CS60010: Deep Learning

Spring 2023

Sudeshna Sarkar

Module 1 Part C

Probability and Information Theory

Sudeshna Sarkar

11 Jan 2023

Probablity

- Intuition:
  - In a process, several outcomes are possible
  - When the process is repeated a large number of times, each outcome occurs with a ***relative frequency***, or ***probability***
- Probability arises in two contexts
  - In actual repeated experiments
    - Example: You record the color of 1,000 cars driving by. 57 of them are green. You estimate the probability of a car being green as 57/1,000 = 0.057.
  - In idealized conceptions of a repeated process
    - Example: You consider the behavior of an unbiased six-sided die. The expected probability of rolling a 5 is 1/6 = 0.1667.
    - Example: You need a model for how people's heights are distributed. You choose a normal distribution to represent the expected relative probabilities.

- There are different sources of uncertainty:
  1. Inherent stochasticity in the system being modeled
     - For example, most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic
  2. Incomplete observability
     - Even deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behavior of the system
  3. Incomplete modeling
     - When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model’s predictions
     - E.g., discretization of real-numbered values, dimensionality reduction, etc.

Random variables

- A ***random variable*** X is a variable that can take on different values
  - Example: X = rolling a die
    - Possible values of X comprise the **sample space**, or **outcome space**, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
    - We denote the event of "seeing a 5" as $\{X = 5\}$ or $X = 5$
    - The probability of the event is $P(\{X = 5\})$ or $P(X = 5)$
    - Also, $P(5)$ can be used to denote the probability that X takes the value of 5
- A ***probability distribution*** is a description of how likely a random variable is to take on each of its possible states
  - A compact notation is common, where $P(X)$ is the probability distribution over the random variable X
    - Also, the notation $\mathrm{X} \sim P(X)$ can be used to denote that the random variable X has probability distribution $P(X)$
- Random variables can be discrete or continuous
  - Discrete random variables have finite number of states: e.g., the sides of a die
  - Continuous random variables have infinite number of states: e.g., the height of a person

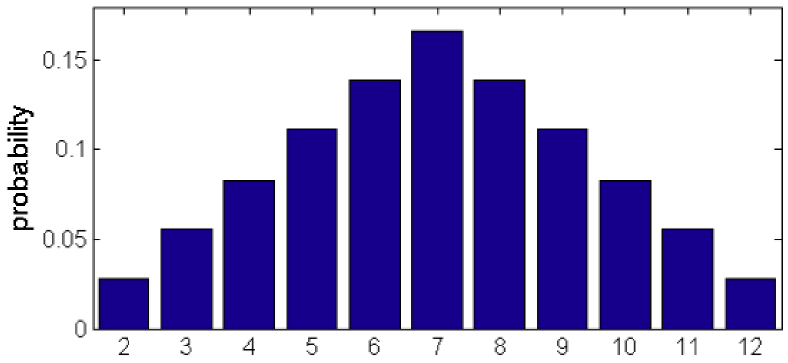
Axioms of probability

- The probability of an event $\mathcal{A}$ in the given sample space $\mathcal{S}$, denoted as $P(\mathcal{A})$, must satisfies the following properties:
  - Non-negativity
    - For any event $\mathcal{A} \in \mathcal{S}$, $P(\mathcal{A}) \geq 0$
  - All possible outcomes
    - Probability of the entire sample space is 1, $P(\mathcal{S}) = 1$
  - Additivity of disjoint events
    - For all events $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{S}$ that are mutually exclusive ($\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$), the probability that both events happen is equal to the sum of their individual probabilities, $P(\mathcal{A}_1 \cup \mathcal{A}_2) = P(\mathcal{A}_1)+P(\mathcal{A}_2)$

- The probability of a random variable $P(X)$ must obey the axioms of probability over the possible values in the sample space $\mathcal{S}$

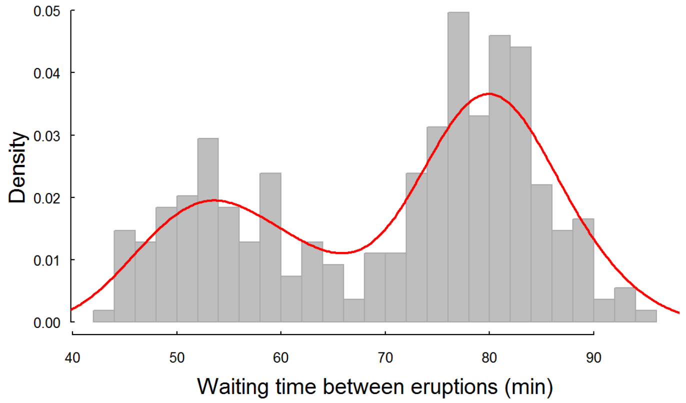
Discrete Variables

- A probability distribution over discrete variables may be described using a ***probability mass function*** (PMF)
  - E.g., sum of two dice



- A probability distribution over continuous variables may be described using a ***probability density function*** (PDF)
  - E.g., waiting time between eruptions of Old Faithful
  - A PDF gives the probability of an infinitesimal region with volume δX
  - To find the probability over an interval [*a*, *b*], we can integrate the PDF as follows:

$$P(X \in [a,b]) = \int_a^b P(X)dX$$

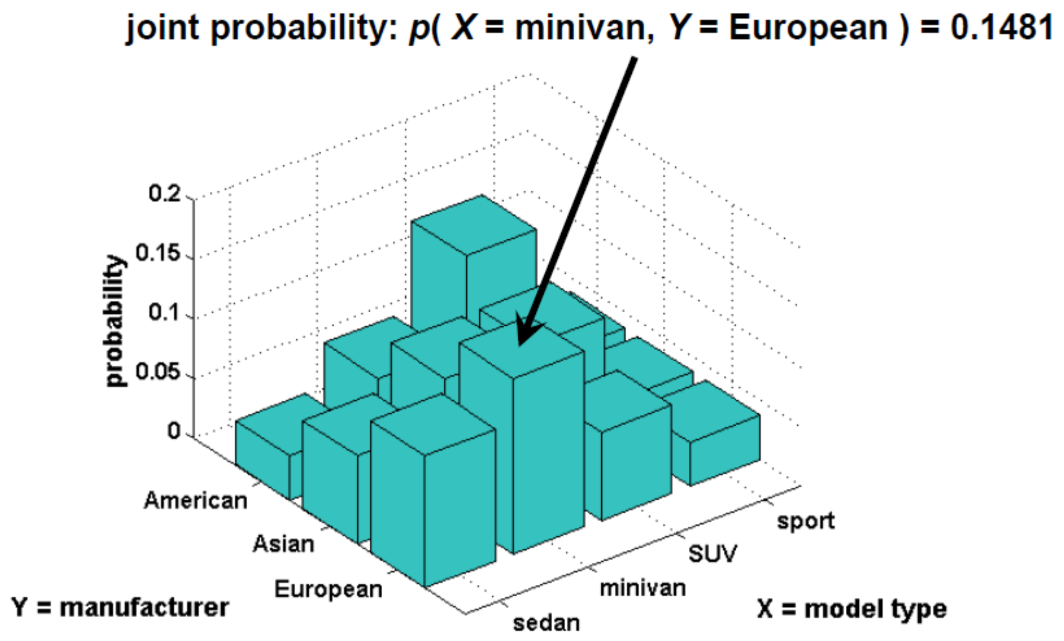


Multivariate Random Variables

- We may need to consider several random variables at a time
- Probability distributions defined over multiple random variables
  - These include joint, conditional, and marginal probability distributions
- The individual random variables can also be grouped together into a random vector, because they represent different properties of an individual statistical unit
- A ***multivariate random variable*** is a vector of multiple random variables $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$

Joint Probability Distribution

- Probability distribution that acts on many variables at the same time is known as a ***joint probability distribution***
- Given any values *x* and *y* of two random variables X and Y, what is the probability that X = *x* and Y = *y* simultaneously?
  - $P(X = x, Y = y)$ denotes the joint probability
  - We may also write $P(x, y)$ for brevity



Marginal Probability Distribution

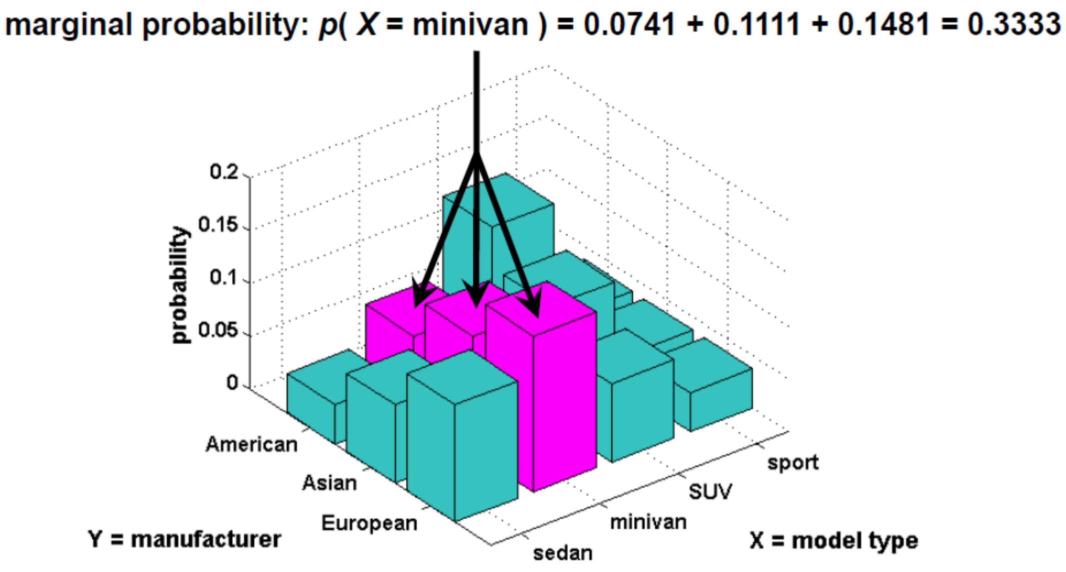
Marginal probability distribution is the probability distribution of a single variable

- It is calculated based on the joint probability distribution $P(X, Y)$ using the sum rule:

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

- For continuous random variables, the summation is replaced with integration,

$$P(X = x) = \int P(X = x, Y = y)\, dy$$

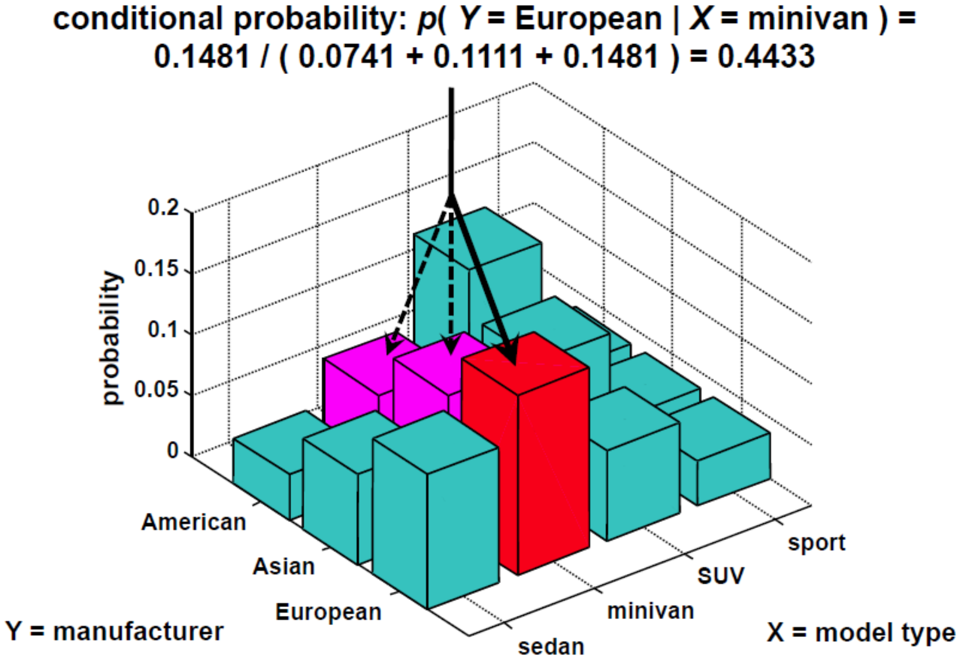


Slide credit: Jeff Howbert — Machine Learning Math Essentials

Conditional Probability Distribution

Conditional probability distribution is the probability distribution of one variable provided that another variable has taken a certain value.

$$P(X = x | Y = y) = \frac{P(X = x,\ \ Y = y)}{P(Y = y)}$$



Slide credit: Jeff Howbert — Machine Learning Math Essentials

Chain rule of probability

$$P\left(x^{(1)}, \cdots, x^{(n)}\right) = P\left(x^{(1)}\right) \prod_{i=2}^{n} P\left(x^{(i)} | x^{(1)}, \cdots, x^{(i-1)}\right)$$

Bayes' Theorem / Bayes' Rule

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- $P(X)$, the prior probability, the initial degree of belief for X
- $P(X|Y)$, the posterior probability, the degree of belief after incorporating the knowledge of Y
- $P(Y|X)$, the likelihood of Y given X
- P(Y), the evidence

Independence

- Two random variables X and Y are ***independent*** if the occurrence of Y does not reveal any information about the occurrence of X. *Denoted* $X \perp Y$
- Therefore, we can write: $P(X|\,Y) = P(X)$
  - Also note that for independent random variables:

$$\forall x \in \mathrm{X}, y \in \mathrm{Y}, \qquad p(\mathrm{X} = x, \mathrm{Y} = y) = p(\mathrm{X} = x)p(\mathrm{Y} = y)$$

$$P(X,Y) = P(X)P(Y)$$

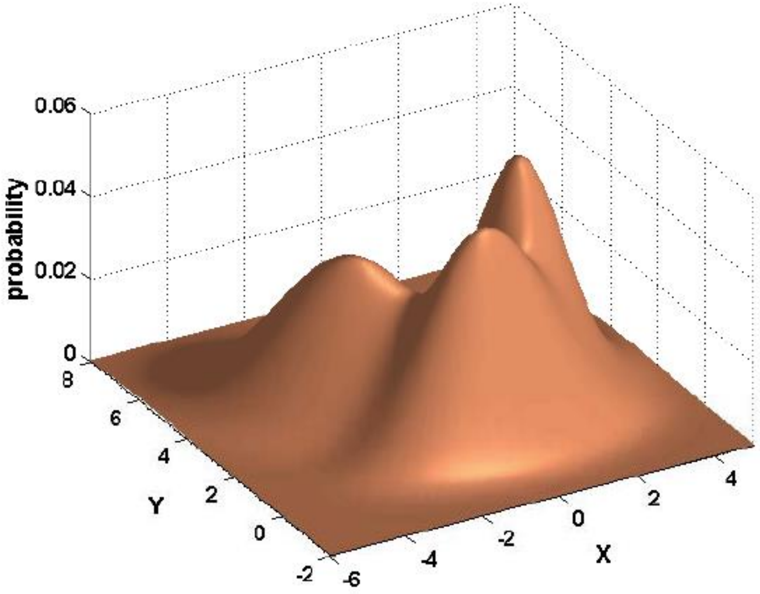
- Two random variables X and Y are ***conditionally independent*** given another random variable Z *denoted* This is denoted as $X \perp Y|Z$ if and only if

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

Slide credit: Jeff Howbert — Machine Learning Math Essentials

Continuous Multivariate Distributions

- Same concepts of joint, marginal, and conditional probabilities apply for continuous random variables
- The probability distributions use integration of continuous random variables, instead of summation of discrete random variables
  - Example: a three-component Gaussian mixture probability distribution in two dimensions



Slide credit: Jeff Howbert — Machine Learning Math Essentials

Expected Value

- The ***expected value*** or ***expectation*** of a function $f(X)$ with respect to a probability distribution $P(X)$ is the average (mean) when X is drawn from $P(X)$
- For a discrete random variable *X*, it is calculated as

$$\mathbb{E}_{X\sim P}[f(X)] = \sum_{X} P(X)f(X)$$

- For a continuous random variable *X*, it is calculated as

$$\mathbb{E}_{X\sim P}[f(X)] = \int P(X)f(X)\,dX$$

- When the identity of the distribution is clear from the context, we can write $\mathbb{E}_X[f(X)]$
- If it is clear which random variable is used, we can write just $\mathbb{E}[f(X)]$

Slide credit: Jeff Howbert — Machine Learning Math Essentials

Expectation

- Linearity of expectations:

$$\mathbb{E}_X[\alpha f(X) + \beta g(X)] = \alpha \mathbb{E}_X[f(X))] + \beta \mathbb{E}_X[g(X))]$$

Variance

- ***Variance*** gives the measure of how much the values of the function $f(X)$ deviate from the expected value as we sample values of X from $P(X)$

$$\mathrm{Var}\big(f(X)\big) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2]$$

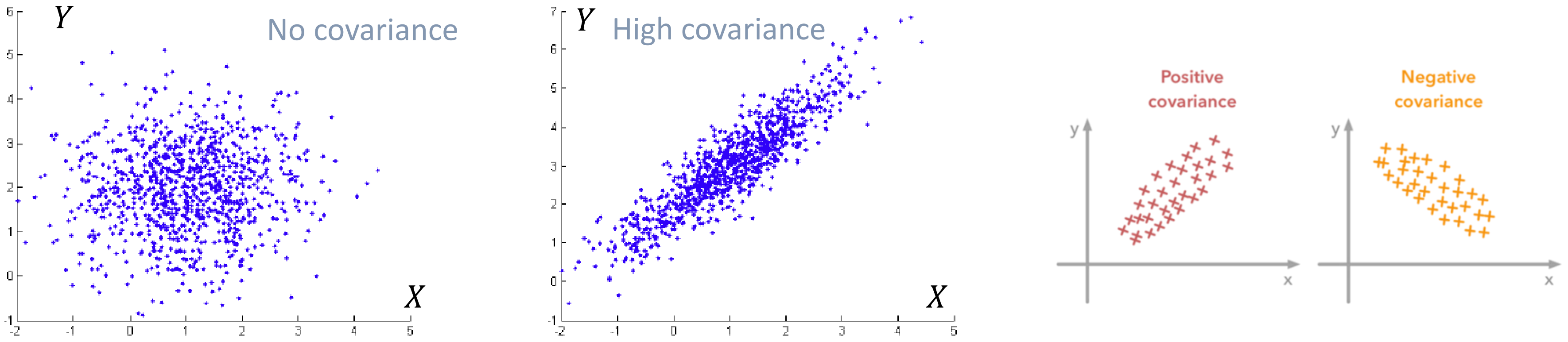
- When the variance is low, the values of $f(X)$ cluster near the expected value
- Variance is commonly denoted with σ^2
- The square root of the variance is the ***standard deviation***
  - Denoted $\sigma = \sqrt{\mathrm{Var}(X)}$

Covariance

- ***Covariance*** gives the measure of how much two random variables are linearly related to each other

$$\mathrm{Cov}\big(f(X), g(Y)\big) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])]$$

- The covariance measures the tendency for X and Y to deviate from their means in same (or opposite) directions at same time



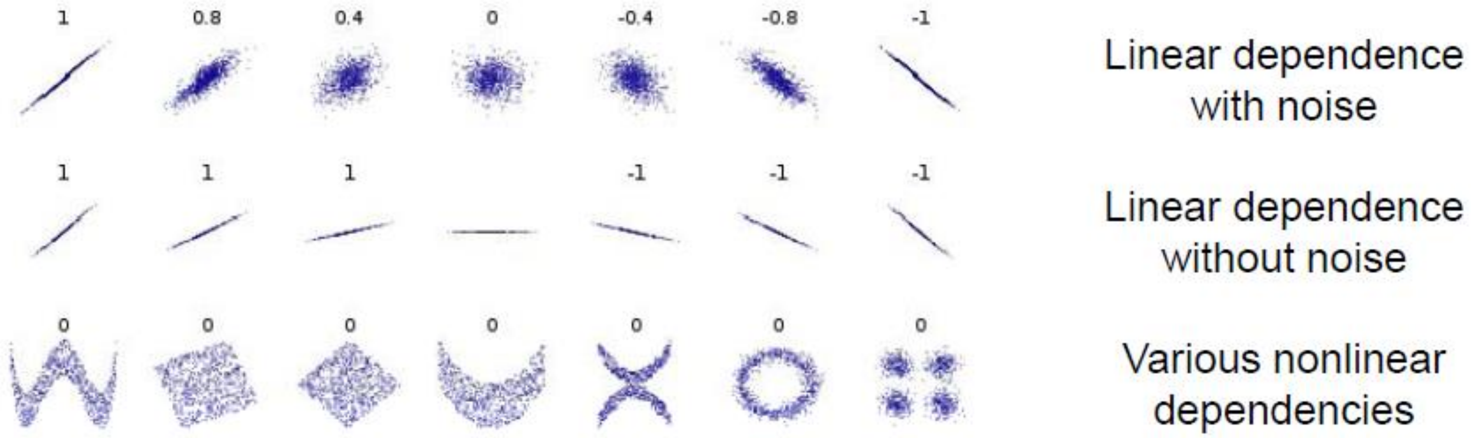
Picture from: Jeff Howbert — Machine Learning Math Essentials

Correlation

- ***Correlation coefficient*** is the covariance normalized by the standard deviations of the two variables

$$\mathrm{corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

  - It is also called Pearson's correlation coefficient and it is denoted $\rho(X, Y)$
  - The values are in the interval $[-1, 1]$
  - It only reflects linear dependence between variables, and it does not measure non-linear dependencies between the variables



Picture from: Jeff Howbert — Machine Learning Math Essentials

Covariance Matrix

- ***Covariance matrix*** of a multivariate random variable $\mathbf{X}$ with states $\mathbf{x} \in \mathbb{R}^{\boldsymbol{n}}$ is an $n \times n$ matrix, such that

$$\mathrm{Cov}(\mathbf{X})_{i,j} = \mathrm{Cov}\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

- I.e.,

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_\mathbf{1}) & \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_\mathbf{2}) & \cdots & \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_{\boldsymbol{n}}) \\ \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_\mathbf{1}) & & & \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_{\boldsymbol{n}}) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(\mathbf{x}_n, \mathbf{x}_\mathbf{1}) & \mathrm{Cov}(\mathbf{x}_n, \mathbf{x}_\mathbf{2}) & \cdots & \mathrm{Cov}(\mathbf{x}_{\boldsymbol{n}}, \mathbf{x}_{\boldsymbol{n}}) \end{bmatrix}$$

- The diagonal elements of the covariance matrix are the variances of the elements of the vector

$$\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_i) = \mathrm{Var}(\mathbf{x}_i)$$

- Also note that the covariance matrix is symmetric, since $\mathrm{Cov}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathrm{Cov}\left(\mathbf{x}_j, \mathbf{x}_i\right)$

Probability Distributions

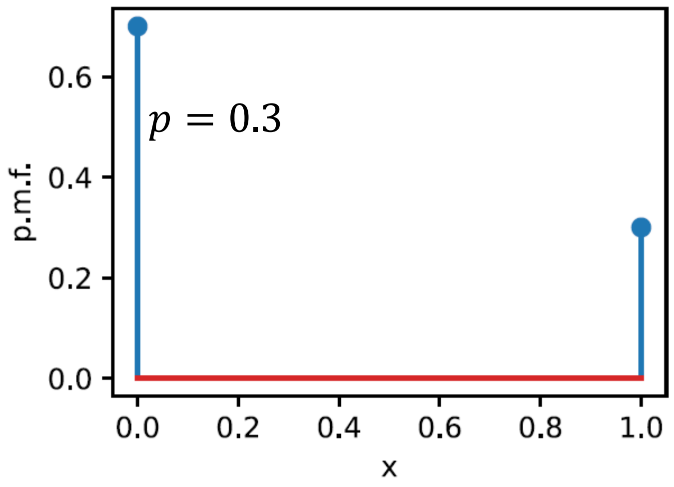
Bernoulli distribution $X \sim Bernoulli(p)$

- Binary random variable X with states $\{0, 1\}$
- The random variable can encodes a coin flip which comes up 1 with probability ϕ and 0 with probability $1 - \phi$

$$P(X = x) = \phi^x (1 - \phi)^{1-x}$$

$$\mathbb{E}_X[\mathrm{X}] = \phi$$

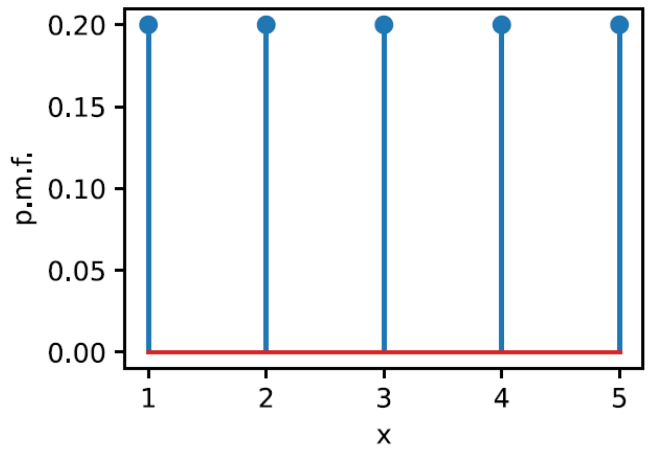
$$Var_X(X) = \phi(1 - \phi)$$



Picture from: http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html

Probability Distributions

- ***Uniform distribution*** $X \sim U(n)$
  - The probability of each value $i \in \{1,2,\dots,n\}$ is $p_i = \frac{1}{n}$
  - Notation:
  - Figure: $n = 5,\ p = 0.2$

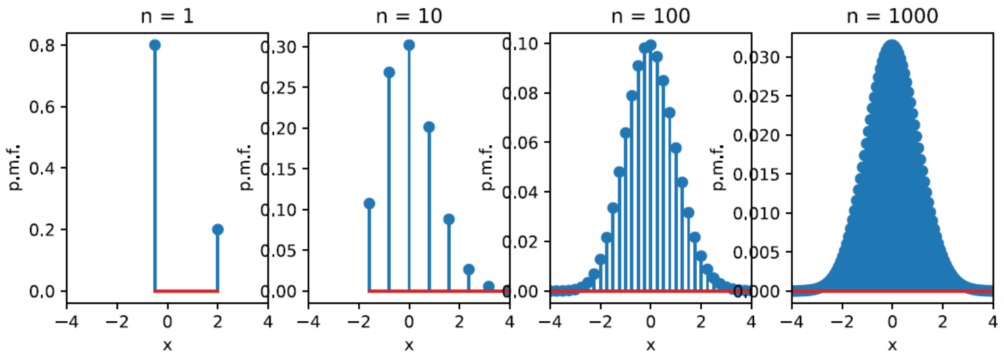


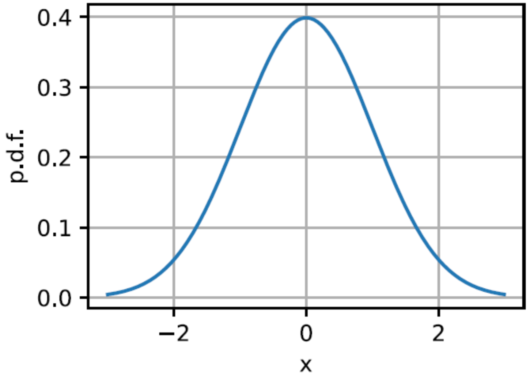
Picture from: http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html

Gaussian Distribution

- Referred to as normal distribution or informally bell-shaped distribution
- Defined with the mean μ and variance σ^2
- Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$
- For a random variable X with n independent measurements, the density is

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





Picture from: http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html

Gaussian Distribution

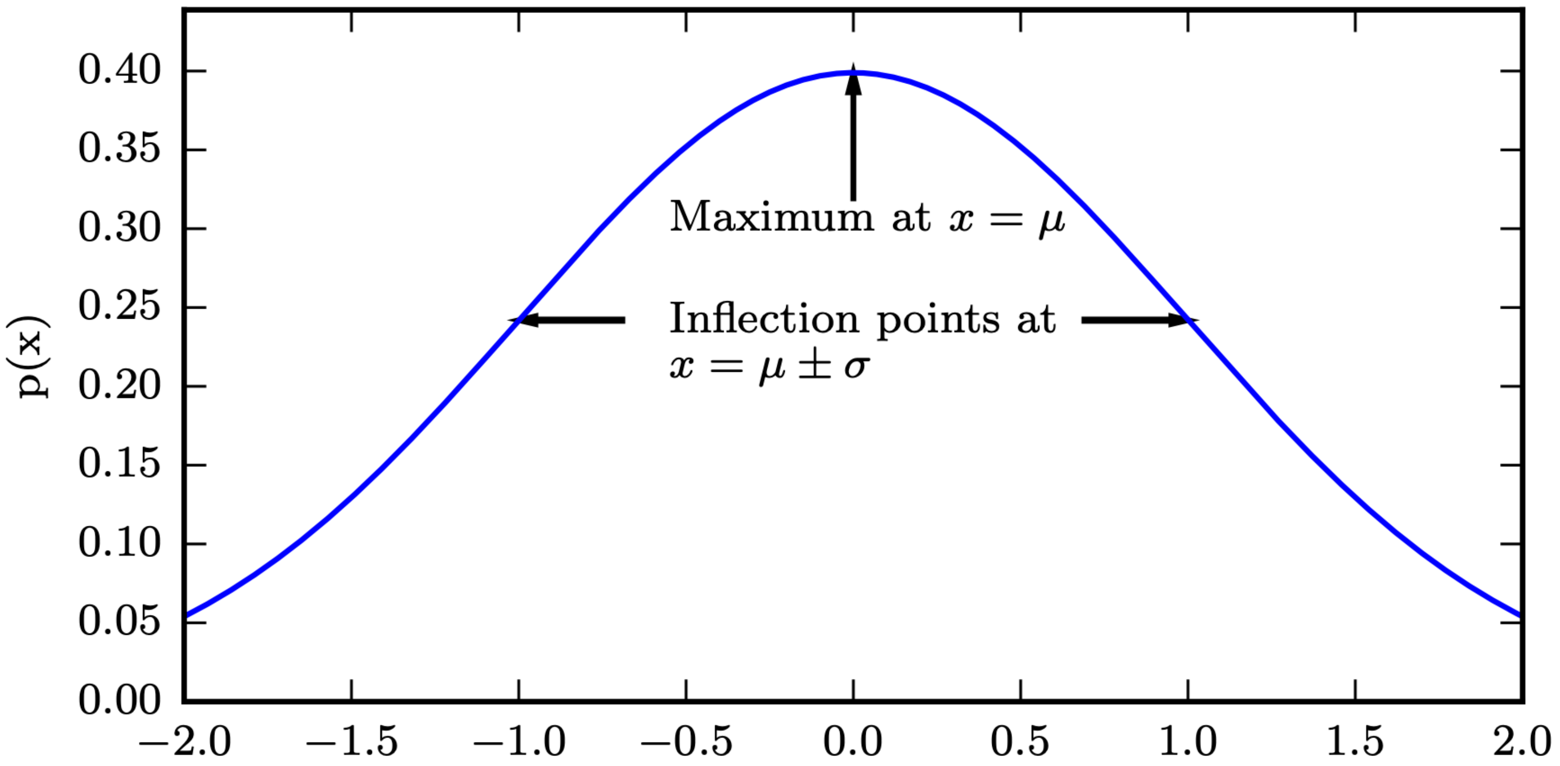
Parametrized by variance:

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{3.21}$$

Parametrized by precision:

$$\mathcal{N}(x;\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{1}{2}\beta(x-\mu)^2\right). \tag{3.22}$$

Gaussian Distribution



Multivariate Gaussian

Parametrized by covariance matrix:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

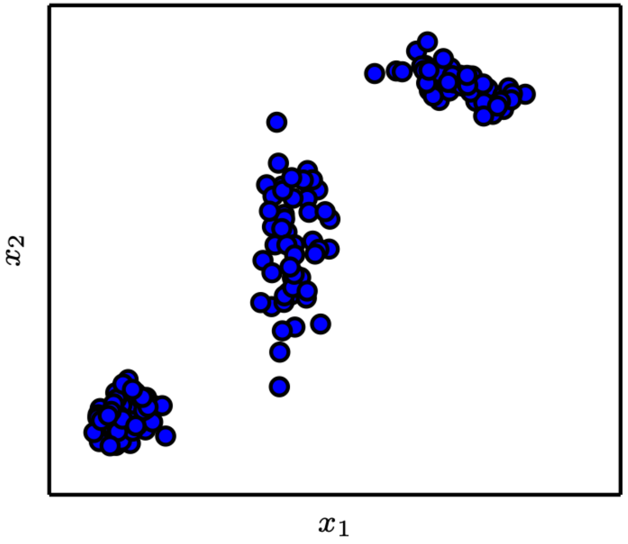
Parametrized by precision matrix

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\beta} (\boldsymbol{x}-\boldsymbol{\mu})\right). \quad ($$

Mixture Distributions

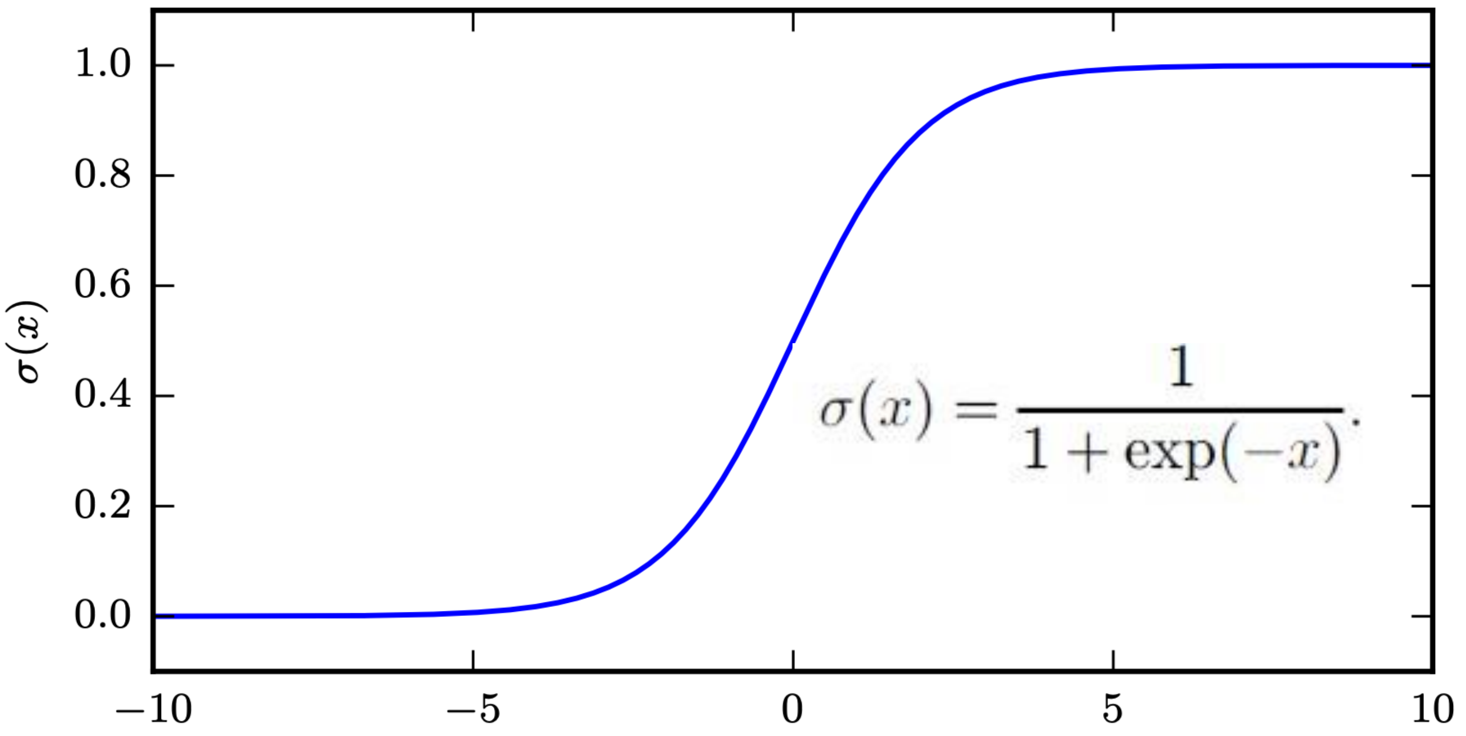
$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i) \tag{3.29}$$

Gaussian mixture with three components



(Goodfellow 2016)

Logistic Sigmoid



Commonly used to parametrize Bernoulli distributions

Softplus Function

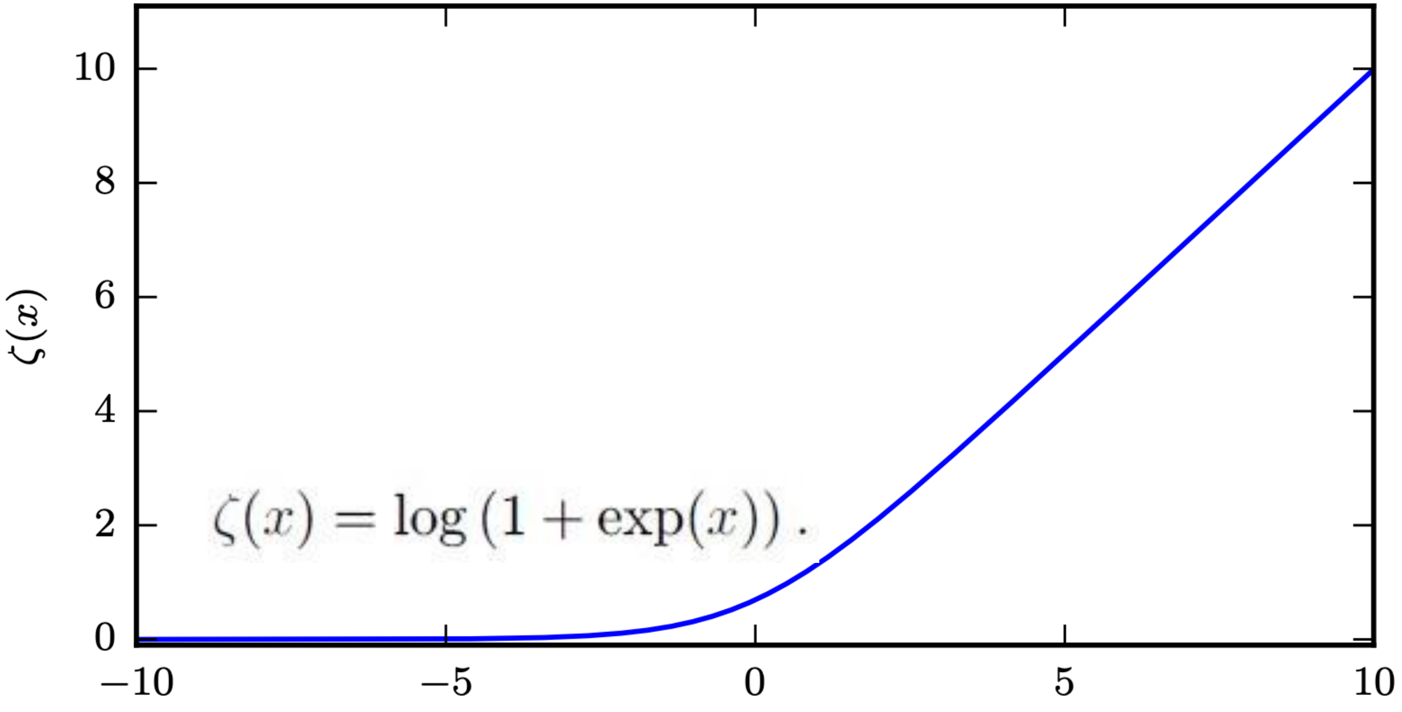


Figure 3.4: The softplus function.

A smoothed version of $x^+ = \max(0, x)$.

Useful properties

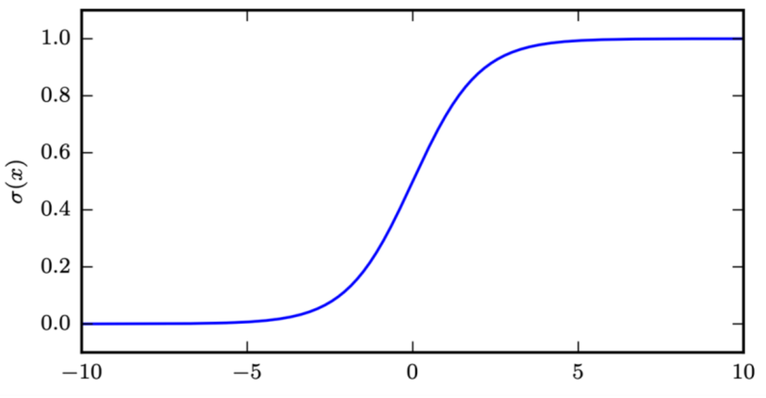


Figure 3.3: The logistic sigmoid function.

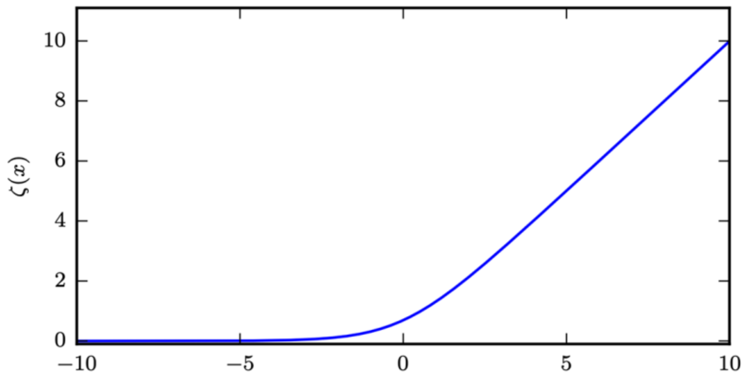


Figure 3.4: The softplus function.

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

$$1 - \sigma(x) = \sigma(-x)$$

$$\log \sigma(x) = -\varsigma(-x)$$

$$\frac{d}{dx}\varsigma(x) = \sigma(x)$$

$$\forall x \in (0,1), \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$

$$\forall x > 0, \varsigma^{-1}(x) = \log(\exp(x) - 1)$$

$$\varsigma(x) = \int_{-\propto}^{x} \sigma(y)dy$$

$$\varsigma(x) - \varsigma(-x) = x$$

Information Theory

Information Theory

- Information theory studies encoding, decoding, transmitting, and manipulating information
  - provides fundamental language for discussing the information processing in computer systems

Information Theory

- Learning that an unlikely event has occurred is more informative that learning that a likely event has occurred!
- Which statement has more information?
    - “The sun rose this morning”
    - “There was a solar eclipse this morning”

- Independent events should have additive information:
    - Finding out that a tossed coin has come up heads twice has two time more information that finding out that a tossed coin has come up heads one time!

Self-Information

Self-information of an event x

$$I(x) = -\log P(x)$$

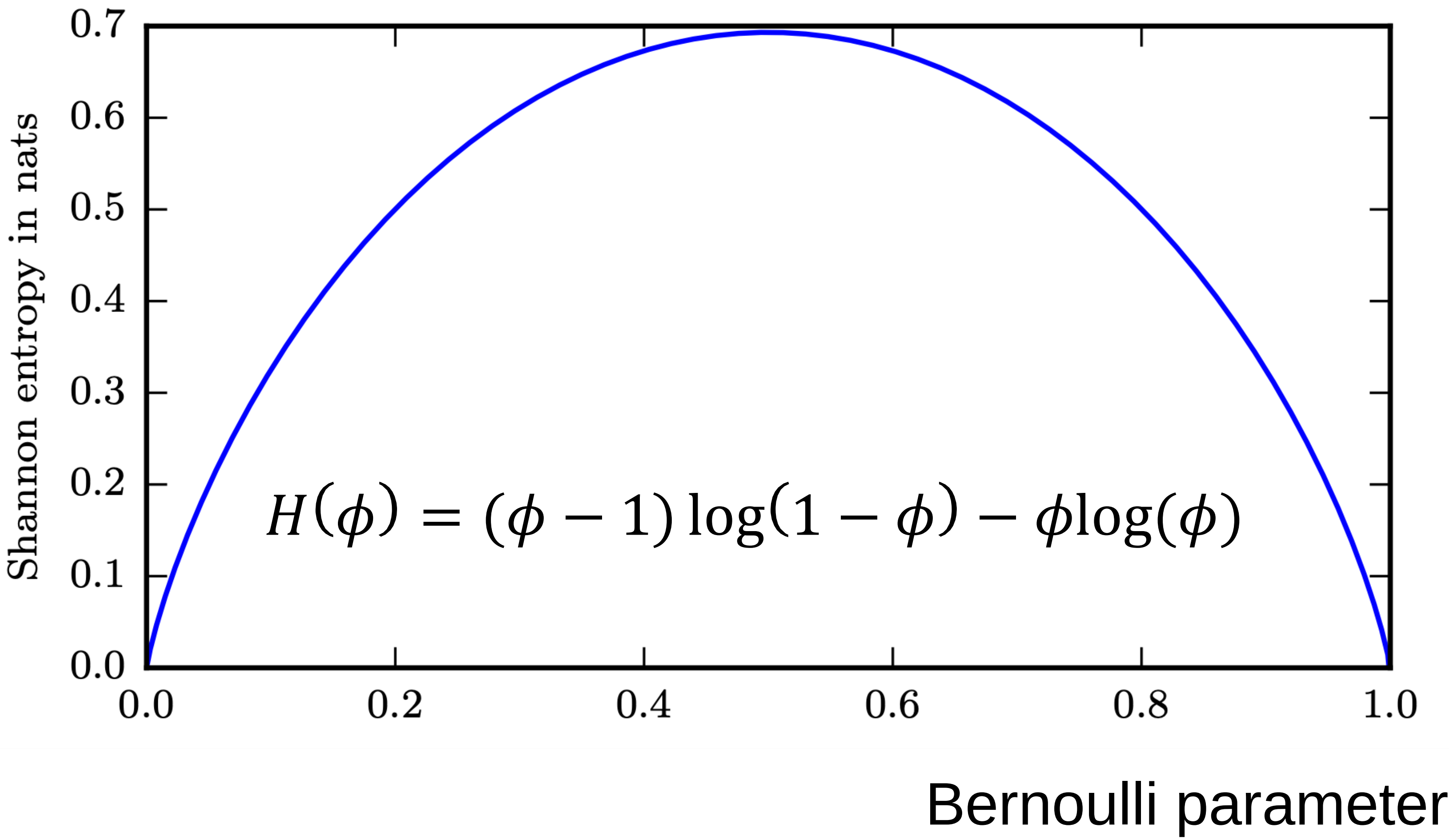
We can quantify the amount of uncertainty in an entire probability distribution using the Shannon entropy.

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)].$$

Entropy is a lower bound on the number of bits needed on average to encode symbols drawn from a distribution P.

- Distributions that are nearly deterministic have low entropy
- Distributions that are nearly uniform have high entropy

Entropy of a Bernoulli Variable



Entropy

$$H(X) = \mathbb{E}_{X\sim P}[I(X)] = -\mathbb{E}_{X\sim P}\,[\log P(X)]$$

- Based on the expectation definition $\mathbb{E}_{X\sim P}[f(X)] = \sum_X P(X) f(X)$, we can rewrite the entropy as

$$H(X) = -\sum_X P(X) \log P(X)$$

- If X is a continuous random variable that follows a probability distribution P with a probability density function $P(X)$, the entropy is

$$H(X) = -\int_X P(X) \log P(X)\, dX$$

  - For continuous random variables, the entropy is also called differential entropy

Kullback-Leibler Divergence

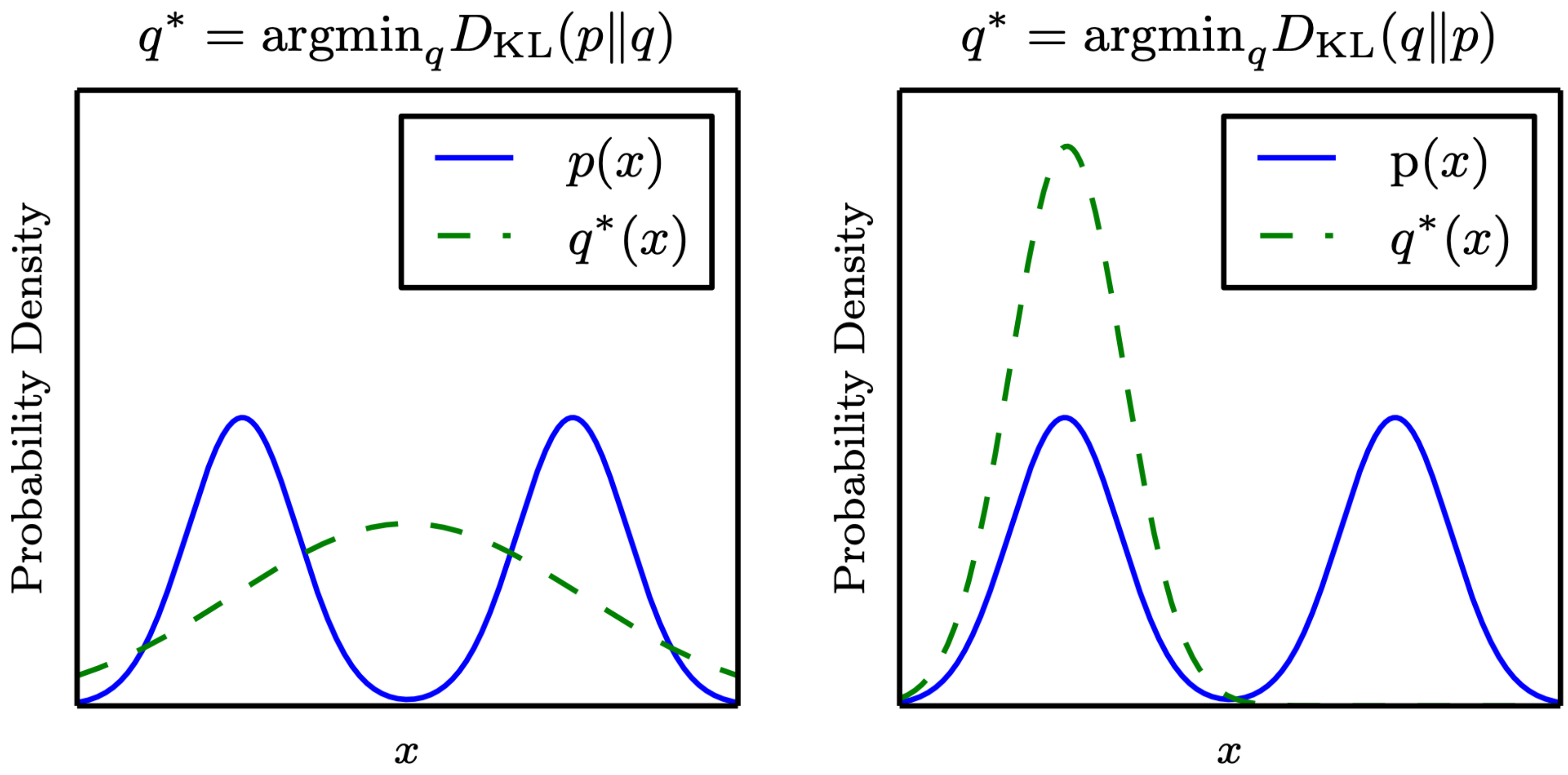
KL divergence:

$$D_{\mathrm{KL}}(P \| Q) = \mathbb{E}_{\mathrm{x} \sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x} \sim P}\left[\log P(x) - \log Q(x)\right]. \tag{3.50}$$

- KL-divergence is the extra amount of information needed to send a message containing symbols drawn from P, when we use a code designed to minimize the length of messages containing symbols drawn from Q
  - KL-divergence is non-negative
  - KL-divergence = 0 if P and Q are the same distribution
- It can be used as a distance measure between distributions
- But it is not a true distance measure since it is not symmetric:

The KL Divergence is Asymmetric

Mixture of two Gaussians for P, One Gaussian for Q



(Goodfellow 2016)

Cross-entropy

$$H(P,Q) = H(P) + D_{KL}(P||Q)$$

$$H(P,Q) = -\mathbb{E}_{x\sim P} \log P(x) + \mathbb{E}_{x\sim P} \log P(x) - \mathbb{E}_{x\sim P} \log Q(x)$$

$$H(P,Q) = -\mathbb{E}_{x\sim P} \log Q(x)$$

Minimzing the cross entropy with respect to Q is equivalent to minimize the KL divergence!

Remark: usually we consider 0 log 0 = 0

Maximum Likelihood

- Cross-entropy is closely related to the ***maximum likelihood*** estimation
- In ML, we want to find a model with parameters θ that maximize the probability that the data is assigned the correct class, i.e., $\mathrm{argmax}_\theta\, P(\mathrm{model} \mid \mathrm{data})$
  - For the classification problem from previous page, we want to find parameters θ so that for the data examples $\{x_1, x_2, \dots, x_n\}$ the probability of outputting class labels $\{y_1, y_2, \dots, y_n\}$ is maximized
  - From Bayes' theorem, $\mathrm{argmax}\, P(\mathrm{model} \mid \mathrm{data})$ is proportional to $\mathrm{argmax}\, P(\mathrm{data} \mid \mathrm{model})$

$$P(\theta | x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n | \theta)\, P(\theta)}{P(x_1, x_2, \dots, x_n)}$$

  - This is true since $P(x_1, x_2, \dots, x_n)$ does not depend on the parameters θ
  - Also, we can assume that we have no prior assumption on which set of parameters θ are better than any others
- Recall that $P(\mathrm{data}|\mathrm{model})$ is the likelihood, therefore, the maximum likelihood estimate of θ is based on solving

$$\arg\max_\theta P(x_1, x_2, \dots, x_n | \theta)$$

Maximum Likelihood

- For a total number of *n* observed data examples $\{x_1, x_2, \dots, x_n\}$, the predicted class labels for the data example x_i is $\hat{\mathbf{y}}_i$
  - Using the multinoulli distribution, the probability of predicting the true class label $\mathbf{y}_i = \{y_{i1}, y_{i2}, \dots, y_{ik}\}$ is $\mathcal{P}(x_i \mid \theta) = \prod_j \hat{y}_{ij}{}^{y_{ij}}$
- Assuming that the data examples are independent, the likelihood of the data given the model parameters θ can be written as

$$\mathcal{P}(x_1, x_2, \dots, x_n|\theta) = \mathcal{P}(x_1|\theta) \cdots \mathcal{P}(x_n \mid \theta) = \prod_j \hat{y}_{1j}{}^{y_{1j}} \cdot \prod_j \hat{y}_{2j}{}^{y_{2j}} \cdots \prod_j \hat{y}_{nj}{}^{y_{nj}} = \prod_i \prod_j \hat{y}_{ij}{}^{y_{ij}}$$

$$\log \mathcal{P}(x_1, x_2, \dots, x_n|\theta) = \log\left(\prod_i \prod_j \hat{y}_{ij}{}^{y_{ij}}\right) = \sum_i \sum_j y_{ij} \log \hat{y}_{ij}$$

- Thus, maximizing the likelihood is the same as minimizing the cross-entropy