Probabilistic Modeling
(From Prior to Posterior)

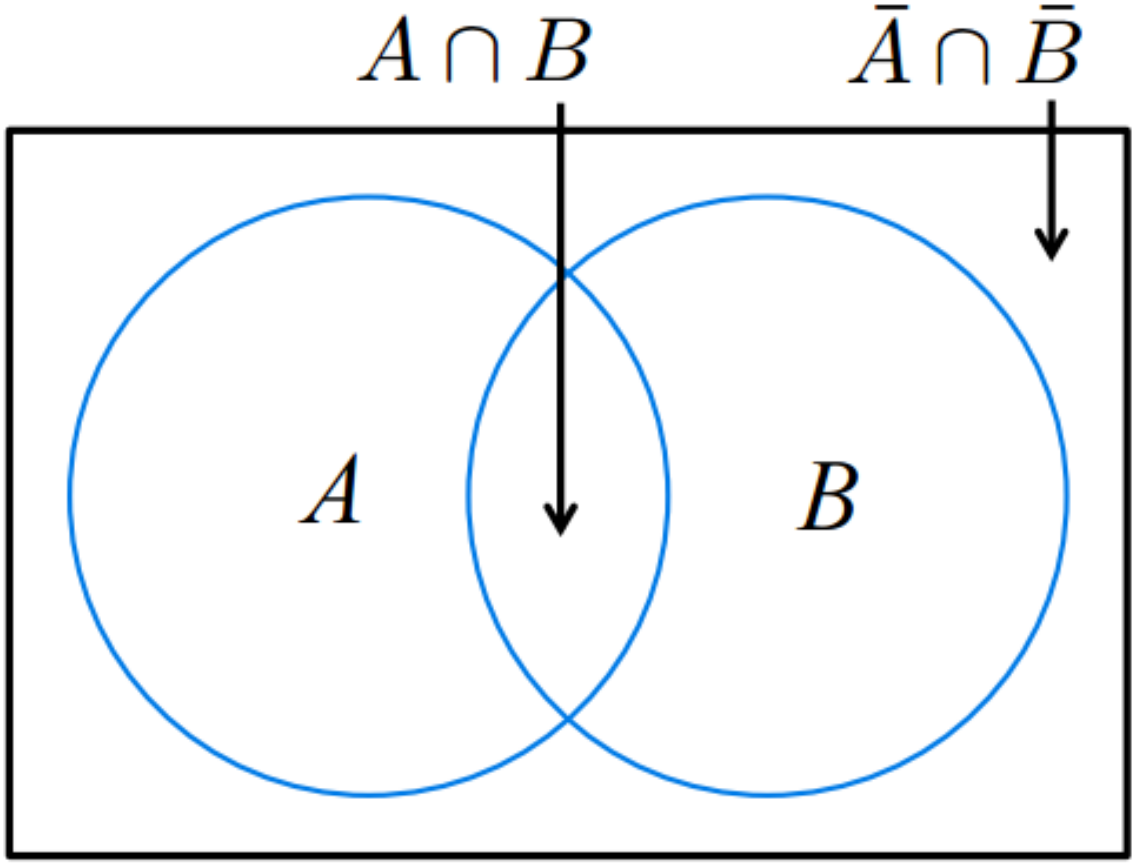
Santosh Chapaneri

Probability theory is nothing but common sense reduced to calculation.
- Pierre-Simon Laplace (1749-1827)

Attributed to:

Refresher – Axioms of Probability

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$



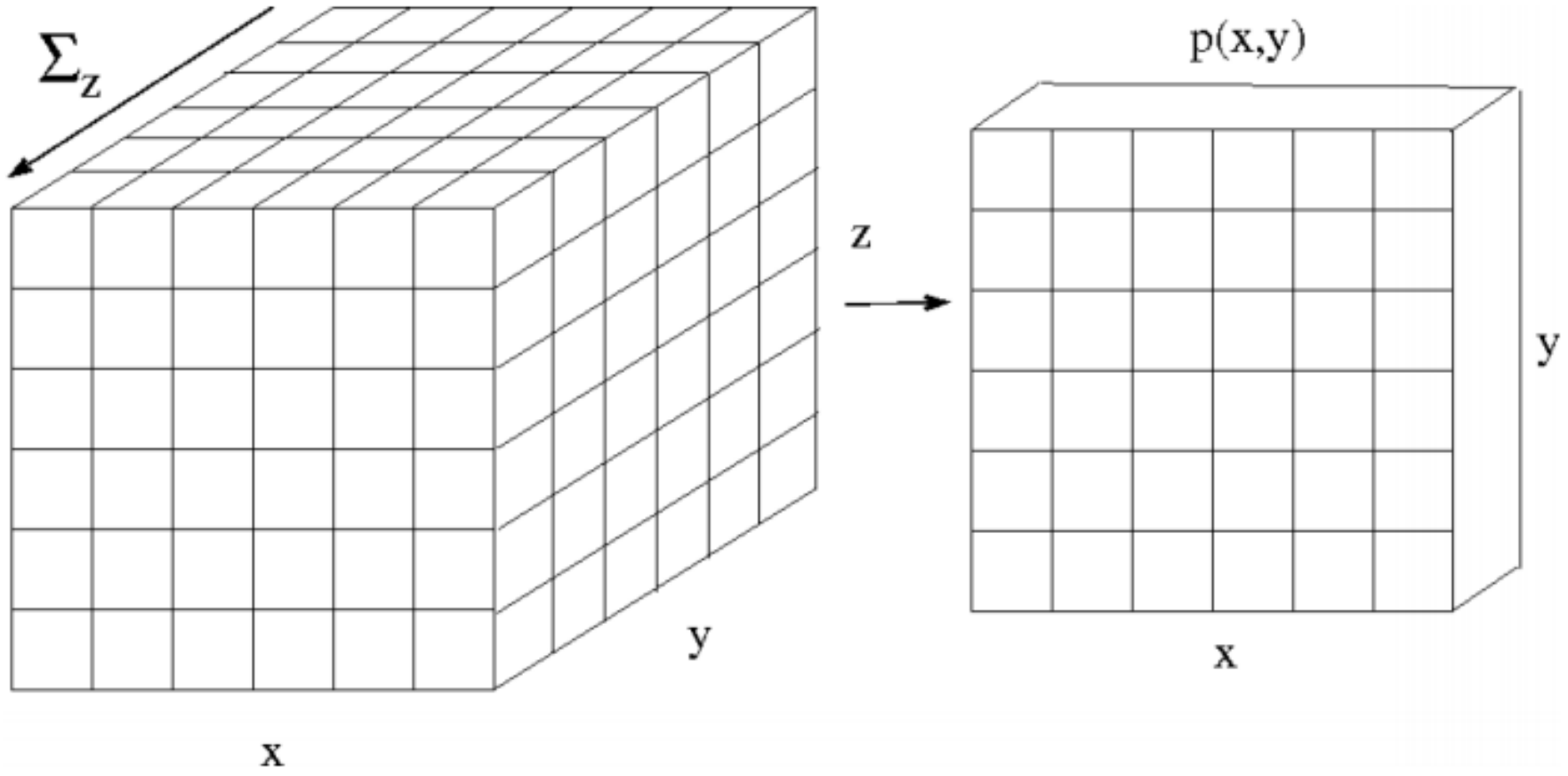
$$0 \leq p(A) \leq 1$$

$$p(\bar{A}) = 1 - p(A)$$

$$p(A \cap B) = p(A \mid B)p(B)$$

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

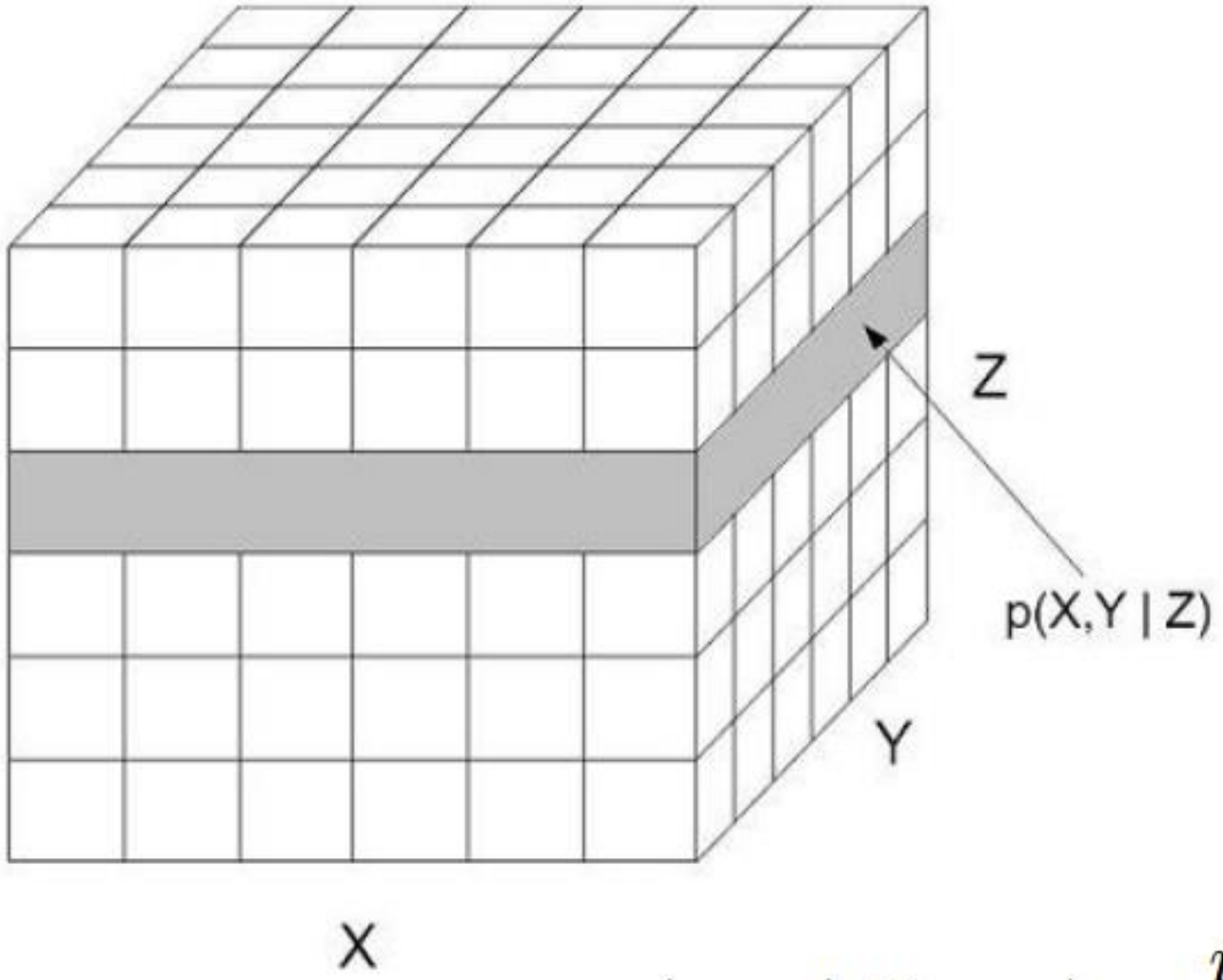
Refresher – Marginal Distributions



$$p(x,y) = \sum_{z \in \mathcal{Z}} p(x,y,z)$$

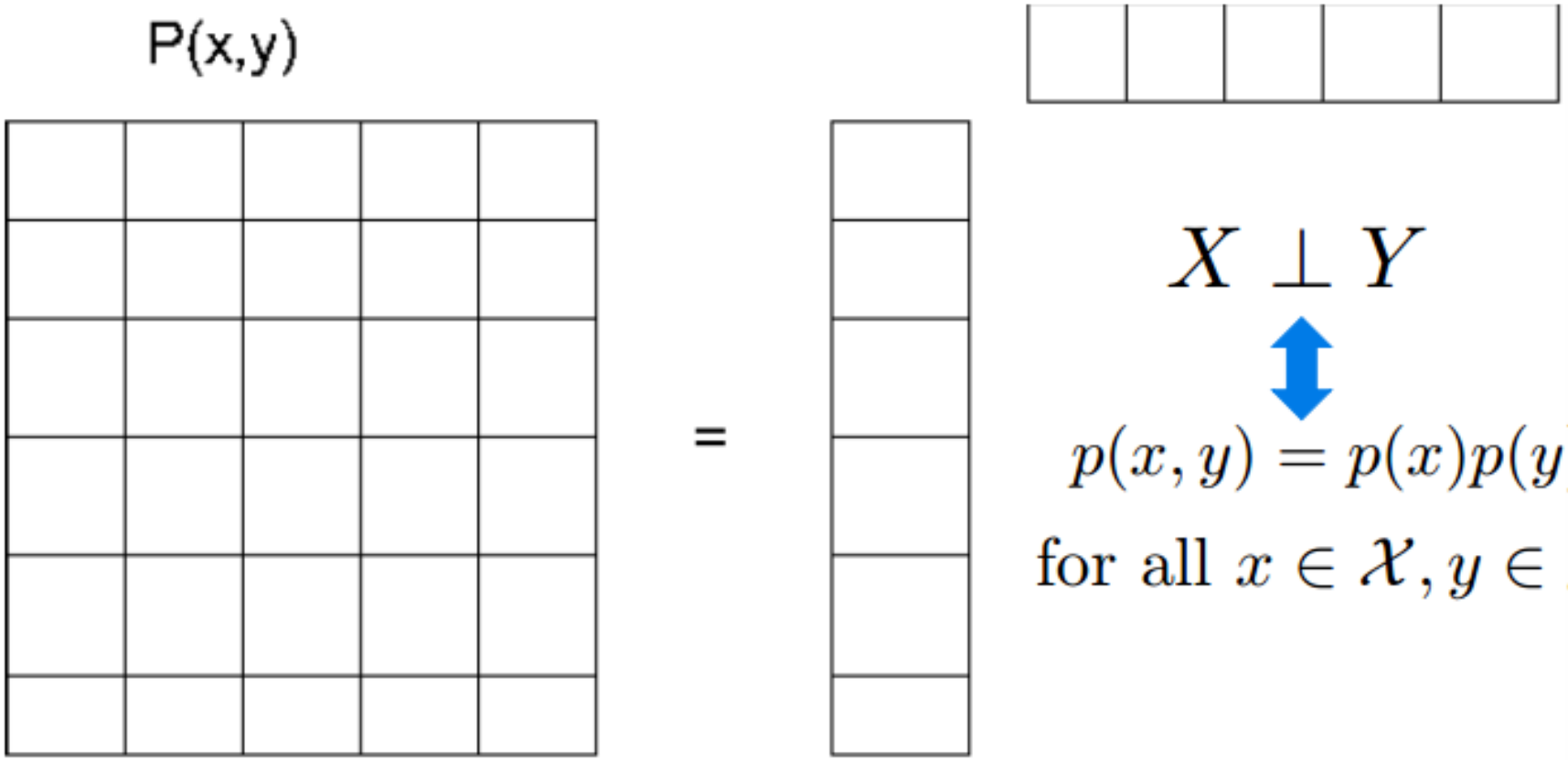
$$p(x) = \sum_{y \in \mathcal{Y}} p(x,y)$$

Refresher – Conditional Distributions



$$p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)}$$

Refresher – Independence



$$X \perp Y$$

$$p(x, y) = p(x)p(y)$$

$$\text{for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

Equivalent conditions on conditional probabilities:

$$p(x \mid Y = y) = p(x) \text{ and } p(y) > 0 \text{ for all } y \in \mathcal{Y}$$

$$p(y \mid X = x) = p(y) \text{ and } p(x) > 0 \text{ for all } x \in \mathcal{X}$$

Refresher – Baye's Theorem

$$p(x, y) = p(x)p(y \mid x) = p(y)p(x \mid y)$$

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(y \mid x)p(x)}{\sum_{x' \in \mathcal{X}} p(x')p(y \mid x')}$$
$$\propto p(y \mid x)p(x)$$

$X \longrightarrow$ unknown parameters we would like to infer

$Y = y \longrightarrow$ observed data available for learning

$p(x) \longrightarrow$ prior distribution (domain knowledge)

$p(y \mid x) \longrightarrow$ likelihood function (measurement model)

$p(x \mid y) \longrightarrow$ posterior distribution (learned information)

Refresher – Discrete Distributions

Bernoulli Distribution: Single toss of a (possibly biased) coin

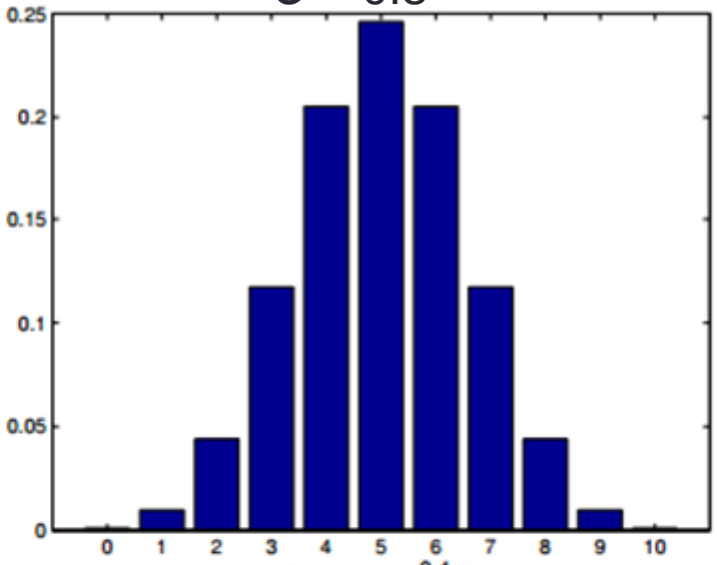
$$\mathcal{X} = \{0, 1\}$$
$$0 \leq \theta \leq 1$$
$$\text{Ber}(x \mid \theta) = \theta^{\delta(x,1)}(1-\theta)^{\delta(x,0)}$$

Binomial Distribution: Toss a single (possibly biased) coin *n* times, and record the number *k* of times it comes up heads

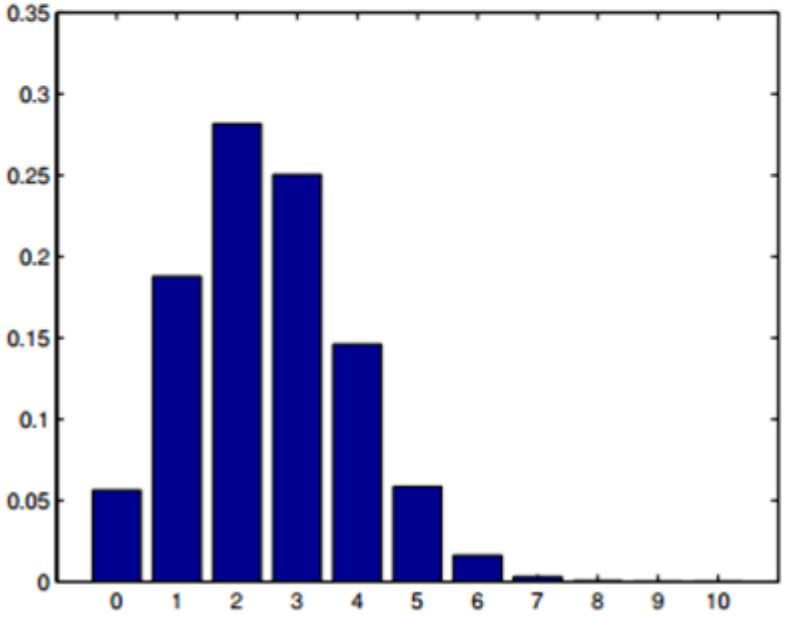
$$\mathcal{K} = \{0, 1, 2, \ldots, n\}$$
$$0 \leq \theta \leq 1$$
$$\text{Bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \qquad \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Refresher – Binomial Distribution

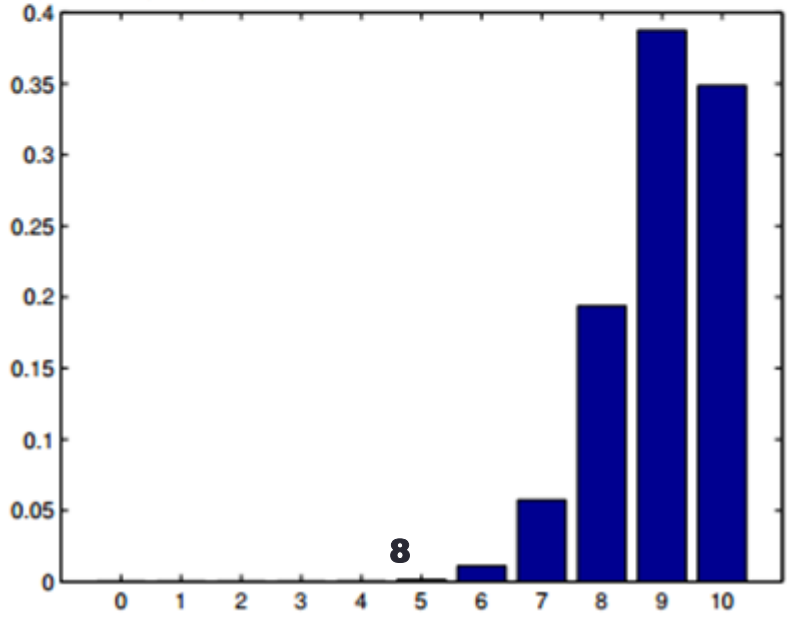
Θ = 0.5



Θ = 0.25



Θ = 0.9

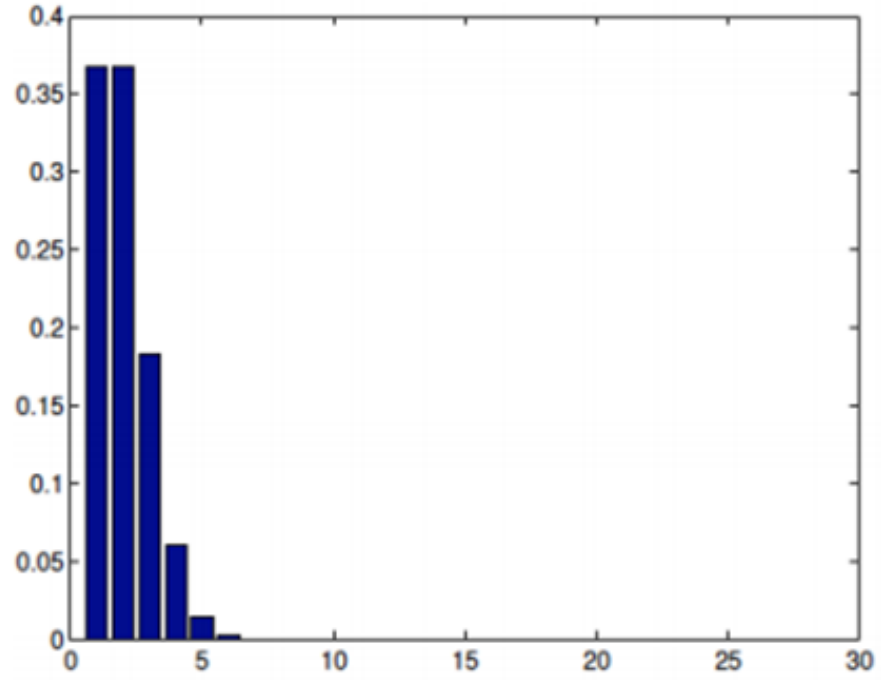


Refresher – Poisson Distribution

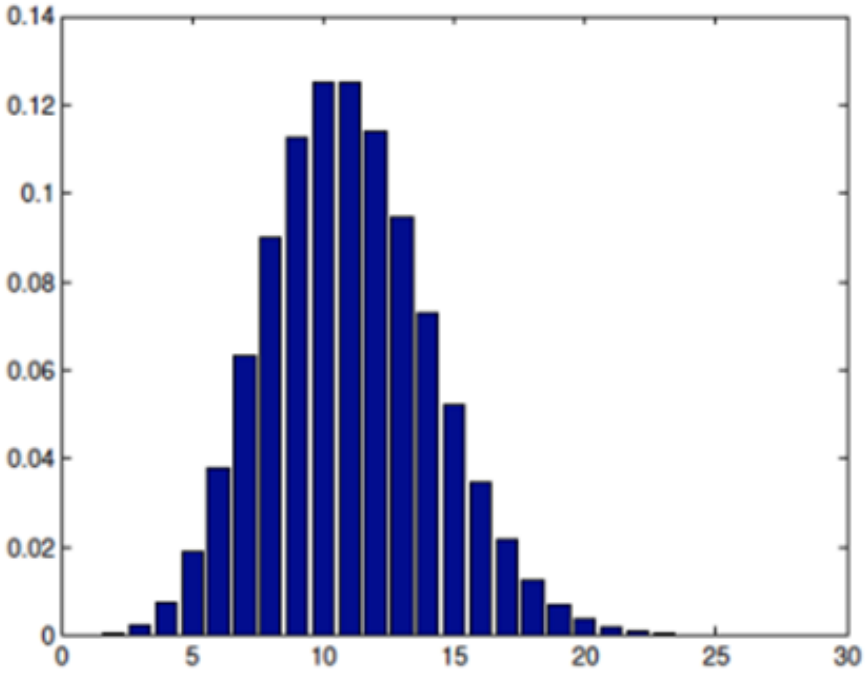
$$\mathcal{X} = \{0, 1, 2, 3, \ldots\}$$

$$\text{Poi}(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!} \qquad \theta > 0$$

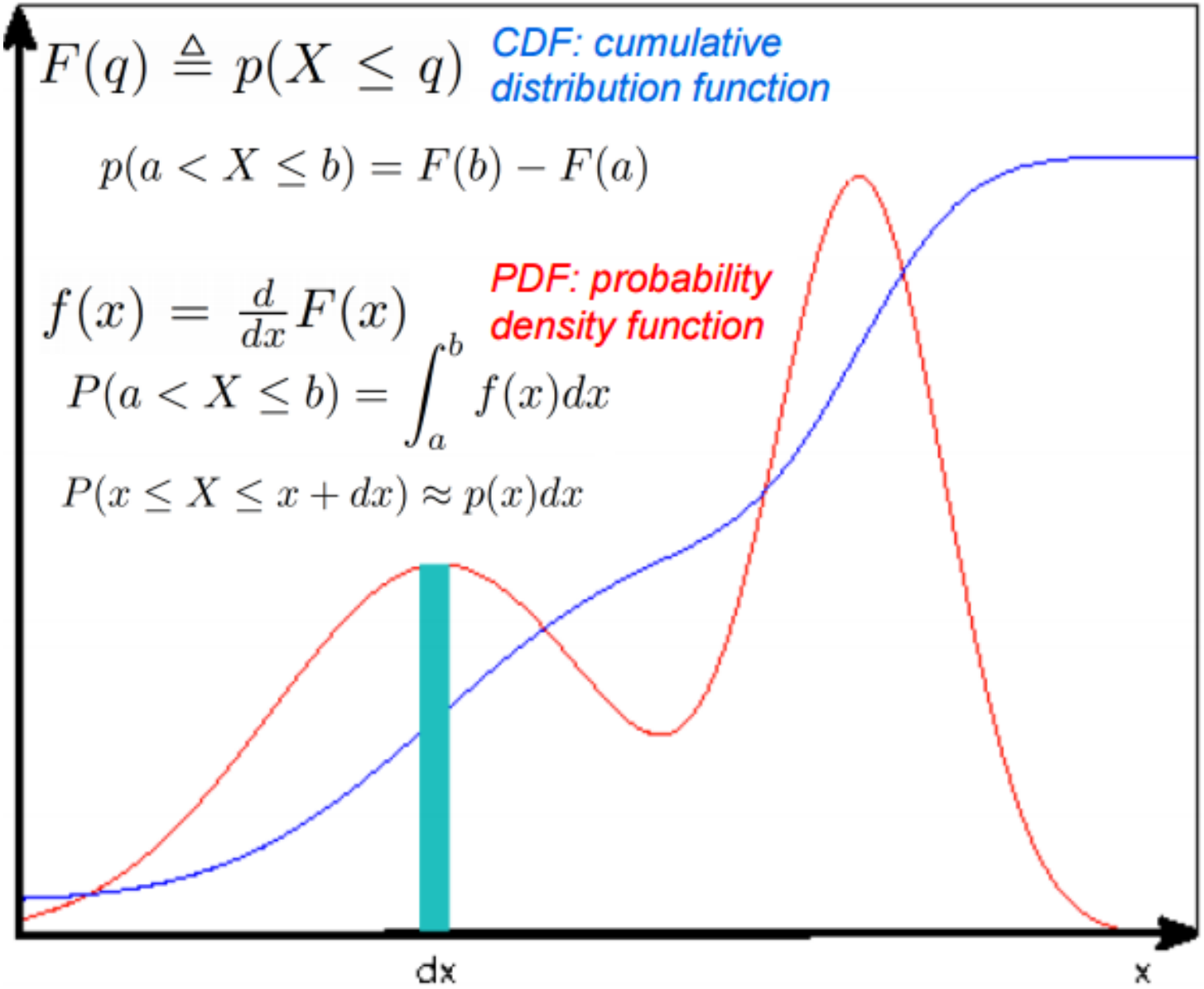
Θ = 1



Θ = 10



Refresher – Continuous Distribution



Refresher – Moments of Random Variables

Mean

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\ p(x)dx \qquad \mathbb{E}[X] \triangleq \sum_{x\in\mathcal{X}} x\ p(x)$$

Variance

$$\text{var}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx = \mathbb{E}\left[X^2\right] - \mu^2$$

$$\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$$ *second moment*

$$\text{std}[X] \triangleq \sqrt{\text{var}[X]}$$ *standard deviation*

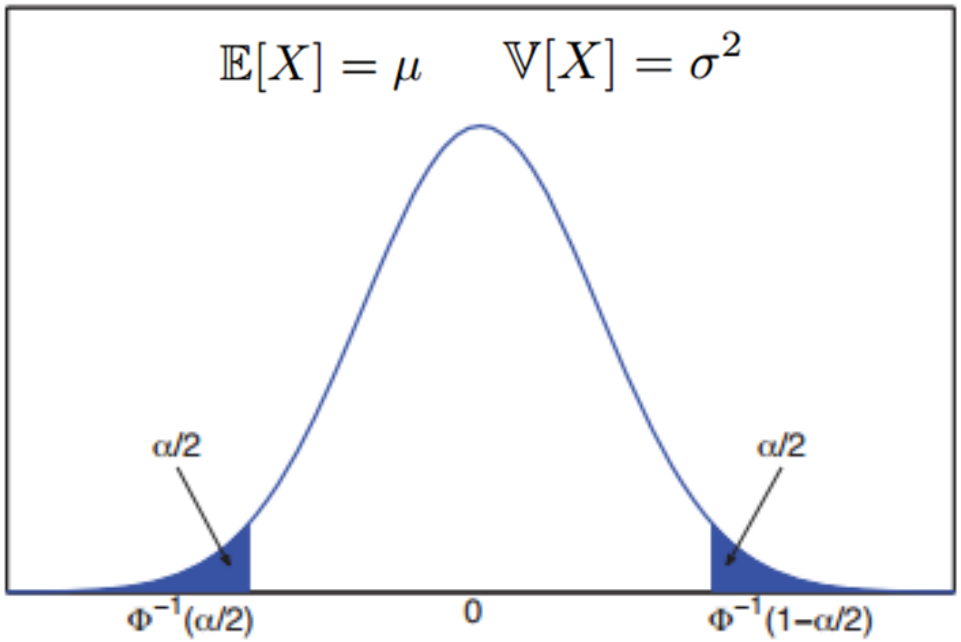
Moments & Conditional Moments

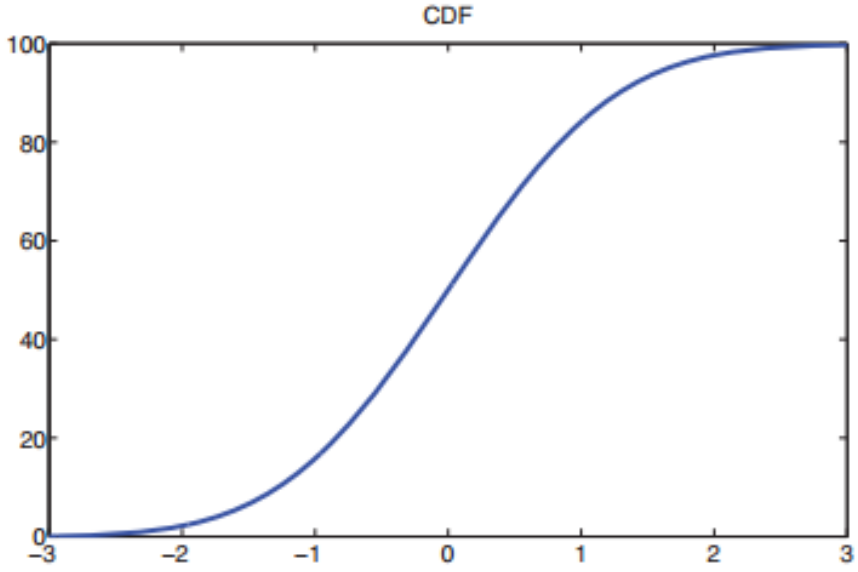
$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x)p(x)\ dx \qquad \mathbb{E}[g(X) \mid Y = y] = \int_{\mathcal{X}} g(x)p(x \mid y)\ dx$$

Refresher – Gaussian (Normal) Distribution

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

$$\Phi(x \mid \mu, \sigma^2) = \int_{-\infty}^{x} \mathcal{N}(z \mid \mu, \sigma^2)\, dz$$





$$(\Phi^{-1}(0.025), \Phi^{-1}(0.975)) = (-1.96, 1.96)$$

***α* quantile** of CDF Φ:

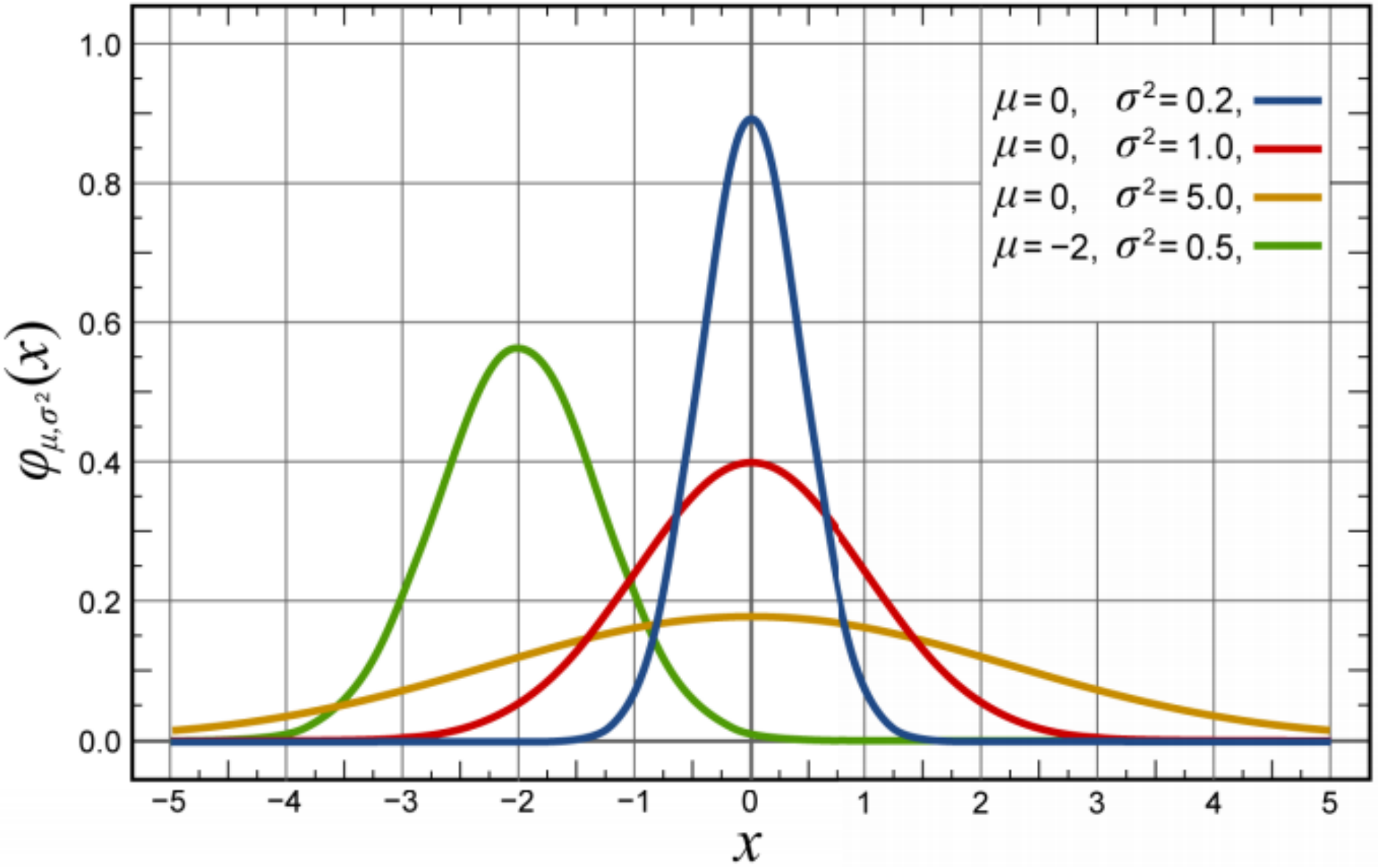
$$P(X \leq x_\alpha) = \alpha$$

95% confidence interval =

$$(\mu - 1.96\sigma, \mu + 1.96\sigma)$$

$\Phi^{-1}(0.5)$ = median
$\Phi^{-1}(0.25)$ = lower quantile
$\Phi^{-1}(0.75)$ = upper quantile

Refresher – Gaussian (Normal) Distribution

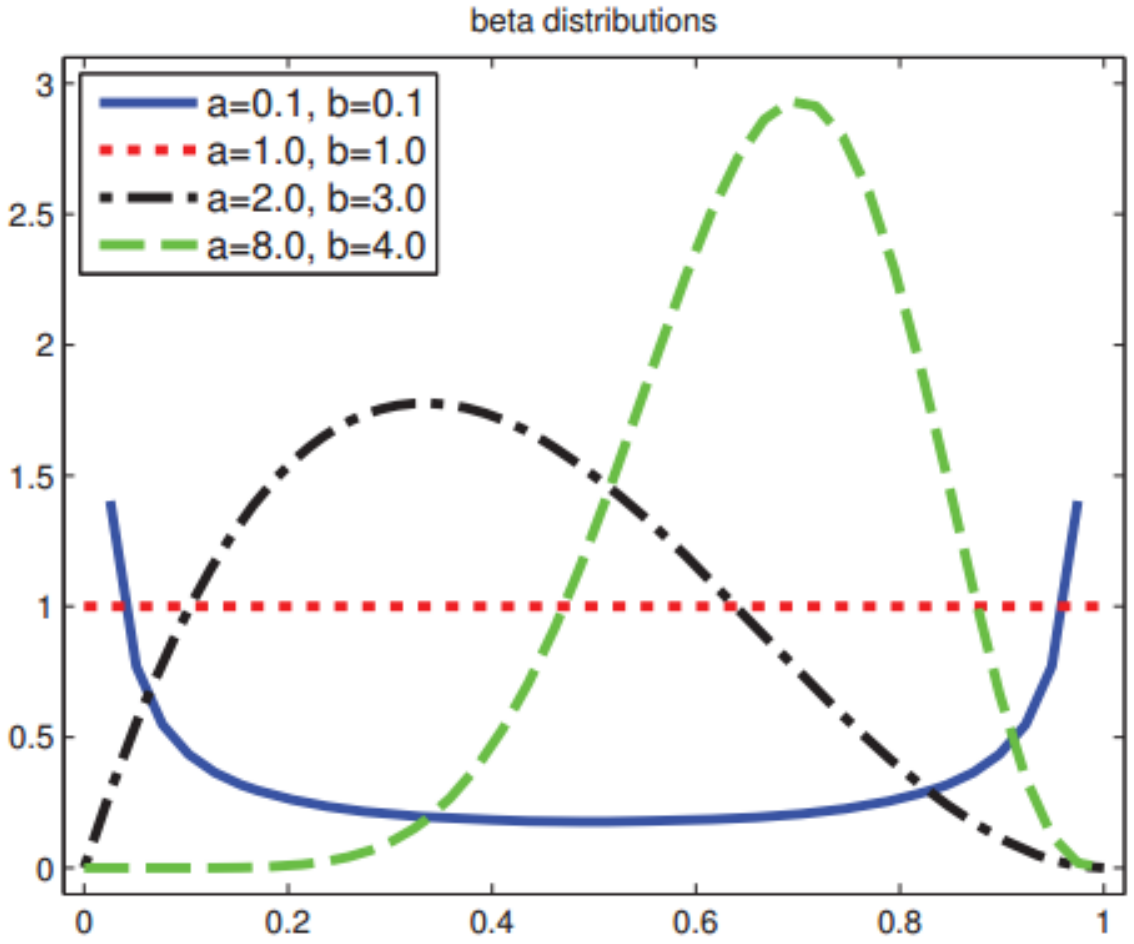


Beta Distribution

$$\text{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

(Beta function)



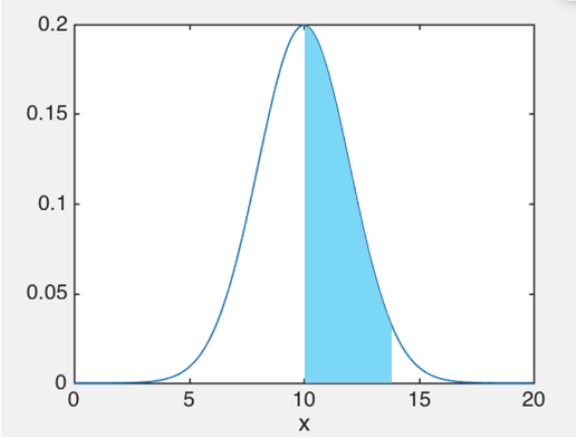
$$\text{mean} = \frac{a}{a+b}$$

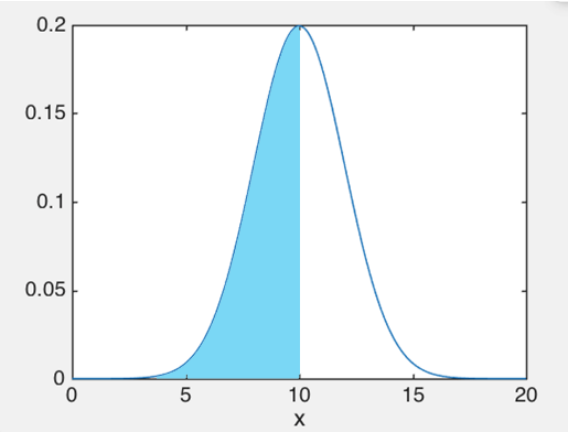
$$\text{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

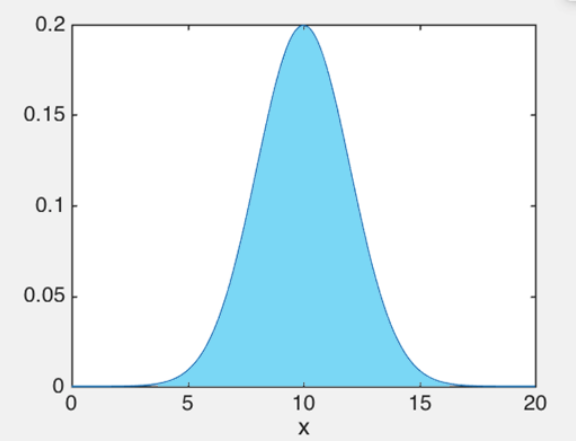
Intuitive meaning of PDF

- p(5.31) = 0.06 and p(5.92) = 0.03 implies ➔ when a value of **X** is sampled, you are 2 times as likely to find that **X** is "very close" to 5.31 than that **X** is "very close" to 5.92.
- p(*a*) = 0.06 and p(*b*) = 0.03 implies ➔ when a value of **X** is sampled, you are 2 times as likely to find that **X** is "very close" to *a* than that **X** is "very close" to *b*.
- p(*a*) = 2*z* and p(*b*) = *z* implies ➔ when a value of **X** is sampled, you are 2 times as likely to find that **X** is "very close" to *a* than that **X** is "very close" to *b*.
- p(*a*) = αz and p(*b*) = *z* implies ➔ when a value of **X** is sampled, you are α times as likely to find that **X** is "very close" to *a* than that **X** is "very close" to *b*.

Intuitive meaning of CDF

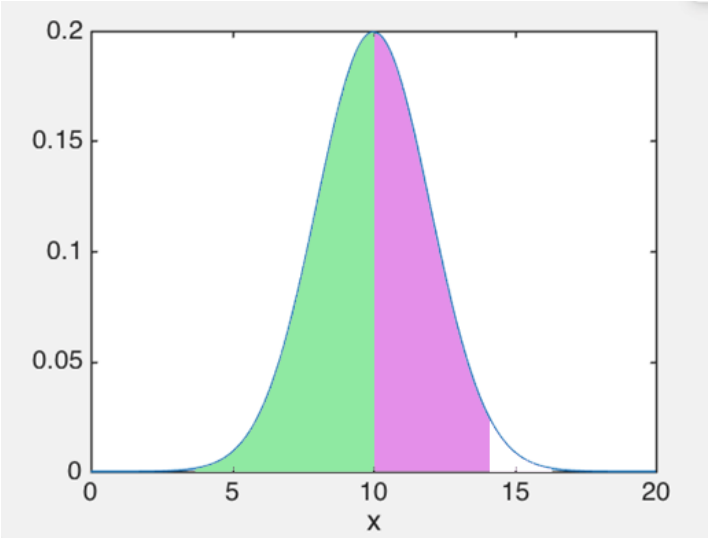






$$P(10 \pounds X \pounds 14) \qquad P(X \pounds 10) \qquad P(X \pounds 20)$$

$$P(10 \pounds X \pounds 14) = P(X \pounds 14) - P(X \pounds 10)$$



Transformations of Random Variables

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\mathrm{cov}[\mathbf{y}] = \mathrm{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

Central Limit Theorem:

Consider *N* IID RV's with pdf's p(x_i) each with mean μ and variance σ^2.

$$S_N = \sum_{i=1}^{N} X_i$$

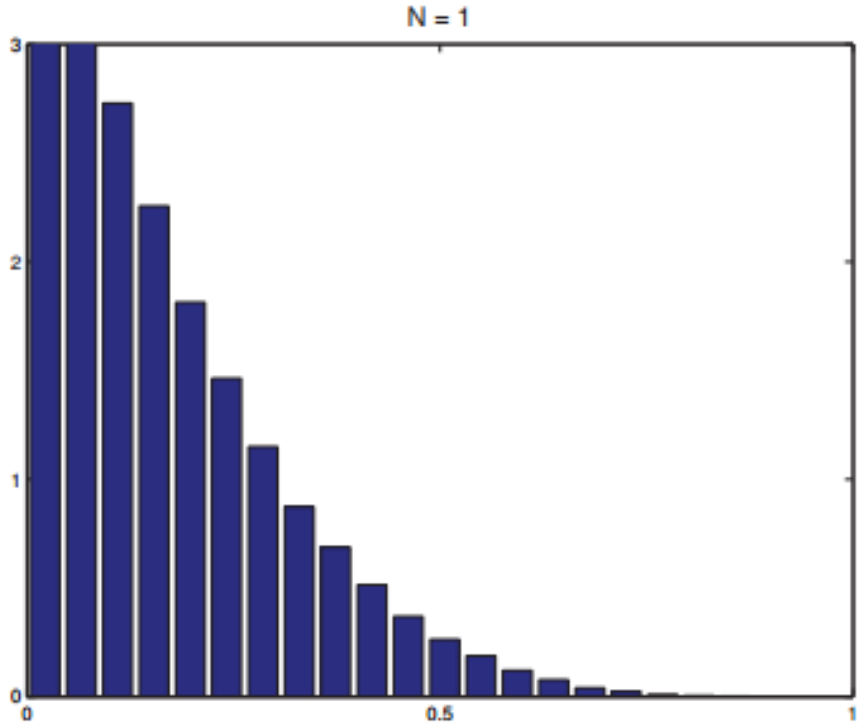
As *N* increases, the CDF of S_N approaches

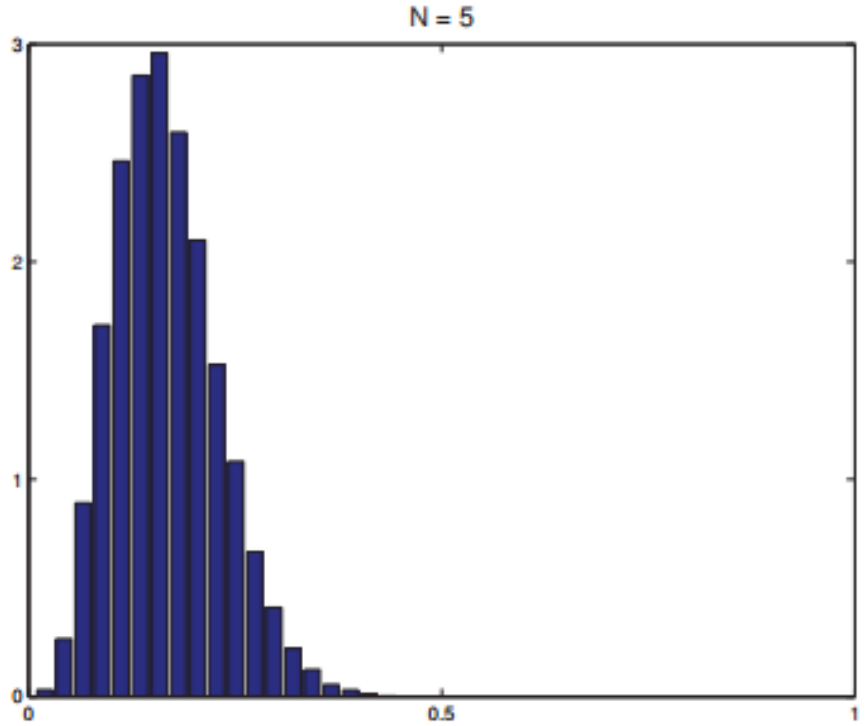
$$p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right)$$

$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}}$$

converges to standard normal

Transformations of Random Variables





Exponential Family of Distributions

A density $f(x|\theta)$ belongs to the exponential family if it has form:

$$f(x \mid \theta) = h(x)g(\theta)\exp\left(\eta(\theta)^T T(x) - A(\theta)\right)$$

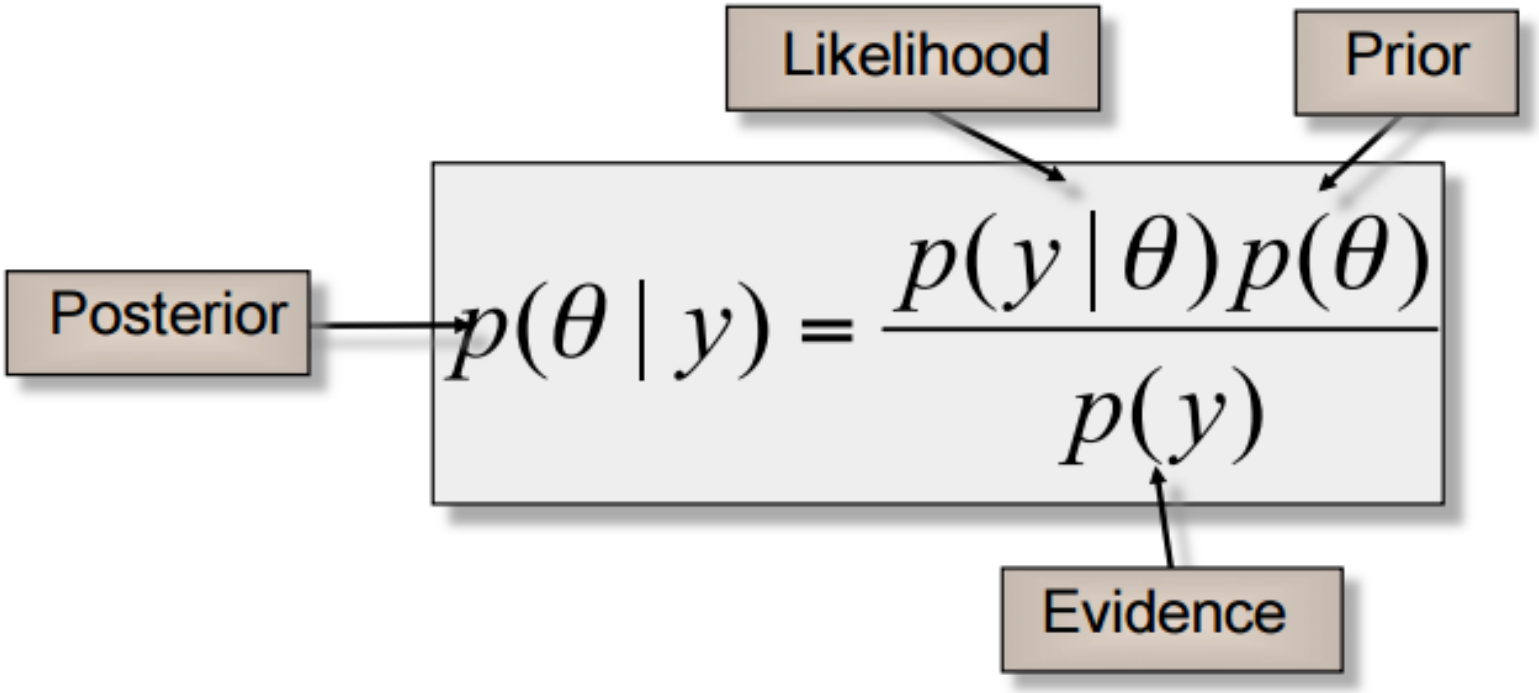
Eg: Gaussian, Bernoulli, Binomial, Poisson, Exponential

Q: Prove that Bernoulli $X \sim Be(\pi)$ is of Exponential family:

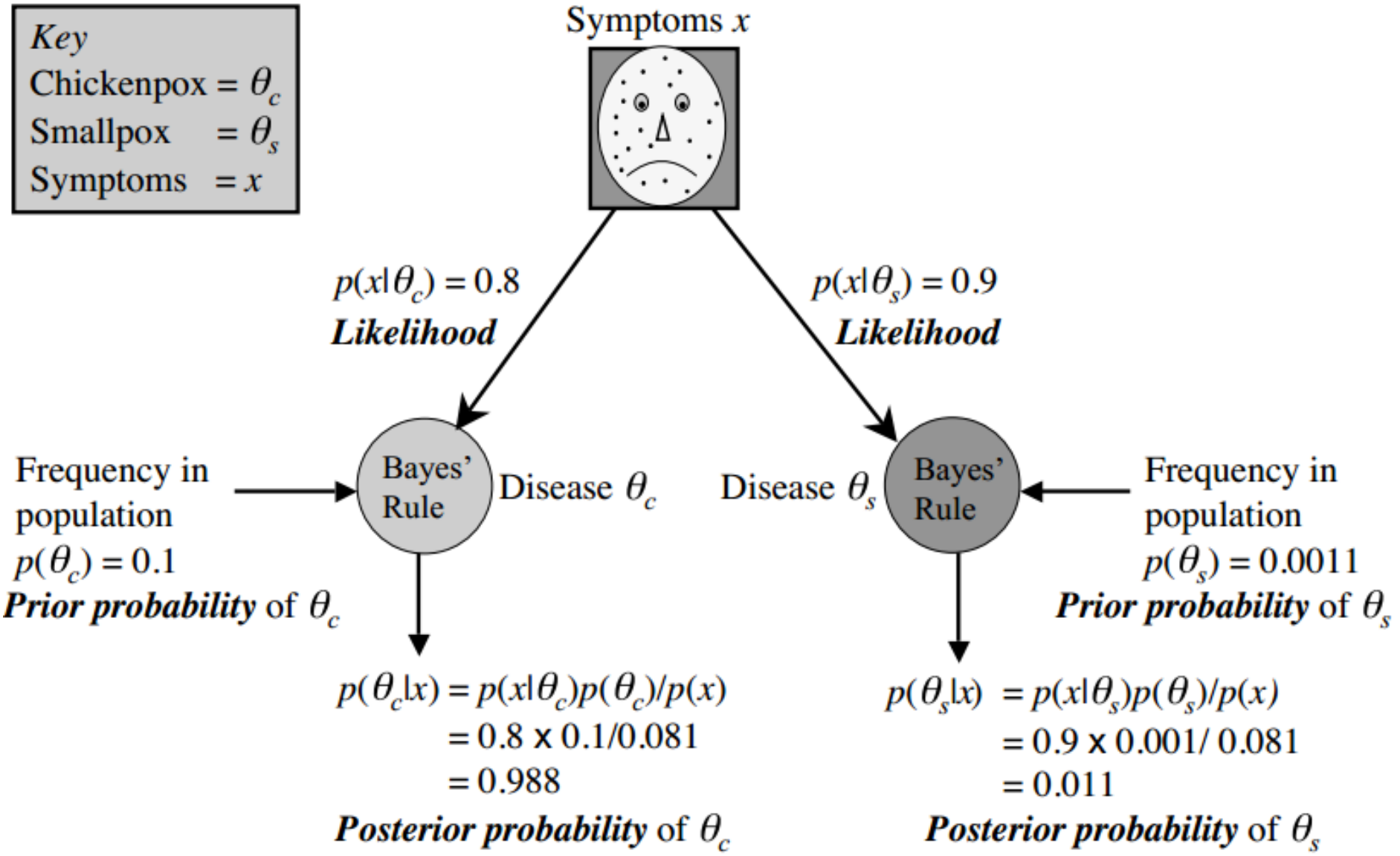
$$\begin{aligned} Be(x \mid \pi) &= \pi^x (1-\pi)^{1-x} \\ &= \exp\left[\log\left(\pi^x (1-\pi)^{1-x}\right)\right] \\ &= \exp\left[x \log \pi + (1-x)\log(1-\pi)\right] \\ &= \exp\left[x(\log \pi - \log(1-\pi)) + \log(1-\pi)\right] \\ &= \exp\left[x \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right] \end{aligned}$$

$$\begin{aligned} h(x) &= 1 \\ T(x) &= x \\ \eta &= \log\left(\frac{\pi}{1-\pi}\right) \\ g &= \log\left(\frac{1}{1-\pi}\right) \end{aligned}$$

Baye's Theorem Applications



Baye's Theorem – Poxy Diseases



Baye's Theorem – Eye Witness

A witness is 90% certain that a certain customer committed the crime. There were 20 people in the bar . . .
Would you convict the person?

- Everyone is presumed innocent until proven guilty:

$$p(X = \text{guilty}) = 1/20$$

- Eyewitness

$$p(Y = \text{eyewitness identifies} | X = \text{guilty}) = 0.9$$
$$\text{and } p(Y = \text{eyewitness identifies} | X = \text{not guilty}) = 0.1$$

Bayes Rule

$$\Pr(X|Y) = \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.1 \cdot 0.95} = 0.3213 = 32\%$$

But most judges would convict him anyway . . .

Baye's Theorem – Aristotle's Deduction

From Logic: "If *A* is true, then *B* is true", one may deduce that "If *B* is false, then *A* is false".

Q: How to reason this probabilistically?

$$p(A = fa \mid B = fa) = 1 - p(A = tr \mid B = fa)$$

$$= 1 - \frac{p(B = fa \mid A = tr)\, p(A = tr)}{p(B = fa \mid A = tr)\, p(A = tr) + p(B = fa \mid A = fa)\, p(A = fa)}$$

$$= 1 - 0 = 1$$

$$p(B = fa \mid A = tr) = 1 - p(B = tr \mid A = tr) = 1 - 1 = 0$$

Baye's Theorem – 9/11 Attacks

Most of us would have assigned almost no probability to terrorists crashing planes into buildings in New York.

But we recognized that a terror attack was an obvious possibility once the first plane hit the World Trade Center.

And we had no doubt we were being attacked once the second tower was hit.

Baye's theorem can replicate this result ==>

Baye's Theorem – 9/11 Attacks

PRIOR PROBABILITY		
Initial estimate of how likely it is that terrorists would crash planes into Manhattan skyscrapers.	x	0.005%
A NEW EVENT OCCURS: FIRST PLANE HITS WORLD TRADE CENTER		
Probability of plane hitting if terrorists are attacking Manhattan skyscrapers.	y	100%
Probability of plane hitting if terrorists are *not* attacking Manhattan skyscrapers (i.e. an accident).	z	0.008%
POSTERIOR PROBABILITY		
Revised estimate of probability of terror attack, given first plane hitting World Trade Center.	$\frac{xy}{xy + z(1-x)}$	38%

Baye's Theorem – 9/11 Attacks

PRIOR PROBABILITY		
Revised estimate of probability of terror attack, given first plane hitting World Trade Center.	x	38%
A NEW EVENT OCCURS: SECOND PLANE HITS WORLD TRADE CENTER		
Probability of plane hitting if terrorists are attacking Manhattan skyscrapers.	y	100%
Probability of plane hitting if terrorists are *not* attacking Manhattan skyscrapers (i.e. an accident).	z	0.008%
POSTERIOR PROBABILITY		
Revised estimate of probability of terror attack, given second plane hitting World Trade Center.	$\frac{xy}{xy + z(1-x)}$	99.99%

Conjugate Priors

- **Conjugacy**: posterior distribution p($\theta|x$) is in the same family as the prior distribution p(θ)

- Example: Binomial data and Beta prior

$$Bin(x \mid \theta) \qquad P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

$$Be(\theta \mid \alpha_1, \alpha_2) \qquad P(\theta) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) + \Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{(\alpha_2 - 1)}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Conjugate Priors

$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

$$= \binom{n}{x} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) + \Gamma(\alpha_2)} \theta^{x+\alpha_1-1}(1-\theta)^{(n-x+\alpha_2-1)}$$

$$P(\theta|x) \propto \theta^{x+\alpha_1-1}(1-\theta)^{(n-x+\alpha_2-1)}$$

- This is simply a **Beta posterior distribution**

$$Be(x+\alpha_1, n-x+\alpha_2)$$

Conjugate Priors – Coin Toss Example

- You have a coin that when flipped ends up head with probability θ and ends up tail with probability (1 - *θ*).
- Trying to estimate *θ*, you flip the coin **14** times. It ends up head **10** times.
- What is the probability of: "*In the next two tosses we will get two heads in a row*"?
- **Would you bet on "yes"?**
- With **Frequentist** approach: *θ* is 10/14, i.e., θ ≈ 0.714.
- In this case, the probability of two heads is 0.7142^2 ≈ 0.51 and it makes sense to bet for the event. Therefore, the frequentist will bet "yes"!

Conjugate Priors – Coin Toss Example

- **Black Swan Paradox** with Frequentist approach: If we observe 0 heads in 14 trials, then θ is 0/14 = 0.

- Analagous to “if I see only white swans, it implies no black swans exist!” => **sparse data problem**.

- Overcome this problem with Bayesian approach

Conjugate Priors – Coin Toss Example

- **Bayesian** (probabilistic) approach:
- Due to **binomial** data, Likelihood function is

$$P(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}$$

$$P(data|\theta) = \binom{14}{10}\theta^{10}(1-\theta)^4$$

- Putting a prior on θ (**Beta**):

$$P(\theta) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)+\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{(\alpha_2-1)}$$

- **Posterior distribution** is: $Be(x+\alpha_1, n-x+\alpha_2)$

$$Be(10+\alpha_1, 4+\alpha_2)$$

Conjugate Priors – Coin Toss Example

- Bayesian (probabilistic) approach:
- **Posterior predictive distribution** is

$$Pr\{HH|data\} = \int_0^1 Pr\{HH|\theta\} \cdot P(\theta|data)d\theta$$

$$= \frac{1}{B(10+\alpha_1, 4+\alpha_2)} \int_0^1 \theta^2 \theta^{10+\alpha_1-1}(1-\theta)^{4+\alpha_2-1}$$

- When $\alpha_1=\alpha_2=1$, this is 0.485.
- So, the Bayesian will bet “no”!

Conjugate Priors – Language Modeling

- Language Modeling with **Dirichlet-multinomial** model: predicting which words might occur next in a sequence
- **Goal**: Given a past sequence of words, how can we predict which one is likely to come next?
- Suppose we observe the following sequence:

```
Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow
```

- Consider (a small) vocabulary of words:

```
mary lamb little big fleece white black snow rain unk
1    2    3      4   5      6     7     8    9    10
```

- Encoding the message:

```
1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8
```

Conjugate Priors – Language Modeling

- Histogram of word counts:

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

- Using a Dirichlet prior, the posterior predictive is

$$p(\tilde{X} = j|D) = E[\theta_j|D] = \frac{\alpha_j + N_j}{\sum_{j'} \alpha_{j'} + N_{j'}} = \frac{1 + N_j}{10 + 17}$$

- With $\alpha_j = 1$, we get

$$p(\tilde{X} = j|D) = (3/27, 5/27, 5/27, 1/27, 2/27, 2/27, 1/27, 2/27, 1/27, 5/27)$$

- **Modes**: $X = 2$ (“lamb”), $X = 3$ (“little”) and $X = 10$ (“unk”).
- **Note**: words “big”, “black” and “rain” are predicted to occur with non-zero probability in the future, even though they have never been seen before!

Correlation

Covariance:

$$\operatorname{cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$= \begin{pmatrix} \operatorname{var}[X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_d] \\ \operatorname{cov}[X_2, X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_d, X_1] & \operatorname{cov}[X_d, X_2] & \cdots & \operatorname{var}[X_d] \end{pmatrix}$$

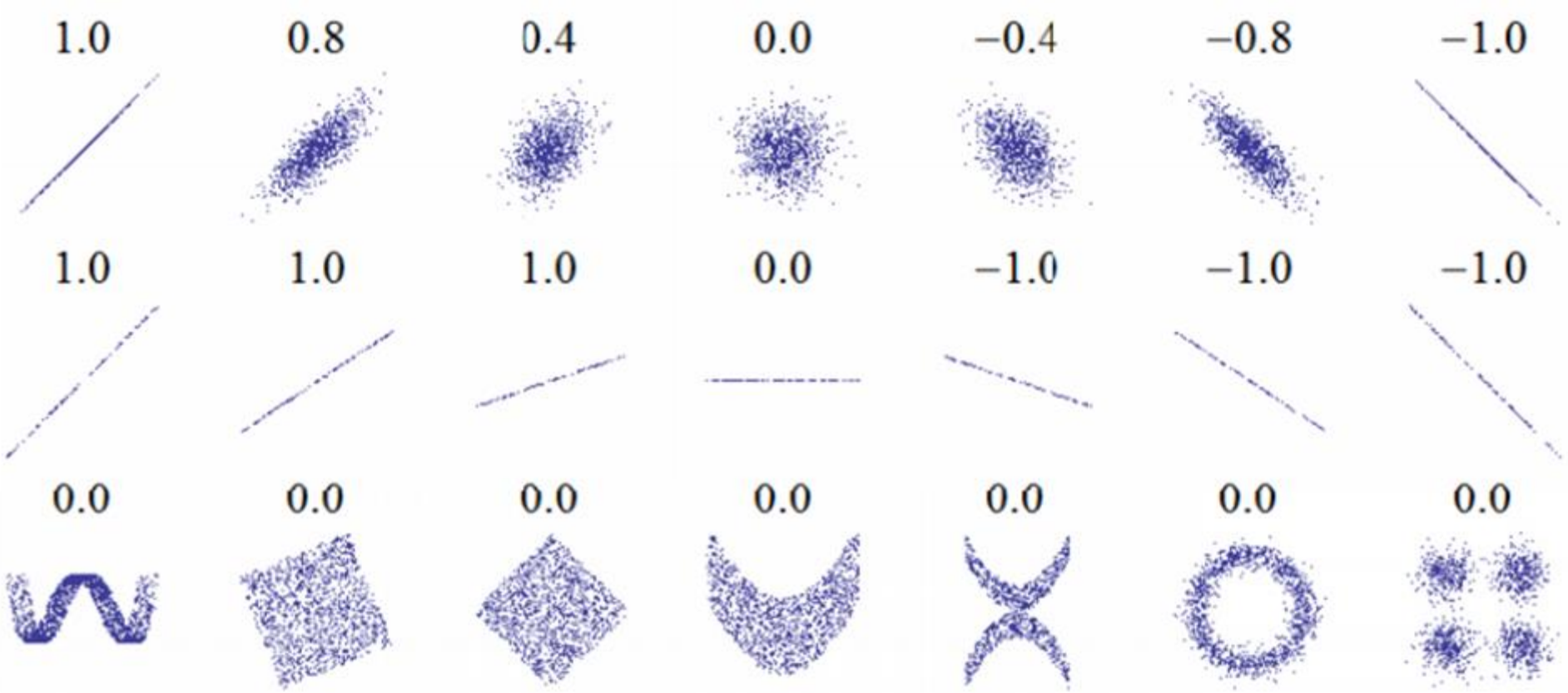
Correlation:

$$\operatorname{corr}[X, Y] \triangleq \frac{\operatorname{cov}[X, Y]}{\sqrt{\operatorname{var}[X] \operatorname{var}[Y]}} \qquad -1 \leq \operatorname{corr}[X, Y] \leq 1$$

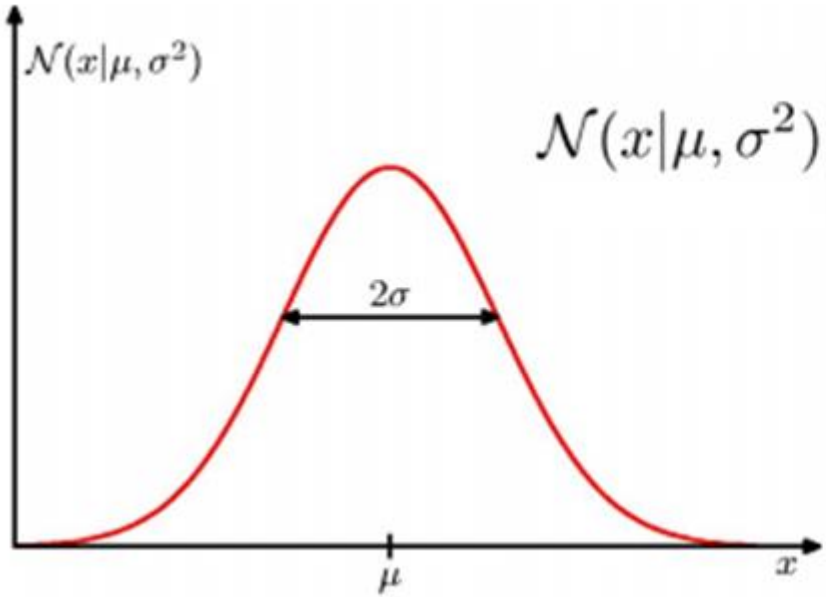
Independence:

$$p(X, Y) = p(X)p(Y) \Rightarrow \operatorname{cov}[X, Y] = 0 \Leftrightarrow \operatorname{corr}[X, Y] = 0$$

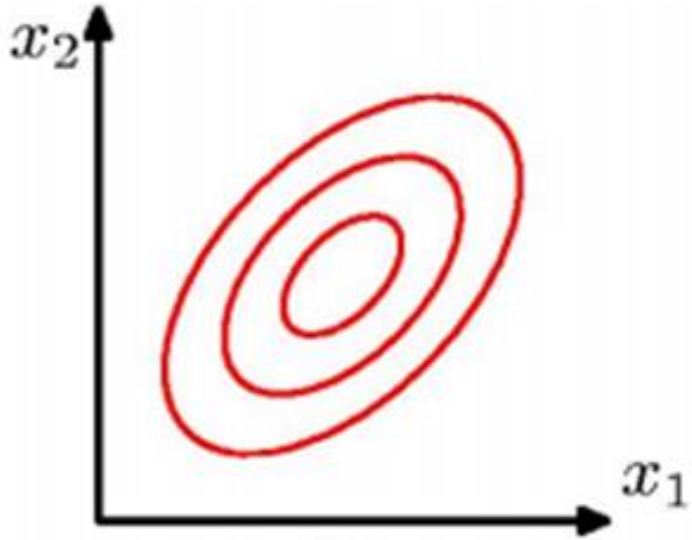
Correlation



Multi-variate Gaussians

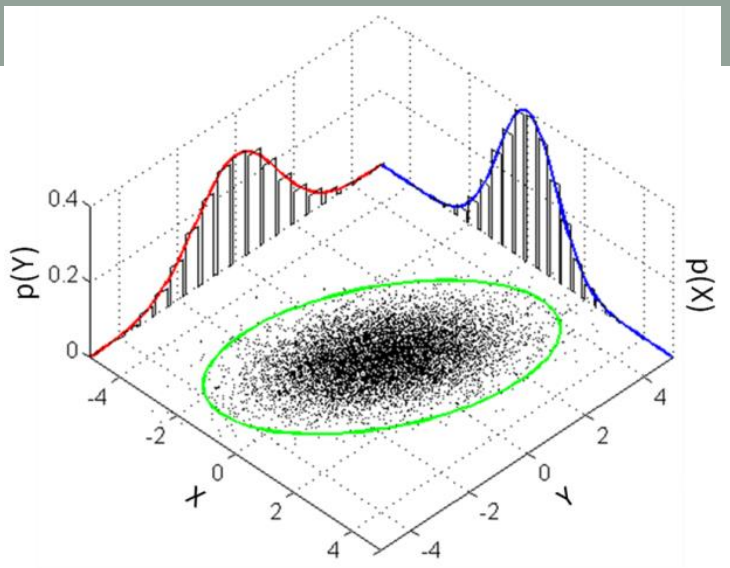


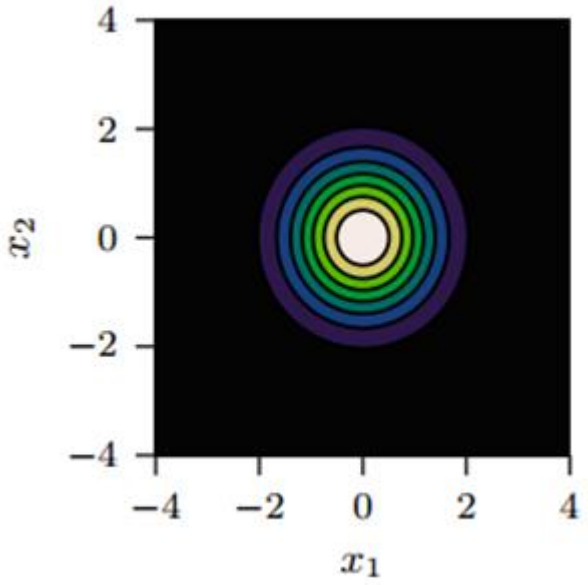
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



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

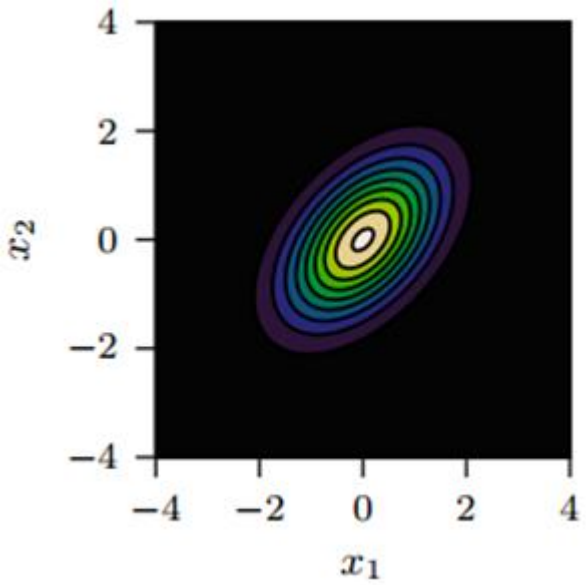
2-D Gaussians





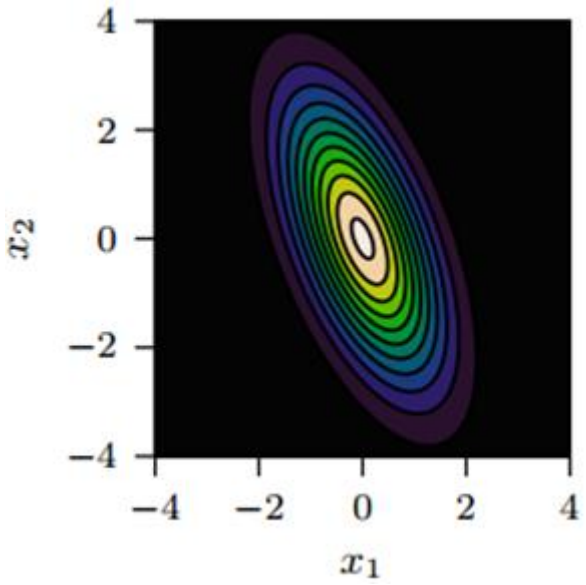
(a) $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\boldsymbol{\mu} = \mathbf{0}$



(b) $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$

$\boldsymbol{\mu} = \mathbf{0}$



(c) $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}$

$\boldsymbol{\mu} = \mathbf{0}$

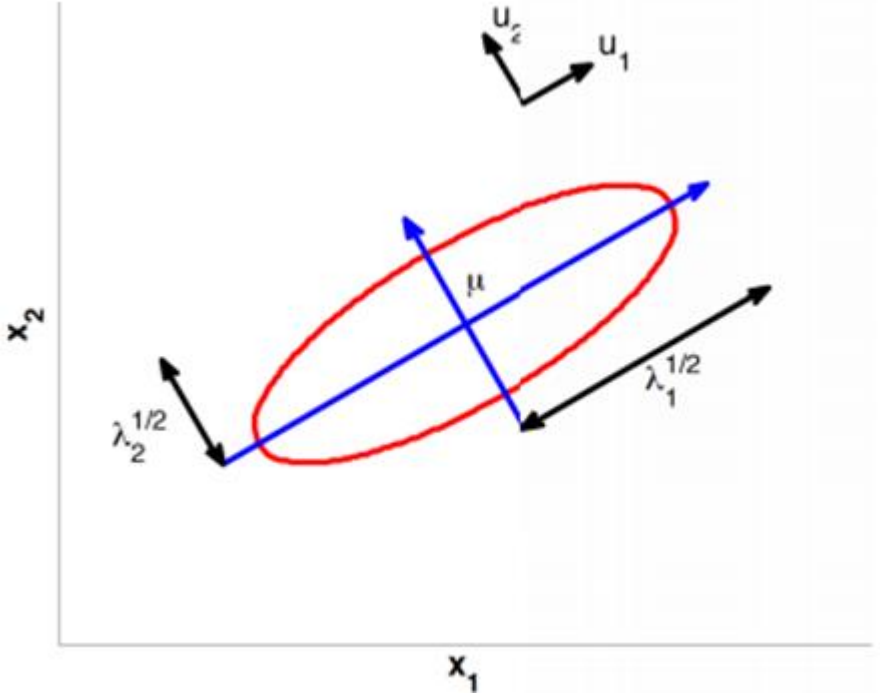
Gaussian Geometry

- Eigenvalues and eigenvectors:

$$\Sigma u_i = \lambda_i u_i, i = 1, \ldots, d$$

- For a *symmetric* matrix:

$$\lambda_i \in \mathbb{R} \qquad u_i^T u_i = 1 \qquad u_i^T u_j = 0$$



Eigen-decomposition:

$$\Sigma = U \Lambda U^T = \sum_{i=1}^{d} \lambda_i u_i u_i^T$$

$$\Sigma^{-1} = U \Lambda^{-1} U^T = \sum_{i=1}^{d} \frac{1}{\lambda_i} u_i u_i^T$$

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

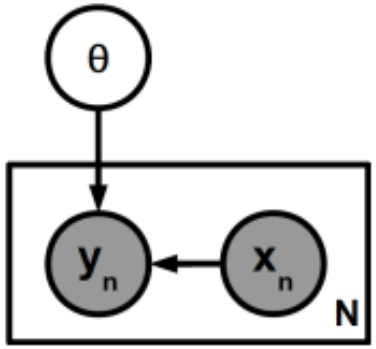
$$y_i = u_i^T (x - \mu)$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

Probabilistic Supervised Learning

- Consider regression/classification scenarios.
- Training data $\{(x_1, y_1), \ldots, (x_N, y_N)\}$
- **Goal**: Learn a function to predict outputs ***y*** from inputs ***x***
- **Solution**: Model the output as a probability distribution

$$y_1, \ldots, y_N \sim p(y|x, \theta)$$

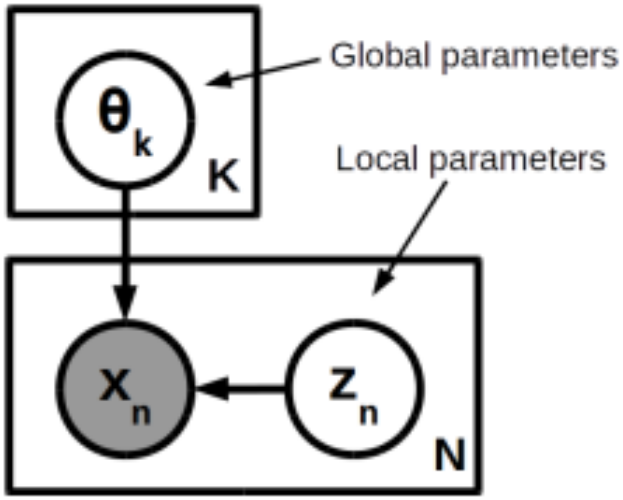


- Learning => estimating parameter θ for given data
- **Advantage**: can make probabilistic predictions for new data

$$p(y_*|x_*, \theta)$$

Probabilistic Unsupervised Learning

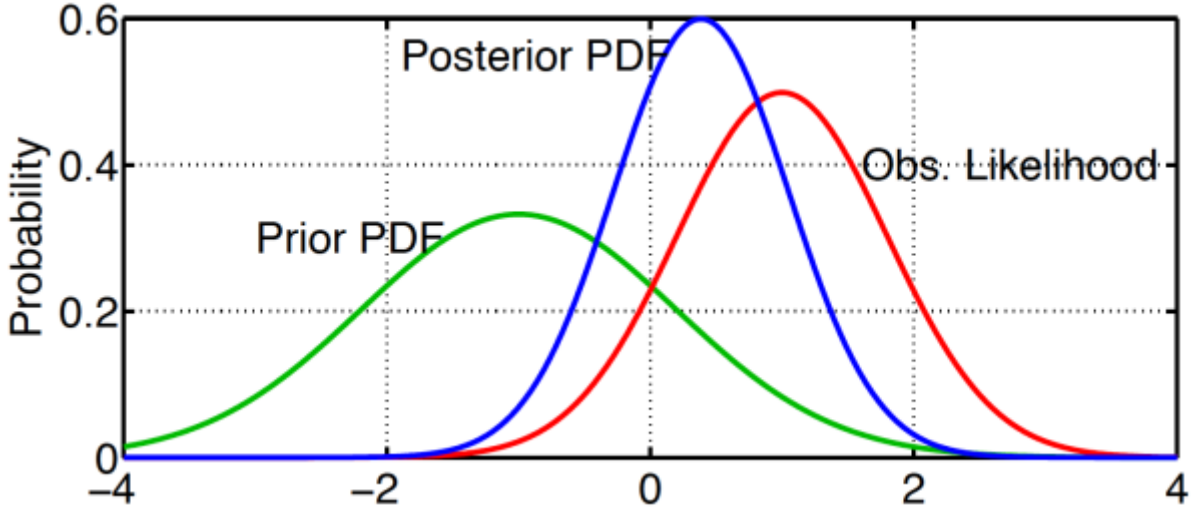
- Consider **clustering** or **dimensionality reduction** problems.
- Each data point $\boldsymbol{x}_n$ assumed to be generated via some latent variable $\boldsymbol{z}_n$ and parameters θ



- **Clustering**: $\boldsymbol{z}_n$ denotes with cluster $\boldsymbol{x}_n$ belongs to
- **Dim. Reduction**: $\boldsymbol{z}_n$ is compressed representation of $\boldsymbol{x}_n$
- **Learning**: estimating parameters θ and latent variables $\boldsymbol{z}_n$ given data

Benefits of Probabilistic Modeling

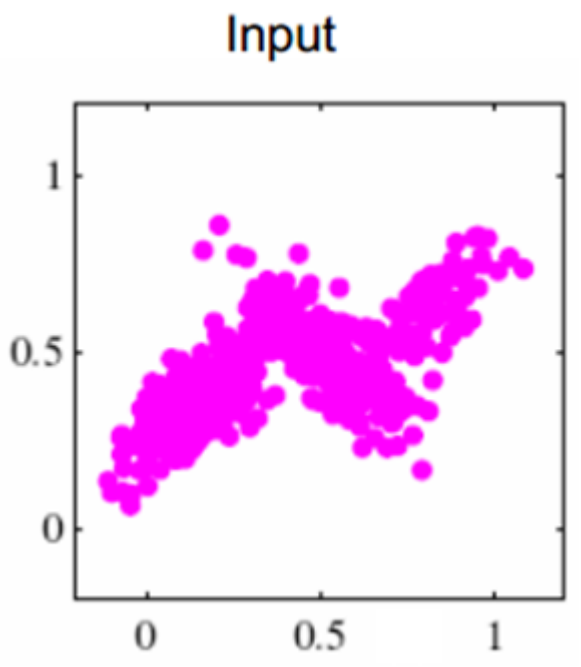
- Can get estimate of **uncertainty** in parameter estimates via the posterior distribution

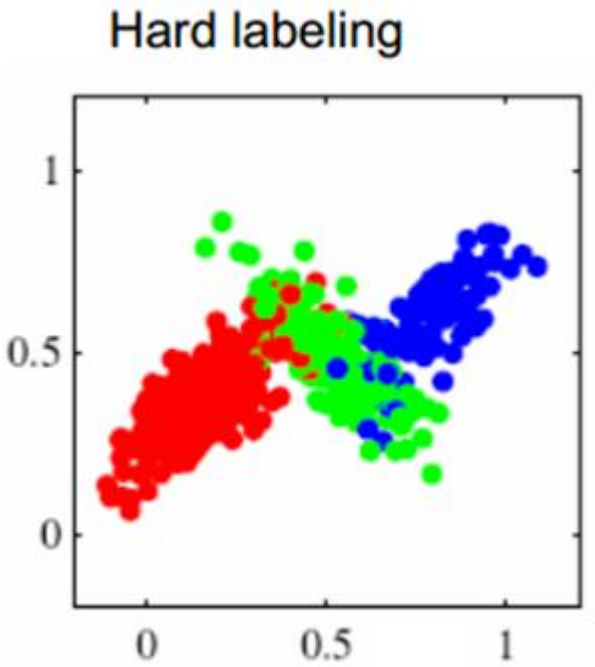


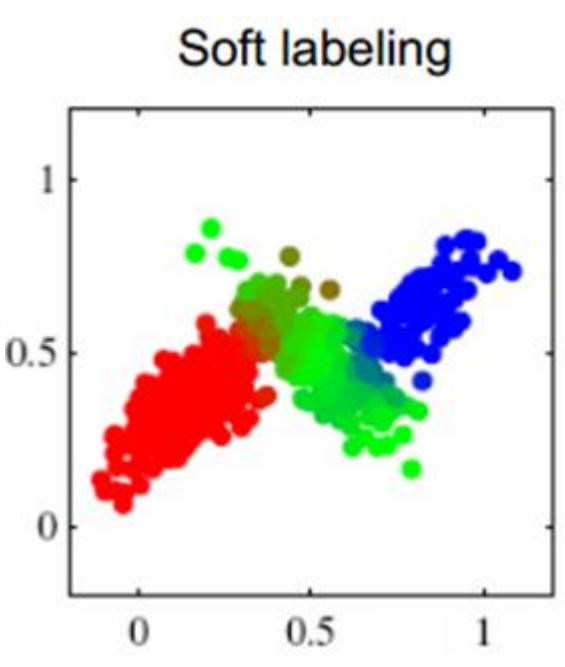
- Instead of point estimate, we know the **full posterior** distribution => **generative model** (knowing how data was generated)

Applications – Mixture Modeling

Desired output







Applications – Mixture Modeling

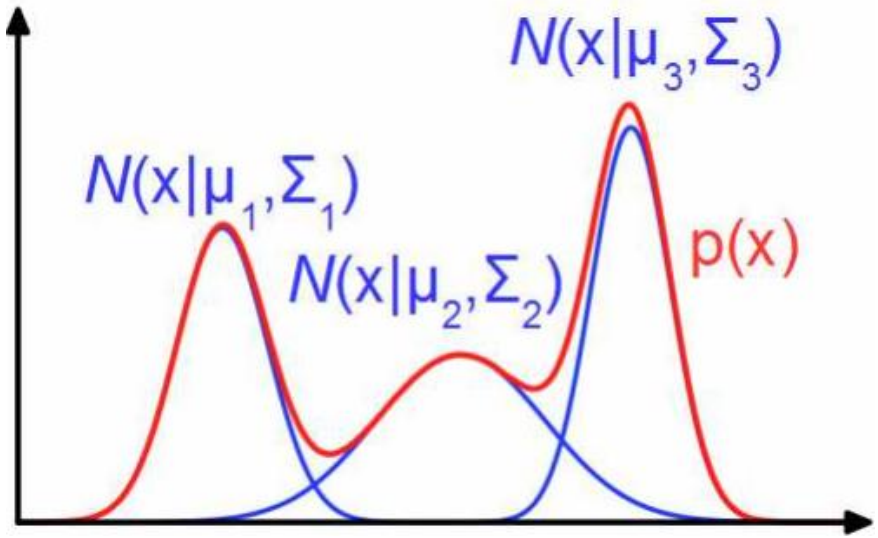
Data distribution $p(x)$ assumed to be a weighted sum of K distributions

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\theta_k)$$

where π_k's are the mixing weights: $\sum_{k=1}^{K} \pi_k = 1, \quad \pi_k \geq 0$ (intuitively, π_k is the proportion of data generated by the k-th distribution)

Gaussian Mixture Model (GMM): component distributions are Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$



Applications – Mixture Modeling

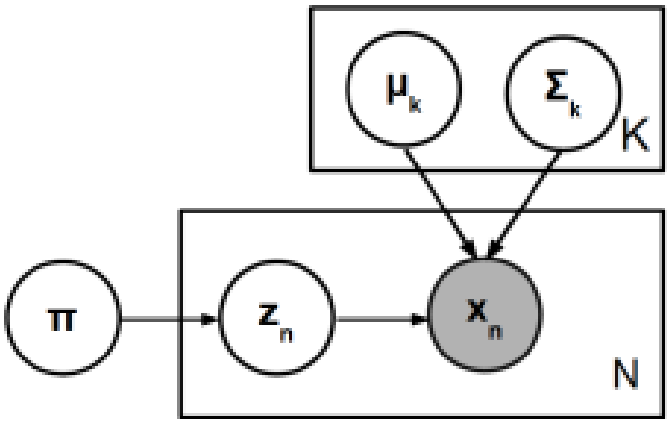
Can think of the data $\{x_1, x_n, \ldots, x_N\}$ using a "generative story"

- For each example x_n, first choose its cluster assignment $z_n \in \{1, 2, \ldots, K\}$ as

$$z_n \sim \text{Multinoulli}(\pi_1, \pi_2, \ldots, \pi_K)$$

- Now generate x from the Gaussian with id z_n

$$x_n | z_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$$



Shaded nodes: Observed

White nodes: Unknowns

$$p(z_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

$p(z_{nk} = 1) = \pi_k$ is the prior probability of x_n going to cluster k

Applications – Mixture Modeling

Initialize parameters $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and mixing weights $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$, and alternate between the following steps until convergence:

- Given current estimates of $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\boldsymbol{\pi}$
  - Estimate the posterior probabilities of cluster assignments

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \qquad \forall n, k$$

- Given the current estimates of cluster assignment probabilities $\{\gamma_{nk}\}$
  - Estimate the mean of each Gaussian

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} \boldsymbol{x}_n \qquad \forall k, \text{where } N_k = \sum_{n=1}^N \gamma_{nk}$$

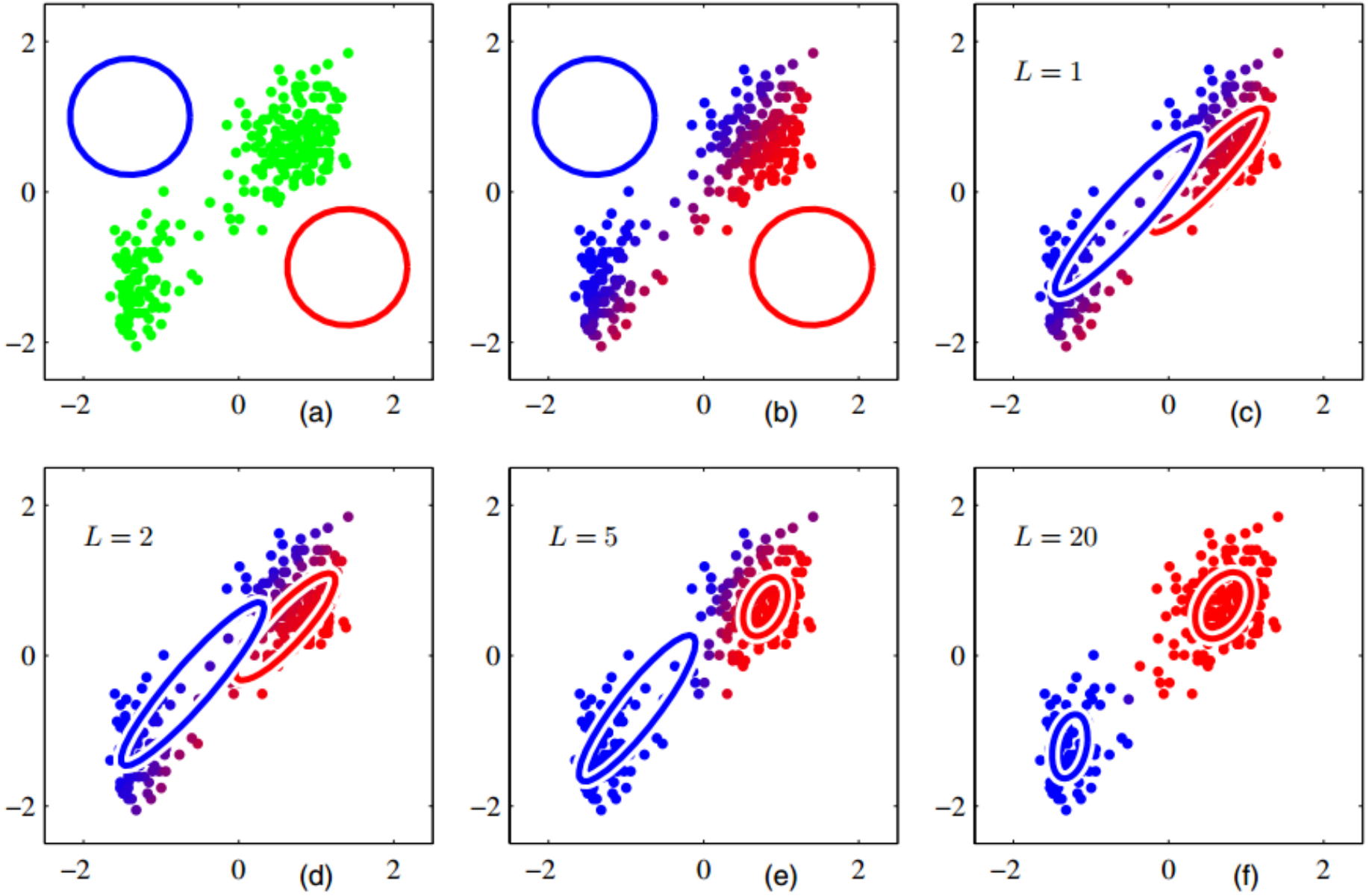
  - Estimate the covariance matrix of each Gaussian

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^\top \qquad \forall k$$

  - Estimate the mixing proportion of each Gaussian

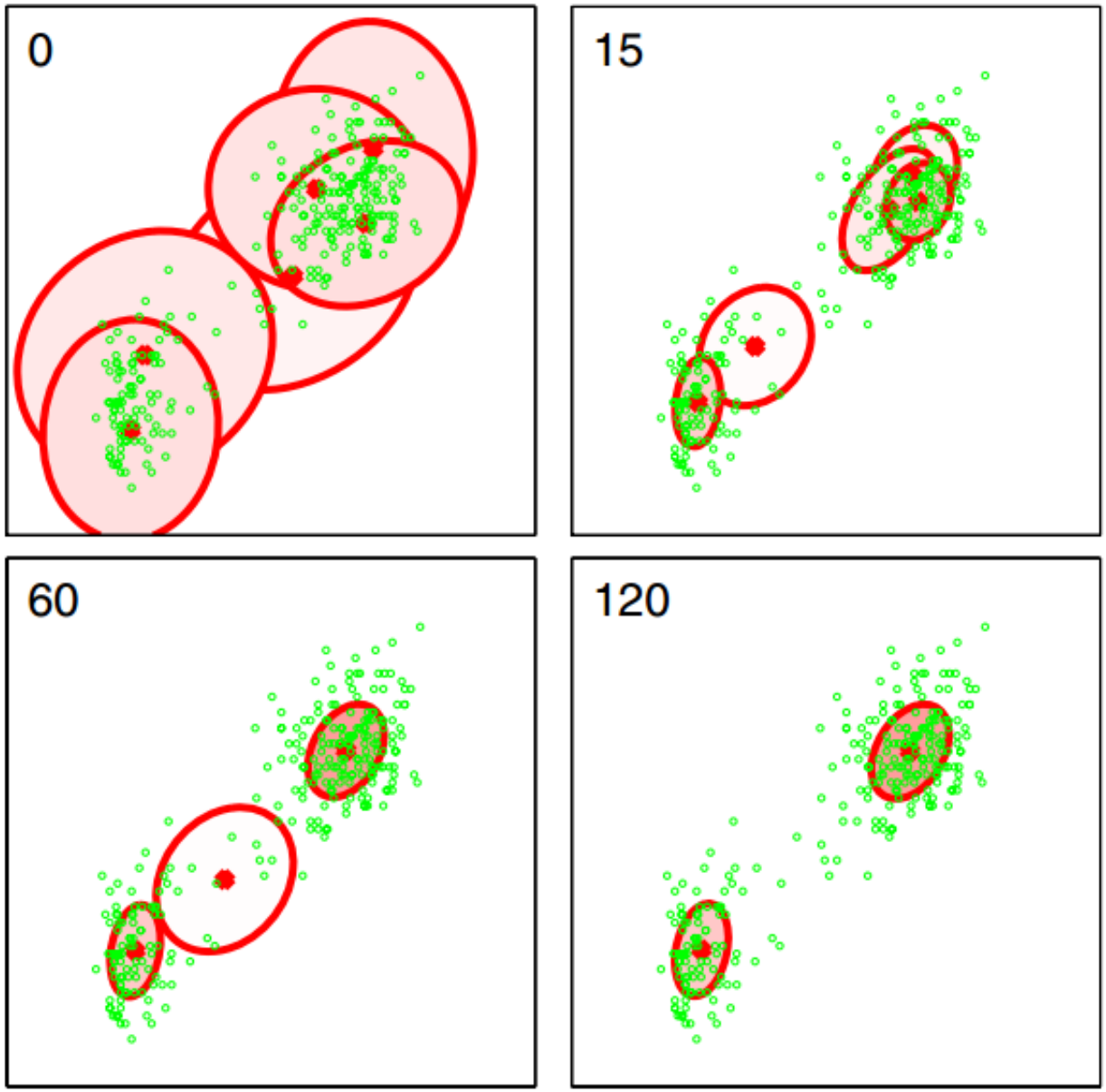
$$\pi_k = \frac{N_k}{N} \qquad \forall k$$

Applications – Mixture Modeling



Applications – Mixture Modeling

Variational Bayesian Inference:



Applications – Linear Regression

Given: N training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, features: $\boldsymbol{x}_n \in \mathbb{R}^D$, response $y_n \in \mathbb{R}$

$\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]^\top$: $N \times D$ feat. matrix, $\mathbf{Y} = [y_1 \dots y_N]^\top$: $N \times 1$ resp. vector

Probabilistic view: responses are generated via a probabilistic model

Assume a "noisy" linear model with regression weight vector $\boldsymbol{w} \in \mathbb{R}^D$:

$$y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$$

Gaussian noise: $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$, β: precision (inverse variance) of Gaussian

$$y_n \sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$$

Goal: Learn regression weight vector $\boldsymbol{w}$ to predict y_* for a new $\boldsymbol{x}_*$

Applications – Linear Regression

For Gaussian response y_n

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \sqrt{\frac{\beta}{2\pi}} \exp\left\{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right\}$$

Thus the likelihood (assuming i.i.d. responses) or *probability* of data:

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right\}$$

Log-likelihood (ignoring constants w.r.t. $\boldsymbol{w}$)

$$\log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) \propto -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

Applications – Linear Regression (MLE)

MLE: Find the $\boldsymbol{w}$ that maximizes the (log) likelihood $\log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w})$

$$\arg\max_{\boldsymbol{w}} \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \arg\min_{\boldsymbol{w}} -\log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

Taking derivative w.r.t. $\boldsymbol{w}$ and setting to zero, we get

$$\boldsymbol{w}_{MLE} = \left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top\right)^{-1} \sum_{n=1}^{N} y_n \boldsymbol{x}_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

- $\mathbf{X}^\top \mathbf{X}$ may be ill-conditioned (not invertible)
- "Uncontrolled" $\boldsymbol{w}$ can lead to overfitting (thus need regularization)

Applications – Linear Regression (MAP)

Assume zero-mean spherical Gaussian prior on weights $\boldsymbol{w} = [w_1 \; w_2 \ldots \; w_D]$

$$p(\boldsymbol{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I}_D) = \left(\frac{\lambda}{2\pi}\right)^{D/2} \exp(-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}) = \left(\frac{\lambda}{2\pi}\right)^{D/2} \exp(-\frac{\lambda}{2}||\boldsymbol{w}||^2)$$

The posterior distribution on $\boldsymbol{w}$: $p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})$

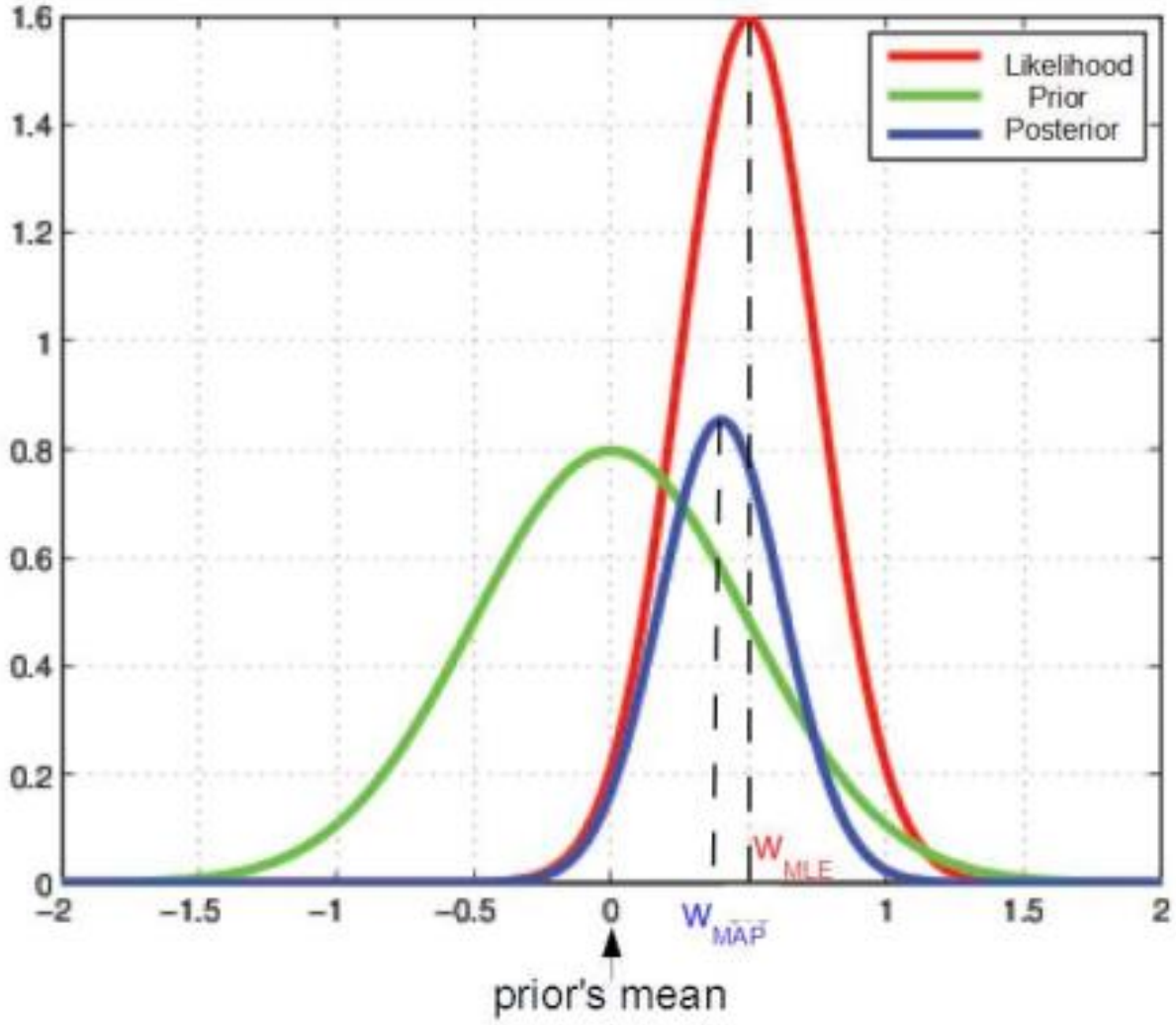
The (log) posterior: $\log p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) = \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) + \log p(\boldsymbol{w})$.

$$\log p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) \quad \propto \quad -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 - \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} \quad \text{(ignoring constants w.r.t } \boldsymbol{w})$$

$$\boldsymbol{w}_{MAP} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1} \sum_{n=1}^{N} y_n \boldsymbol{x}_n = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\top \mathbf{Y}$$

Applications – Linear Regression (MAP)

$\boldsymbol{w}_{MAP}$ is a compromise between prior's mean and $\boldsymbol{w}_{MLE}$



Applications – Linear Regression (MAP)

MLE Objective

$$\arg\max_{w} \log p(\mathbf{Y}|\mathbf{X}, w) = \arg\min_{w} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - w^\top x_n)^2$$

MLE solution

$$\boxed{w_{MLE} = (\sum_{n=1}^{N} x_n x_n^\top)^{-1} \sum_{n=1}^{N} y_n x_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}$$

MAP Objective

$$\arg\max_{w} \log p(w|\mathbf{X}, \mathbf{Y}) \propto \arg\max_{w} \log p(\mathbf{Y}|\mathbf{X}, w) p(w) = \arg\min_{w} \sum_{n=1}^{N} (y_n - w^\top x_n)^2 + \frac{\lambda}{\beta} w^\top w$$

MAP solution

$$\boxed{w_{MAP} = (\sum_{n=1}^{N} x_n x_n^\top + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \sum_{n=1}^{N} y_n x_n = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \mathbf{Y}}$$

Applications – Linear Regression (Bayesian)

MLE/MAP only provide a point estimate of $\boldsymbol{w}$ (no estimate of uncertainty)

Infer the full posterior of $\boldsymbol{w}$: $$p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})}{p(\mathbf{Y}|\mathbf{X})}$$

Since the likelihood and the prior, both, are Gaussian, the posterior will also be Gaussian (due to conjugacy)

$$p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\text{where} \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \mathbf{Y})$$

$$\boldsymbol{\Sigma} = (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1}$$

Applications – Linear Regression (Bayesian)

MLE/MAP only use a point estimate ($\boldsymbol{w}_{MLE}/\boldsymbol{w}_{MAP}$) for making prediction

Fully Bayesian approach of making predictions is via the predictive posterior

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \mathbf{Y}) = \int_{\boldsymbol{w}} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}) p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) d\boldsymbol{w}$$

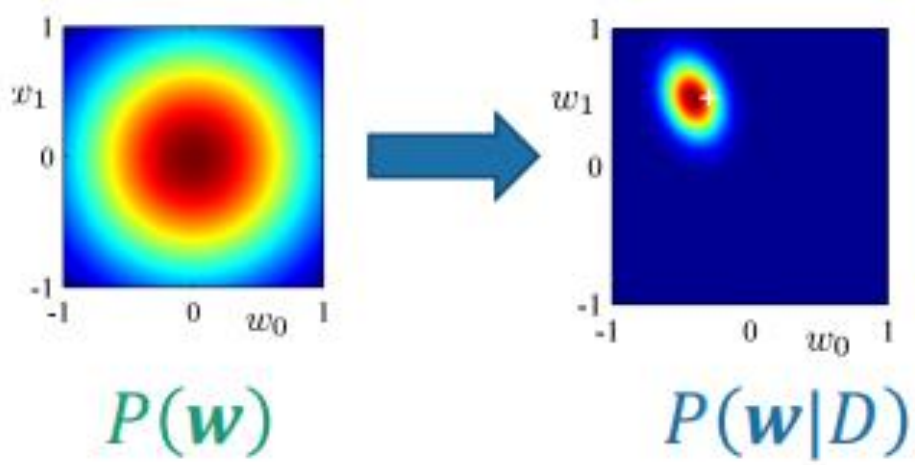
(Predictive Posterior)

Predictive Posterior: Don't use a single $\boldsymbol{w}$ to make predictions but average $p(y_*|\boldsymbol{x}_*, \boldsymbol{w})$ over all possible $\boldsymbol{w}$'s (each weighted by its posterior probability)

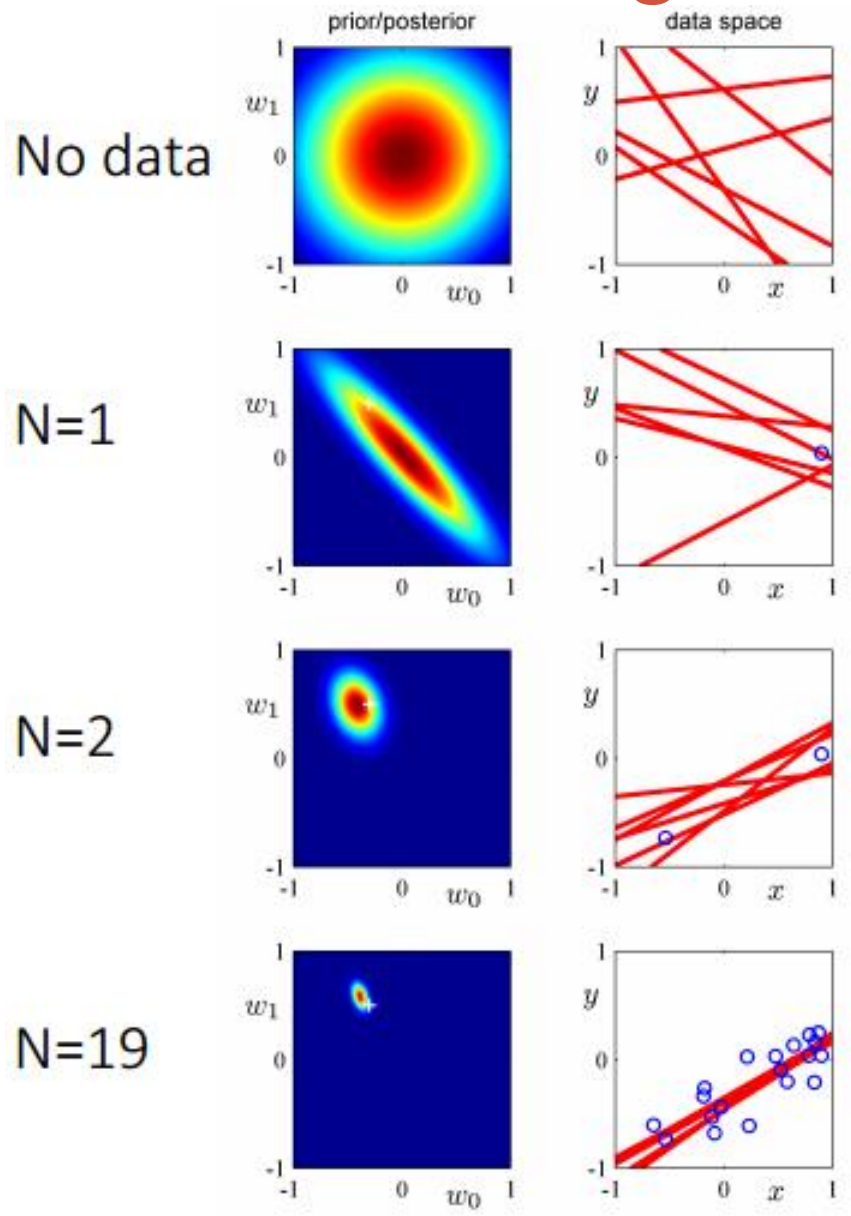
$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma} \boldsymbol{x}_*)$$

Applications – Linear Regression (Bayesian)

We now have the **posterior distribution** instead of a **single best value**.
It contains our knowledge about the compatibility of all possible solutions with our data and assumptions.

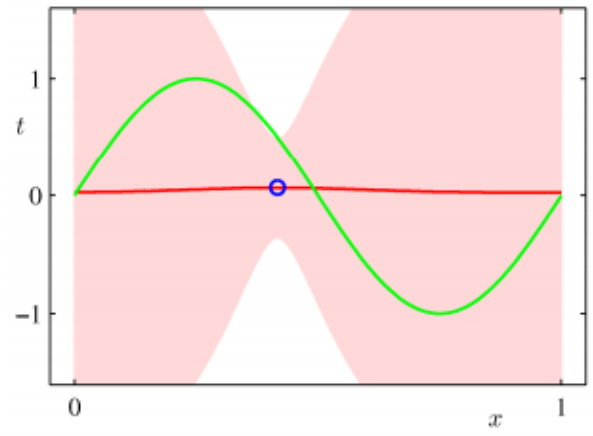


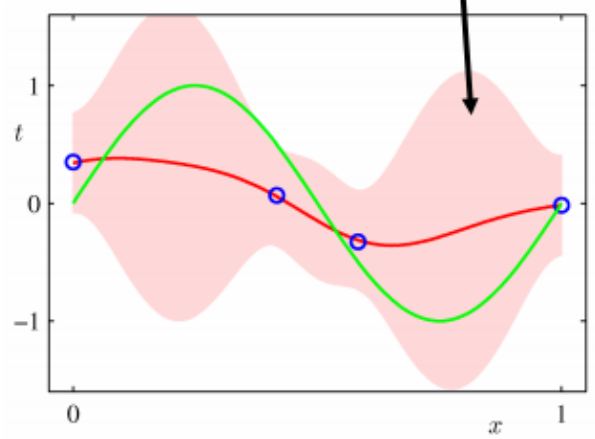
Applications – Linear Regression (Bayesian)

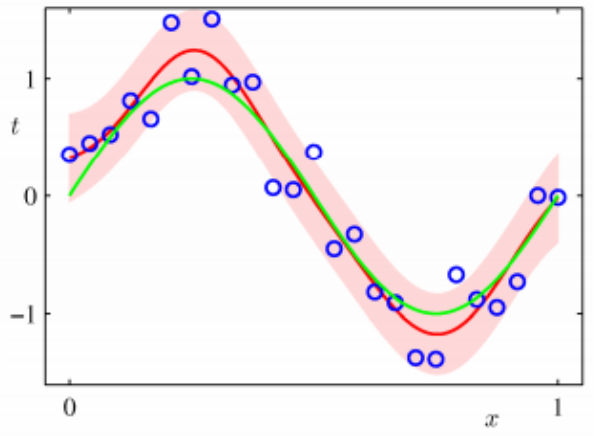


Applications – Polynomial Fitting (Bayesian)

Tells us about the outcome's certainty!







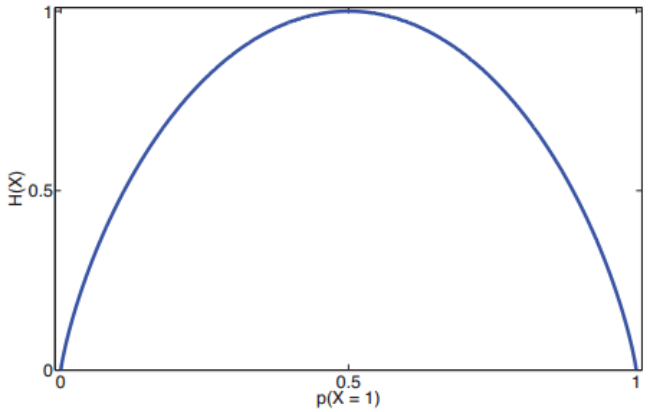
Applications – Information Theory

- Entropy $H[p(\mathbf{x})]$ of distribution $p(\mathbf{x})$ → non-negative measure of amount of "uncertainty"

$$H[p(\mathbf{x})] = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

- For binary RV's, p(X=1) = θ and p(X=0) = 1 – θ; so, entropy is

$$\begin{aligned} \mathbb{H}(X) &= -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \\ &= -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)] \end{aligned}$$



Applications – Information Theory

- Relative Entropy or **Kullback-Leibler (KL) divergence**: measure of divergence between two distributions *p* and *q*

$$KL(p \,\|\, q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$$

$$KL_{Ber}(p \,\|\, q) = p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)}$$

- Since KL is not a distance, it is not symmetric. To use it as a distance measure, use **Jensen-Shannon divergence**:

$$JS(p_1, p_2) = 0.5KL(p_1 \,\|\, q) + 0.5KL(p_2 \,\|\, q)$$

$$q = 0.5p_1 + 0.5p_2$$

Applications – Monte Carlo Integration

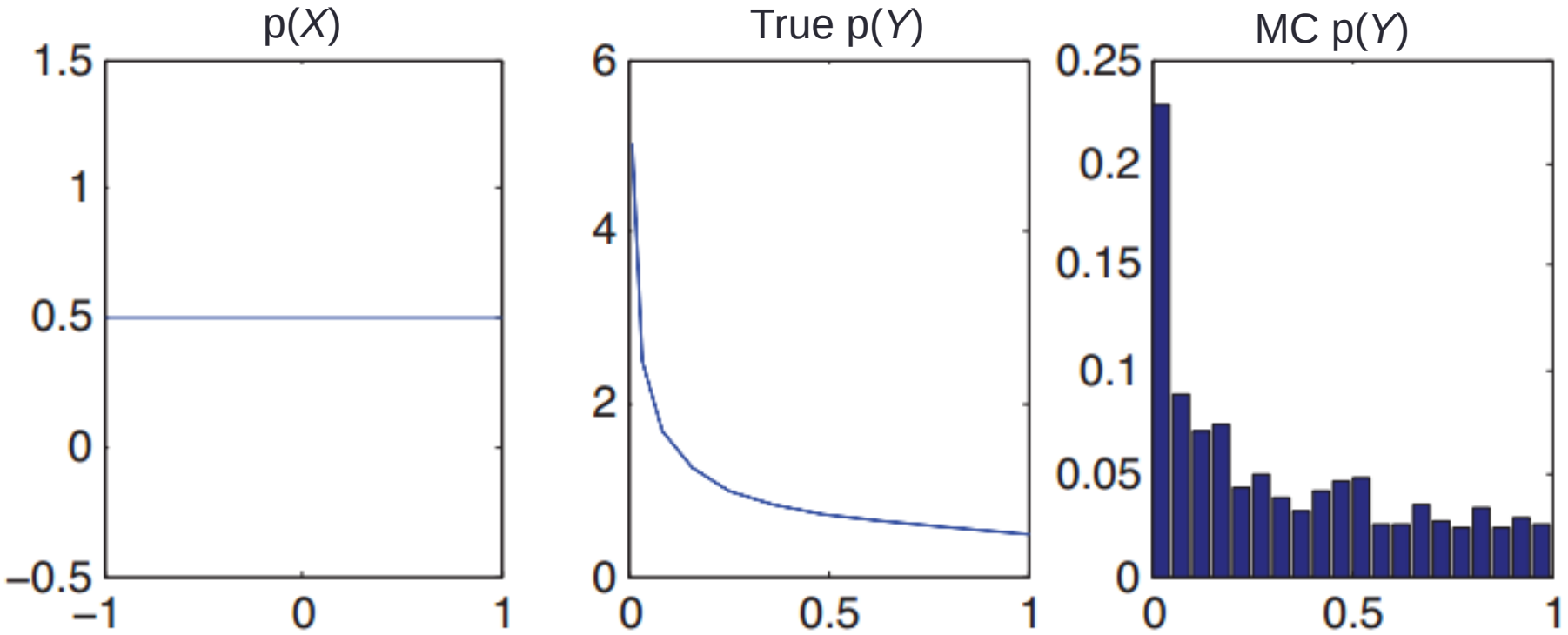
- Used to approximately calculate an integral analytically
- Generate *S* samples from the distribution, call them $x_1, \ldots, x_S$
- Given the samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$

- Eg: Using Monte Carlo, we can approximate the expected value of any function of RV: Simply draw samples and compute arithmetic mean!

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

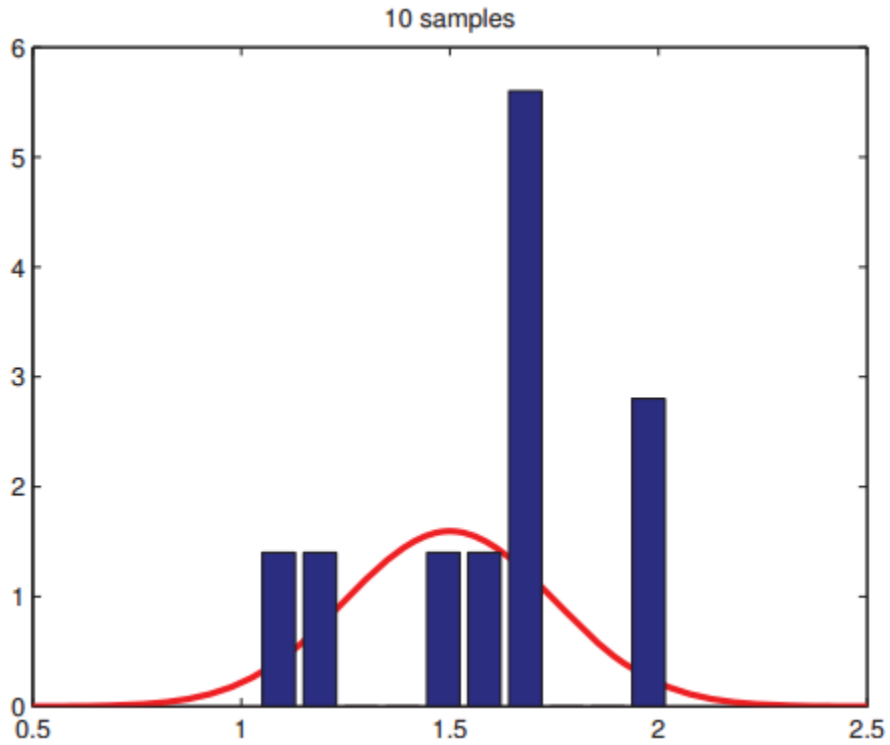
Applications – Monte Carlo Integration

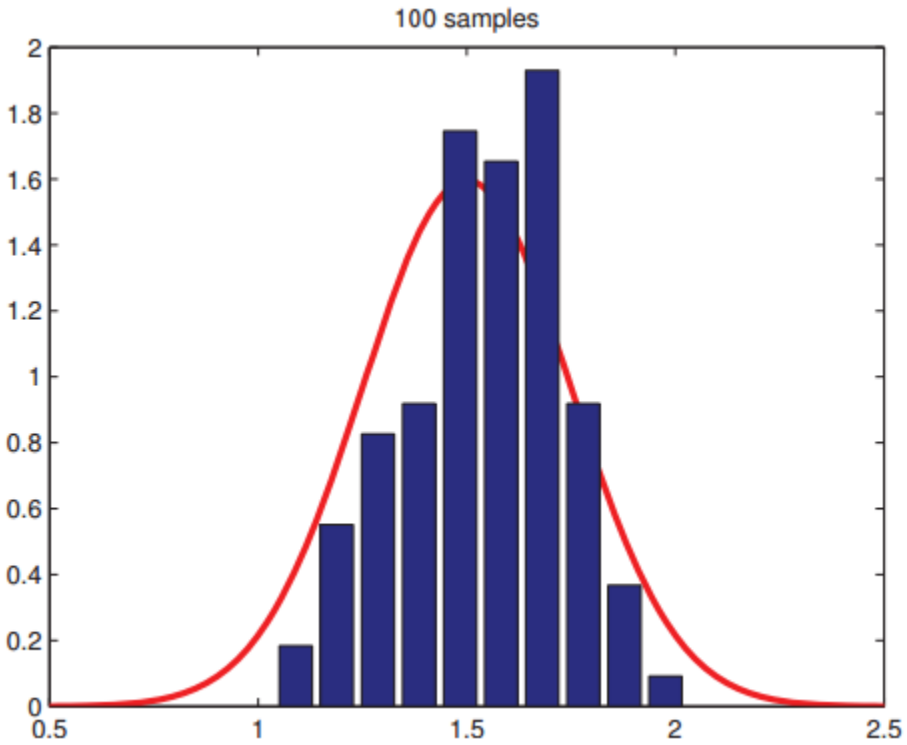
- Eg: Change of Variables using Monte Carlo
- Let $X \sim \text{Unif}(-1, 1)$, $Y = X^2$.
- Approximate p(Y) by drawing samples from p(X), squaring them and computing the resulting empirical distribution.



Applications – Monte Carlo Integration

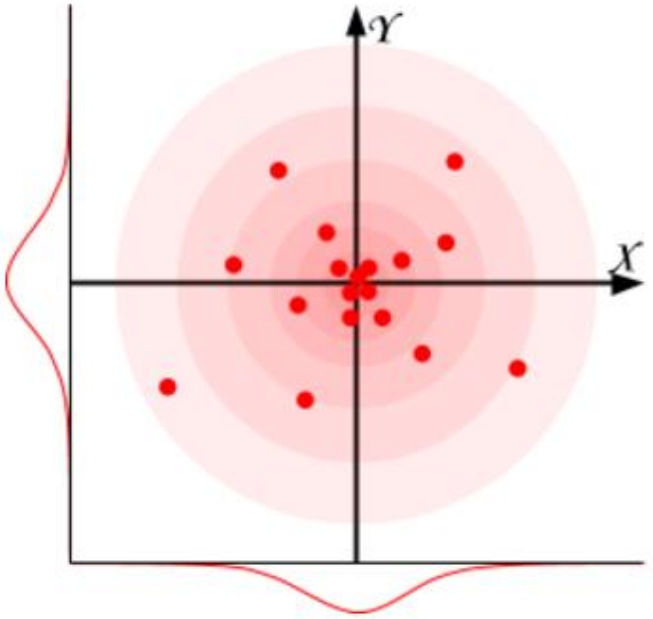
- Accuracy of Monte Carlo approximation increases as number of samples increases





Applications – Gaussian random numbers

- **From Gaussian to Uniform:**
- Consider two IID Gaussian RVs: X ~ N(0,1), Y ~ N(0,1); sample x and y from these:



- Joint probability

$$P(x,y) = P(x)P(y) =$$

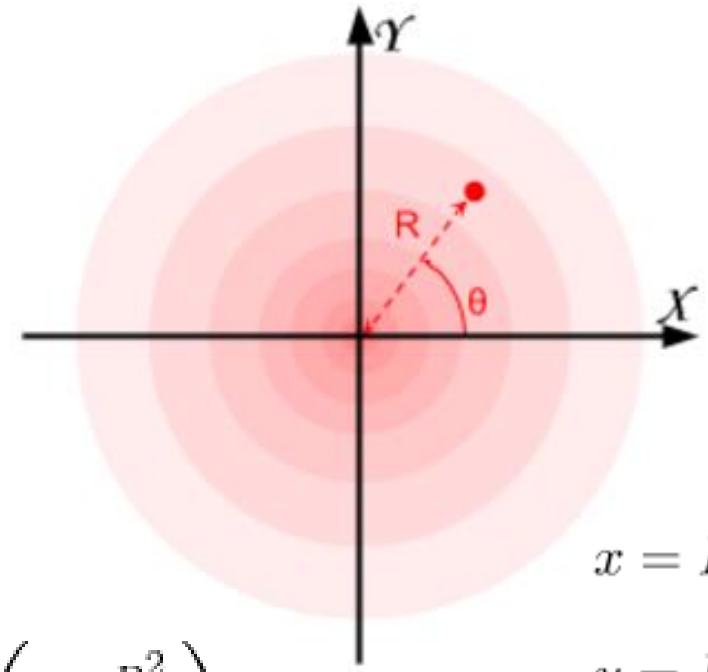
$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

Applications – Gaussian random numbers

- Switch to polar co-ordinates:

$$R = \sqrt{x^2 + y^2}$$

$$\theta = arctan\left(\frac{y}{x}\right)$$



$$x = Rcos\,(\theta)$$

$$y = Rsin\,(\theta)$$

$$\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}} = \frac{1}{2\pi}e^{-\frac{R^2}{2}} = \left(\frac{1}{2\pi}\right)\left(e^{-\frac{R^2}{2}}\right)$$

$$R^2 \sim Exp\left(\frac{1}{2}\right)$$

$$\theta \sim Unif(0, 2\pi) = 2\pi Unif(0, 1)$$

$$Exp(\lambda) = \frac{-log(Unif(0,1))}{\lambda}$$

$$R \sim \sqrt{-2log\,(Unif(0,1))}$$

Applications – Gaussian random numbers

- **From Uniform to Gaussian:**
- a) Generate two random numbers u_1, u_2 ~ Unif(0,1)
- b) Use these to find radius and angle

$$R = \sqrt{-2log\,(u_1)} \qquad \theta = 2\pi u_2$$

- c) Convert from polar to Cartesian co-ordinates:

$$(Rcos\theta, Rsin\theta)$$

- This is known as **Box-Muller method**

Applications – Gaussian random numbers

- To **sample from multi-variate Gaussian N(μ, $\mathbf{\Sigma}$)**:
  - **Cholesky decomposition** of covariance matrix $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T$, where L is lower triangular matrix
  - **Sample** from $X \sim N(0, \mathbf{I})$ using Box-Muller method
  - **Change of variable**: $Y = \mu + \mathbf{L}X$

- What did we achieve:
  - mean(Y) = μ
  - cov(Y) = $\mathbf{L}$ cov(X) $\mathbf{L}^T = \mathbf{L}\,\mathbf{I}\,\mathbf{L}^T = \mathbf{\Sigma}$

Summary: Frequentist v/s Bayesian

"If I flip this coin, the probability that it will come up heads is 0.5."

• **Frequentist Interpretation**: If we flip this coin many times, it will come up heads about half the time. "*Probabilities are the expected frequencies of events over repeated trials*".

• **Bayesian Interpretation**: I believe that my next toss of this coin is equally likely to come up heads or tails. "*Probabilities quantify subjective beliefs about single events*".

References

1) S. Ross, **Introduction to Probability and Statistics for Engineers and Scientists** (Fourth Edition), Academic Press, Boston, 2009.

2) C. Bishop, **Pattern Recognition and Machine Learning**, Springer, 2006.

3) K. Murphy, **Machine Learning: A Probabilistic Perspective**, MIT Press, 2012.

4) D. Barber, **Bayesian Reasoning and Machine Learning** , Cambridge University Press, 2012.

5) L. Devroye, L. Györfi, and G. Lugosi, **A Probabilistic Theory of Pattern Recognition**, Springer-Verlag New York, 1996

6) J. Stone, **Bayes' Rule: A Tutorial Introduction to Bayesian Analysis,** Sebtel Press, 2013.