Bandits, Experts, and Games

CMSC 858G    Fall 2016
University of Maryland

Intro to Probability*

Alex Slivkins
Microsoft Research NYC

* Many of the slides adopted from Ron Jin and Mohammad Hajiaghayi
Outline

- Basics: “discrete” probability
- Basics: “continuous” probability
- Concentration inequalities
Random events

- **Experiment**: e.g.: toss a coin twice
- **Sample space**: possible outcomes of an experiment
  - \( S = \{HH, HT, TH, TT\} \)
- **Event**: a subset of possible outcomes
  - \( A = \{HH\}, B = \{HT, TH\} \)
  - complement \( \bar{A} = \{HT, TH, TT\} \)
  - disjoint (mutually exclusive) events: if \( A \cap B = \emptyset \).
- Shorthand:
  - \( AB \) for \( A \cap B \)
- For now: assume finite #outcomes
Definition of Probability

- **Probability of an outcome** $u$:
  a number assigned to $u$, $\Pr(u) \geq 0$
  - Two coin tosses: $\{\text{HH, HT, TH, TT}\}$
    each outcome has probability $\frac{1}{4}$.
  - Axiom: $\sum_{u \in S} \Pr(u) = 1$

- **Probability of an event** $A \subset S$:
  a number assigned to event: $\Pr(A) = \sum_{u \in A} \Pr(u)$

- **Probability space**:
  - sample space $S$
  - probability $\Pr(u)$ for each outcome $u \in S$
Joint Probability

- For events A and B, the **joint probability** $\Pr(AB)$ (also written as $\Pr(A \cap B)$) is the probability that both events happen.

- Example: $A = \{HH\}$, $B = \{HT, TH\}$, what is the joint probability $\Pr(AB)$?

Zero
Independence

- Two events $A$ and $B$ are independent if

$$Pr(AB) = Pr(A) \cdot Pr(B)$$

“Occurrence of $A$ does not affect the probability of $B$”

- **Prop:** $Pr(\overline{AB}) = Pr(\overline{A}) \cdot Pr(B)$

- **Proof:**

$$Pr(AB) + Pr(\overline{AB}) = Pr(B)$$

$$Pr(\overline{AB}) = Pr(B) - Pr(AB)$$

$$= Pr(B) - Pr(A) \cdot Pr(B)$$

$$= Pr(B) \cdot (1 - Pr(A)) = Pr(B) \cdot Pr(\overline{A}).$$

- Events $\{A_i\}$ are **mutually independent** in case

$$Pr(\bigcap_i A_i) = \prod_i Pr(A_i)$$
Independence: examples

- Recall A and B are independent if \( \Pr(AB) = \Pr(A)\Pr(B) \)

- Example: Medical trial
  - 4000 patients
  - choose one patient unif. at random: each patient chosen w/ prob 1/4000
  - A = {the patient is a Woman}
  - B = {drug fails}

\[
\begin{array}{|c|c|c|}
\hline
 & \text{Women} & \text{Men} \\
\hline
\text{Success} & 200 & 1800 \\
\hline
\text{Failure} & 1800 & 200 \\
\hline
\end{array}
\]

- Is event A independent from event B?
  - \( \Pr(A) = 0.5, \Pr(B) = 0.5, \Pr(AB) = \frac{9}{20} \)
Independence: examples

- Consider the experiment of tossing a coin twice

- Examples: is event A independent from event B?
  - A = \{HT, HH\} = \{Coin1=H\}, B = \{HT\}
  - A = \{HT\}, B = \{TH\}

- Disjoint ≠ Independence

- If A is independent from B, B is independent from C, is A independent from C?
  
  Not necessarily, say A=C
Conditional probability

- If $A$ and $B$ are events with $\Pr(A) > 0$, the \textit{conditional probability of $B$ given $A$} is

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

- Example: medical trial

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>200</td>
<td>1800</td>
</tr>
<tr>
<td>Failure</td>
<td>1800</td>
<td>200</td>
</tr>
</tbody>
</table>

Choose one patient at random
- $A = \{\text{Patient is a Woman}\}$
- $B = \{\text{Drug fails}\}$
- $\Pr(B \mid A) = 18/20$
- $\Pr(A \mid B) = 18/20$

- If $A$ is independent from $B$, $\Pr(A \mid B) = \Pr(A)$
Conditional Independence

- Event A and B are *conditionally independent given C* if
  \[ \Pr(AB|C) = \Pr(A|C) \Pr(B|C) \]
- Events \( \{A_i\} \) are conditionally mutually independent given C if
  \[ \Pr(\bigcap_i A_i|C) = \prod_i \Pr(A_i|C) \]
Conditional Independence (cont’d)

- Example: three events A, B, C
  - Pr(A) = Pr(B) = Pr(C) = 1/5
    - Pr(AC) = Pr(BC) = 1/25, Pr(AB) = 1/10
    - Pr(ABC) = 1/125
  - Are A, B independent? 1/5 * 1/5 ≠ 1/10
  - Are A, B conditionally independent given C?
    - Pr(A|C) = (1/25)/(1/5) = 1/5,
    - Pr(B|C) = (1/25)/(1/5) = 1/5
    - Pr(AB|C) = (1/125)/(1/5) = 1/25 = Pr(A|C)Pr(B|C)

- A and B are independent
  ≠ A and B are conditionally independent
Random Variable

- **Experiment**: e.g.: toss a coin twice
  - sample space $S$ and probability $\Pr(\cdot)$
- A *random variable* $X$ assigns a number to every outcome
  - $X = \#\text{heads}$
  - “function from sample space to numbers”
  - shorthand: RV for “random variable”
- **Distribution** of $X$ assigns probability $\Pr(X = x)$ to every $x \in \mathbb{R}$
  - *probability mass function* (pmf) $f_X(x) = \Pr(X = x)$
- **Support** of $X$ is the set of all $x \in \mathbb{R}$ for which $f_X(x) > 0$
Random Variable: Example

- Experiment: three rolls of a die. Let $X$ be the sum of #dots on the three rolls.
- What are the possible values for $X$?
- $\Pr(X = 3) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216}$,
- $\Pr(X = 5) = ?$
Expectation

- Expectation of random variable $X$
  \[ E[X] = \sum_x x \Pr(X = x) \]
  - weighted average of numbers in the support
- Nice properties:
  - $E[c] = c$ for any constant $c$.
  - Linear: $E[\alpha X] = \alpha E[X]$ for any $\alpha \in \mathbb{R}$
  - Monotone: if $X \leq Y$ with prob. 1, then $E[X] \leq E[Y]$
Conditional expectation

- **Conditional expectation** of RV $X$ given event $A$:

$$E[X|A] = \sum_{x \in \text{support}} x \Pr(X = x | A)$$

- same formula as $E[X]$, but with conditional probabilities

- expectation of $X$ in a “conditional” probability space
  - same sample space as before
  - all probabilities conditioned on $A$

- same nice properties as before
Variance

- **Variance** of RV $X$:
  \[ Var(X) = E((X - E[X])^2) = E(X^2) - (E[X])^2 \]
  - characterizes how much $X$ spreads away from its expectation

- **Nice properties:**
  - $Var(X) \geq 0$
  - $Var(X + c) = Var(X)$ for any constant $c$
  - $Var(\alpha X) = \alpha^2 Var(X)$ for any $\alpha \in \mathbb{R}$

- **Standard deviation** $\sigma(X) = \sqrt{Var(X)}$

- NB: variance can be infinite!
  - $X = 2^i$ with probability $2^{-i}$, for each $i = 1,2,3,\ldots$. 
Uniform distribution

- choose “uniformly at random” (u.a.r.)
  - sample space: $K$ items
  - same probability $\frac{1}{K}$ for each item.

- (discrete) uniform distribution
  - random variable $X$ can take $K$ possible values
  - all values have the same probability $\frac{1}{K}$
Bernoulli & Binomial

- **Bernoulli** distribution
  - success with probability $p$, failure otherwise
  - **Bernoulli** RV $X$ (a.k.a. 0-1 RV):
    \[
    \Pr(X = 1) = p \quad \text{and} \quad \Pr(X = 0) = 1 - p
    \]
  - $E[X] = p$, $\text{Var}(X) = E[X^2] - E[X]^2 = p - p^2$

- **Binomial distribution**
  - $X = \#$ successes in $n$ draws of a Bernoulli distribution
  - $X_i \sim \text{Bernoulli}(p), \ i = 1 \ldots n$
    
    \[X = \sum_{i=1}^{n} X_i, \ X \sim \text{Bin}(p, n)\]
  - $E[X] = np$, $\text{Var}(X) = np(1-p)$
Independent RVs

- Two random variables $X$ and $Y$ on the same experiment
  - outcomes of two coin tosses
- Joint distribution: $f_{X,Y}(x, y) = \Pr(X = x, Y = y)$
- $X$ and $Y$ are **independent** if for all $x, y \in \mathbb{R}$
  $$f_{X,Y}(x, y) = \Pr(X = x) \Pr(Y = y)$$
  - equiv.: if events \{X = x\} and \{Y = y\} are independent
- Basic properties:
  $$E[XY] = E[X] E[Y]$$
  $$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$
- RVs $X, Y, Z, \ldots$ **mutually independent** if
  $$\Pr(X = x, Y = y, Z = z, \ldots) = \Pr(X = x) \Pr(Y = y) \Pr(Z = z) \ldots$$
- Shorthand: **IID** for “independent and identically distributed”
Outline

- Basics: “discrete” probability
- Basics: “continuous” probability
- Concentration inequalities
Infinitely many outcomes

- experiments can have infinitely many outcomes
  - all finite sequences of coin tosses
  - countably many outcomes => same treatment as before
- experiments can have "continuously" many outcomes
  - throw a dart randomly into a unit interval
    Outcomes: all numbers in \([0,1]\)
  - infinite sequence of coin tosses
    Outcomes: infinite binary sequences

- Sample space \(S\): set of all possible outcomes
  - Events: subsets of \(S\)
- Probabilities assigned to events, not to individual outcomes!
Definition of Probability

- **Probability of an event**: a number assigned to event $\Pr(A)$
  - Axiom 1: $0 \leq \Pr(A) \leq 1$
  - Axiom 2: $\Pr(S) = 1$, $\Pr(\emptyset) = 0$
  - Axiom 3: For any two events $A$ and $B$,
    $$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(AB)$$

- Corollaries
  - $\Pr(\overline{A}) = 1 - \Pr(A)$
  - For every sequence of disjoint events
    $$\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$$
**Probability space**

- *Probability space* consists of three things:
  - sample space $S$
  - set of events $\mathcal{F}$ (where each event is a subset of $S$)
  - probability $Pr(A)$ for each event $A \in \mathcal{F}$

- $\mathcal{F}$ is the set of events that "we care about"
  - OK to care about some, but not all events
    ($\mathcal{F}$ does not have to include all events)
  - $\mathcal{F}$ must satisfy some formal properties ("$\sigma$-algebra")
    to make probability well-defined
Random variable $X$

- **Experiment**: infinite sequence of coin tosses
  - sample space: infinite binary sequences $(b_1, b_2, ...)$

- A **random variable** $X$ assigns a number to every outcome
  - $X = 0.b_2b_4b_6 ... \in [0,1]$
  - “function from sample space to numbers”

- **Distribution** of $X$: assigns probability to every interval:
  - $\Pr(a \leq X \leq b)$
    - cumulative distribution function (cdf)
    - $F_X(x) = \Pr(X \leq x)$
Continuous vs discrete

- **“Continuous”** random variable $X$:
  - each possible value happens with zero probability
  - “throw a dart randomly into a unit interval”

- **“Discrete”** random variable $Y$:
  - each possible value happens with positive probability
  - #heads in two coin tosses
  - NB: may happen even if #outcomes is infinite, e.g.:
    \[
    \Pr(Y = i) = 2^{-i}, \quad i = 1, 2, 3, \ldots
    \]

- RVs can be neither “continuous” nor “discrete”! E.g., max($X, Y$)
**Probability density function (pdf)**

- **Pdf** for random variable $X$ is a function $f_X(x)$ such that
  \[ \Pr(a \leq X \leq b) = \int_a^b f_X(x) \, dx \]
  - not guaranteed to exist (but exists in many useful cases)

- **Support** of $X = \{ \text{all } x \text{ such that } f_X(x) > 0 \}$
  - How to define “support” if pdf does not exist? E.g.:
    - $Y$ is discrete random variable, and
      $Z = X$ with probability $\frac{1}{2}$, and $Z = Y$ otherwise.
    - Then $\text{support}(Z) = \text{support}(X) \cup \text{support}(Y)$
Expectation

- If pdf $f_X$ exists, then expectation is
  \[ E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) \, dx \]

- General definition (for any random variable)
  - Lebesgue integral of $X$ with respect to measure $\Pr(\cdot)$
  - no need to know what it is, for this course

- Same nice properties as in the discrete case
Uniform distribution

● Informally:
  ➢ “Throw a random dart into an interval \([a,b]\) ”
  ➢ “each number has the same probability ”

● Formally:
  ➢ sample space: all numbers in \([a,b]\)
  ➢ probability density function: \(f_X(x) = 1/(b - a)\)
  ➢ equivalently:
    \[
    \Pr(a' \leq X \leq b') = (b' - a')/(b - a)
    \]
    for every interval \([a', b'] \subset [a, b]\)
Independent RVs

- Two random variables $X$ and $Y$ on the same experiment
  - “two throws of a dart into a unit interval”
- **Joint distribution** of $X$ and $Y$
  - assigns probability $Pr(X \in I, Y \in J)$, for any two intervals $I, J$
- $X$ and $Y$ are independent if for all intervals $I, J$
  - $Pr(X \leq x, Y \leq y) = Pr(X \leq x) \cdot Pr(Y \leq y)$
  - equivalently: if events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent
- Random variables $X, Y, Z, \ldots$ *mutually independent* if
  - $Pr(X \leq x, Y \leq y, Z \leq z, \ldots) = Pr(X \leq x) \cdot Pr(Y \leq y) \cdot Pr(Z \leq z) \ldots$
Normal (Gaussian) Distribution

- Random variable $X \sim N(\mu, \sigma^2)$ defined by pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right)$$

  - two parameters: expectation $\mu$ and variance $\sigma^2$
  - “standard normal distribution”: $N(0,1)$

- Nice properties:
  - If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
  - Central Limit Theorem (informally): If $Y_1, \ldots, Y_n$ are IID RVs with finite variance, their average converges to a normal distribution as $n \to \infty$
Outline

- Basics: “discrete” probability
- Basics: “continuous” probability
- Concentration inequalities
Concentration inequalities

- **Setup:** $X_1, \ldots, X_n$ random variables.
  (not necessarily identically distributed)
  $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is the average, and $\mu = \mathbb{E}[\bar{X}]$

- **Strong Law of Large Numbers:**
  $$\Pr \left( \bar{X} \xrightarrow{n} \mu \right) = 1$$

- **Want:** $\bar{X}$ is *concentrated* around $\mu$ when $n$ is large,
  i.e. that $|\bar{X} - \mu|$ is small with high probability.
  
  $\Pr(|\bar{X} - \mu| \leq "\text{small}") \geq 1 - "\text{small}"

  such statements are called “concentration inequalities”
Hoeffding Inequality (HI)

- High-prob. event: \( \mathcal{E}_{\alpha,T} = \left\{ |\bar{X} - \mu| \leq \sqrt{\frac{\alpha \log T}{n}} \right\} \), \( \alpha \geq 0 \)

- **HI**: Assume \( X_i \in [0,1] \) for all \( i \). Then
  \[
  \Pr(\mathcal{E}_{\alpha,T}) \geq 1 - 2T^{-2\alpha}.
  \]

  - \( \alpha = 2 \) suffices for most applications in this course. 
    T controls probability; can be the time horizon in MAB
  - this is a convenient re-formulation of HI for our purposes 
    more “flexible” and “generic” formulation exists

- “Chernoff Bounds”: special case when \( X_i \in \{0,1\} \)

- Relevant notation: \( r = \sqrt{\frac{\alpha \log T}{n}} \) “confidence radius 
  \[ [\mu - r, \mu + r] \) “confidence interval”
Hoeffding Inequality (extensions)

- **Recall:** \( \mathcal{E}_{\alpha, T} = \left\{ |\bar{X} - \mu| \leq \sqrt{\frac{\alpha \log T}{n}} \right\}, \alpha \geq 0 \)

- **"HI for intervals":** Assume \( X_i \in [a_i, b_i] \) for all \( i \). Then
  \[
  \Pr(\mathcal{E}_{\alpha \beta, T}) \geq 1 - 2T^{-2\alpha}, \text{ where } \beta = \frac{1}{n} \sum_{i=1}^{n} (b_i - a_i)^2.
  \]

- **"HI for small variance":**
  Assume \( X_i \in [0,1] \) and \( \text{Var}(X_i) \leq \nu \) for all \( i \). Then
  \[
  \Pr(\mathcal{E}_{\alpha \nu, T}) \geq 1 - 2T^{-\alpha/4}.
  \]
  as long as \( n \) is large enough: \( \frac{n}{\log n} \geq \frac{\alpha}{9\nu} \).

- **"HI for Gaussians":**
  Assume \( X_i \) is Gaussian with variance \( \leq \nu \). Then
  \[
  \Pr(\mathcal{E}_{\alpha \nu, T}) \geq 1 - 2T^{-\alpha/2}.
  \]
Concentration for non-independent RVs

- **Setup:** $X_1, \ldots, X_n$ independent random variables in $[0,1]$ (not necessarily independent or identically distributed)
  $$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$  is the average.

- **Assume:** there is a number $\mu_i \in [0,1]$ such that
  $$E(X_i | X_1 \in J_1, \ldots, X_{i-1} \in J_{i-1}) = \mu_i$$  for any intervals $J_1, \ldots, J_{i-1} \subset \mathbb{R}$.

- Let $\mathcal{E}_\alpha = \left\{ |\bar{X} - \mu| \leq \sqrt{\frac{\alpha \log T}{n}} \right\}$, $\alpha \geq 0$.

- Then (corollary from “Azuma-Hoeffding inequality”)
  $$\Pr(\mathcal{E}_{\alpha,T}) \geq 1 - 2T^{-\alpha/2}$$
Resources

• A survey on concentration inequalities by Fan Chung and Linyuan Lu (2010)

• Another survey on concentration inequalities by Colin McDiarmid (1998).

• Wikipedia
  - Hoeffding inequality
  - Azuma-Hoeffding inequality