HUST
ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Expectations and parameter estimation

Dam Quang Tuan

Expectation

- Expectation of a random variable tells the expected or average value it takes
- Expectation of a discrete random variable $X \in S_X$ having PMF $p(X)$

$$\mathbb{E}[X] = \sum_{x \in S_X} x p(x)$$

Probability that $X = x$

- Expectation of a continuous random variable $X \in S_X$ having PDF $p(X)$

$$\mathbb{E}[X] = \int_{x \in S_X} x p(x) dx$$

Probability density at $X = x$

Note that this exp. is w.r.t. the distribution $p(f(X))$ of the r.v. $f(X)$

- The definition applies to functions of r.v. too (e.g.., $\mathbb{E}[f(X)]$)

Often the subscript is omitted but do keep in mind the underlying distribution

- Exp. is always w.r.t. the prob. dist. $p(X)$ of the r.v. and often written as $\mathbb{E}_p[X]$

Expectation: intuition

- Recall the following probability distribution of patient arrivals:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

$$\sum_{i=1}^{5} x_i\, p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$

Non-trivial discrete probabilities

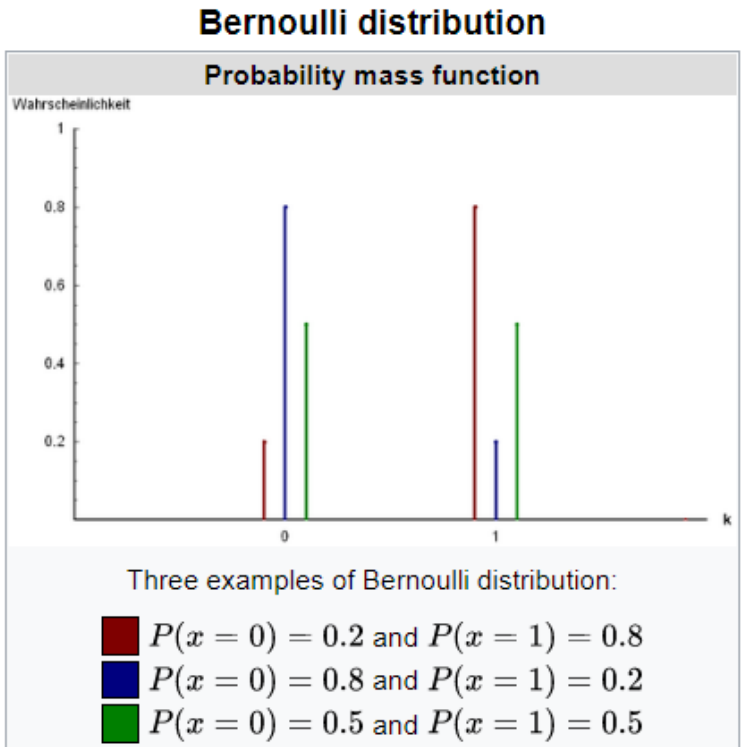
Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Probability **Bernoulli distribution**

A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial — a random experiment that has only two outcomes (usually called a “Success” or a “Failure”). For example, the probability of getting a heads (a “success”) while flipping a coin is 0.5. The probability of “failure” is 1 – P (1 minus the probability of success, which also equals 0.5 for a coin toss). It is a special case of the binomial distribution for n = 1. In other words, it is a binomial distribution with a single trial (e.g. a single coin toss).

The probability mass function f of this distributio

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$ [2]



Probability **Bernoulli distribution**

A Bernoulli trial is one of the simplest experiments you can conduct. It's an experiment where you can have one of two possible outcomes. For example, "Yes" and "No" or "Heads" and "Tails." A few examples:

Coin tosses: record how many coins land heads up and how many land tails up.
Births: how many boys are born and how many girls are born each day.
Rolling Dice: the probability of a roll of two die resulting in a double six.

What are the 3 conditions for a Bernoulli trial?
The 3 conditions for a Bernoulli trial are:
1. Each trial has only two possible outcomes: True/False, Yes/No, Success/Failure, etc.
2. The trials are independent. They do not influence each other.
3. The probabilities of success and failure do not change. They remain the same for all trials.

Probability: Binomial distribution

- An event *A* has probability *p* of occurring in a single trial. Find the probability that *A* occurs exactly *k* times in determined location, $k \leq n$ in *n* trials.

$$P_0(\omega) = \underbrace{P(A)P(A)\cdots P(A)}_{k} \quad \underbrace{P(\overline{A})P(\overline{A})\cdots P(\overline{A})}_{n-k}$$

$$= p^k q^{n-k}.$$

- P{ A occurs exactly *k* time in *n* trials } = $C^k_n p^k q^{n-k}$
  - Bernoulli formula.

Probability: Binomial distribution

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

$$n_{C_x} p^x (1-p)^x$$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

Probability: Binomial distribution

- Let *X* represent a Binomial r.v, then we have
- (1) $$P(k_1 \le X \le k_2) = \sum_{k=k_1}^{k_2} P_n(k) = \sum_{k=k_1}^{k_2} \binom{n}{k} p^k q^{n-k}.$$

- Since the binomial coefficient $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ grows quite rapidly with *n*, it is difficult to compute probability P for large *n*. In this context, two approximations are extremely useful.

A discrete distribution: binomial

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called "success" and "failure"
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin

Binomial distribution

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this exact arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails) $=(1/2)^3 \times (1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1 \times (1/2)^3 \times (1/2)^1 = (1/2)^3 \times (1/2)^2$

Binomial distribution

In fact, $(1/2)^3 \times (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 +$ ….. for as many unique arrangements as there are—but how many are there??

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

$_5C_3 = 5!/3!2! = 10$

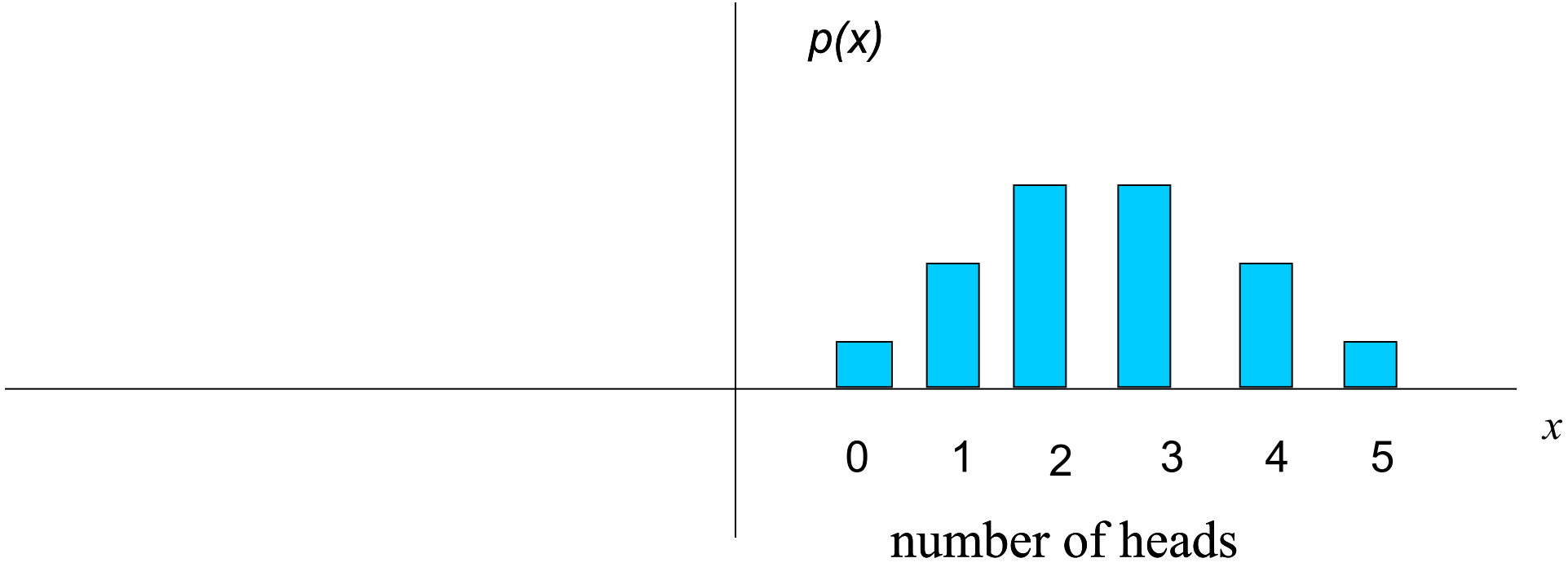
Outcome	Probability
THHHT	$(1/2)^3 \times (1/2)^2$
HHHTT	$(1/2)^3 \times (1/2)^2$
TTHHH	$(1/2)^3 \times (1/2)^2$
HTTHH	$(1/2)^3 \times (1/2)^2$
HHTTH	$(1/2)^3 \times (1/2)^2$
HTHHT	$(1/2)^3 \times (1/2)^2$
THTHH	$(1/2)^3 \times (1/2)^2$
HTHTH	$(1/2)^3 \times (1/2)^2$
HHTHT	$(1/2)^3 \times (1/2)^2$
THHTH	$(1/2)^3 \times (1/2)^2$
10 arrangements	$\times (1/2)^3 \times (1/2)^2$

The probability of each unique outcome (note: they are all equal)

$$\therefore \text{P(3 heads and 2 tails)} = \binom{5}{3} \times P(heads)^3 \times P(tails)^2 = 10 \times (½)^5 = 31.25\%$$

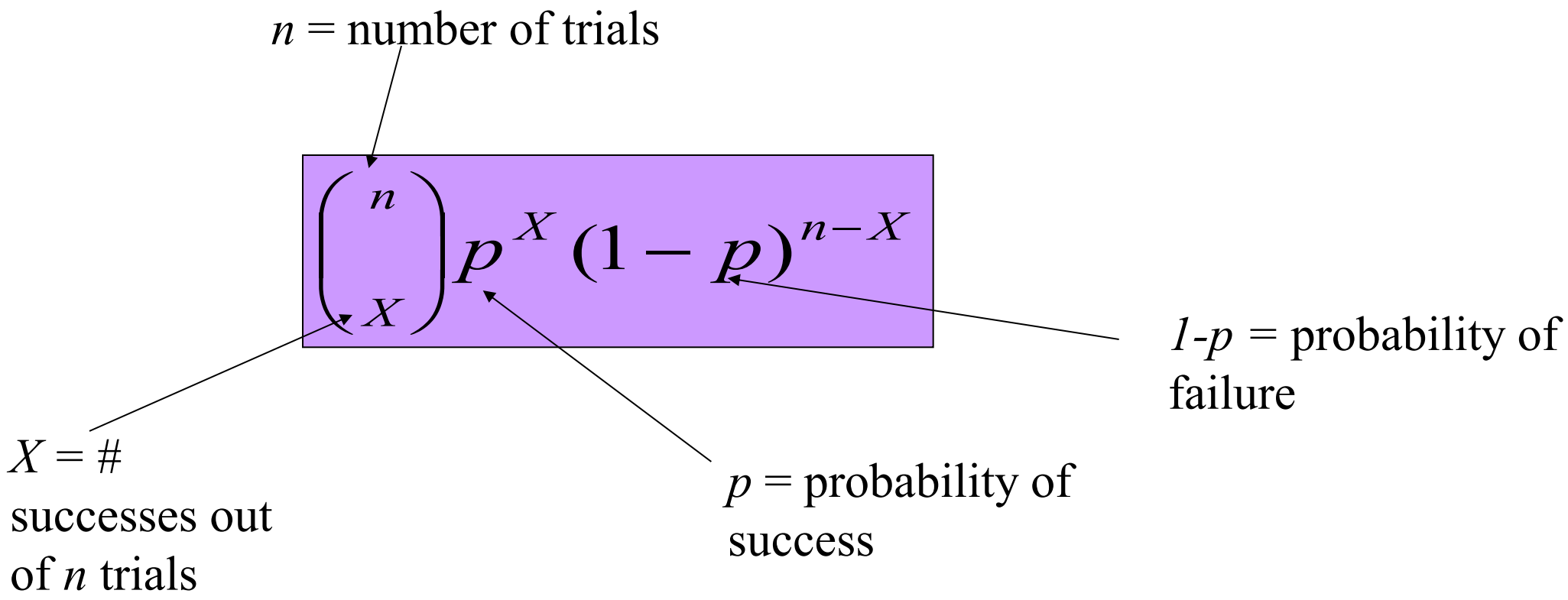
Binomial distribution function:

X= the number of heads tossed in 5 coin tosses



Binomial distribution, generally

Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in *n* independent trials, then the probability of exactly *X* "successes"=



Probability

Bernoulli Distribution	Binomial Distribtuion
Bernoulli distribution is used when we want to model the outcome of a single trial of an event.	If we want to model the outcome of multiple trials of an event, Binomial distribution is used.
It is represented as X ~ Bernoulli (p). Here, p is the probability of success.	It is denoted as X ~ Binomial (n, p). Where n is the number of trials.
Mean, E[X] = p	Mean, E[X] = np
Variance, Var[X] = p(1-p)	Variance, Var[X]= np(1-p)
Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then the Bernoulli distribution can be used to model the passing or failing in such an exam.	Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then if we want to find the probability that a student will pass in exactly 4 out of 5 exams, we use the Binomial Distribution.

Common Probability Distributions

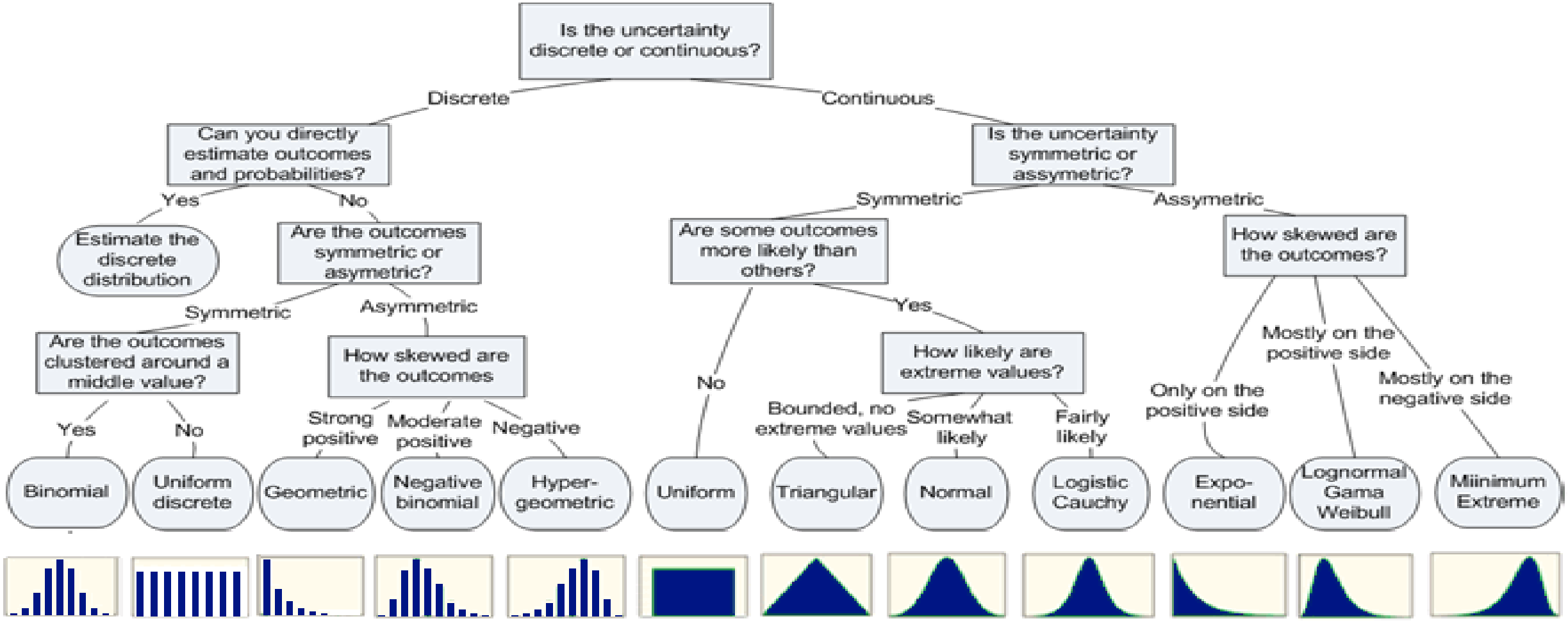
Important: We will use these extensively to model data as well as parameters of models

- Some common discrete distributions and what they can model
  - **Bernoulli:** Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
  - **Binomial:** Bounded non-negative integers, e.g., # of heads in n coin tosses
  - **Multinomial/multinoulli:** One of K (>2) possibilities, e.g., outcome of a dice roll
  - **Poisson:** Non-negative integers, e.g., # of words in a document
- Some common continuous distributions and what they can model
  - **Uniform:** numbers defined over a fixed range
  - **Beta:** numbers between 0 and 1, e.g., probability of head for a biased coin
  - **Gamma:** Positive unbounded real numbers
  - **Dirichlet:** vectors that sum of 1 (fraction of data points in different clusters)
  - **Gaussian:** real-valued numbers or real-valued vectors

The Law of Large Numbers

And

Central Limit Theorem


Is the uncertainty discrete or continuous?
Discrete
Continuous
Can you directly estimate outcomes and probabilities?
Is the uncertainty symmetric or assymetric?
Yes
No
Symmetric
Assymetric
Estimate the discrete distribution
Are the outcomes symmetric or asymetric?
Are some outcomes more likely than others?
How skewed are the outcomes?
Symmetric
Asymmetric
Yes
No
Are the outcomes clustered around a middle value?
How skewed are the outcomes
How likely are extreme values?
Only on the positive side
Mostly on the positive side
Mostly on the negative side
Yes
No
Strong positive
Moderate positive
Negative
Bounded, no extreme values
Somewhat likely
Fairly likely
Binomial
Uniform discrete
Geometric
Negative binomial
Hyper-geometric
Uniform
Triangular
Normal
Logistic Cauchy
Expo-nential
Lognormal Gama Weibull
Miinimum Extreme

Expectation: A Few Rules

X and Y need not be even independent. Can be discrete or continuous

- Expectation of sum of two r.v.'s: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Proof is as follows
  - Define $Z = X + Y$

$$\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot p(Z = z) \qquad \text{s.t. } z = x + y \text{ where } x \in S_X \text{ and } y \in S_Y$$

$$= \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) \cdot p(X = x, Y = y)$$

$$= \sum_x \sum_y x \cdot p(X = x, Y = y) + \sum_x \sum_y y \cdot p(X = x, Y = y)$$

$$= \sum_x x \sum_y p(X = x, Y = y) + \sum_y y \sum_x p(X = x, Y = y)$$

$$= \sum_x x \cdot p(X = x) + \sum_y y \cdot p(Y = y)$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Used the rule of marginalization of joint dist. of two r.v.'s

Expectation: A Few Rules (Contd)

- Expectation of a scaled r.v.: $\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$
- Linearity of expectation: $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$
- (More General) Lin. of exp.: $\mathbb{E}[\alpha f(X) + \beta g(Y)] = \alpha\mathbb{E}[f(X)] + \beta\mathbb{E}[g(Y)]$
- Exp. of product of two independent r.v.'s: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Law of the Unconscious Statistician (LOTUS): Given an r.v. X with a known prob. dist. $p(X)$ and another random variable $Y = g(X)$ for some function g

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{y \in S_Y} y p(y) = \sum_{x \in S_X} g(x) p(x)$$

- Rule of iterated expectation: $\mathbb{E}_{p(X)}[X] = \mathbb{E}_{p(Y)}[\mathbb{E}_{p(X|Y)}[X|Y]]$

α is a real-valued scalar

α and β are real-valued scalars

f and g are arbitrary functions.

Requires finding $p(Y)$

Requires only $p(X)$ which we already have

Variance and Covariance

- Variance of a scalar r.v. tells us about its spread around its mean value $\mathbb{E}[X] = \mu$

$$\text{var}[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2$$

- Standard deviation is simply the square root is variance
- For two scalar r.v.'s X and Y, the covariance is defined by

$$\text{cov}[X, Y] = \mathbb{E}[\{X - \mathbb{E}[X]\}\{Y - \mathbb{E}[Y]\}] = \mathbb{E}[XY] - \mathbb{E}[\text{X}]\mathbb{E}[\text{Y}]$$

- For two vector r.v.'s X and Y (assume column vec), the covariance matrix is defined by

$$\text{cov}[X, Y] = \mathbb{E}[\{X - \mathbb{E}[X]\}\{Y^\top - \mathbb{E}[Y^\top]\}] = \mathbb{E}[XY^\top] - \mathbb{E}[\text{X}]\mathbb{E}[Y^\top]$$

- Cov. of components of a vector r.v. X: $\text{cov}[X] = \text{cov}[X, X]$
- Note: The definitions apply to functions of r.v. too (e.g., $\text{var}[f(X)]$)
- Note: Variance of sum of independent r.v.'s: $\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$

Transformation of Random Variables

- Suppose $Y = f(X) = AX + b$ be a linear function of a vector-valued r.v. X (A is a matrix and b is a vector, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\text{cov}[X] = \Sigma$, then for the vector-valued r.v. Y

$$\mathbb{E}[Y] = \mathbb{E}[AX + b] = A\mu + b$$

$$\text{cov}[Y] = \text{cov}[AX + b] = A\Sigma A^{\top}$$

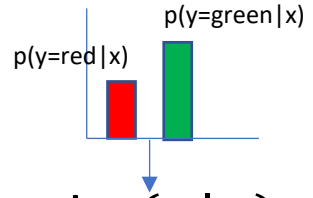
- Likewise, if $Y = f(X) = a^{\top}X + b$ be a linear function of a vector-valued r.v. X (a is a vector and b is a scalar, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\text{cov}[X] = \Sigma$, then for the scalar-valued r.v. Y

$$\mathbb{E}[Y] = \mathbb{E}[a^{\top}X + b] = a^{\top}\mu + b$$

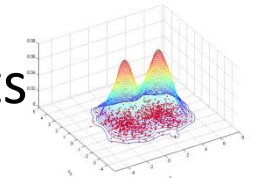
$$\text{var}[Y] = \text{var}[a^{\top}X + b] = a^{\top}\Sigma a$$

Probabilistic ML: Some Motivation

- In many ML problems, we want to model and reason about data probabilistically
- At a high-level, this is the density estimation view of ML, e.g.,



  - Given input-output pairs $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$ estimate the conditional $p(y|\boldsymbol{x})$
  - Given inputs $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, estimate the distribution $p(\boldsymbol{x})$ of the inputs



  - Note 1: These dist. will depend on some parameters θ (to be estimated), and written as

$$p(y|\boldsymbol{x}, \theta) \quad \text{or} \quad p(\boldsymbol{x}|\theta)$$

  - Note 2: These dist. sometimes assumed to have a specific form, but sometimes not
- Assuming the form of the distribution to be known, the goal in estimation is to use the observed data to estimate the parameters of these distributions

Probabilistic Modeling: The Basic Idea

- Assume N observations $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$, generated from a presumed prob. model

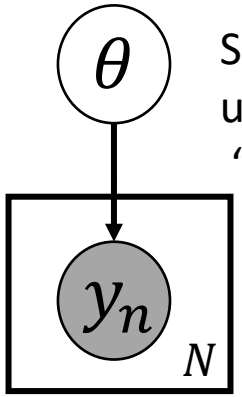
$$y_n \sim p(y|\theta) \qquad \forall n$$

(assumed independently & identically distributed (i.i.d.))

- Here $p(y|\theta)$ is a conditional distribution, conditioned on params θ (to be learned)
  - Note: θ may be fixed unknown or an unknown random variable (we will study both cases)

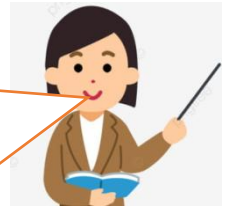
$y_n \sim p(y|\theta) \quad \forall n$

The parameters θ may themselves depend on other unknown/known parameters (called hyperparameters), which may depend on other unknowns, and so on. ☺ This is essentially "hierarchical" modeling (will see various examples later)



Such diagrams are usually called the "plate notation"

The Predictive dist. tells us how likely each possible value of a new observation y_* is. Example: if y_* denotes the outcome of a coin toss, then what is $p(y_* = "head"|\boldsymbol{y})$, given N previous coin tosses $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$



- Some of the tasks that we may be interested in
  - Parameter estimation: Estimating the unknown parameters θ (and other unknowns θ depends on)
  - Prediction: Estimating the **predictive distribution** of new data, i.e., $p(y_*|\boldsymbol{y})$ - this is also a conditional distribution (conditioned on past data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$, as well as θ and other

Parameter Estimation in Probabilistic Models

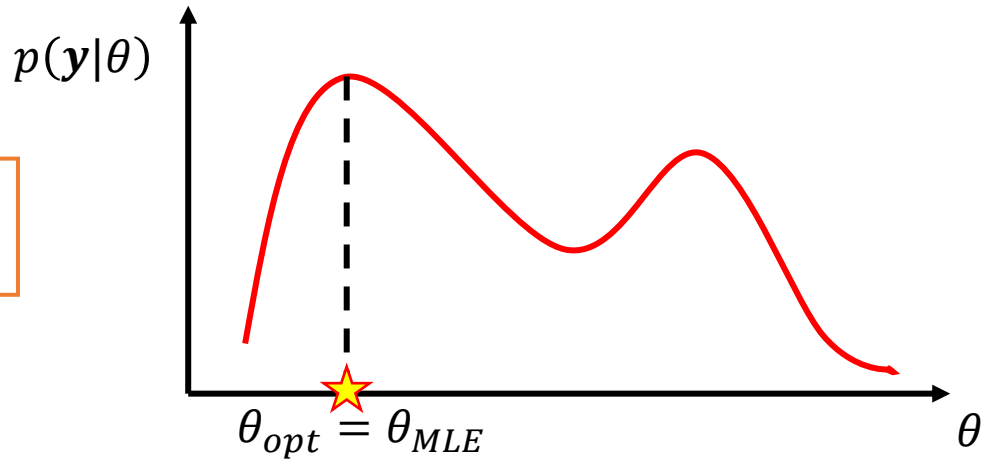
- Since data is assumed to be i.i.d., we can write down its total probability as

$$p(\boldsymbol{y}|\theta) = p(y_1, y_2, \ldots, y_N|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$

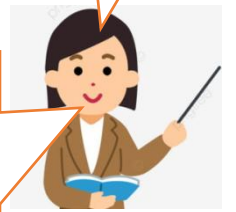
This now is an optimization problem essentially (θ being the unknown)

- $p(\boldsymbol{y}|\theta)$ called "likelihood" - probability of observed data as a function of params θ

How do I find the best θ ?



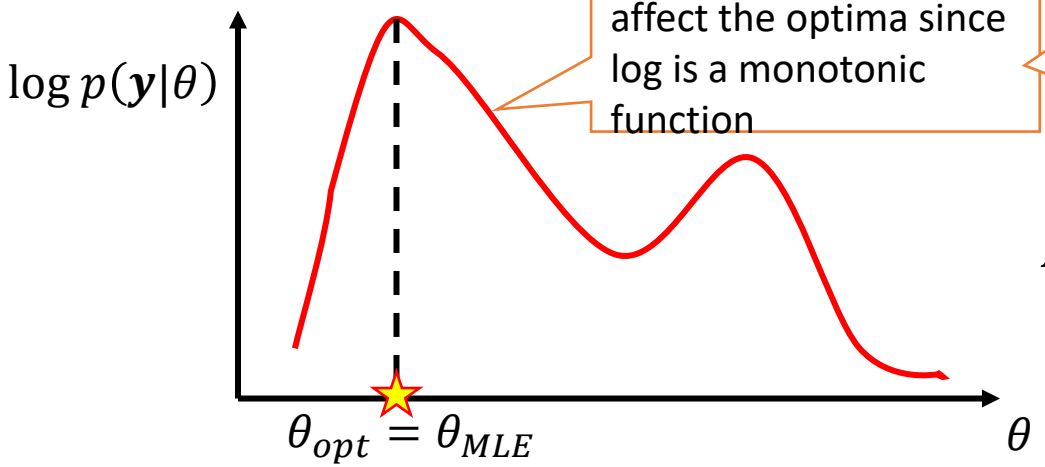
Well, one option is to find the θ that maximizes the likelihood (probability of the observed data) – basically, which value of θ makes the observed data most likely to have come from the assumed distribution $p(y|\theta)$ --- Maximum Likelihood Estimation (MLE)



- In parameter estimation, the goal is to find the "best" θ, given observed data $\boldsymbol{y}$
- Note: Instead of finding single best, sometimes may be more informative to learn a distribution for θ (can tell us about uncertainty in our estimate of θ – more later)

Maximum Likelihood Estimation (MLE)

- The goal in MLE is to find the optimal θ by maximizing the likelihood
- In practice, we maximize the log of the likelihood (log-likelihood in short)



Taking log doesn't affect the optima since log is a monotonic function

Leads to simpler algebra/calculus, and also yields better numerical stability when implementing it on computer (dealing with log of probabilities)

$$LL(\theta) = \log p(\boldsymbol{y}|\theta) = \log \prod_{n=1}^{N} p(y_n|\theta)$$

$$= \sum_{n=1}^{N} \log p(y_n|\theta)$$

- Thus the MLE problem is

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, LL(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(y_n|\theta)$$

- This is now an optimization (maximization problem). Note: θ may have constraints

Maximum Likelihood Estimation (MLE)

Negative Log-Likelihood (NLL)

- The MLE problem can also be easily written as a minimization problem

$$\theta_{MLE} = \operatorname{argmax}_{\theta} \sum_{n=1}^{N} \log p(y_n|\theta) = \operatorname{argmin}_{\theta} \sum_{n=1}^{N} -\log p(y_n|\theta)$$

- Thus MLE can also be seen as minimizing the negative log-likelihood (NLL)

$$\theta_{MLE} = \operatorname{argmin}_{\theta} NLL(\theta)$$

Indeed. It may overfit. Several ways to prevent it: Use regularizer or other strategies to prevent overfitting. Alternatives, use "prior" distributions on the parameters θ that we are trying to estimate (which will kind of act as a regularizer as we will see shortly)

Such priors have various other benefits as we will see later

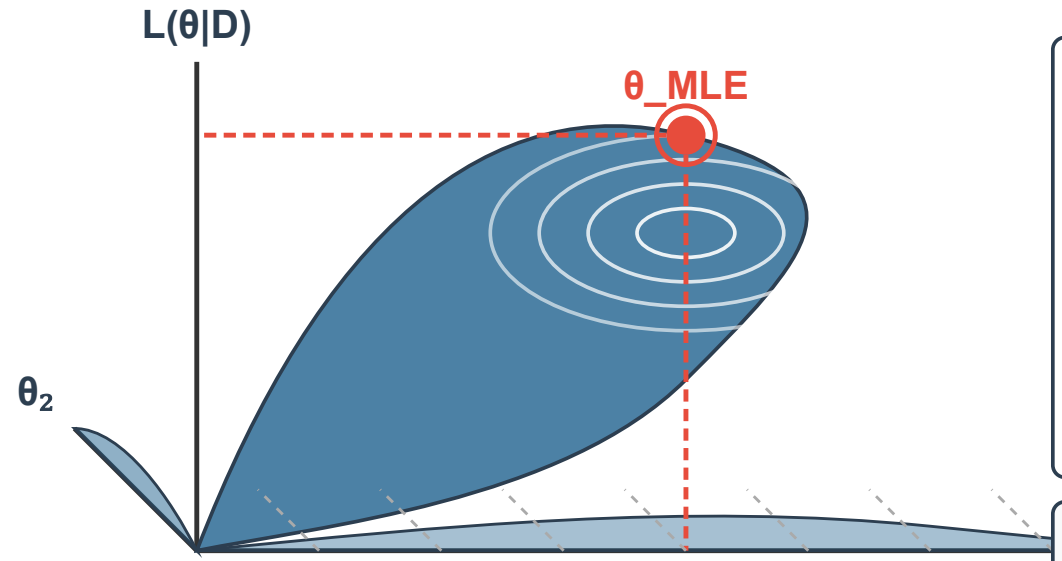
- NLL is analogous to a loss function
  - The negative log-lik $(-\log p(y_n|\theta))$ is akin to the loss on each data point
- Thus doing MLE is akin to minimizing training loss

Does it mean MLE could overfit? If so, how to prevent this?

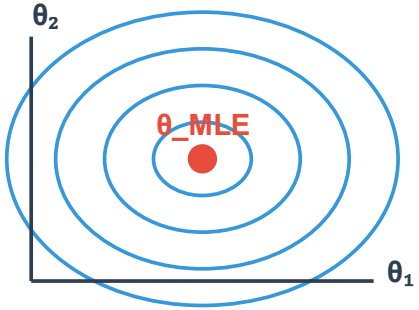
Maximum Likelihood Estimation (MLE)

Geometric Interpretation of MLE

Finding the parameter values that maximize the likelihood function



Contour View (Top-Down)



MLE Optimization:

- Start with initial guess
- Follow gradient uphill
- Stop at maximum point

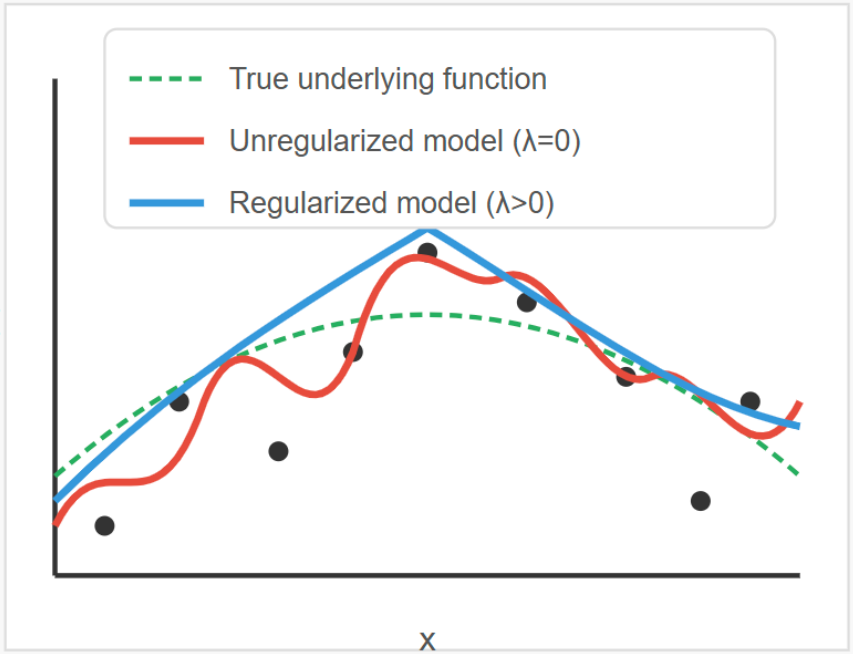
Key Insights:

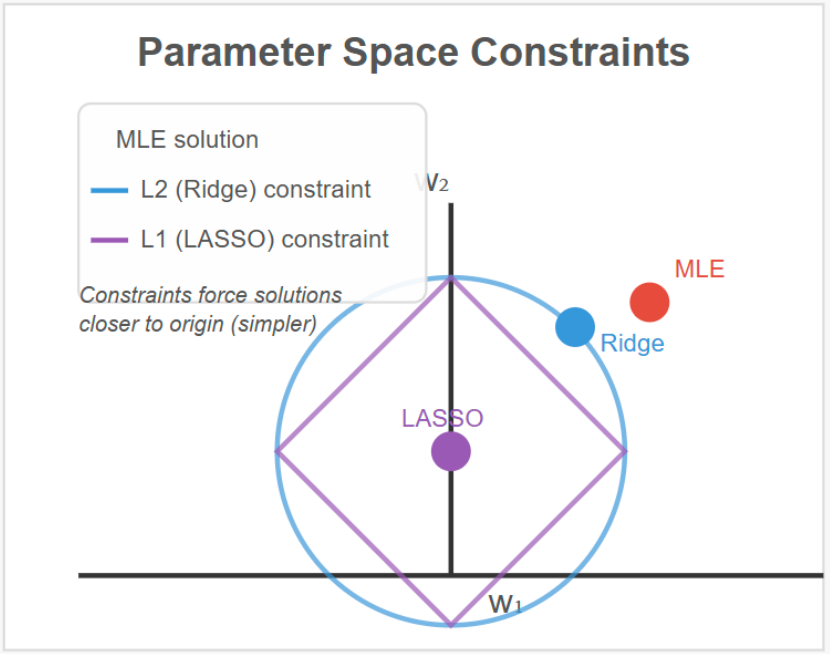
- MLE finds the parameter values (θ_1, θ_2) at the peak of the likelihood surface
- Mathematically equivalent to finding where $\nabla L(\theta|D) = 0$ or minimizing -log $L(\theta|D)$

Maximum Likelihood Estimation (MLE)

Regularization Solutions for MLE

Controlling complexity to prevent overfitting





Regularized MLE Objective Functions:

Ridge: $\theta\_MLE = \text{argmin} [NLL(\theta) + \lambda\|\theta\|^2_2]$

LASSO: $\theta\_MLE = \text{argmin} [NLL(\theta) + \lambda\|\theta\|_1]$

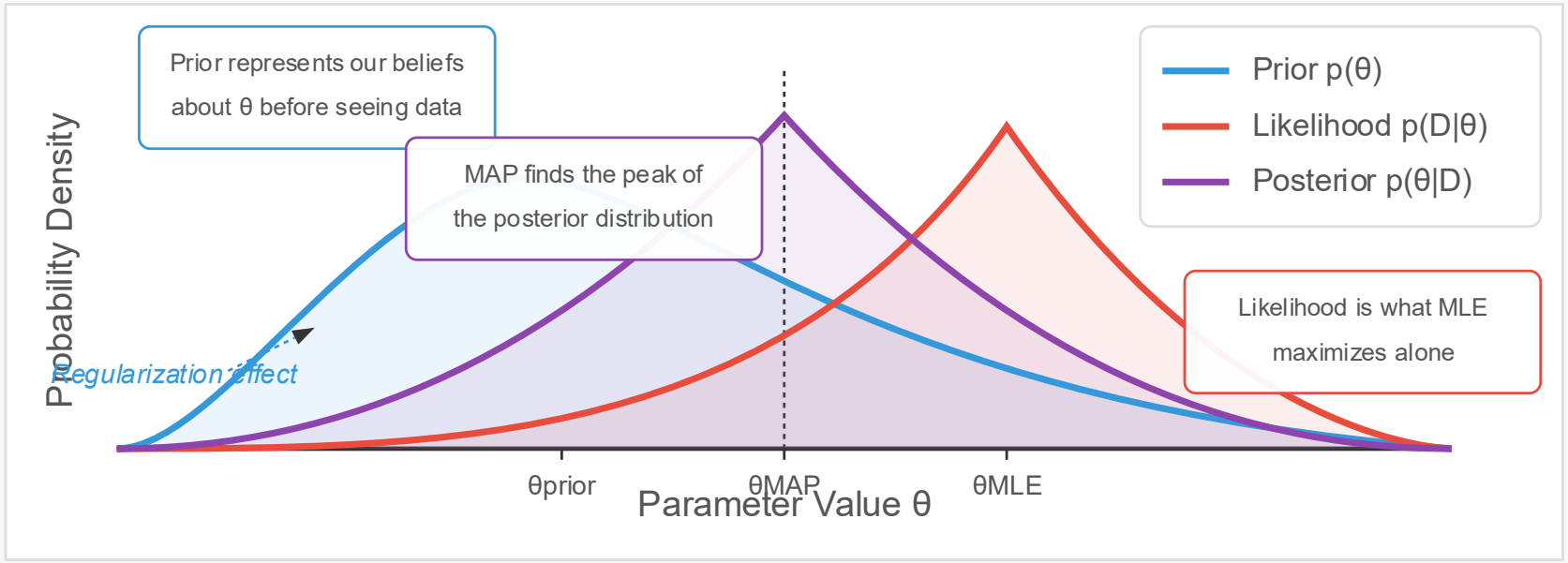
Elastic Net: $NLL(\theta) + \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|^2_2$

λ controls the strength of regularization: larger values = simpler models

Maximum Likelihood Estimation (MLE)

Bayesian Perspective on MLE

From MLE to Maximum A Posteriori (MAP) Estimation



Bayes' Theorem

$p(\theta|D) \propto p(D|\theta) \times p(\theta)$

Posterior ∝ Likelihood × Prior

MAP estimation with a Gaussian prior is equivalent to L2 regularized MLE

MLE: An Example

- Consider a sequence of N coin toss outcomes (observations)
- Each observation y_n is a binary random variable. Head: $y_n = 1$, Tail: $y_n = 0$
- Each y_n is assumed generated by a **Bernoulli distribution** with param $\theta \in (0,1)$

Probability of a head

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Here θ the unknown param (probability of head). Want to estimate it using MLE
- Log-likelihood: $\sum_{n=1}^{N} \log p(y_n|\theta) = \sum_{n=1}^{N} [y_n \log \theta + (1-y_n)\log(1-\theta)]$
- Maximizing log-lik (or minimizing NLL) w.r.t. θ will give a closed form expression

Take deriv. set it to zero and solve. Easy optimization

I tossed a coin 5 times – gave 1 head and 4 tails. Does it means θ = 0.2?? The MLE approach says so. What is I see 0 head and 5 tails. Does it mean θ = 0?

$$\theta_{MLE} = \frac{\sum_{n=1}^{N} y_n}{N}$$

Thus MLE solution is simply the fraction of heads! Makes intuitive sense!

Indeed – if you want to trust MLE solution. But with small number of training observations, MLE may overfit and may not be reliable. We will soon see better alternatives that use prior distributions!

References

CS771: Intro to Machine Learning (Fall 2021), Nisheeth Srivastava, IIT Kanpur