$-measurable.

Probability space

Definition 1.8 (Probability space)

Let $(\Omega, \mathcal{H})$ be a measurable space. A probability measure $\mathbb{P}$ a measure on $(\Omega, \mathcal{H})$ such that $\mathbb{P}(\Omega) = 1$. A probaiblity space is a triplet $(\Omega, \mathcal{H}, \mathbb{P})$. Ω is called the sample space, and the elements ω are called outcomes. A subset of outcomes $A \in \mathcal{H}$ are called events.

Random variables

Definition 1.9 (Random variables)

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ be a measurable space. A $\mathcal{H}/\mathcal{E}-$measurable mapping is a random variable, satisfying

$$\forall A \in \mathcal{E}, \quad X^{-1}(A) := \{\omega \in \Omega | X(\omega) \in A\} \in \mathcal{H}. \tag{3}$$

We say X is a E-valued random variable.

Definition 1.10 (Distribution)

Let X be a random variable on $(E, \mathcal{E})$. The distribution or law of X is

$$\forall A \in \mathcal{E}, \quad \mu(A) = \mathbb{P}(X^{-1}(A)) := \mathbb{P}\{X \in A\}. \tag{4}$$

The probability space $(\Omega, \mathcal{H}, \mathbb{P})$ is often called as the background probability space, and the measure space $(E, \mathcal{E}, \mu)$ defined with X is called the induced probability space.

Random variables - examples

- Consider flipping two coins. The background probability space is
  - Sample space is $\{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$,
  - Set of events $\mathcal{H}$ (e.g., $\{\text{HH}, \text{HT}, \text{TH}\}$).
  - Probability measure $\mathbb{P}$
- Define a random variable X as $X(\omega) =$ number of heads in ω.

$$X(\text{HH}) = 2,\ X(\text{HT}) = 1,\ X(\text{TH}) = 1,\ X(\text{TT}) = 0. \tag{5}$$

- The distribution is then defined as

$$\mu(\{2\}) = \mathbb{P}(X^{-1}(\{2\})) = \mathbb{P}(\{\text{HH}\}), \tag{6}$$
$$\mu(\{1\}) = \mathbb{P}(X^{-1}(\{1\})) = \mathbb{P}(\{\text{HT, TH}\}), \tag{7}$$
$$\mu(\{0\}) = \mathbb{P}(X^{-1}(\{0\})) = \mathbb{P}(\{\text{TT}\}). \tag{8}$$

Definition 1.11 (Distribution function)
Let X be a random variable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The distribution function (a.k.a. Cumulative Distribution Function (CDF)) of X is defined as

$$F(x) := \mu((-\infty, x]) = \mathbb{P}\{X \leq x\}. \tag{9}$$

Definition 1.12 (Probability density function)
If $F(x)$ can be written as

$$F(x) = \int_{-\infty}^{x} f(x)\lambda(\mathrm{d}x), \tag{10}$$

where $\lambda(\mathrm{d}x)$ is a Lebesgue measure, $f(x)$ is called the density or Probability Density Function (PDF) of X.

Definition 1.13 (Probability mass function)
Let X be a random variable on a measure space $(E, \mathcal{E}, \mu)$ with $\mathcal{E}$ being discrete ($\mathcal{E}$ contains only singleton sets, i.e., $\{x\}$ for $x \in E$). The Probability Mass Function (PMF) is the density of X with respect to the counting measure ν (i.e., $\nu(A) =$ number of elements in A).

$$\mu(A) = \int_A f(x)\nu(\mathrm{d}x) = \sum_{x \in A} f(x). \tag{11}$$

See supplementary for more rigorous definition of PDF and PMF.

Expectation

Definition 1.14 (Expectation)

Let X be a random variable on $(E, \mathcal{E})$ with distribution μ. Then, expectation of X is defined as

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_\Omega X(\omega)\mathrm{d}\mathbb{P}. \tag{12}$$

Theorem 1.1 (Law of the unconscious statistician (LOTUS))

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_E x\mu(\mathrm{d}x). \tag{13}$$

If there exists a density $f(x)$ w.r.t. the Lebesgue measure,

$$\mathbb{E}[X] = \int_E xf(x)\mathrm{d}x. \tag{14}$$

Convergence of random variables

Sequence of random variables

- We often encounter with sequences of random variables $(X_n)_{n\geq 1}$; for example, for i.i.d. random variables $X_1, \ldots, X_n$,

$$\bar{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n}, \tag{15}$$

- What does it mean for such sequence to converge to something?

Definition 2.1 (Almost sure convergence)

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}$ be a sequence of random variables, and X be a random variable defined on it.
$(X_n)_{n\geq 1}$ is said to be almost surely convergent to X if

$$\mathbb{P}\Big\{ \lim_{n\to\infty} X_n = X \Big\} = \mathbb{P}\Big(\Big\{\omega \,\Big|\, \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1, \tag{16}$$

and denoted as $X_n \overset{\text{a.s.}}{\to} X$.

Convergence in probability

Definition 2.2 (Convergence in probability)
A sequence of $\mathbb{R}$-valued random variables $(X_n)_{n\geq 1}$ is said to converge in probability to X if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \varepsilon\} = 0, \tag{17}$$

and denoted as $X_n \xrightarrow{\mathrm{p}} X$.

Definition 2.3 (Convergence in distribution)

Let $(X_n)_{n\geq 1}$ be a sequence of $\mathbb{R}$-valued random variables with distribution functions $(F_n)_{n\geq 1}$. Let X be a $\mathbb{R}$-valued random variable with distribution F. $(X_n)_{n\geq 1}$ is said to converge in distribution to X if for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{18}$$

and denoted as $X_n \xrightarrow{\mathrm{d}} X$.

Relationship between different types of convergence

Proposition 2.1

$X_n \stackrel{\text{a.s.}}{\to} X$ *implies* $X_n \stackrel{\text{p}}{\to} X$.

Proposition 2.2

$X_n \stackrel{\text{p}}{\to} X$ *implies* $X_n \stackrel{\text{d}}{\to} X$.

See the supplementary for the counter-examples of converse statements.

Useful theorems

Theorem 2.3 (Arithemetic operations)

If $X_n \stackrel{\text{a.s.}}{\to} X$ *and* $Y_n \stackrel{\text{a.s.}}{\to} Y$*,*

1. $X_n + Y_n \stackrel{\text{a.s.}}{\to} X + Y$.
2. $X_n - Y_n \stackrel{\text{a.s.}}{\to} X - Y$.
3. $X_n Y_n \stackrel{\text{a.s.}}{\to} XY$.
4. $X_n / Y_n \stackrel{\text{a.s.}}{\to} X/Y$ *provided that* Y_n *and* Y *are nonzero almost surely.*

These also hold for convergence in probability.

Theorem 2.4 (Continuous mapping theorem)
Let $(X_n)_{n\geq 1}$ be a sequence of $\mathbb{R}$-valued random variables and $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then,

1. $X_n \overset{\text{a.s.}}{\to} X \implies f(X_n) \overset{\text{a.s.}}{\to} f(X)$.
2. $X_n \overset{\text{p}}{\to} X \implies f(X_n) \overset{\text{p}}{\to} f(X)$.
3. $X_n \overset{\text{d}}{\to} X \implies f(X_n) \overset{\text{d}}{\to} f(X)$.

Theorem 2.5 (Slutsky's theorem)
Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be a sequence of random variables such that $X_n \xrightarrow{\mathrm{d}} X$ and $Y_n \xrightarrow{\mathrm{p}} c$ for some constant c. Then,

1. $X_n + Y_n \xrightarrow{\mathrm{d}} X + c$.
2. $X_n - Y_n \xrightarrow{\mathrm{d}} X - c$.
3. $X_n Y_n \xrightarrow{\mathrm{d}} cX$.
4. $X_n / Y_n \xrightarrow{\mathrm{d}} X/c$ *provided that* Y_n *and* c *are nonzero almost surely.*

Useful theorems

Theorem 2.6 (Weak law of large numbers)
Let $(X_n)_{n\geq 1}$ *be a sequence of i.i.d. random variables with* $\mathbb{E}[|X|] = \mu < \infty$. *Then,*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{p}} \mu. \tag{19}$$

Theorem 2.7 (Strong law of large numbers)
Let $(X_n)_{n\geq 1}$ *be as above.*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{a.s.}} \mu. \tag{20}$$

Useful theorems

Theorem 2.8 (Central limit theorem)

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{21}$$

Exponential families and conjugate priors

Before we proceed - about notations

In classical statistical context, we write

- A random variable is uppercase X.
- A realization of a random variable is written with lowercase x.
- A probability is written with $\mathbb{P}$ (or $\Pr$ or $\mathbf{P}$, ...).
- A distribution is written as μ, and its PDF is written as f (or other notations for functions).

However, in modern ML context (and for the rest of our course), we will write

- A random variable and its realization are written as lowercase x.
- A PDF (or PMF) is simply written as $p(x)$.

Definition 3.1 (Exponential families)

A random variable X belongs to exponential families if its PDF (or PMF) is written as

$$p(x|\eta) = \exp(T(x)^\top \eta - \mathbb{1}^\top A(\eta) - B(x)), \tag{22}$$

where $T(x)$ is sufficient statistics, η is a natural parameter, and $B(x)$ is a base measure.

Exponential families - Bernoulli distribution

A Bernoulli distribution on $\{0, 1\}$ with probability θ has a PMF

$$p(x|\theta) = \theta^x (1-\theta)^{1-x} = \exp\left(x \log \frac{\theta}{1-\theta} + \log(1-\theta)\right). \tag{23}$$

Hence, Bernoulli distributoin is an exponential family with

$$T(x) = x, \tag{24}$$

$$\eta = \log \frac{\theta}{1-\theta} \tag{25}$$

$$A(\eta) = -\log(1-\theta) = \log(1+e^{\eta}), \tag{26}$$

$$B(x) = 0. \tag{27}$$

Exponential families - Gamma distribution

A gamma distribution on $(0, \infty)$ with parameters $a > 0, b > 0$ has a PDF

$$p(x|a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, \tag{28}$$

where $\Gamma(\cdot)$ is a Gamma function,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t. \tag{29}$$

Gamma distribution is an exponential family because

$$p(x|a, b) = \exp\left(\begin{bmatrix} \log x \\ x \end{bmatrix}^\top \begin{bmatrix} a - 1 \\ -b \end{bmatrix} - \log \Gamma(a) + a \log b \right). \tag{30}$$

Exponential families - Gamma distribution

$$
\begin{aligned}
T(x) &= [\log x, x]^\top && (31)\\
\eta &= [a-1, -b]^\top && (32)\\
A(\eta) &= \log\Gamma(a) - a\log b && (33)\\
&= \log\Gamma(\eta_1+1) - (\eta_1+1)\log(-\eta_2) && (34)
\end{aligned}
$$

Exponential families - Gaussian distribution

A multivariate Gaussian distribution on $\mathbb{R}^d$ with mean μ and covariance Σ has a PDF

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}. \qquad (35)$$

This can be written as

$$\exp\left(\begin{bmatrix} x \\ xx^\top \end{bmatrix}^\top \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix} - \mathbb{1}^\top \begin{bmatrix} \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\ \frac{1}{2}\log|\Sigma| \end{bmatrix} - \frac{d}{2}\log 2\pi \right), \qquad (36)$$

where the matrices inside vectors are implicitly vectorized.

Exponential families - Gaussian distribution

$$
\begin{aligned}
T(x) &= [x, xx^\top]^\top, && (37)\\
\eta &= \left[\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right]^\top && (38)\\
A(\eta) &= \left[\frac{1}{2}\mu^\top\Sigma^{-1}\mu, \frac{1}{2}\log|\Sigma|\right]^\top && (39)\\
&= \left[-\frac{1}{4}\eta_1^\top\eta_2^{-1}\eta_1, -\frac{1}{2}\log|-2\eta_2|\right] && (40)\\
B(x) &= \frac{d}{2}\log 2\pi. && (41)
\end{aligned}
$$

Nice properties of exponential families

Note that

$$\int f(x|\eta)\mathrm{d}x = \int \exp(T(x)^\top \eta - \mathbb{1}^\top A(\eta) - B(x))\mathrm{d}x$$
$$= \int \frac{\exp(T(x)^\top \eta - B(x))}{\exp(\mathbb{1}^\top A(\eta))}\mathrm{d}x = 1. \tag{42}$$

Hence we have

$$\exp(\mathbb{1}^\top A(\eta)) = \int \exp(T(x)^\top \eta - B(x))\mathrm{d}x. \tag{43}$$

Taking the derivative w.r.t. η on both sides gives (check by yourself)

$$\mathbb{E}[T(x)] = \frac{\partial \mathbb{1}^\top A(\eta)}{\partial \eta}. \tag{44}$$

Nice properties of exponential families

Taking the derivative again gives (check by yourself)

$$\mathrm{Cov}(T(x_i), T(x_j)) = \frac{\partial^2 \mathbb{1}^\top A(\eta)}{\partial \eta_i \partial \eta_j}. \tag{45}$$

Example: gamma distribution:

$$\mathbb{E}[T(x)] = \mathbb{E}[[\log x, x]^\top] = [\psi(a) - \log b, a/b] \tag{46}$$

$$\mathrm{Var}(T(x)) = [\mathrm{Var}(\log x), \mathrm{Var}(x)] = [\psi'(a), a/b^2], \tag{47}$$

where $\psi(z) = \frac{\mathrm{d} \log \Gamma(z)}{\mathrm{d} z}$ is the digamma function.

Conjugate priors for exponential famillies

- Given a likelihood $p(x|\theta)$, we choose a prior $p(\theta)$. Then, for some specific choice of priors, $p(\theta|x)$ remains the same distribution as $p(\theta)$. Such $p(\theta)$ is called to be a conjugate prior of $p(x|\theta)$.
- List of conjugate priors: *https://en.wikipedia.org/wiki/Conjugate_prior*.

Conjugate priors example: Beta-Bernoulli

Bernoulli likelihood:

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}. \tag{48}$$

Prior on θ as beta distribution with parameter $a, b > 0$:

$$p(\theta) = \mathrm{Beta}(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}. \tag{49}$$

Conjugate priors example: Beta-Bernoulli

Assume we have observed $X = \{x_1, \ldots, x_n\}$. By Bayes' rule, the posterior is written as

$$p(\theta|X) \propto p(X, \theta) = \theta^{n_1+a-1}(1-\theta)^{n_0+b-1}, \tag{50}$$

where $n_1 = \sum_{i=1}^{n} \mathbb{1}_{\{x=1\}}$ and $n_0 = n - n_1$. Hence, we can see that

$$p(\theta|X) = \mathrm{Beta}(\theta|a + n_1, b + n_0). \tag{51}$$

Conjugate priors for exponential families

Exponential families have conjugate priors which are also in exponential families. A conjugate prior for a likelihood

$$p(x|\eta) = \exp(T(x)^\top \eta - \mathbb{1}^\top A(\eta) - B(x)) \tag{52}$$

has a form

$$\begin{aligned} p(\eta; \chi, \nu) &= \exp(\eta^\top \chi - \nu^\top A(\eta) - \mathbb{1}^\top C(\chi, \nu)) \\ &= \exp\left(\begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix}^\top \begin{bmatrix} \chi \\ \nu \end{bmatrix} - \mathbb{1}^\top C(\chi, \nu) \right). \end{aligned} \tag{53}$$

$[\eta, -A(\eta)]^\top$ is the sufficient statistics (corresponds to $T(x)$), $[\chi, \nu]^\top$ is the natural parameter (corresponds to η), and $C(\chi, \nu)$ is the log-partition function (corresponds to $A(\eta)$).

Conjugate priors for exponential families

Assume we have data $X = \{x_i\}_{i=1}^n$.

$$
\begin{aligned}
p(X, \eta) &= p(\eta) \prod_{i=1}^{n} p(x_i|\eta) \\
&= \exp\left(\begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix}^\top \begin{bmatrix} \chi + \sum_{i=1}^n T(x_i) \\ \nu + n\mathbb{1} \end{bmatrix} \right. \\
&\qquad \left. - \mathbb{1}^\top C(\chi, \nu) - \sum_{i=1}^{n} B(x_i) \right). \qquad (54)
\end{aligned}
$$

Can you recognize the posterior (and the marginal likelihood $p(X)$?)

Conjugate priors for exponential families

$$p(\eta|X) = \exp\left(\begin{bmatrix}\eta \\ -A(\eta)\end{bmatrix}^\top \begin{bmatrix}\chi + \sum_{i=1}^n T(x_i) \\ \nu + n\mathbb{1}\end{bmatrix} - \mathbb{1}^\top C\left(\chi + \sum_{i=1}^n T(x_i), \nu + n\mathbb{1}\right)\right) \tag{55}$$

$$p(X) = \exp\left(\mathbb{1}^\top C\left(\chi + \sum_{i=1}^n T(x_i), \nu + n\mathbb{1}\right) - \mathbb{1}^\top C(\chi, \nu) - \sum_{i=1}^n B(x_i)\right). \tag{56}$$

Example - Beta-Bernoulli

Bernoulii likelihood:

$$p(x|\eta) = \exp\left(x \log \frac{\theta}{1-\theta} + \log(1-\theta)\right). \tag{57}$$

Beta prior:

$$p(\eta) = \exp\left(\begin{bmatrix} \log \frac{\theta}{1-\theta} \\ \log(1-\theta) \end{bmatrix}^\top \begin{bmatrix} a-1 \\ a+b-2 \end{bmatrix} + \log \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)},\right) \tag{58}$$

with

$$\chi = a - 1, \tag{59}$$

$$\nu = a + b - 2, \tag{60}$$

$$C(\chi, \nu) = \log \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}. \tag{61}$$

Example - Beta-Bernoulli

The posterior and marginal are

$$\begin{aligned} p(\eta|X) &\propto \exp\left(\begin{bmatrix} \log \frac{\theta}{1-\theta} \\ \log(1-\theta) \end{bmatrix}^\top \begin{bmatrix} a + \sum_{i=1}^n x_i - 1 \\ a + b + n - 2 \end{bmatrix} \right) \\ &= \mathrm{Beta}(x|a + n_1, b + n_0). \end{aligned} \tag{62}$$

$$p(X) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+n_1)\Gamma(b+n_0)}{\Gamma(a+b+n)}. \tag{63}$$

Supplementary

Definition 4.1 (σ-finite measure)
Let $(E, \mathcal{E}, \mu)$ be a measure space. μ is said to be σ-finite if E can covered with countable unions of measure finite sets, i.e., there exists $(A_n)_{n\geq 1}$ such that

$$\bigcup_{n\geq 1} A_n = X \text{ and } \mu(A_n) < \infty \text{ for } n \geq 1. \tag{64}$$

A measure space with σ-finite measure is said to be a σ-finite measure space.

Examples of σ-finite measures include

- Lebesgue measures.
- Counting measures.

Radon-Nikodym derivative

Definition 4.2 (Absolute continuity)

Let μ and λ be a measure on a measurable space $(E, \mathcal{E})$. μ is said to be absolute continuous w.r.t. ν if for any $A \in \mathcal{E}$,

$$\lambda(A) = 0 \Longrightarrow \mu(A) = 0, \tag{65}$$

and denote as $\mu \ll \lambda$.

Radon-Nikodym derivative

Theorem 4.1 (Radon-Nikodym)
Let μ and λ be two measures on $(E, \mathcal{E})$ and assume λ is σ-finite. If $\mu \ll \lambda$, there exists a nonnegative measurable function $f : E \to [0, \infty)$ satisfying

$$\mu(A) = \int_A f(x)\lambda(\mathrm{d}x). \tag{66}$$

The function f is unique, and called the Radon-Nikodym derivative of μ w.r.t. ν, and denoted as $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$.

Radon-Nikodym derivative

- The PDF of a $\mathbb{R}$-valued random variable is the Radon-Nikodym derivative of the CDF w.r.t. the Lebesgue measure.
- The PMF of a discrete random variable is the Radon-Nikodym derivative of the distribution (measure μ) w.r.t. the counting measure.
- In general, the density of a random variable on a measure space with Borel sets is the Radon-Nikodym derivative of the distribution μ w.r.t. the reference measure λ.

Sequence of sets

Let $(A_n)_{n\geq 1}$ be a sequence of sets. Define

$$\limsup_n A_n := \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m, \tag{67}$$

$$\liminf_n A_n := \bigcup_{n\geq 1} \bigcap_{m\geq n} A_m. \tag{68}$$

Interpretation (check it by yourself):

- $x \in \limsup_n A_n$ means that x belongs to infinitely many of $(A_n)_{n\geq 1}$.
- $x \in \liminf_n A_n$ means that x belongs to all but finitely many of $(A_n)_{n\geq 1}$.

Lemma 1 (Borel-Cantelli)
Let $(A_n)_{n\geq 1}$ be a sequence of events in a probability space. Then,

$$\sum_{n\geq 1} \mathbb{P}(A_n) < \infty \implies \mathbb{P}\Big(\limsup_n A_n\Big) = 0. \tag{69}$$

Lemma 2 (Second Borel-Cantelli)
Let $(A_n)_{ngeq1}$ be a sequence of independent events in a probability space. Then,

$$\sum_{n\geq 1} \mathbb{P}(A_n) = \infty \implies \mathbb{P}\Big(\limsup_n A_n\Big) = 1. \tag{70}$$

Theorem 4.2 (Borel-Cantelli for proving almost-sure convergence)

Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be a random variable on a common probability space. Then,

$$\forall \varepsilon > 0, \ \sum_{n\geq 1} \mathbb{P}(|X_n - X| > \varepsilon) < \infty \implies X_n \overset{\text{a.s.}}{\to} X. \tag{71}$$

Proof.

Define an event $A_n(\varepsilon) = \{\omega \mid |X_n(\omega) - X(\omega)| > \varepsilon\}$. By Borel-Cantelli lemma, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\limsup_n A_n(\varepsilon)\Big) = 0. \tag{72}$$

Proof Cont.

Now consider the event $A = \{\omega \mid \lim_n X_n(\omega) = X_n(\omega)\}$. We have to show that $\mathbb{P}(A) = 1$. $\omega \in A$ implies that for any $\varepsilon > 0$ there exists n such that $\omega \in A_m^c(\varepsilon)$ for all $m \geq n$, i.e.,

$$\omega \in \bigcap_{\varepsilon>0} \bigcup_{n\geq 1} \bigcap_{m\geq n} A_m^c(\varepsilon) = \left(\bigcup_{\varepsilon>0} \limsup_n A_n(\varepsilon) \right)^c. \tag{73}$$

Hence, we have

$$\begin{aligned} \mathbb{P}(A) &= 1 - \mathbb{P}\left(\bigcup_{\varepsilon>0} \limsup_n A_n(\varepsilon) \right) \\ &\geq 1 - \sum_{\varepsilon>0} \mathbb{P}\left(\limsup_n A_n(\varepsilon) \right) = 1, \end{aligned} \tag{74}$$

as desired. □

Checking almost-sure convergence

Theorem 4.3 (Borel-Cantelli to disproving almost-sure convergence)

Let $(X_n)_{n\geq 1}$ be a sequence of independent random variables and X be a random variable on a common probability space. Then,

$$\forall \varepsilon > 0, \sum_{n\geq 1} \mathbb{P}(|X_n - X| > \varepsilon) = \infty \implies X_n \overset{\text{a.s.}}{\nrightarrow} X. \tag{75}$$

Prove it by yourself!

Converge in probability but not almost surely

Let $(X_n)_{n\geq 1}$ be a sequence of random variables with distribution

$$\mathbb{P}(X_n = n) = \frac{1}{n}, \quad \mathbb{P}(X_n = 0) = 0. \tag{76}$$

Then $X_n \xrightarrow{\mathrm{p}} 0$ but not $X_n \xrightarrow{\mathrm{a.s.}} 0$. Show it by yourself (Hint: use Theorem 4.3).

Converge in distribution but not in probability

Let $\Omega = \{0, 1\}$ be a sample space with probability measure $\mathbb{P}(\{0\}) = 1/2$ and $\mathbb{P}(\{1\}) = 1/2$. Define a sequence of random variables $(X_n)_{n \geq 1}$ as $X_n(0) = 0$ and $X_n(1) = 1$ for all $n \geq 1$. Define also X as $X(0) = 1$ and $X(1) = 0$. Then, it is easy to check that $F_n = F$, but $|X_n(0) - X_n(0)| = 1$ for all n so does not converge in probability.