# Getting Started: PML Basics

CS772A: Probabilistic Machine Learning

Piyush Rai

# Plan Today

- Quick refresher of basics of probability/statistics ideas for PML
  - Random variables
  - Probability distributions and their probabilities
  - Commonly used probability distributions
- Basics of probability modeling of data
  - Parameter estimation in probabilistic models
  - Prediction in probabilistic models

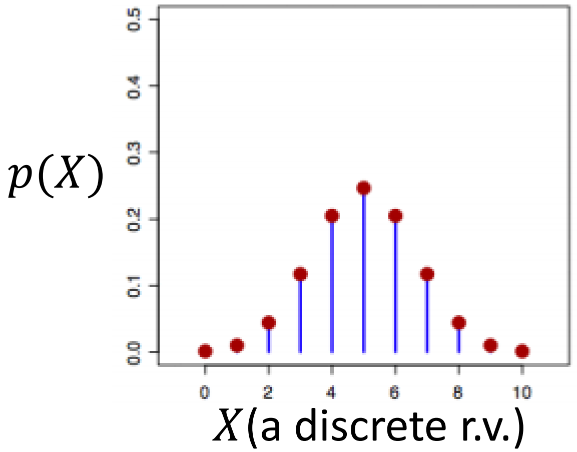
# Prob/Stats Refresher

# Random Variables

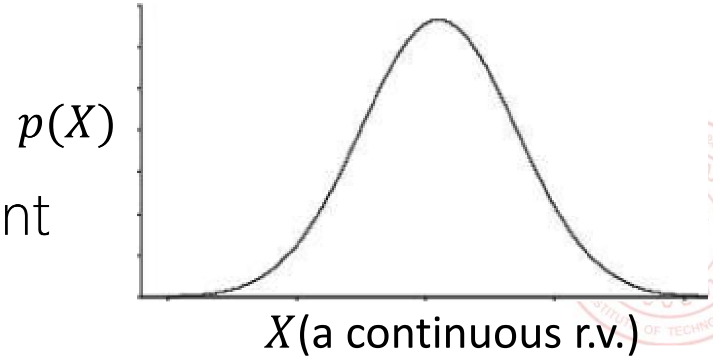
- Informally, a random variable (r.v.) $X$ denotes possible outcomes of an event

- Can be discrete (i.e., finite many possible outcomes) or continuous

- Some examples of discrete r.v.
  - $X \in \{0, 1\}$ denoting outcomes of a coin-toss
  - $X \in \{1, 2, \ldots, 6\}$ denoting outcome of a dice roll



- Some examples of continuous r.v.
  - $X \in (0, 1)$ denoting the bias of a coin
  - $X \in \mathbb{R}$ denoting heights of students in a class
  - $X \in \mathbb{R}$ denoting time to get to your hall from the department

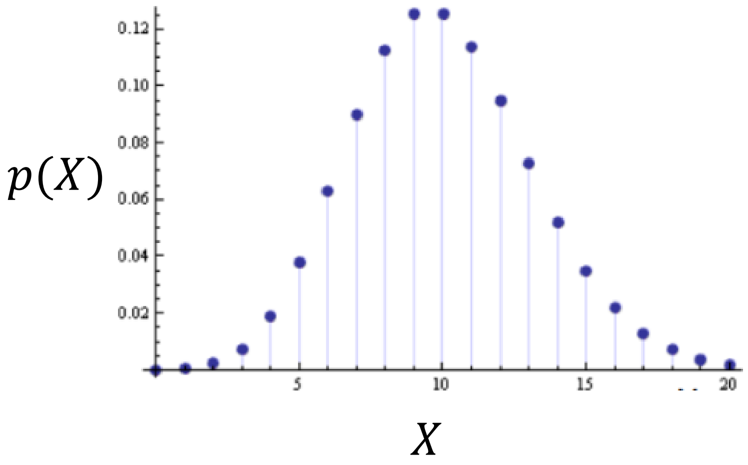


# Discrete Random Variables

- For a discrete r.v. $X$, $p(x)$ denotes $p(X = x)$ - probability that $X = x$

- $p(X)$ is called the probability mass function (PMF) of r.v. $X$
  - $p(x)$ or $p(X = x)$ is the value of the PMF at $x$

$$p(x) \geq 0$$
$$p(x) \leq 1$$
$$\sum_x p(x) = 1$$



# Continuous Random Variables

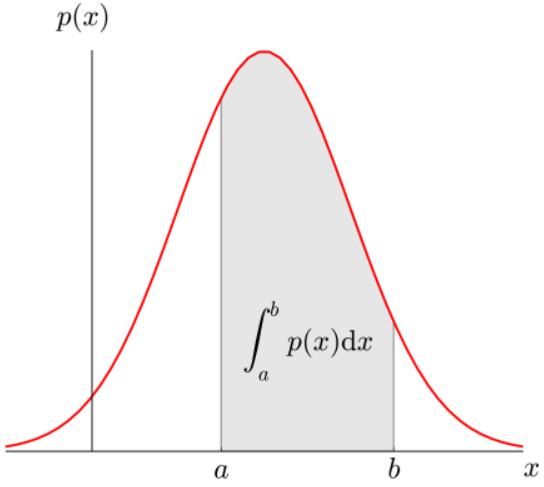
- For a continuous r.v. $X$, a *probability* $p(X = x)$ or $p(x)$ is meaningless
- For cont. r.v., we talk in terms of prob. within an interval $X \in (x, x + \delta x)$
  - $p(x)\delta x$ is the prob. that $X \in (x, x + \delta x)$ as $\delta x \to 0$
  - $p(x)$ is the probability density at $X = x$

Yes, probability density at a point $x$ can very well be larger than 1. The integral however must be equal to 1

$$p(x) \geq 0$$

~~$p(x) \leq 1$~~

$$\int p(x)dx = 1$$



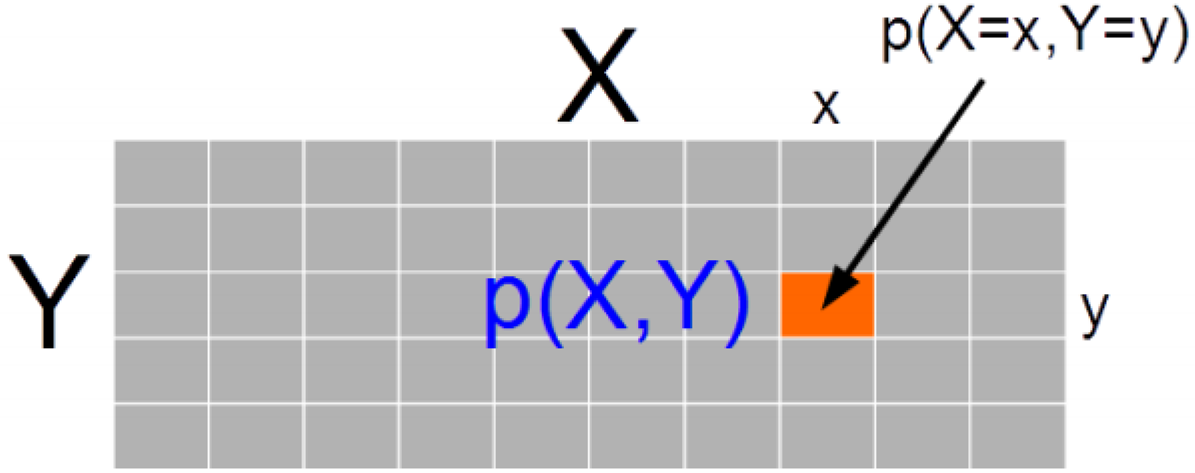
# A word about notation

- $p(.)$ can mean different things depending on the context
- $p(X)$ denotes the distribution (PMF/PDF) of an r.v. $X$
- $p(X = x)$ or $p_X(x)$ or simply $p(x)$ denotes the prob. or prob. density at value $x$
  - Actual meaning should be clear from the context (but be careful)
- Exercise same care when $p(.)$ is a specific distribution (Bernoulli, Gaussian, etc.)
- The following means generating a random sample from the distribution $p(X)$

$$x \sim p(X)$$

# Joint Probability Distribution

- Joint prob. dist. $p(X,Y)$ models probability of co-occurrence of two r.v. $X,Y$
- For discrete r.v., the joint PMF $p(X,Y\ )$ is like a table (that sums to 1)



For 3 r.v.'s, we will likewise have a "cube" for the PMF. For more than 3 r.v.'s too, similar analogy holds

$$\sum_x \sum_y p(X=x, Y=y) = 1$$

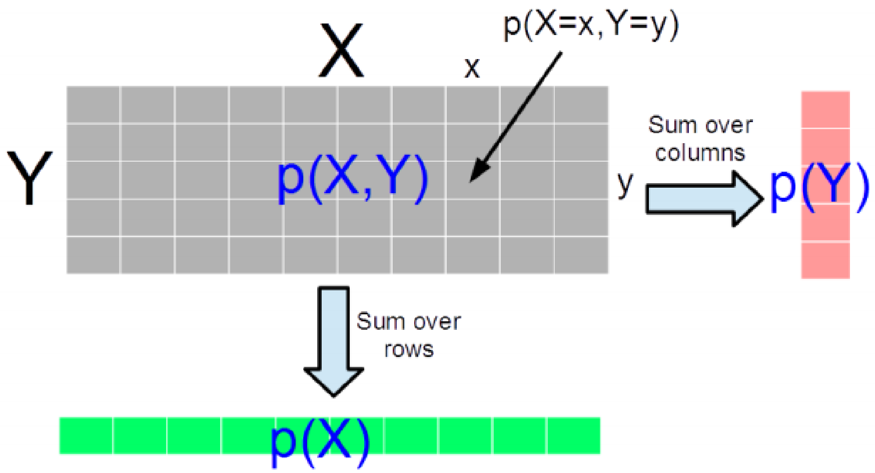
- For two continuous r.v.'s $X$ and $Y$, we have joint PDF $p(X,Y)$

$$\int_x \int_y p(X=x, Y=y)dxdy = 1$$

For more than two r.v.'s, we will likewise have a multi-dim integral for this property

# Marginal Probability Distribution

- Consider two r.v.'s X and Y (discrete/continuous – both need not of same type)
- Marg. Prob. is PMF/PDF of one r.v. accounting for all possibilities of the other r.v.
- For discrete r.v.'s, $p(X) = \sum_y p(X, Y = y)$ and $p(Y) = \sum_x p(X = x, Y)$
- For discrete r.v. it is the sum of the PMF table along the rows/columns



The definition also applied for two <u>sets</u> of r.v.'s and marginal of one set of r.v.'s is obtained by summing over all possibilities of the second set of r.v.'s

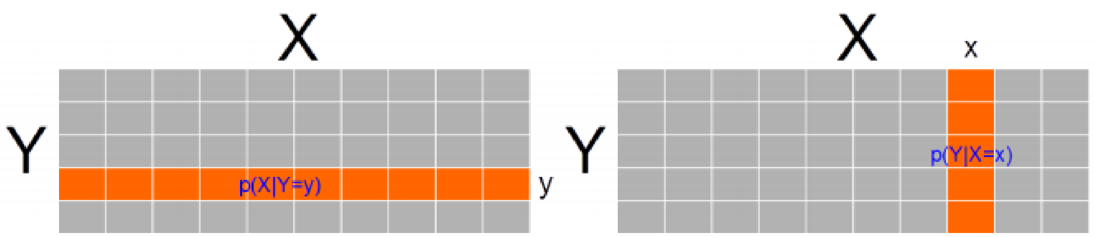
For discrete r.v.'s, marginalization is called summing over, for continuous r.v.'s, it is called "integrating out"

- For continuous r.v.'s, $p(X) = \int_y p(X, Y = y) dy, \quad p(Y) = \int_x p(X = x, Y) dx$

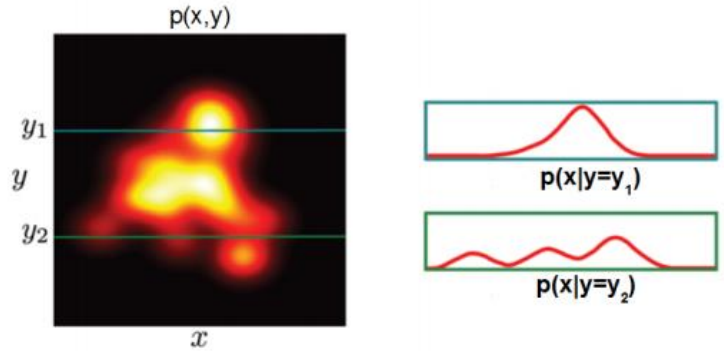
# Conditional Probability Distribution

- Consider two r.v.'s $X$ and $Y$ (discrete/continuous – both need not of same type)
- Conditional PMF/PDF $p(X|Y)$ is the prob. dist. of one r.v. $X$, fixing other r.v. $Y$
- $p(X|Y = y)$ or $p(Y|X = x)$ like taking a slice of the joint dist. $p(X,Y)$

Discrete Random Variables



Continuous Random Variables



- Note: A conditional PMF/PDF may also be conditioned on something that is not the value of an r.v. but some fixed quantity in general

We will see cond. dist. of output $\boldsymbol{y}$ given weights $\boldsymbol{w}$ (r.v.) and features $\boldsymbol{X}$ written as $p(y|w,X)$

# Some Basic Rules

- **Sum Rule:** Gives the marginal probability distribution from joint probability distribution

$$\text{For discrete r.v.: } p(X) = \textstyle\sum_Y p(X,Y)$$

$$\text{For continuous r.v.: } p(X) = \textstyle\int_Y p(X,Y)dY$$

- **Product Rule:** $p(X,Y) = p(Y|X)p(X) = p(X|Y)p(Y)$

- **Bayes' rule:** Gives conditional probability distribution (can derive it from product rule)

$$\boxed{p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}}$$

$$\text{For discrete r.v.: } p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$

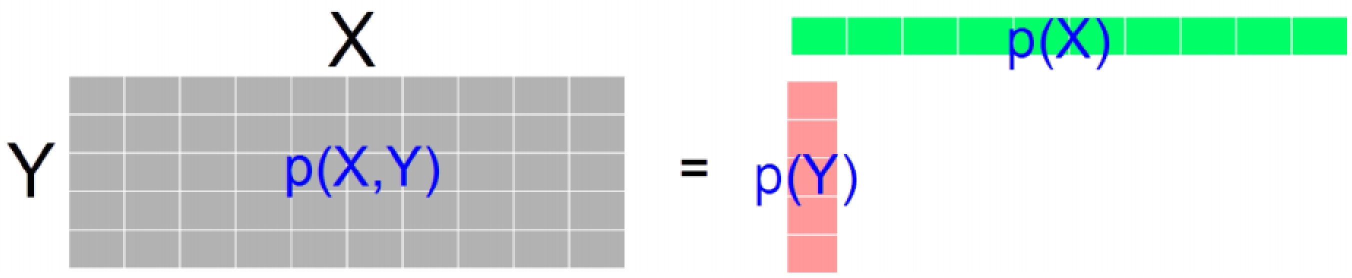
$$\text{For continuous r.v.: } p(Y|X) = \frac{p(X|Y)p(Y)}{\int_Y p(X|Y)p(Y)dY}$$

- **Chain Rule:** $p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2|X_1)\ldots p(X_N|X_1, \ldots, X_{N-1})$

# Independence

- $X$ and $Y$ are independent when knowing one tells nothing about the other

$$\begin{aligned} p(X|Y=y) &= p(X) \\ p(Y|X=x) &= p(Y) \\ p(X,Y) &= p(X)p(Y) \end{aligned}$$



- The above is the marginal independence $(X \perp\!\!\!\perp Y)$

- Two r.v.'s $X$ and $Y$ may not be marginally indep but may be given the value of another r.v. $Z$

$$p(X,Y|Z=z) = p(X|Z=z)p(Y|Z=z) \qquad X \perp\!\!\!\perp Y|Z$$

# 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: 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 $Z = X + Y$

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

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

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

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

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

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

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

# Expectation: A Few Rules (Contd)

$\alpha$ is a real-valued scalar

$\alpha$ and $\beta$ are real-valued scalars

$f$ and $g$ are arbitrary functions.

- 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 $g$

Requires finding $p(Y)$

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

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

LOTUS also applicable for continuous r.v.'s

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

# Variance and Covariance

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

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

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

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

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

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

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

Important result

# Entropy

- Entropy of a continuous/discrete distribution $p(X)$

$$
\begin{aligned}
H(p) &= -\int p(X)\log p(X)dX \\
H(p) &= -\sum_{k=1}^{K} p(X=k)\log p(X=k)
\end{aligned}
$$

- In general, a peaky distribution would have a smaller entropy than a flat distribution
- Note that the KL divergence can be written in terms of expetation and entropy terms

$$KL(p||q) = \mathbb{E}_{p(X)}[-\log q(X)] - H(p)$$

- Some other definition to keep in mind: conditional entropy, joint entropy, mutual information, etc.

# KL Divergence

- Kullback–Leibler divergence between two probability distributions $p(X)$ and $q(X)$

$$KL(p||q) \quad = \quad \int p(X)\log\frac{p(X)}{q(X)}dX = -\int p(X)\log\frac{q(X)}{p(X)}dX \qquad \text{(for continuous distributions)}$$

$$KL(p||q) \quad = \quad \sum_{k=1}^{K} p(X=k)\log\frac{p(X=k)}{q(X=k)} \qquad \text{(for discrete distributions)}$$

- It is non-negative, i.e., $KL(p||q) \geq 0$, and zero if and only if $p(X)$ and $q(X)$ are the same
- For some distributions, e.g., Gaussians, KL divergence has a closed form expression
- KL divergence is not symmetric, i.e., $KL(p||q) \neq KL(q||p)$

# Common Probability Distributions

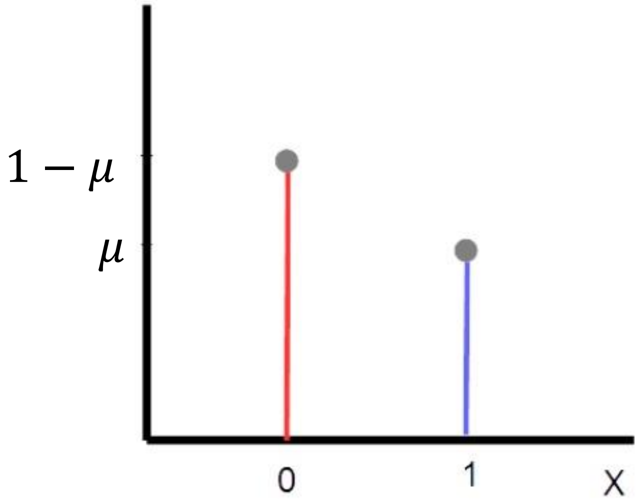
Important: We will use these extensively to model data as well as parameters of models

- 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 classes/clusters)
  - **Gaussian:** real-valued numbers or real-valued vectors

# Discrete Distributions

# Bernoulli Distribution

- Distribution over a binary random variable $X \in \{0,1\}$, e.g., outcome of a coin-toss
- Defined by probability parameter $\mu \in (0,1)$ s.t. $\mu = p(X = 1)$
- The probability mass function (PMF) of Bernoulli is



$$p(X = x|\mu) = \mu^x (1-\mu)^{1-x}$$

- Expectation: $\mathbb{E}[X] = \mu$

- Variance: var$[X] = \mu(1-\mu)$

# Binomial Distribution

- Distribution over number of successes $m$ in $N$ trials, e.g., number of heads in $N$ coin tosses

- Defined by a parameter $\mu \in (0,1)$, probability of success of each trial

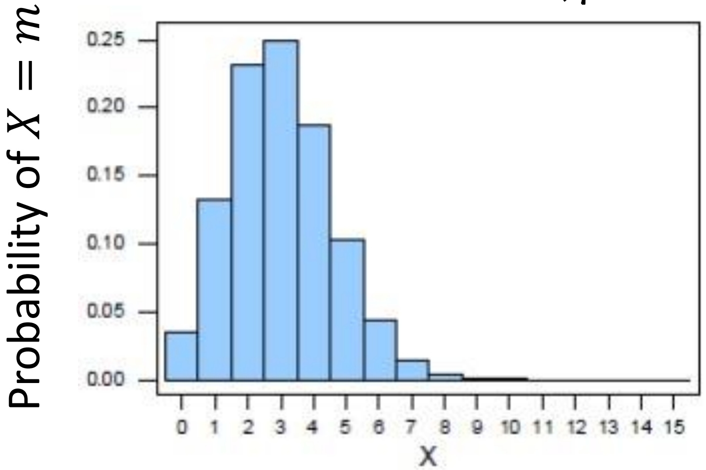
- The probability mass function (PMF) of Binomial is

$$p(X = m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Expectation: $\mathbb{E}[X] = N\mu$

- Variance: $\text{var}[X] = N\mu(1-\mu)$

Binomial with $N = 15, \mu = 0.2$



# Multinoulli Distribution

- Generalization of Bernoulli distribution for discrete/categorical variable $X$ taking one of $K > 2$ outcomes, e.g., outcome of a single dice roll

- Note: If $X = \boldsymbol{i}$, we can also use a one-hot vector of length $K$ to denote $X$

$$X = [0, 0, \dots, 0, 1, 0, \dots, 0, 0]$$

Vector of all zeros except the $i^{th}$ entry $x_i$ which is 1; all other $x_j$, for $j \neq i$ are 0

Probability of the $i^{th}$ outcome

- Multinoulli is defined by $K$ params $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_K]$, $\mu_i \in (0,1)$ and $\sum_{i=1}^{K} \mu_i = 1$

- The PMF of Multinoulli is

$$p(X|\boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{x_i}$$

- Expectation: $\mathbb{E}[x_i] = \mu_i$, variance: $\text{var}[x_i] = \mu_i(1 - \mu_i)$

# Multinomial Distribution

E.g., same dice rolled $N$ times

- Generalization of multinomial for a $K$ outcome trial repeated $N > 1$ times
- Defines distribution of random var. $X$ denoting counts of each possible outcome
- Can use a vector of length $K$ to denote $X$

The $i^{th}$ entry $x_i$ denotes the number of times we had outcome $i$

$\sum_{i=1}^{K} x_i = N$

$$X = [x_1, x_2, \dots, x_i, \dots, x_{K-1}, x_K]$$

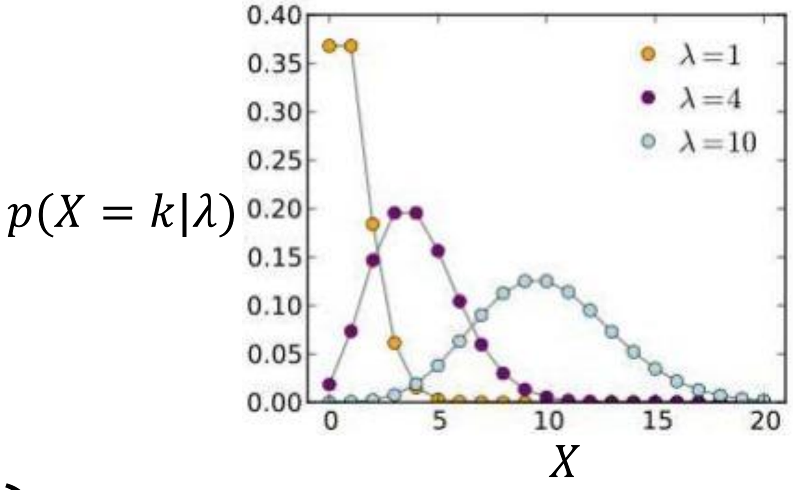
- Multinomial is defined by $K$ params $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_K]$, $\mu_i \in (0,1)$ and $\sum_{i=1}^{K} \mu_i = 1$
- The PMF of Multinomial is

$$p(X|\boldsymbol{\mu}) = \binom{N}{x_1, x_2, \dots, x_K} \prod_{i=1}^{K} \mu_i^{x_i}$$

- Expectation: $\mathbb{E}[x_i] = N\mu_i$, variance: $\text{var}[x_i] = N\mu_i(1 - \mu_i)$
- Multinomial can also be viewed as a generalization of Binomial for $K > 2$ outcomes

# Poisson Distribution

- Distribution a non-negative integer (count) random variable $X$, e.g., number of events in a fixed interval of time



- Defined by a non-negative rate parameter $\lambda$

- The PMF of Poisson is

$$p(X = k|\mu) = \frac{\lambda^k \exp(-\lambda)}{k!} \qquad (k = 0,1,2,\ldots)$$

- Expectation: $\mathbb{E}[X] = \lambda$, variance: $\text{var}[X] = \lambda$

# Continuous Distributions

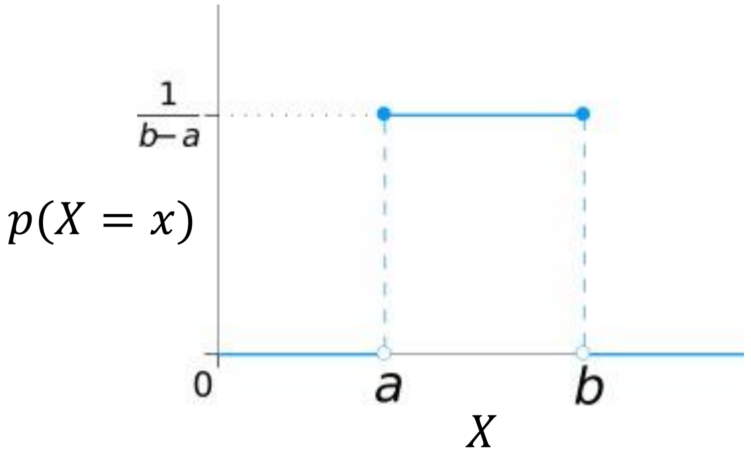
# Uniform Distribution

- Distribution over a uniformly distributed random variable in interval $[a, b]$

- The probability density function (PDF) is

Recall that since $X$ is continuous, this is not the probability of $X = x$ but probability of $X \in (x, x + \delta x)$ where $\delta x$ is very small

$$p(X = x|\mu) = \frac{1}{(b-a)}$$



- Expectation: $\mathbb{E}[X] = \frac{(a+b)}{2}$

- Variance: $\text{var}[X] = \frac{(b-a)^2}{12}$

# Beta Distribution

- Distribution over a random var. $\pi \in (0,1)$, e.g., probability of head for a coin

- Defined by two parameters $\alpha, \beta > 0$. They control the shape of the distribution

- The probability density function (PDF) is

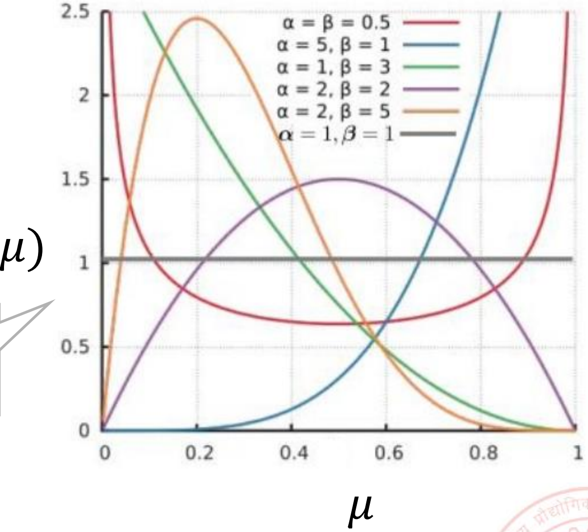
$$p(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

$\Gamma$ denotes the gamma function: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\exp(-t)dt$

- Expectation: $\mathbb{E}[\pi] = \frac{\alpha}{\alpha+\beta}$

- Variance: $\text{var}[\pi] = \frac{\alpha+\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Also equivalent to a uniform distribution for $\alpha = 1, \beta = 1$



# Dirichlet Distribution

- Distribution over a random non-neg vector $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_K]$ that sums to 1, e.g., vector of probabilities of a dice roll showing each of the $K$ faces

$$0 \leq \pi_i \leq 1, \qquad \forall i = 1,2,\dots,K, \qquad \sum_{i=1}^{K} \pi_i = 1$$

- Equivalent to a distribution over the $K - 1$ dimensional simplex

- Defined by $K$ non-negative parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_K]$

These parameters control the shape of the Dirichlet distribution

- The PDF is

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \pi_i^{\alpha_i - 1}$$

Dirichlet is like a $K$-dimensional generalization of the Beta distribution

$\hat{\alpha}_i = \frac{\alpha_i}{\alpha_0}$

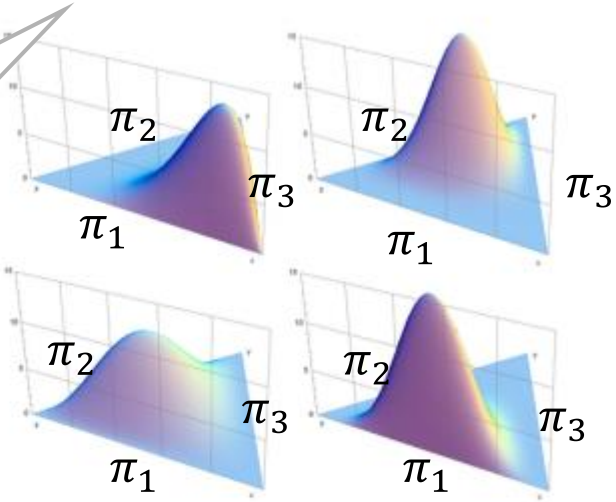
- Expectation: $\mathbb{E}[\pi_i] = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_i}$, variance: $\text{var}[\pi_i] = \frac{\hat{\alpha}_i(1-\hat{\alpha}_i)}{(\alpha_0+1)}$ where $\alpha_0 = \sum_{i=1}^{K} \alpha_i$

# Dirichlet Distribution (contd)

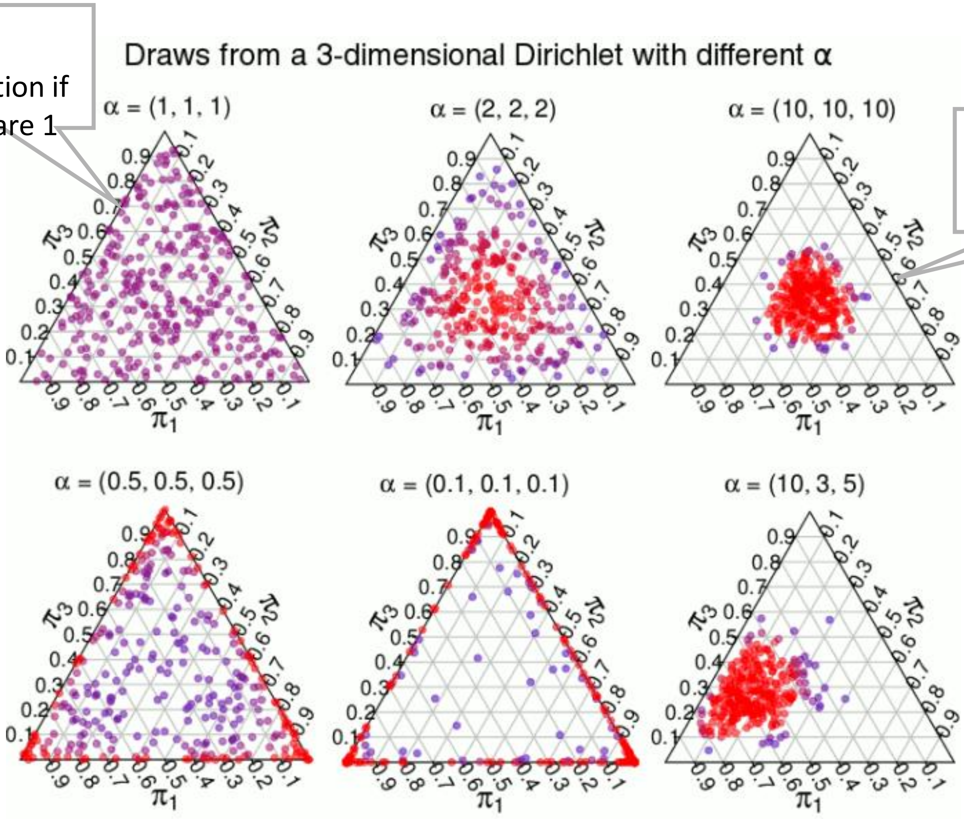
- Shape of the Dirichlet distribution $(K=3)$, as $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\dots,\alpha_K]$ varies
- Each point within the two-dim $(K-1)$, simplices (triangles) below is a random probability vector $\boldsymbol{\pi}=[\pi_1,\pi_2,\pi_3]$ of length 3, drawn from the Dirichlet

Visualizations of PDFs of some 3-dim Dirichlet distributions (each generated using a different conc. Param vector

$\boldsymbol{\alpha}$ controls the shape of the Dirichlet (just like Beta distribution's hyperparameters)



Like a uniform distribution if all $\alpha_k$'s are 1



All $\alpha_k$'s large results in peak around the center of the simplex

More red means we will get more points from that region when drawing random $\boldsymbol{\pi}$ vectors from the Dirichlet

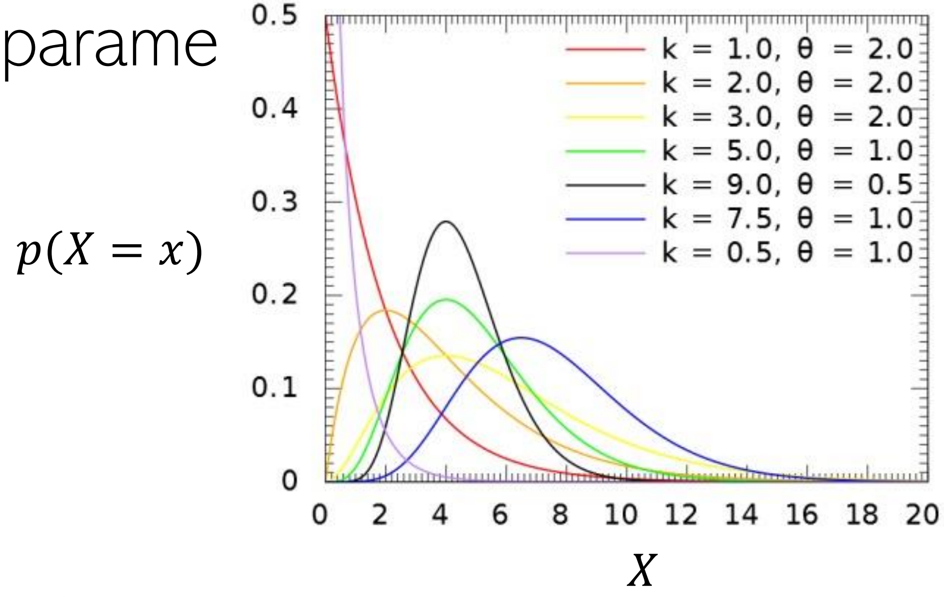
# Gamma Distribution

- Distribution over non-negative random variable $X > 0$, e.g., time between phone-calls at a call center

- Defined by a shape parameters $k$ and a scale parame

- The PDF is

$$p(X = x|k, \theta) = \frac{x^{k-1}\exp(-\frac{x}{\theta})}{\theta^k \Gamma(k)}$$

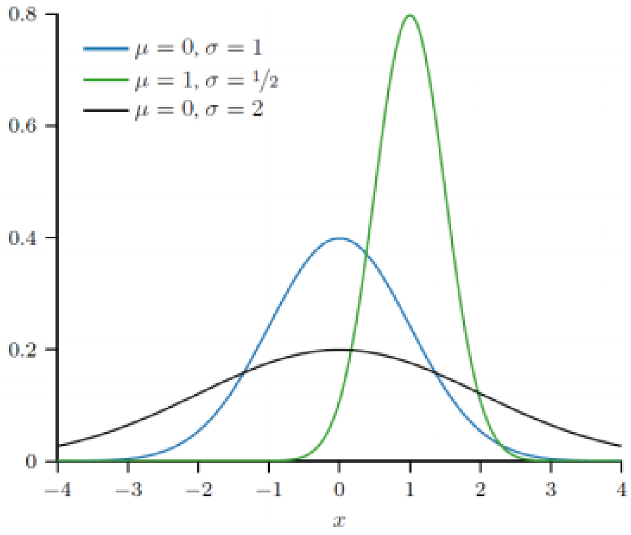


- Expectation: $\mathbb{E}[X] = k\theta$, variance: $\text{var}[X] = k\theta^2$
- Note: Sometimes, the gamma distribution can also be defined in another parameterization (shape and inverse scale $(1/\theta)$)

# Gaussian Distribution (Univariate)

- Distribution over real-valued scalar random variables $X \in \mathbb{R}$, e.g., height of students in a class
- Defined by a scalar mean $\mu$ and a scalar variance $\sigma^2$

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



- Mean: $\mathbb{E}[X] = \mu$
- Variance: $\text{var}[X] = \sigma^2$
- Inverse of variance is called precision: $\beta = \frac{1}{\sigma^2}$.

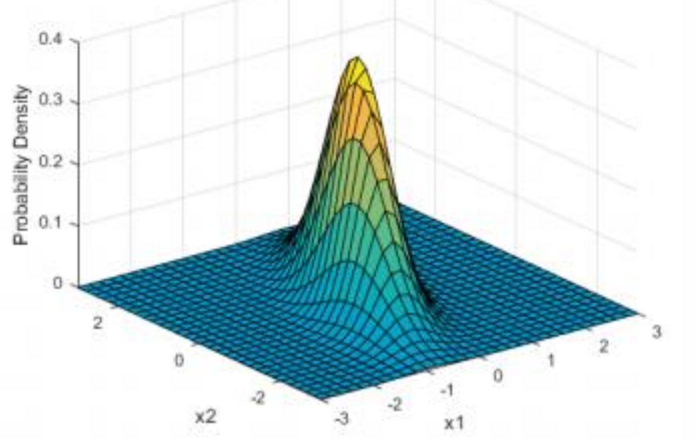
Gaussian PDF in terms of precision

$$\mathcal{N}(X = x|\mu, \beta) = \sqrt{\frac{\beta}{2\pi}}\exp\left[-\frac{\beta}{2}(x-\mu)^2\right]$$

# Gaussian Distribution (Multivariate)

- Distribution over real-valued vector random variables $\boldsymbol{X} \in \mathbb{R}^D$
- Defined by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a covariance matrix $\boldsymbol{\Sigma}$

A two-dimensional Gaussian

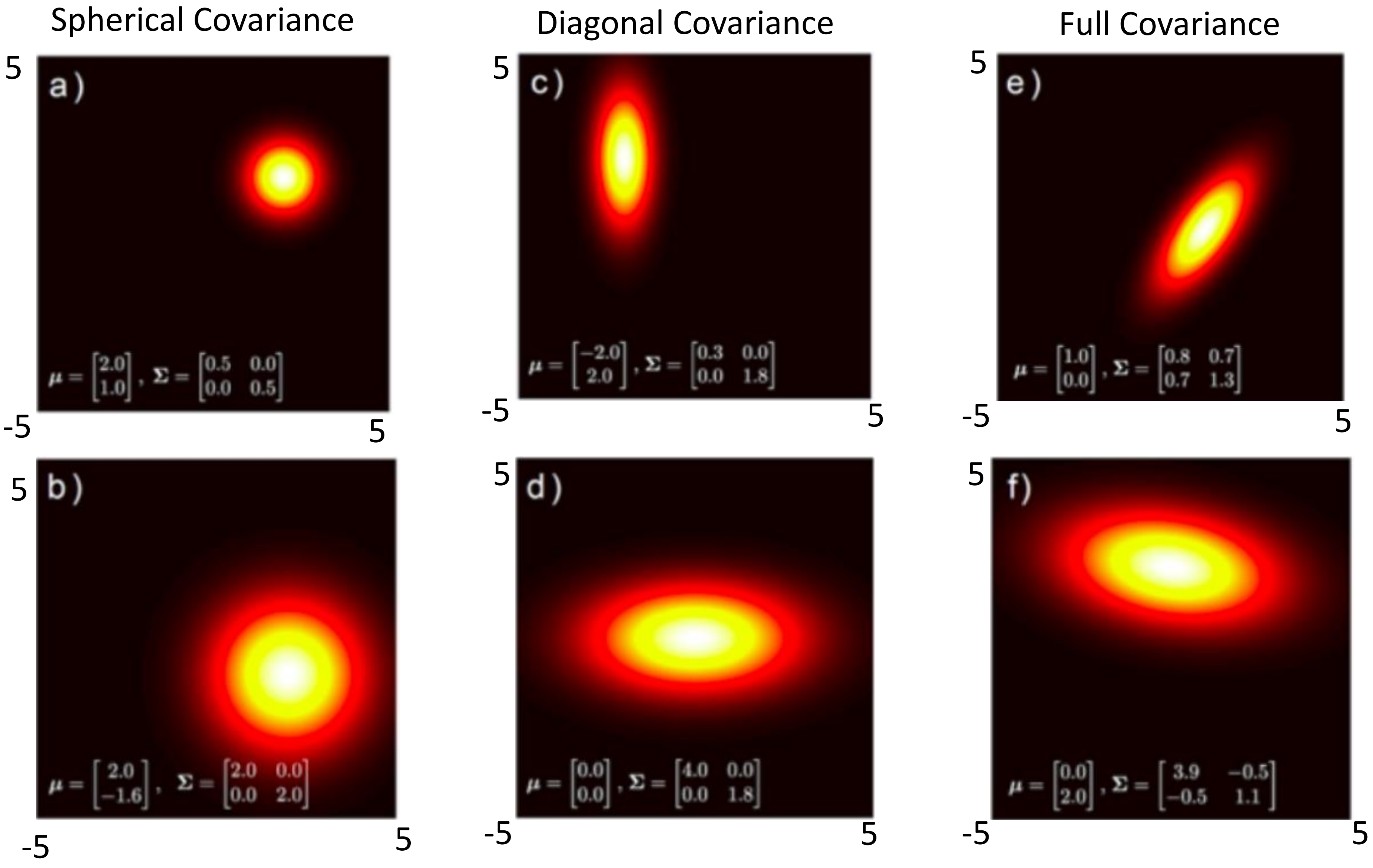


$$\mathcal{N}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp[-(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]$$

- Note: The cov. matrix $\boldsymbol{\Sigma}$ must be symmetric and PSD
  - All eigenvalues are positive
  - $\boldsymbol{z}^\top \boldsymbol{\Sigma} \boldsymbol{z} \geq \boldsymbol{0}$ for any real vector $\boldsymbol{z}$

- The covariance matrix also controls the shape of the Gaussian

# Covariance Matrix for Multivariate Gaussian



Spherical Covariance

Diagonal Covariance

Full Covariance

Spherical: Equal spreads (variances) along all dimensions

Diagonal: Unequal spreads (variances) along all directions but still axis-parallel

Full: Unequal spreads (variances) along all directions and also spreads along oblique directions

# Multivariate Gaussian: Marginals and Conditionals

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ with $\Lambda = \Sigma^{-1}$. Suppose

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

- The marginal distribution is simply

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa})$$

- The conditional distribution is given by

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

**Thus marginals and conditionals of Gaussians are Gaussians**

# 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 $\mathbf{cov}[X] = \Sigma$, then for the vector-valued r.v. $Y$

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

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

- 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 $\mathbf{cov}[X] = \Sigma$, then for the scalar-valued r.v. $Y$

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

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

# Probabilistic Modeling

# Probabilistic Modeling of Data: The Setup

- We are given some training data $\mathcal{D}$
- For supervised learning, $\mathcal{D}$ contains $N$ input-label pairs $(\boldsymbol{x}_i, y_i)_{i=1}^{N}$
- For unsupervised learning, $\mathcal{D}$ contains $N$ inputs $(\boldsymbol{x}_i)_{i=1}^{N}$
- Other settings are also possible (e.g., semi-sup., reinforcement learning, etc)
- Our goal is to estimate the distribution (and thus $\boldsymbol{\theta}$) using training data
- Once the distribution is estimated, we can do things such as
  - Predict labels of new inputs, along with our confidence in these predictions
  - Generate new data with similar properties as training data
  - .. and a lot of other useful tasks, e.g., detecting outliers

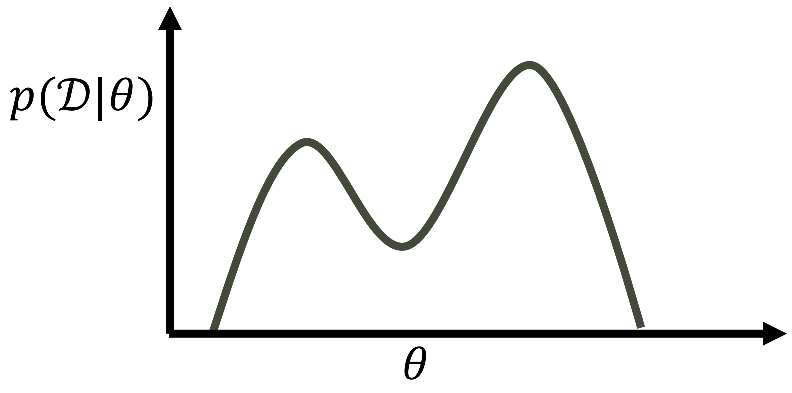
# Data/Observation Model

- Observation model is usually defined by a suitable probability distribution

$$p(\mathcal{D}|\theta)$$

- The distribution's parameters $\boldsymbol{\theta}$ are unknown and need to be estimated
- The quantity $p(\mathcal{D}|\theta)$ is also referred to as the "likelihood"

It's a function of $\boldsymbol{\theta}$, not a distribution over $\theta$

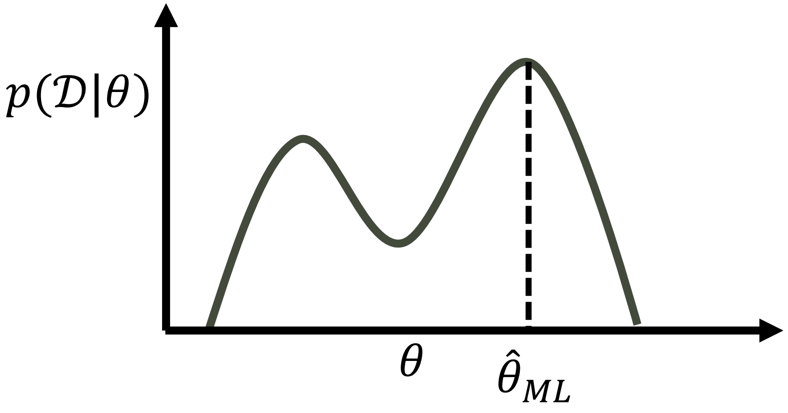


- Likelihood gives us the probability of the observed data $\mathcal{D}$ as a function of $\boldsymbol{\theta}$

# Parameter Estimation: A Simple Approach

- Likelihood itself is a useful quantity to estimate the parameters $\boldsymbol{\theta}$
- Maximum Likelihood (ML) is a popular method

$$\hat{\theta}_{ML} = \arg\max_{\theta} \; \log p(\mathcal{D}|\theta)$$

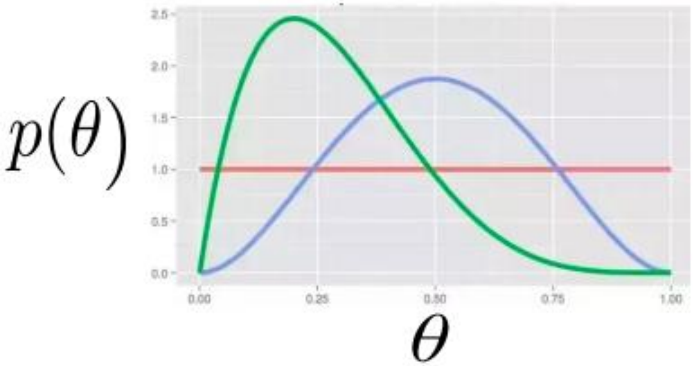


- MLE is akin to minimizing a loss function. Negative log likelihood (NLL) = Loss
- MLE however as a few issues
  - Provides only a point estimate of $\boldsymbol{\theta}$ (thus no uncertainty estimate)
  - Does not allow incorporating prior knowledge about $\boldsymbol{\theta}$
- Using a prior distribution $p(\theta)$ over $\boldsymbol{\theta}$ can help address these issues

# The Prior

- The prior $p(\theta|\alpha)$ plays an important role in probabilistic/Bayesian modeling
  - Here $\alpha$ denotes the parameters ("hyperparameters") of the prior distribution

- Reflects our prior beliefs about possible parameter values <u>before</u> seeing the data

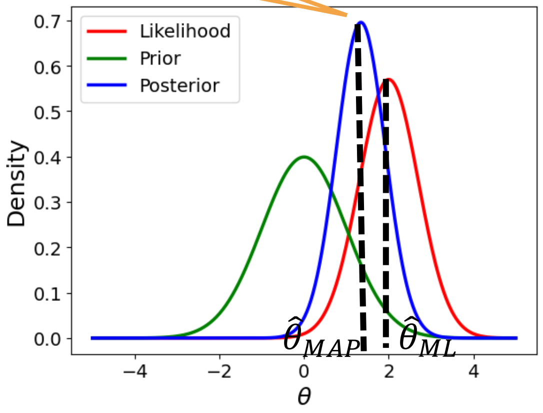


- Can be "subjective" or "objective"
  - Subjective: Prior (our beliefs) derived from past experiments
  - Objective: Prior represents "neutral knowledge" (e.g.. uniform, vague prior)

- Can also be seen as a regularizer (we will see the reason soon)

# Using Prior in Parameter Estimation

Posterior can be seen as a "compromise" between likelihood and the prior



- Can use prior in following ways during parameter estimation
  - Computing the distribution of the parameters conditioned on data

The posterior distribution

$$p(\theta|\mathcal{D},\alpha) = \frac{p(\mathcal{D},\theta|\alpha)}{p(\mathcal{D}|\alpha)} = \frac{p(\mathcal{D}|\theta,\alpha)p(\theta|\alpha)}{\int p(\mathcal{D}|\theta,\alpha)p(\theta|\alpha)d\theta}$$

Assuming $\boldsymbol{\alpha}$ is known so the posterior is conditioned on $\boldsymbol{\alpha}$ as well

$$= \frac{p(\mathcal{D}|\theta)p(\theta|\alpha)}{\int p(\mathcal{D}|\theta)p(\theta|\alpha)d\theta} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Given $\boldsymbol{\theta}$, the data is conditionally independent of the prior's hyperparameters $\boldsymbol{\alpha}$ so $p(\mathcal{D}|\theta,\alpha) = p(\mathcal{D}|\theta)$

An important quantity. More on this later

  - Computing the (MAP) maximum-a-posteriori estimate (report maxima of the posterior)

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \ \log p(\theta|\mathcal{D},\alpha) = \underset{\theta}{\operatorname{argmax}} \left[\log p(\mathcal{D}|\theta) + \log p(\theta|\alpha)\right]$$

$$= \underset{\theta}{\operatorname{argmin}} \ \left[NLL(\theta) - \log p(\theta|\alpha)\right]$$

We only need to find the maxima of the posterior

Note that computing MAP estimate does not require computing the posterior ☺

Akin to a regularizer added to the loss

The regularizer hyperparameter is part of prior

# Parameter Estimation: Summary of approaches

- Usually one of the following approach taken
  1. A single "best" **point estimate** of the parameters by optimizing an objective function

$$\hat{\theta} = \text{argmax}_{\theta}\, f(\mathcal{D};\theta)$$

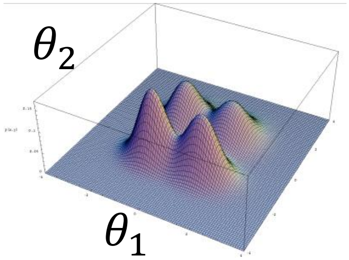
$f$ can be log-likelihood (for MLE) or log-posterior (for MAP)

  2. A distribution over the parameters (conditioned on observed data $\mathcal{D}$)

The posterior distribution

$$p(\theta|\mathcal{D})$$

$\theta = [\theta_1, \theta_2]$

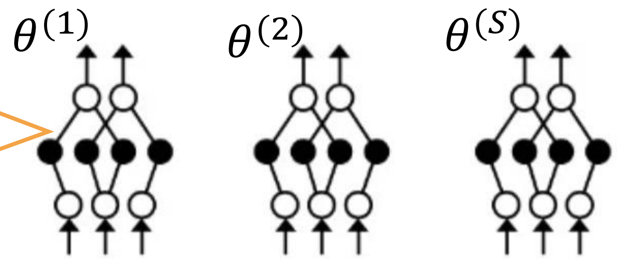


  3. A set/ensemble of point estimates of parameters (applying approach 1 multiple times)

Computing multiple point estimates, each using a different subset of the training data

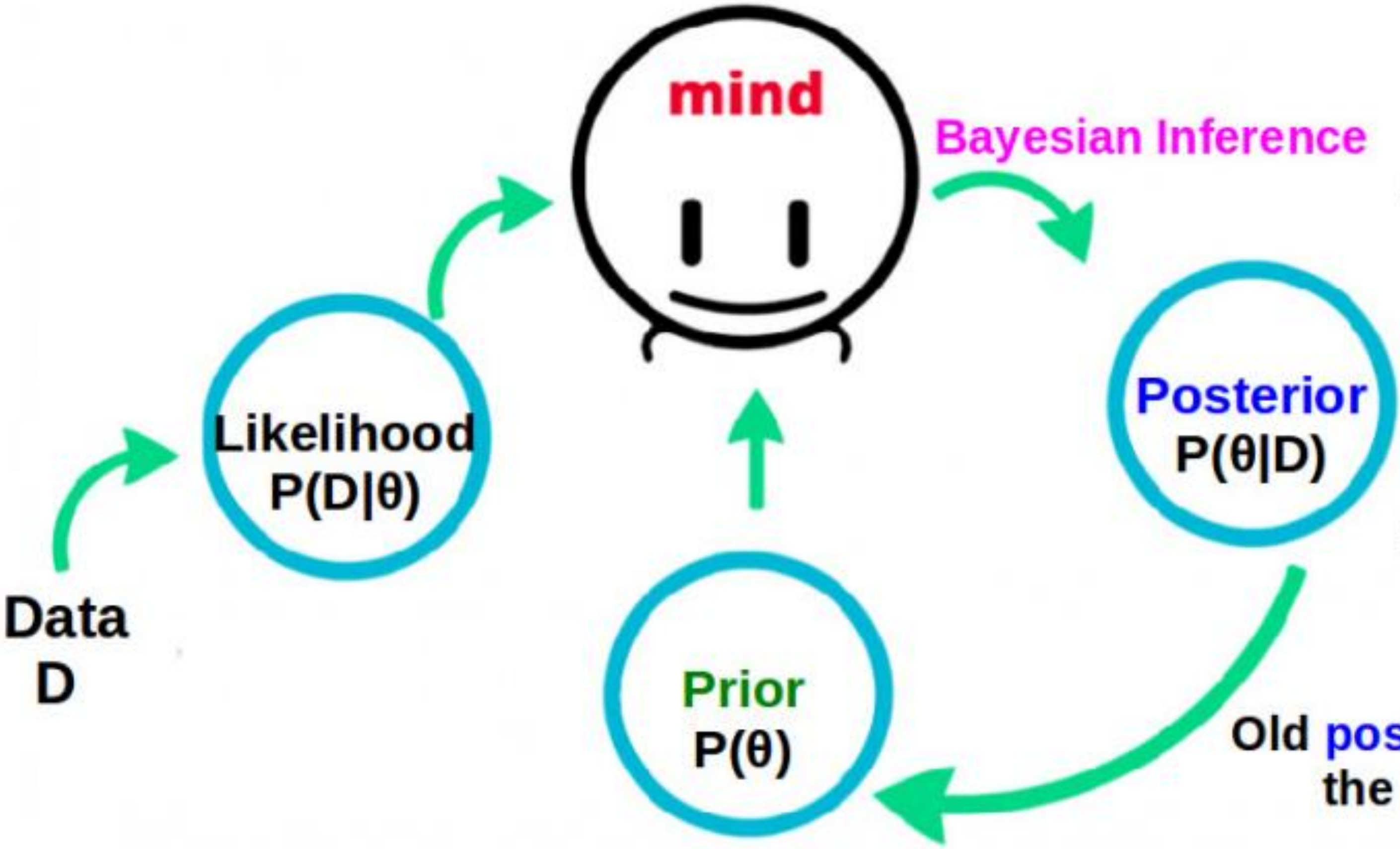
$$\{\hat{\boldsymbol{\theta}}(\mathcal{D}') : \mathcal{D}' \sim p^*\}$$

Ensemble (e.g., training a deep neural net with multiple different initializations)



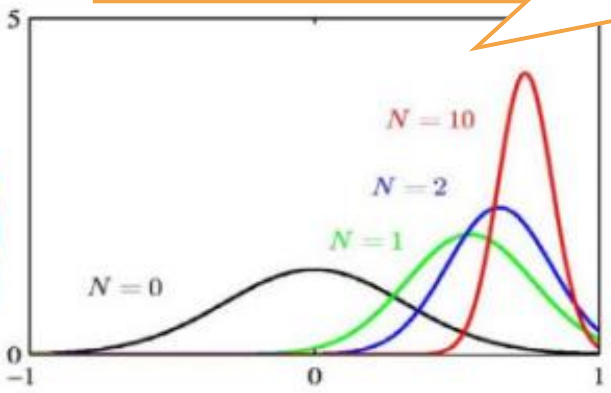
# An Important Aspect: Posterior Updates

- Posterior updates in Bayesian inference can naturally be done in an online fashion



Old posterior becomes the new prior

Also, the posterior's spread/variance gets smaller as we use more and more data to infer it



# Making Predictions: The Predictive Distribution

- If we have computed $p(\theta|\mathcal{D})$ then the predictive distribution can be defined as

$$
\begin{aligned}
p(\mathcal{D}_*|\mathcal{D}) = \int p(\mathcal{D}_*, \theta|\mathcal{D})\, d\theta \; &= \int p(\mathcal{D}_*|\theta, \mathcal{D}) p(\theta|\mathcal{D})\, d\theta \\
&= \int p(\mathcal{D}_*|\theta) p(\theta|\mathcal{D})\, d\theta \\
&= \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}_*|\theta)]
\end{aligned}
$$

- If we don't have $p(\theta|\mathcal{D})$ but a set/ensemble of estimates $\{\theta^{(i)}\}_{i=1}^{S}$ then

$$p(\mathcal{D}_*|\mathcal{D}) \approx \frac{1}{S}\sum_{i=1}^{S} p\left(\mathcal{D}_*\middle|\theta^{(i)}\right)$$

- If we only have a single point estimate $\hat{\theta}$ then

$$p(\mathcal{D}_*|\mathcal{D}) \approx p\left(\mathcal{D}_*\middle|\hat{\theta}\right)$$

Faster to compute since no expectation/averaging required but less robust because it only considers a single best estimate of $\theta$