Revision on Discrete Probability

Robert M. Gower

November 8, 2019

1 Probability measures

Let Ω be a *finite* set of something we are interested in. For instances, say we are interested in the result of tossing a coin n times. Let H be "heads" and T be "tails". In this case $\Omega = \{H, T\}^n$. In particular, we would like to be able to answer questions such as "what is the chance that all n tosses are head"? Or "what is the chance there are an even number of tails"? That is, we would like to assign a probability to subsets of Ω.

To this end we define a probability measure $\mathbb{P}$ with

$$\mathbb{P} : C \subset \Omega \quad \mapsto \quad \mathbb{P}[C] \in [0, 1].$$

This probability map satisfies the following axioms

- **Total probability law** $\mathbb{P}[\Omega] = 1$.
- **Countably Additive** For sets $C_i \subset \Omega$ that are disjoint we have that

$$\cap_{i=1}^n C_i = \emptyset \quad \Longrightarrow \quad \mathbb{P}[\cup_{i=1}^n C_i] = \sum_{i=1}^n \mathbb{P}[C].$$

With this rule we can calculate the probability of the union of two sets. Indeed let $C, B \subset \Omega$ and note that

$$B \cup C = (B \setminus C) \bigcup C.$$

Furthermore since

$$B = (B \setminus C) \bigcup (B \cap C),$$

we have by taking the probability that

$$\mathbb{P}[B] = \mathbb{P}[B \setminus C] + \mathbb{P}[B \cap C]$$

and consequently

$$\mathbb{P}[B \cup C] = \mathbb{P}[B \setminus C] + \mathbb{P}[C] = \mathbb{P}[B] + \mathbb{P}[C] - \mathbb{P}[B \cap C].$$

Remark 1.1. This definition of a probability measure is only possible because Ω is a finite set. If Ω had infinitely many elements, then the probability measure is not defined over all subsets of Ω, but rather a collection of subset known as a σ–algebra.

2 Discrete Random Variables

Our finite set Ω can be anything from results of coin tosses to different types of fruits. But dealing with such strange objects makes it hard. It is much easier to deal with real numbers. In particular, if we want to do some measurements or calculations based on the outcome of an experiment that gives $\omega \in \Omega$ we need numbers. Because of this, we often deal with real functions over Ω. Let

$$\begin{aligned} X : \Omega &\mapsto \mathbb{R} \\ \omega \in \Omega &\mapsto X(\omega), \end{aligned}$$

be this real valued function. We further suppose that X has a discrete number of output given by

$$X(\Omega) = \{x_1, \ldots, x_n\}.$$

Now we would like to know the probability of each output occurring, say

$$\mathbb{P}[X = x_i] = p_i, \quad \text{for } i = 1, \ldots, n.$$

First note that $\mathbb{P}[X = x_i]$ is well defined because

$$\{X = x_i\} \stackrel{\text{def}}{=} \{\omega \in \Omega : X(\omega) = x_i\} = X^{-1}(x_i),$$

is a subset of Ω. That is $\{X = x_i\}$ is the set of all elements in Ω such that X maps them to x_i. Though awkward, often the curly brackets in $\{X = x_i\}$ are dropped, and we use $X = x_i$ instead. With or with no brackets, bare in mind that this $X = x_i$ denotes a set of points in Ω. For short hand, we write $\mathbb{P}[X = x_i] = p_i$ as simply $x_i \sim p_i$.

We refer to this X as a *random variable* because it takes on real values (thus the variable part) and we are unsure of its value (thus the random part).

To better illustrate this definition of random variable, consider the following example.

Example with sequence of coin tosses. Random variables are used to model uncertainty. For instance, consider a sequence of n coin tosses where $\Omega = \{H, T\}^n$ where H is "Heads" and T is "Tails". Say we want to measure "the number of Heads observed". We could use a random variable X for this. Let X be a function with image $\{0, \ldots, n\}$. Now let us say that this is a biased coin that returns H with probability q and T with probability $1 - q$. In this case X is a random variable with

$$\mathbb{P}[X = i] = \binom{n}{i} q^i (1 - q)^{n-i} \text{ for } i = 1, \ldots, n.$$

Using our previous notation this means that $i \sim \binom{n}{i} q^i (1-q)^{n-i}$. That is the outcome i will appear with probability $\binom{n}{i} q^i (1-q)^{n-i}$.

Joint and conditional probability. Given two random variables X and Y we use $\mathbb{P}[X = x_i \text{ and } Y = y_j]$ to denote the joint probability of both $X = x_i$ and $Y = y_j$ occurring. Said in another way, it is the probability of an element in $X^{-1}(x_i) \cap Y^{-1}(y_j)$ occurring. Consequently we have that

$$\sum_i \mathbb{P}[X = x_i \text{ and } Y = y_j] \;=\; \mathbb{P}[Y = y_j]\,. \tag{1}$$

This last equation can be derived using

$$\mathbb{P}[Y = y_j] = \mathbb{P}[\cup_i \{X = x_i\}, Y = y_j] = \sum_{i=1}^{n} \mathbb{P}[X = x_i, Y = y_j]$$

where we used that $\{X = x_i\}$ and $\{X = x_j\}$ are disjoint for $i \neq j$. Indeed, since if $\omega \in \{X = x_i\} \cap \{X = x_j\}$ then $X(\omega) = x_i$ and $X(\Omega) = x_j$, which is only possible if $x_i = x_j$.

Conditional probability. We define the *conditional probability* of $X = x_i$ given $Y = y_j$ as

$$\mathbb{P}[X = x_i \,|\, Y = y_j] \;=\; \frac{\mathbb{P}[X = x_i \text{ and } Y = y_j]}{\mathbb{P}[Y = y_j]}\,, \tag{2}$$

if $\mathbb{P}[Y = y_j] \neq 0$, otherwise $\mathbb{P}[X = x_i \,|\, Y = y_j] = 0$. Intuitively, this conditional probability is the probability of $X = x_i$ occurring given that we have observed $Y = y_j$ occurring.

Independence We say X and Y are independent if

$$\mathbb{P}[X = x_i \text{ and } Y = y_j] \;=\; \mathbb{P}[X = x_i]\,\mathbb{P}[Y = y_j]\,, \quad \text{for all } i, j \in \{1, \ldots, n\}.$$

Expectation. We define the expectation of X as the average of the elements in the image of X as weighted by their probability of occurring

$$\mathbb{E}[X] \;\stackrel{\text{def}}{=}\; \sum_{i=1}^{n} x_i \mathbb{P}[X = x_i]\,.$$

It is now straightforward to show that expectation is a *Linear operator.* That is, let X and Y be two random variables and let $\alpha, \beta \in \mathbb{R}$. It follows that

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]\,.$$

Exe: Prove this!

Conditional Expectation. Let X and Y be two random variables. We define the conditional expectation of X on $Y = y_j$ as

$$\mathbb{E}[X \mid Y = y_j] \stackrel{\text{def}}{=} \sum_{i=1}^{n} x_i \mathbb{P}[X = x_i \mid Y = y_j]. \tag{3}$$

In other words, it is the expected value of X but given that $Y = y_j$ has occurred. We can extend this definition to define the following random variable

$$\mathbb{E}[X \mid Y] : y_i \quad \mapsto \quad \mathbb{E}[X \mid Y = y_i].$$

That is $\mathbb{E}[X \mid Y]$ is a function that takes elements from the domain of Y and maps to $\mathbb{R}$.

Ex. 1 — . Show that $\mathbb{E}[\cdot \mid Y]$ is a linear operator. That is, let X, Y and Z be random variables and let $\alpha, \beta \in \mathbb{R}$. Show that

$$\mathbb{E}[\alpha X + \beta Y \mid Z] = \alpha \mathbb{E}[X \mid Z] + \beta \mathbb{E}[Y \mid Z]$$

Answer (Ex. 1) — Note that $\mathbb{P}[\alpha X + \beta Y = \alpha x_i + \beta y_j] =$

$$\begin{aligned}\mathbb{E}[\alpha X + \beta Y \mid Z] &= \sum_{j,k} (\alpha x_j + \beta y_k) \mathbb{P}[X = x_j, Y = y_k \mid Z] \\ &= \alpha \sum_{j,k} x_j \mathbb{P}[X = x_i, Y = y_k \mid Z] + \beta \sum_{j,k} y_k \mathbb{P}[X = x_j, Y = y_k \mid Z].\end{aligned}$$

Now by the definition of conditional probability

$$\begin{aligned}\sum_{j,k} x_j \mathbb{P}[X = x_j, Y = y_k \mid Z = z_i] &= \sum_j \frac{x_j}{\mathbb{P}[Z = z_i]} \sum_k \mathbb{P}[X = x_j, Y = y_k, Z = z_i] && \text{(Using (2))} \\ &= \sum_j \frac{x_j}{\mathbb{P}[Z = z_i]} \mathbb{P}[X = x_j, Z = z_i] && \text{(Using (1))} \\ &= \sum_j x_j \mathbb{P}[X = x_j \mid Z = z_i]. \quad \square\end{aligned}$$

Ex. 2 — Let X and Y be two random variables. Prove the following *tower rule* otherwise known as the *rule of total expectation*:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] \;=\; \mathbb{E}[X] \tag{4}$$

Answer (Ex. 2) —

$$\begin{aligned}\mathbb{E}\left[\sum_i x_i \mathbb{P}\left[X = x_i \mid Y\right]\right] &= \sum_j \sum_i x_i \mathbb{P}\left[X = x_i \mid Y = y_j\right] \mathbb{P}\left[Y = y_j\right] \\ &= \sum_j \sum_i x_i \mathbb{P}\left[X = x_i, Y = y_j\right] \\ &= \sum_i x_i \sum_j \mathbb{P}\left[X = x_i, Y = y_j\right] \\ &= \sum_i x_i \mathbb{P}\left[X = x_i\right] \;=\; \mathbb{E}\left[X\right]. \quad \square\end{aligned}$$

Remark 2.1 (Modern definition of conditional expectation)**.** The property (4) is what inspired the *modern* definition for conditional expectation. Let

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

We say that $\mathbb{E}\left[X \mid Y\right]$ is the conditional expectation of X given Y if it is a random variable defined over the same space as Y that satisfies

$$\mathbb{E}\left[I_A \mathbb{E}\left[X \mid Y\right]\right] \;=\; \mathbb{E}\left[I_A X\right], \quad \forall A \in \{Y^{-1}(B) \;\; : \;\; B \subset \mathcal{B}\},$$

where $\mathcal{B}$ is the *Borel σ–algebra*[a]. This definition generalizes to continuous random variables, and avoid some "dividing by zero" issues with our definition of conditional probability (2).

[a]The Borel σ–algebra, denoted by $\mathcal{B}$, is the collection of all open sets $(a,\ b) \subset \mathbb{R}$, complements of open sets, and finite unions of open sets. Since $(-\infty,\ \infty)$ is also an open set, we have that $\mathbb{R} \in \mathcal{B}$ and its complement $\emptyset \in \mathcal{B}$.

Ex. 3 — Let X and Y be two independent random variables. Let $Z = XY$ be their product. Show that

$$\mathbb{E}\left[Z \mid Y\right] = \mathbb{E}\left[X\right] Y, \tag{5}$$

and consequently

$$\mathbb{E}\left[Z\right] = \mathbb{E}\left[X\right] \mathbb{E}\left[Y\right].$$

Answer (Ex. 3) — From the definition (3) we have

$$\begin{aligned}
\mathbb{E}[Z \mid Y = y_j] &= \sum_{i,k} x_i y_k \mathbb{P}[Z = x_i y_k \mid Y = y_j] \\
&= \sum_{x_i \neq 0, y_k \neq 0} x_i y_k \frac{\mathbb{P}[XY = x_i y_k, Y = y_j]}{\mathbb{P}[Y = y_j]} \quad \text{Definition (2)} \\
&= \sum_{x_i \neq 0} x_i y_j \frac{\mathbb{P}[XY = x_i y_j, Y = y_j]}{\mathbb{P}[Y = y_j]} \qquad (\text{Since } \mathbb{P}[XY = x_i y_k, Y = y_j] = 0 \text{ if } k \neq i) \\
&= \sum_{x_i \neq 0} x_i y_j \frac{\mathbb{P}[X = x_i, Y = y_j]}{\mathbb{P}[Y = y_j]} \\
&= \sum_{x_i \neq 0} x_i y_j \mathbb{P}[X = x_i] \;=\; y_j \mathbb{E}[X]\,. \qquad (6)
\end{aligned}$$

Consequently using (4) we have that

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

3 Random Vectors and their convergence

Random vectors. A random vector $\mathbf{v} = (V_1, \ldots, V_n)^\top$ is a column vector[1] of random variables that share the same domain (indeed the same probability space). That is, $\mathbf{v}$ is a vector valued map taking inputs from say $\omega \in \Omega$ and returning $v(\omega) = (V_1(\omega), \ldots, V_n(\omega)) \in \mathbb{R}^n$.

The definitions and properties we established in the previous section carry over verbatim to random vectors. For instance, let $\mathbf{v}$ and $\mathbf{w}$ be two independent random vectors, that is for $w, v \in \mathbb{R}^n$ we have that

$$\mathbb{P}[\mathbf{v} = v \text{ and } \mathbf{w} = v] = \mathbb{P}[\mathbf{v} = v]\,\mathbb{P}[\mathbf{w} = w]\,.$$

In the case we have that

$$\mathbb{E}[\langle \mathbf{v}, \mathbf{w} \rangle \mid \mathbf{w}] \;=\; \langle \mathbb{E}[\mathbf{v}], \mathbf{w} \rangle\,. \qquad (7)$$

This follows from (5) since

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{v}, \mathbf{w} \rangle \mid \mathbf{w}] \;=\; \mathbb{E}\left[\sum_{i=1}^n V_i W_i \mid \mathbf{w}\right] \quad &= \sum_{i=1}^n \mathbb{E}[V_i W_i \mid \mathbf{w}] \\
= \sum_{i=1}^n \mathbb{E}[V_i]\, W_i \quad &= \langle \mathbb{E}[\mathbf{v}], \mathbf{w} \rangle\,. \qquad (\text{Using (5)})
\end{aligned}$$

To warm up and practice using the random vectors, we now solve the following exercise.

[1] There is now a long standing tradition in optimization that all vectors and gradients are column vectors.

Ex. 4 — Let x_t be a sequence of random vectors such that

$$\mathbb{E}\left[x_t\right] = 0 \in \mathbb{R}^n$$

for $t = 1, \ldots, T$ and

$$\mathbb{E}\left[\|x^t\|^2\right] \leq B,$$

where $B > 0$. Let $\alpha > 0$, $y_0 = 0 \in \mathbb{R}^n$ and y_t defined recursively by

$$y^{t+1} = y^t - \alpha x^t,$$

for $t = 1, \ldots, T$. Show that

$$\frac{\mathbb{E}\left[\|y^{t+1}\|^2\right]}{t+1} \;\leq\; \alpha^2 B. \tag{8}$$

Answer (Ex. 4) — Take squared norm we have that

$$\begin{aligned}\|y^{t+1}\|^2 &= \|y^t - \alpha x^t\|^2 \\ &= \|y^t\|^2 - 2\alpha\left\langle y^t, x^t\right\rangle + \alpha^2\|x^t\|^2\end{aligned}$$

Taking expectation conditioned on y^t and using (7) we have that

$$\begin{aligned}\mathbb{E}\left[\|y^{t+1}\|^2 \mid y^t\right] &= \|y^t\|^2 - 2\alpha\left\langle y^t, \mathbb{E}\left[x^t\right]\right\rangle + \alpha^2\mathbb{E}\left[\|x^t\|^2\right] \\ &\leq \|y^t\|^2 + \alpha^2 B.\end{aligned}$$

Taking expectation again

$$\begin{aligned}\mathbb{E}\left[\|y^{t+1}\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|y^{t+1}\|^2 \mid y^t\right]\right] \qquad\qquad \text{(Using (4))} \\ &\leq \mathbb{E}\left[\|y^t\|^2\right] + \alpha^2 B.\end{aligned}$$

Unrolling the recurrence now gives

$$\mathbb{E}\left[\|y^{t+1}\|^2\right] \leq (t+1)\alpha^2 B. \quad \square$$

3.1 Convergence of random variables

So far in the course you have seen gradient descent. Soon we will see stochastic methods such as stochastic gradient descent. In stochastic methods, all the iterates x^t are random vectors. We would like these iterates to converge to a minimizer x^* of our given optimization problem. Because the x^t's are random we always prove convergence in expectation or probability. In particular, we will focus on two forms of convergence. Let $\rho \in [0,\ 1)$ and let $f : \mathbb{R}^n \to \mathbb{R}_+$. We say that the function values *converge in expectation* at a linear rate ρ if

$$\mathbb{E}\left[f(x^{t+1}) - f(x^*)\right] \;\leq\; (1-\rho)\mathbb{E}\left[f(x^t) - f(x^*)\right]. \tag{9}$$

We say that the iterates *converge in L2* at a linear rate ρ if

$$\mathbb{E}\left[\|x^{t+1}-x^*\|^2\right] \;\leq\; (1-\rho)\mathbb{E}\left[\|x^t-x^*\|^2\right]. \tag{10}$$

Both (9) and (10) are referred to as *linear rates* of convergence since, if we unroll the recurrence and take natural logarithms on both sides of (10), for example, we have that

$$\log\left(\mathbb{E}\left[\|x^t-x^*\|^2\right]\right) \leq t\log(1-\rho)+\log\left(\|x^0-x^*\|^2\right).$$

And so in log scale the iterates converge linearly and proportional to t.

Yet another commonly used form of convergence of random variables is convergence in *probability*, where by, given any $\delta>0$ we have that

$$\lim_{t\to\infty}\mathbb{P}\left[\|x^t-x^*\|^2>\delta\right] \;=\; 0. \tag{11}$$

Some important questions we need to consider are: 1) Which of these forms of convergence is the "strongest"? 2) What about the variance? and 3) how do these forms of convergence translate to an iteration complexity?

We answer the first question in the following lemma.

Lemma 3.1. The convergence in L2 (10) implies the convergence in expectation of the iterates as can be seen through the equality

$$\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right] = \left\|\mathbb{E}\left[x^t-x^*\right]\right\|^2 + \mathbb{E}\left[\left\|x^t-x^*-\mathbb{E}\left[x^t-x^*\right]\right\|^2\right]. \tag{12}$$

Furthermore (10) implies convergence in probability.

Proof. Let $(x^t-x^*)_k$ be a sequence of random vectors that converges to zero according to (10). The convergence in expectation follows from the equality

$$\begin{aligned}\left\|\mathbb{E}\left[x^t-x^*\right]\right\|^2 &= \left\|\mathbb{E}\left[x^t-x^*\right]\right\|^2 + \mathbb{E}\left[\|x^t-x^*\|^2\right]-\mathbb{E}\left[\|x^t-x^*\|^2\right]\\ &= \mathbb{E}\left[\|x^t-x^*\|^2\right]-\left(\mathbb{E}\left[\|x^t-x^*\|^2\right]-2\mathbb{E}\left[\left\langle x^t-x^*,\mathbb{E}\left[x^t-x^*\right]\right\rangle\right]+\left\|\mathbb{E}\left[x^t-x^*\right]\right\|^2\right)\\ &= \mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]-\mathbb{E}\left[\left\|x^t-x^*-\mathbb{E}\left[x^t-x^*\right]\right\|^2\right]. \end{aligned}\tag{13}$$

Indeed, since $\mathbb{E}\left[\left\|x^t-x^*-\mathbb{E}\left[x^t-x^*\right]\right\|^2\right]\geq 0$ we have that

$$\left\|\mathbb{E}\left[x^t-x^*\right]\right\|^2 \leq \mathbb{E}\left[\left\|x^t-x^*\right\|^2\right] \overset{(10)}{\leq} (1-\rho)^t\|x^0-x^*\|^2.$$

Consequently the norm of the expected error converges with rate $\sqrt{1-\rho}$. Finally, let $\delta>0$. Using Markov's inequality we have

$$\mathbb{P}(\|x^t-x^*\|^2\geq\delta) \leq \frac{\mathbb{E}\left[\|x^t-x^*\|^2\right]}{\delta} \overset{(10)}{\leq} \frac{(1-\rho)^k}{\delta}\|x^0-x^*\|^2. \tag{14}$$

Thus $\mathbb{P}(\|x^t-x^*\|^2\geq\delta\|x^0-x^*\|^2)\to 0$ as $t\to\infty$. □

Iteration Complexity Given ϵ how many iterations t are needed before $\mathbb{E}\left[f(x^t) - f(x^*)\right] < \epsilon$ or $\mathbb{E}\left[\|x^t - x^*\|^2\right] < \epsilon$. ? We answer this next in the following exercises.

Ex. 5 — Consider a sequence $(\alpha_t)_t \in \mathbb{R}_+$ that converge to zero according to

$$\alpha_t \leq \frac{C}{t},$$

where $C > 0$. Given an $\epsilon > 0$, show that

$$t \geq \frac{C}{\epsilon} \quad \Rightarrow \alpha_t < \epsilon.$$

We refer to this result as a $O(1/\epsilon)$ iteration complexity.

Answer (Ex. 6) — Follows since

$$\alpha_t \;\leq\; C\frac{1}{t} \;\leq\; C\frac{\epsilon}{C} \;=\; \epsilon.$$

Ex. 6 — Using that

$$\frac{1}{1-\rho}\log\left(\frac{1}{\rho}\right) \geq 1, \tag{15}$$

prove the following lemma.

Lemma 3.2. Consider the sequence $(\alpha_k)_k \in \mathbb{R}_+$ of positive scalars that converges to zero according to

$$\alpha_k \leq \rho^k \, \alpha_0, \tag{16}$$

where $\rho \in [0, 1)$. For a given $1 > \epsilon > 0$ we have that

$$k \geq \frac{1}{1-\rho}\log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \alpha_k \leq \epsilon \, \alpha_0. \tag{17}$$

We refer to this as a $O(\log(1/\epsilon))$ iteration complexity.

Following the introduction, we can write $\alpha^t \overset{\text{def}}{=} \mathbb{E}\left[f(x^t) - f(x^*)\right]$ or $\alpha^t \overset{\text{def}}{=} \mathbb{E}\left[\|x^t - x^*\|^2\right]$. The type of convergence (16) is known as *linear convergence at a rate of* ρ^k.

Answer (Ex. 6) — *Proof.* First note that if $\rho = 0$ the result follows trivially. Assuming $\rho \in (0,\, 1)$, rearranging (16) and applying the logarithm to both sides gives

$$\log\left(\frac{\alpha_0}{\alpha_k}\right) \geq k \log\left(\frac{1}{\rho}\right). \tag{18}$$

Now using (15) and assuming that

$$k \geq \frac{1}{1-\rho} \log\left(\frac{1}{\epsilon}\right), \tag{19}$$

we have that

$$\begin{aligned} \log\left(\frac{\alpha_0}{\alpha_k}\right) &\overset{(18)}{\geq} k \log\left(\frac{1}{\rho}\right) \\ &\overset{(19)}{\geq} \frac{1}{1-\rho} \log\left(\frac{1}{\rho}\right) \log\left(\frac{1}{\epsilon}\right) \\ &\overset{(15)}{\geq} \log\left(\frac{1}{\epsilon}\right) \end{aligned}$$

Applying exponentials to the above inequality gives (17). □