CS145: Probability & Computing

Lecture 21: Review

Figure credits:
*Bertsekas & Tsitsiklis, **Introduction to Probability**, 2008*
*Pitman, **Probability**, 1999*

Outline of different statistical inference methods

I have a sequence of independent random variables $X_1, \ldots, X_n$
from a same distribution with parameter θ

I can ask different questions about the distribution
(statistical inference)

Outline of different statistical inference methods

I have a sequence of independent random variables $X_1, \ldots, X_n$
from a same distribution with parameter θ

I can test n hypothesis on the distribution

$$\theta = 0 \quad ?$$

Hypothesis Testing

Outline of different statistical inference methods

I have a sequence of independent random variables $X_1, \ldots, X_n$
from a same distribution with parameter θ

I can try to estimate the parameter θ

(return a single value)

Parameter Estimation
Maximum Likelihood

e.g. $\hat{\theta} = \text{argmax}_\theta L(\theta) = \text{argmax}_\theta \prod_{i=1}^{n} f(X_i; \theta)$

Outline of different statistical inference methods

I have a sequence of independent random variables $X_1, \ldots, X_n$
from a same distribution with parameter θ

I can find an interval that is likely to contain θ

$$\text{interval } I : \Pr(\theta \notin I) \leq \delta$$

(interval estimation, confidence interval)

Interval Estimation
Monte-Carlo

e.g. $I = \left[\frac{1}{n}\sum_{i=1}^{n} X_i - 2\sigma, \frac{1}{n}\sum_{i=1}^{n} X_i + 2\sigma\right]$

$$\Pr(E(X) \notin I) \leq 1/4$$

(Chebyshev's inequality)

Bayesian Hypothesis Testing

Bayesian Hypothesis Testing

Also known as classification, categorization, or discrimination.

We want to choose between two *mutually exclusive hypotheses*:

- *H=0: Null* hypothesis
- *H=1: Alternative* hypothesis

There is some *prior probability* of each hypothesis:

- Probability of *H=0*: $p_H(0) = q$
- Probability of *H=1*: $p_H(1) = 1 - q$

Observed data *X* has a *likelihood function* under each hypothesis:

- Discrete data: $p_{X|H}(x \mid 0), \quad p_{X|H}(x \mid 1)$
- Continuous data: $f_{X|H}(x \mid 0), \quad f_{X|H}(x \mid 1)$

Formulas on following slides assume discrete X for simplicity.

Loss Functions

We need to formalize the notion of the *cost of a mistake*:

$$L(h, g) = \text{cost of predicting hypothesis } g \text{ when } h \text{ is true.}$$

Properties of standard *loss functions* used for hypothesis testing:

- Assume there is *no loss for correct decisions*:

$$L(0,0) = L(1,1) = 0$$

- **Type I Error:** Positive loss for *false positives* or “false alarms”

$$L(0,1) = \lambda_{01} > 0$$

- **Type II Error:** Positive loss for *false negatives* or “missed detections”

$$L(1,0) = \lambda_{10} > 0$$

- Can encode “utilities” or “rewards” as negative losses

Example: Spam Classification

$p_{X|H}(x \mid h) =$ *Model of words in email: naïve Bayes, Markov chain, …*

Decision	*h=0*: Ham (not spam)	*h=1*: Spam
$g = 0$	$L(0,0) = 0$	$L(1,0) = \lambda_{10} > 0$ ***False negative:*** *A spam email is placed in your Inbox.*
$g = 1$	$L(0,1) = \lambda_{01} > 0$ ***False positive:*** *Some real email is placed in Spam folder.*	$L(1,1) = 0$

Example: Medical Diagnosis

$f_{X|H}(x \mid h) =$ *Results of various laboratory tests, scans, …*

Decision	*h=0*: Healthy	*h=1*: Serious Illness
$g = 0$	$L(0,0) = 0$	$L(1,0) = \lambda_{10} > 0$ ***False negative:*** *Illness goes untreated and you become more sick.*
$g = 1$	$L(0,1) = \lambda_{01} > 0$ ***False positive:*** *Unnecessary painful or costly medical tests.*	$L(1,1) = 0$

Bayesian Decision Theory

We are given both a *probabilistic model* and a *loss function*:

Posterior distribution:

$$p_{H|X}(h \mid x) = \frac{p_{X|H}(x \mid h)p_H(h)}{p_X(x)}$$

Loss function:

$$L(0,1) = \lambda_{01} > 0 \qquad L(1,0) = \lambda_{10} > 0$$

The optimal decision then *minimizes the posterior expected loss*:

$$\delta(x) = \arg\min_g \; E[L(h,g) \mid X = x] = \arg\min_g \sum_{h=0}^{1} L(h,g)p_{H|X}(h \mid x)$$

Likelihood Ratio Tests

Expected loss of guessing hypothesis *h=1*:

$$L(0,1)p_{H|X}(0 \mid x) + L(1,1)p_{H|X}(1 \mid x) = \lambda_{01}p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis *h=0*:

$$L(0,0)p_{H|X}(0 \mid x) + L(1,0)p_{H|X}(1 \mid x) = \lambda_{10}p_{H|X}(1 \mid x)$$

The optimal decision then *minimizes the posterior expected loss*:

$$\delta(x) = \arg\min_g \; E[L(h,g) \mid X = x] = \arg\min_g \sum_{h=0}^{1} L(h,g)p_{H|X}(h \mid x)$$

Likelihood Ratio Tests

Expected loss of guessing hypothesis *h=1*:

$$L(0,1)p_{H|X}(0 \mid x) + L(1,1)p_{H|X}(1 \mid x) = \lambda_{01}p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis *h=0*:

$$L(0,0)p_{H|X}(0 \mid x) + L(1,0)p_{H|X}(1 \mid x) = \lambda_{10}p_{H|X}(1 \mid x)$$

It is optimal to decide *h=1* if and only if:

$$\lambda_{01}p_{H|X}(0 \mid x) \leq \lambda_{10}p_{H|X}(1 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \cdot \left(\frac{\lambda_{01}}{\lambda_{10}}\right) \qquad p_H(0) = q$$

Minimizing Probability of Error

The general *likelihood ratio test* picks *h=1* if and only if:

$$\lambda_{10} p_{H|X}(1 \mid x) \geq \lambda_{01} p_{H|X}(0 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \cdot \left(\frac{\lambda_{01}}{\lambda_{10}}\right)$$

$$p_H(0) = q$$

If *all errors are equally costly* this simplifies: $\lambda_{10} = \lambda_{01} = 1$

$$p_{H|X}(1 \mid x) \geq p_{H|X}(0 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right)$$

Pick hypothesis with larger posterior probability to minimize number of errors

Bivariate Distribution

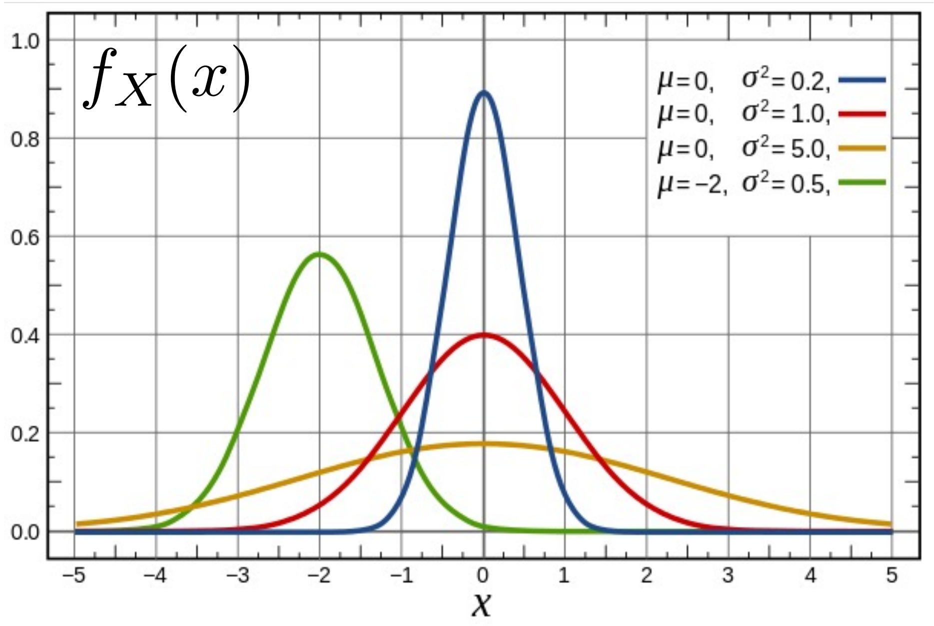
Normal Random Variables

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

$$E[X] = \mu$$

$$\mathrm{Var}[X] = E[(X-\mu)^2] = \sigma^2$$

$$\sqrt{\mathrm{Var}[X]} = \sigma \text{ is the standard deviation}$$



***Theorem**: A linear function of a Gaussian variable is Gaussian!*

$$Y = aX + b$$

$$\boxed{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2}\left(\frac{y-\bar{\mu}}{\bar{\sigma}}\right)^2}$$

$$\bar{\mu} = a\mu + b, \qquad \bar{\sigma} = |a|\sigma$$

Bivariate Normal Distribution

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \qquad f_V(v) = \frac{1}{\sqrt{2\pi}} e^{-v^2/2}$$

- A *bivariate normal distribution* is any joint distribution defined as a *linear function of two independent normal distributions*
- First consider the following particular linear function:

$$X = \sqrt{\frac{1+\rho}{2}} U + \sqrt{\frac{1-\rho}{2}} V \qquad Y = \sqrt{\frac{1+\rho}{2}} U - \sqrt{\frac{1-\rho}{2}} V$$

- The *joint probability density function* of *X* and *Y* equals:

$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2}{2(1-\rho^2)} - \frac{y^2}{2(1-\rho^2)} + \frac{\rho x y}{1-\rho^2}\right\}$$

$$\rho = 0 \Rightarrow f_{XY}(x,y) = \frac{1}{2\pi} \exp\left\{-\frac{x^2}{2} - \frac{y^2}{2}\right\} = f_X(x) f_Y(y) \Rightarrow$$ ***Independence!***

Bivariate Normal Distribution

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2(1-\rho^2)} - \frac{(y-\mu_y)^2}{2\sigma_y^2(1-\rho^2)} + \frac{\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y(1-\rho^2)}\right\}$$

- Coordinate system and units for random variable *X:*

  Mean: $\mu_x = E[X]$ $\qquad P(X \le \mu_x) = P(X \ge \mu_x) = 0.5$

  Standard deviation: $\sigma_x = \sqrt{\mathrm{Var}(X)}$

- Coordinate system and units for random variable *Y:*

  Mean: $\mu_y = E[Y]$ $\qquad P(Y \le \mu_y) = P(Y \ge \mu_y) = 0.5$

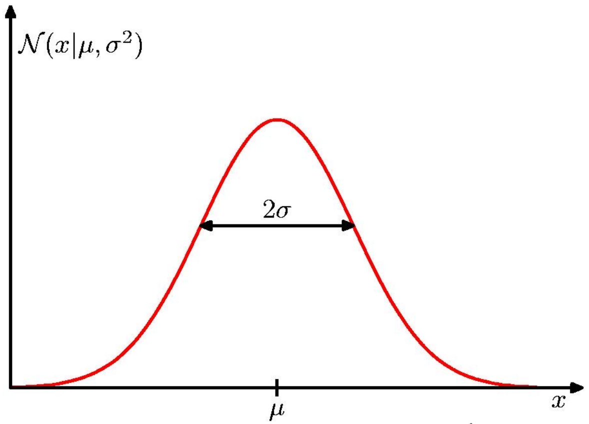
  Standard deviation: $\sigma_y = \sqrt{\mathrm{Var}(Y)}$

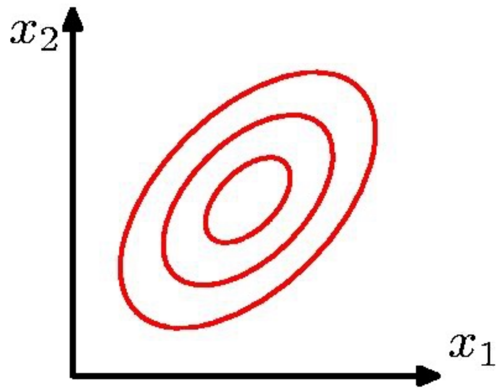
- Dependence between *X, Y* measured by *correlation coefficient:*

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}, \qquad -1 \le \rho \le 1$$

For bivariate variables: X and Y independent if and only if $\rho = 0$

Multivariate Normal Distribution

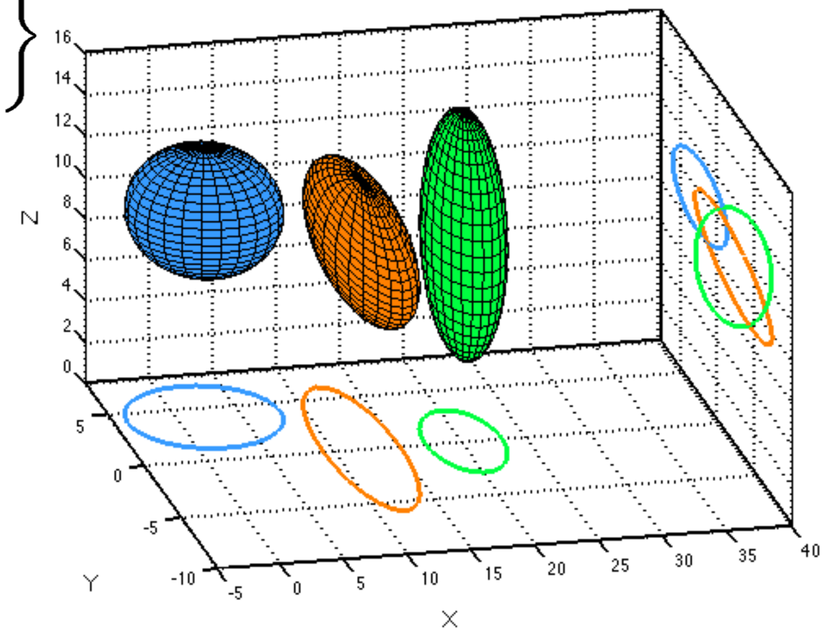




$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

0.5 meter

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



$$\mu = E[X]$$
$$\Sigma = E[(X-\mu)(X-\mu)^T]$$

D-dimensional ellipsoids parameterized by mean vector & covariance matrix

Exercise from the homework

Let $\vec{X} = (X_1, X_2)^T$ be a bivariate normal distribution with $E(X_1) = E(X_2) = 0$, $\mathbb{V}(X_1) = \mathbb{V}(X_2) = 1$, and $\text{Cov}(X, Y) = 0$, i.e.

Let $S = \text{sign}(X_0)$ be a random variable with support $\{-1, 1\}$, where $X_0 \sim \mathcal{N}(0, 1)$ is a standard normal random variable that is independent to X_1 and X_2. The function $\text{sign}(x) = 1$ if $x \geq 0$ and $\text{sign}(x) = -1$ if $x < 0$.

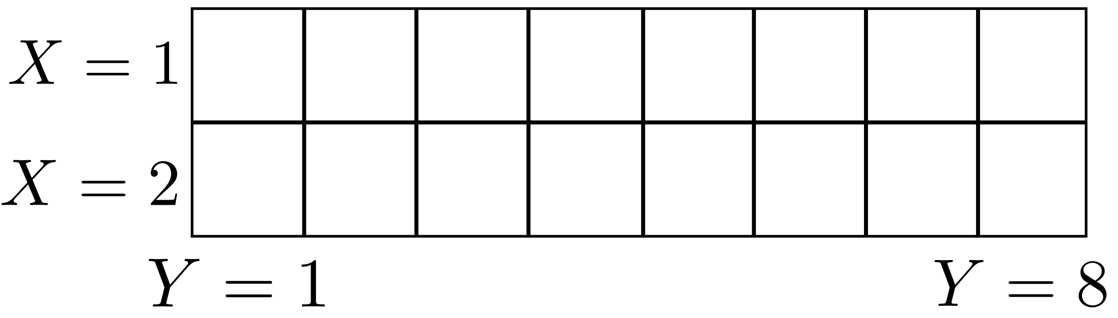
(d) Show that SX_1 and $S|X_1|$ are both normal random variables.

(e) Show that $SX_1 + SX_2$ is a normal random variable.

(f) Is the vector $(X_0, S|X_1|)$ distributed as a bivariate normal?

Conditional Probability and Expectation

Joint Probability Distribution



In this example, N=2 and M=8, and the joint PMF is a 2x8 matrix.

- Consider two random variables *X, Y.*
  Suppose range of X is size N, range of Y is size M.
- The *joint probability mass function* or *joint distribution* of two variables:

$$p_{XY}(x, y) = P(X = x \text{ and } Y = y)$$

$$p_{XY}(x, y) \geq 0, \qquad \sum_x \sum_y p_{XY}(x, y) = 1.$$

- The joint distribution is uniquely specified by *NM-1* numbers

Joint Probability Distribution

Infer discrete X from discrete Y:

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_{Y|X}(y \mid x)}{p_Y(y)}$$

$$p_Y(y) = \sum_x p_X(x)p_{Y|X}(y \mid x)$$

Example:

- $X = 1, 0$: airplane present/not present
- $Y = 1, 0$: something did/did not register on radar

Infer continuous X from continuous Y:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y \mid x)}{f_Y(y)}$$

$$f_Y(y) = \int_x f_X(x)f_{Y|X}(y \mid x)\,dx$$

Example: X: some signal; "prior" $f_X(x)$
Y: noisy version of X
$f_{Y|X}(y \mid x)$: model of the noise

Infer discrete X from continuous Y:

$$p_{X|Y}(x \mid y) = \frac{p_X(x)f_{Y|X}(y \mid x)}{f_Y(y)}$$

$$f_Y(y) = \sum_x p_X(x)f_{Y|X}(y \mid x)$$

Example:

- X: a discrete signal; "prior" $p_X(x)$
- Y: noisy version of X
- $f_{Y|X}(y \mid x)$: continuous noise model

Infer continuous X from discrete Y:

$$f_{X|Y}(x \mid y) = \frac{f_X(x)p_{Y|X}(y \mid x)}{p_Y(y)}$$

$$p_Y(y) = \int_x f_X(x)p_{Y|X}(y \mid x)\,dx$$

Example:

- X: a continuous signal; "prior" $f_X(x)$ (e.g., intensity of light beam);
- Y: discrete r.v. affected by X (e.g., photon count)
- $p_{Y|X}(y \mid x)$: model of the discrete r.v.

Example

- Suppose 90% of hard drives in some laptop computer model have exponentially distributed lifetime param θ_0

$$f_{Y|X}(y \mid 0) = \theta_0 e^{-\theta_0 y} \qquad p_X(0) = 0.9$$

- However, 10% of hard drives have a manufacturing defect that gives them a shorter lifetime $\theta_1 > \theta_0$

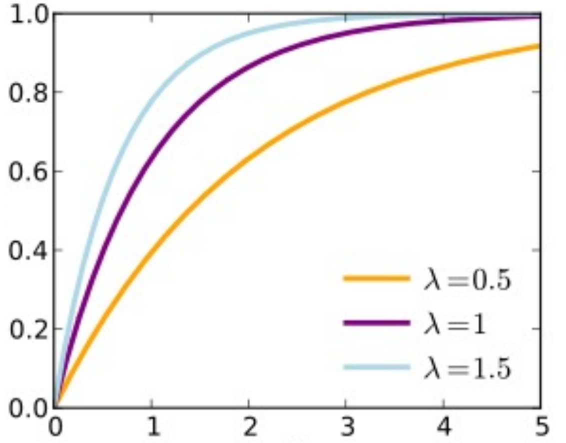
$$f_{Y|X}(y \mid 1) = \theta_1 e^{-\theta_1 y} \qquad p_X(1) = 0.1$$

- If your hard drive *fails after exactly t seconds of operation*, what is the probability it is defective?

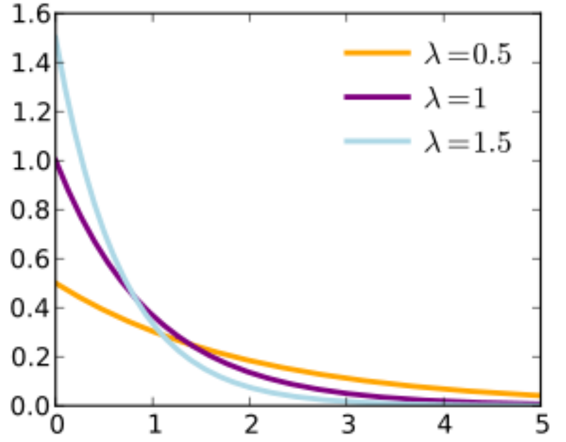
$$P(X = 1 \mid Y = t) = \frac{f_{Y|X}(y \mid 1) p_X(1)}{f_Y(y)}$$

$$= \frac{0.1\theta_1 e^{-\theta_1 t}}{0.1\theta_1 e^{-\theta_1 t} + 0.9\theta_0 e^{-\theta_0 t}}$$

Exponential Distributions:

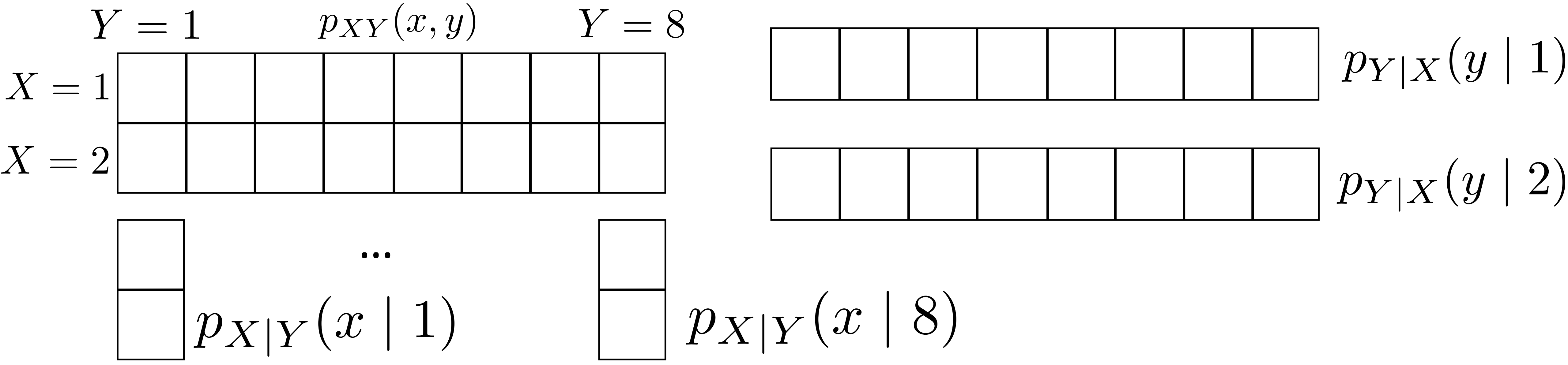


$$F_Y(y) = 1 - e^{-\theta y}$$



$$f_Y(y) = \theta e^{-\theta y}$$

Conditional Expectation



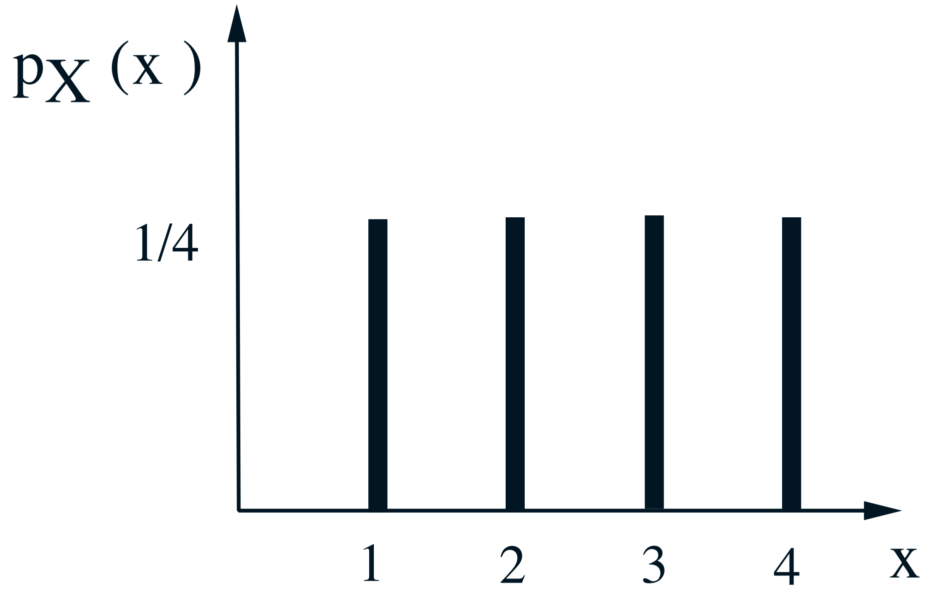
$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}$$

- Given that I observe *Y=y*, the *conditional expectation* of *X* equals

$$E[X \mid Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x \mid y)$$

- If *X* and *Y* are not independent, observing *Y=y* may change the mean of *X*

Conditional Expectation



$$E[X] = 2.5$$

Given $Y = \{X \geq 2\}$ is observed,

$$p_{X|Y}(x \mid y) =$$

$$E[X \mid Y] = 3$$

- Given that I observe *Y=y*, the *conditional expectation* of *X* equals

$$E[X \mid Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x \mid y)$$

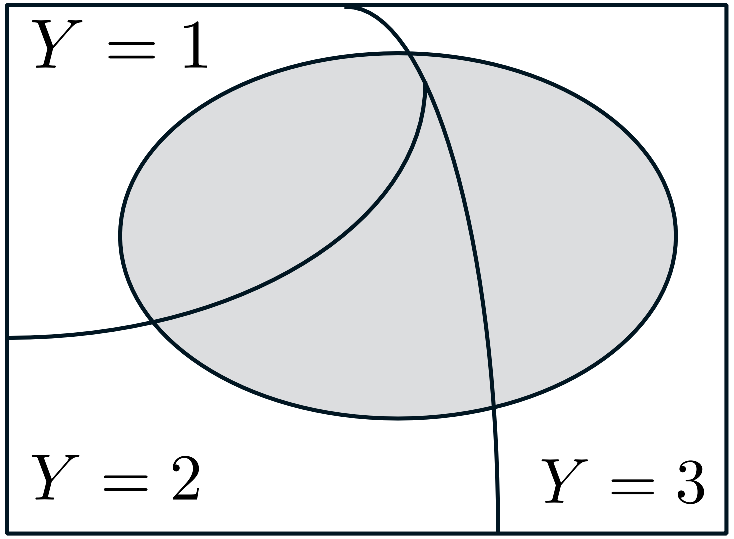
- If *X* and *Y* are not independent, observing *Y=y* may change the mean of *X*

Total Expectation Theorem

	$Y=1$		$Y=3$
$X=0$			
$X=1$			

$p_{XY}(x,y)$

Shaded where X=1,
Unshaded where X=0.



$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x,y)}{p_Y(y)} = \frac{p_{XY}(x,y)}{\sum_{x'} p_{XY}(x',y)}$$

- Applying the definitions of joint, marginal, and conditional distributions:

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x,y) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x \mid y) p_Y(y)$$

$$E[X] = \sum_{y \in \mathcal{Y}} p_Y(y) E[X \mid Y = y]$$

Mean is a weighted average of (possibly simpler) conditional means.

Monte-Carlo

The Weak Law of Large Number

$X_1, X_2, \ldots$ i.i.d.
finite mean μ and variance σ^2

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

sample mean or empirical mean

$$E[M_n] = \mu \qquad \mathrm{Var}[M_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- Chebyshev's inequality bounds distance between the true mean and the "empirical" or "sample" mean:

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) \ \leq \ \frac{\mathrm{Var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

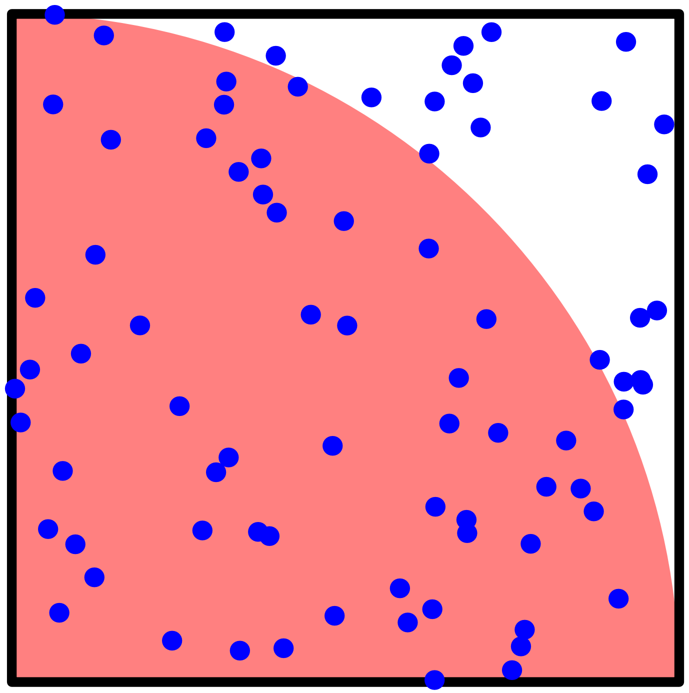
- *The empirical mean converges to the true mean in probability*

$$\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$$

- True even if variance not finite, but proof more challenging.

Monte-Carlo

1. For $i = 1$ to N
    1. Choose X and Y uniformly at random from $[0, 1]$
    2. If $X^2 + Y^2 \leq 1$ then $Z_i = 1$ else $Z_i = 0$.
2. $Z = \sum_{i=1}^{N} 4Z_i$
3. $S = \frac{1}{N} \sum_{i=1}^{N} 4Z_i$



Z_i is a 0-1 r.v. with $Pr(Z_i = 1) = \frac{\pi}{4}$.

$$E[Z_i] = \frac{\pi}{4} \qquad Var[Z_i] = \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)$$

How good is this estimate?

Chebyshev's Inequality:

Theorem

For **any** *random variable* X, *and any* $a > 0$,

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}.$$

$$E[S] = \frac{1}{N} E[4Z_i] = \pi,$$

$$Var[4Z_i] \leq 16, \text{ since } 0 < 4Z_i < 4. \ Var[S] = \frac{16}{N}$$

$$Pr(|S - \pi| \geq \epsilon) \leq \frac{16}{N\epsilon^2}$$

For $N \geq 128,000$,

$$Pr(|S - \pi| \geq 0.05) \leq 0.05$$

How Good is the Estimate?

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_n$ be independent random variables such that for all $1 \le i \le n$, $E[X_i] = \mu$ and $Pr(a \le X_i \le b) = 1$. Then

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \ge \epsilon) \le 2e^{-2n\epsilon^2/(b-a)^2}$$

$E[S] = \frac{1}{N}\sum_{i=1}^{N} E[4Z_i] = \pi$, and $0 \le 4Z_i \le 4$

$P(|S - \pi| \ge \epsilon) \le 2e^{-2n\epsilon^2/4^2}$

For $\epsilon = \sqrt{\frac{8\ln(2/\delta)}{n}}$, $\quad P(|S - \pi| \ge \epsilon) \le \delta$

For $n = 12,000$, $\quad P(|S - \pi]| \ge 0.05) \le 0.05$

Monte-Carlo

$$E[g] = \int g(x) f_X(x)\, dx$$

For many complex models, integral is intractable but we can still:

- *Simulate* the target distribution: $P(X_i \leq x_i) = F_X(x_i)$
- *Evaluate* the target function: $g_i = g(x_i)$

A *Monte Carlo method* uses computer simulation to approximate:

$$E[g] \approx \frac{1}{n} \sum_{i=1}^{n} g(x_i) = M_n \qquad P(X_i \leq x_i) = F_X(x_i)$$

Selecting $x_1, \ldots, x_n$ according to the distribution $F_X(x)$