Probability and Information Theory

Lecture slides for Chapter 3 of *Deep Learning*
www.deeplearningbook.org
Ian Goodfellow
2016-09-26

adapted by m.n. for CMPS 392

Probability

- Sample space Ω: set of all outcomes of a random experiment
- Set of events $\mathcal{F}$: collection of possible outcomes of an experiment.
- Probability measure: $P: \mathcal{F} \rightarrow \mathbb{R}$
  - Axioms of probability
    - $P(A \geq 0)$ for all $A \in \mathcal{F}$
    - $P(\Omega) = 1$
    - If $A_1, A_2, \ldots$ are disjoint events then

$$P\left(\bigcup_i A_i\right) = \sum_i (A_i)$$

(Goodfellow 2016)

Random variable

- Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads.
- Here, the elements of the sample space Ω are 10-length sequences of heads and tails.
- For example, we might have

$$w_0 = < H, H, T, H, T, H, H, T, T, T >$$

- However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails.
- Instead we usually care about real-valued functions of outcomes, such as
  - ❑ the number of heads that appear among our 10 tosses,
  - ❑ or the length of the longest run of tails.
- These functions, under some technical conditions, are known as random variables: $X: \Omega \rightarrow \mathbb{R}$

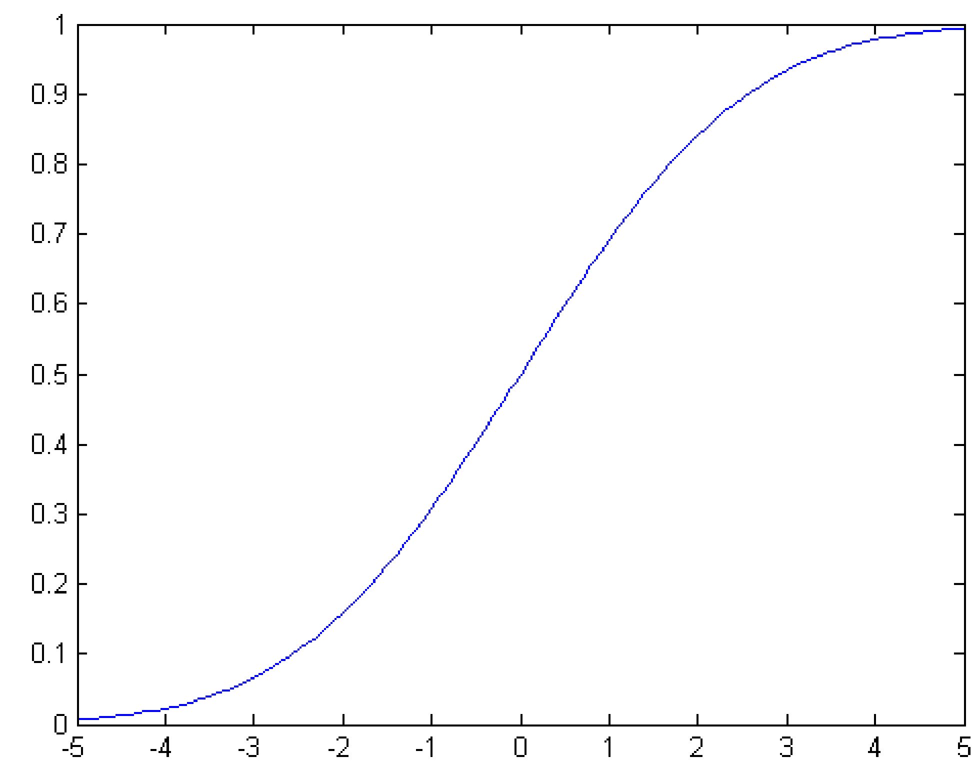
(Goodfellow 2016)

Discrete vs. continuous

- Discrete random variable:
  - ❑ $P(X = k) = P(\{w: X(w) = k\})$
- Continuous random variable:
  - ❑ $P(a \leq \ X \leq b) = P(\{w: a \leq X(w) \leq b\})$

A cumulative distribution function (CDF):

$$P(X \leq x)$$



Probability Mass Function (discrete variable)

- The domain of P must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
- $\sum_{x \in \mathrm{x}} P(x) = 1$. We refer to this property as being **normalized**. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring.

Example: uniform distribution: $$P(\mathrm{x} = x_i) = \frac{1}{k}$$

(Goodfellow 2016)

Probability Density Function (continuous variable)

- The domain of p must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$.

Example: uniform distribution: $u(x; a, b) = \frac{1}{b-a}$.

The pdf at some point x is not the probability of x: $p(x) \neq P(\mathrm{x} = x)$

(Goodfellow 2016)

Computing Marginal Probability with the Sum Rule

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y). \tag{3.3}$$

$$p(x) = \int p(x, y) dy. \tag{3.4}$$

(Goodfellow 2016)

Conditional Probability

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}. \tag{3.5}$$

(Goodfellow 2016)

Chain Rule of Probability

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)})\Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}). \tag{3.6}$$

(Goodfellow 2016)

Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y). \tag{3.7}$$

(Goodfellow 2016)

Conditional Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x}=x, \mathrm{y}=y \mid \mathrm{z}=z) = p(\mathrm{x}=x \mid \mathrm{z}=z)p(\mathrm{y}=y \mid \mathrm{z}=z). \tag{3.8}$$

(Goodfellow 2016)

Expectation

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x), \tag{3.9}$$

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx. \tag{3.10}$$

linearity of expectations:

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}_{\mathrm{x}}[f(x)] + \beta\mathbb{E}_{\mathrm{x}}[g(x)], \tag{3.11}$$

(Goodfellow 2016)

Variance and Covariance

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]. \tag{3.12}$$

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right]. \tag{3.13}$$

Covariance matrix:

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j). \tag{3.14}$$

(Goodfellow 2016)

Bernoulli Distribution

$$P(\mathrm{x} = 1) = \phi \tag{3.16}$$

$$P(\mathrm{x} = 0) = 1 - \phi \tag{3.17}$$

$$P(\mathrm{x} = x) = \phi^x (1 - \phi)^{1-x} \tag{3.18}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi \tag{3.19}$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi) \tag{3.20}$$

(Goodfellow 2016)

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:

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

(Goodfellow 2016)

Gaussian Distribution

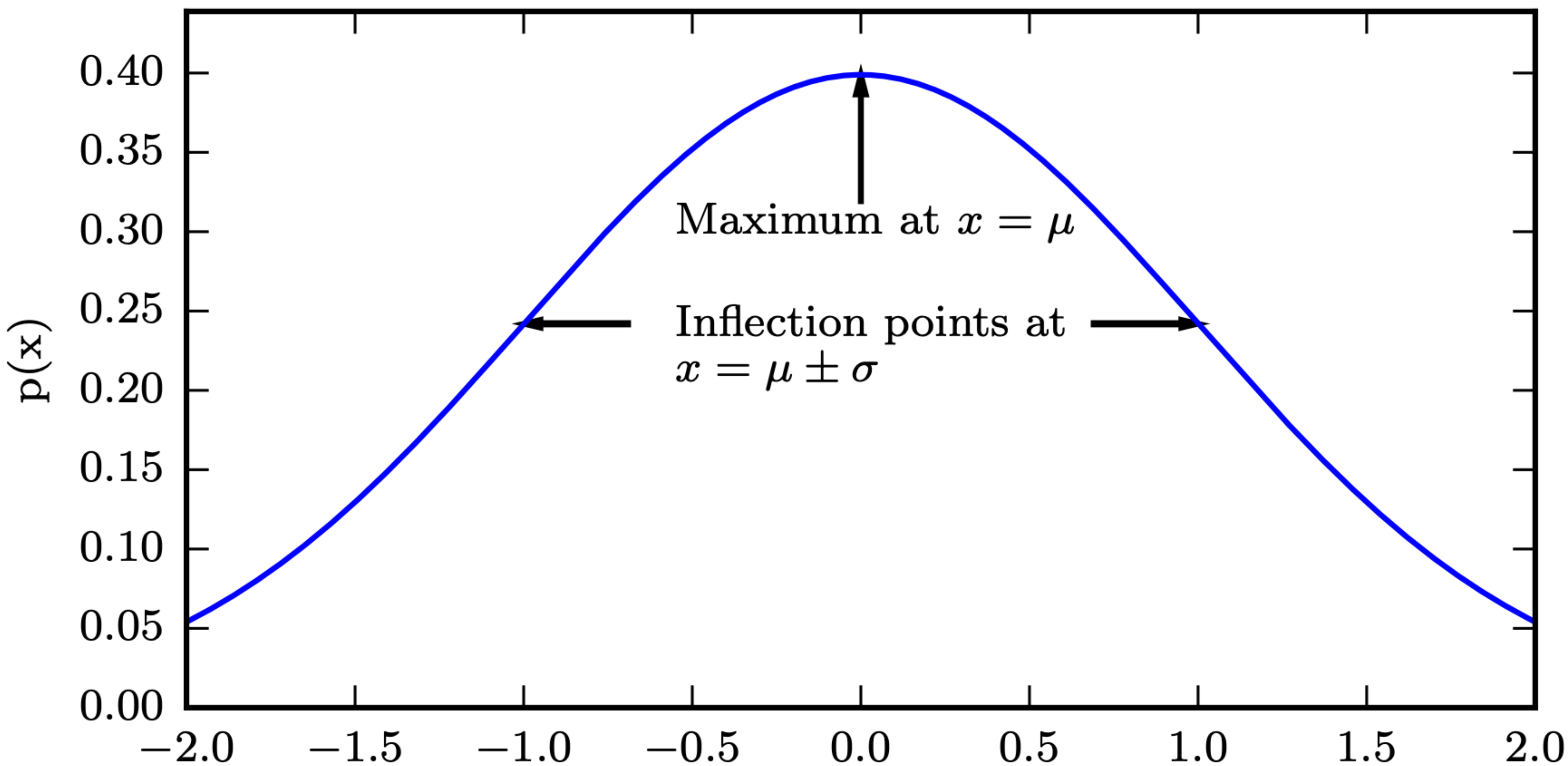


Figure 3.1

(Goodfellow 2016)

Multivariate Gaussian

Parametrized by covariance matrix:

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{3.23}$$

Parametrized by precision matrix:

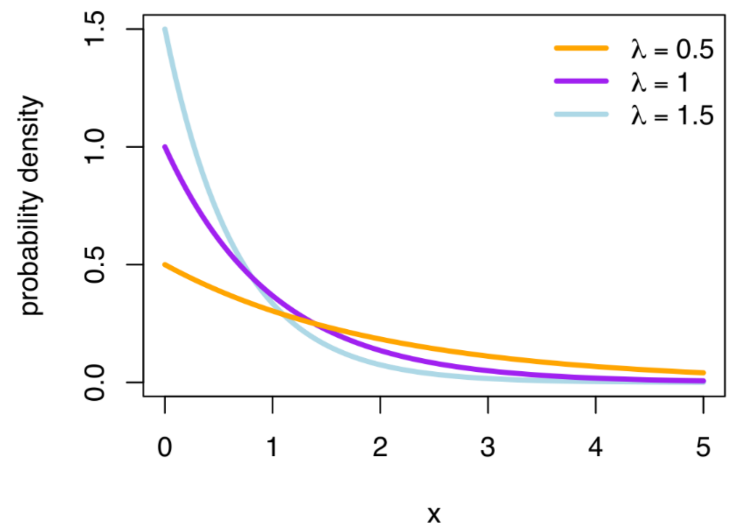
$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{3.24}$$

(Goodfellow 2016)

More Distributions

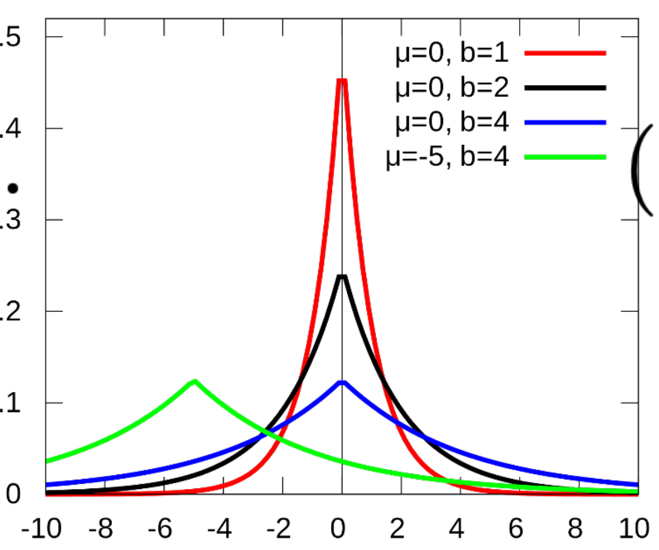
Exponential:

$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp\left(-\lambda x\right) \quad (3.25)$$



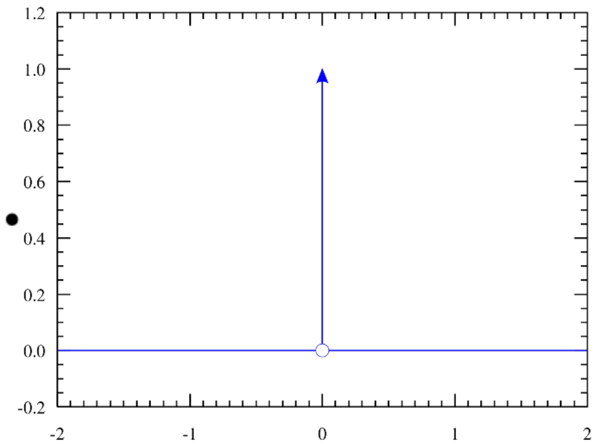
Laplace:

$$\mathrm{Laplace}(x;\mu,\gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right). \quad (3.26)$$



Dirac:

$$p(x) = \delta(x-\mu). \quad (3.27)$$



(Goodfellow 2016)

Empirical Distribution

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}) \tag{3.28}$$

(Goodfellow 2016)

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

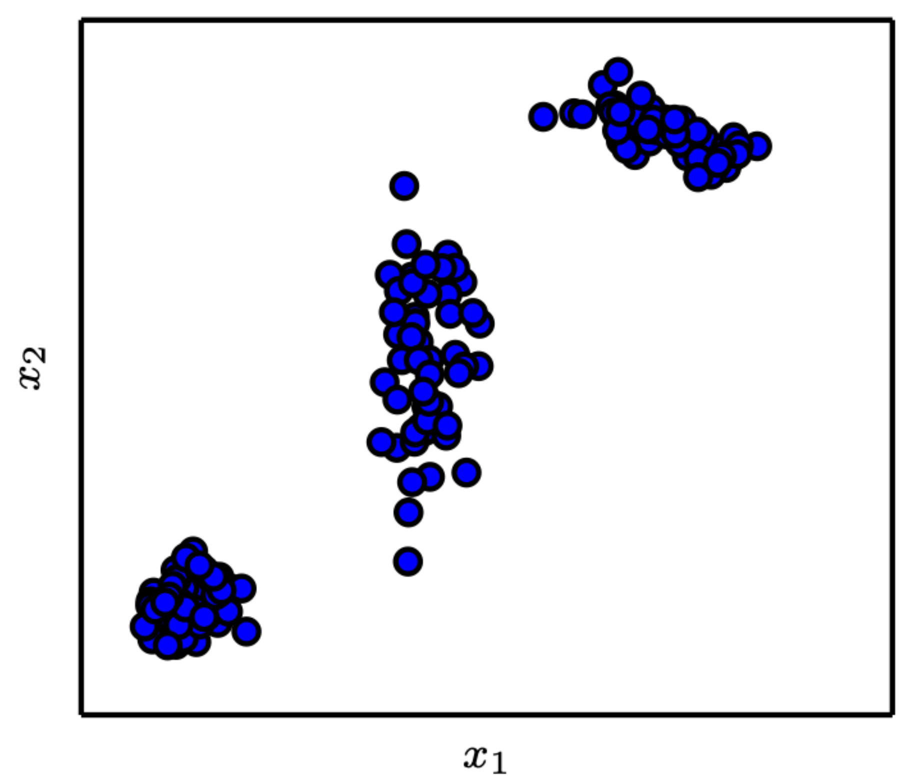


Figure 3.2

(Goodfellow 2016)

Logistic Sigmoid

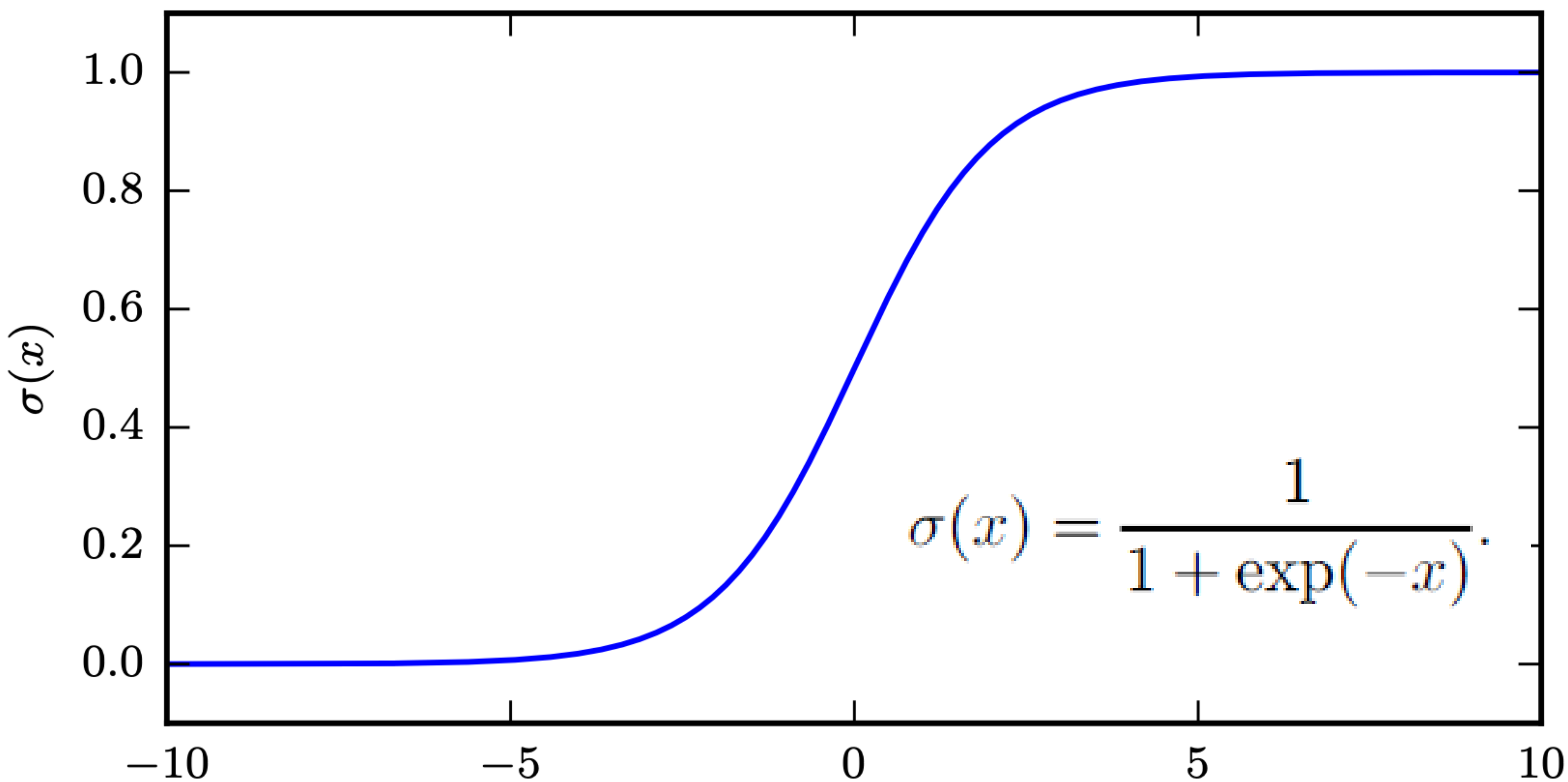


Figure 3.3: The logistic sigmoid function.

Commonly used to parametrize Bernoulli distributions

Softplus Function

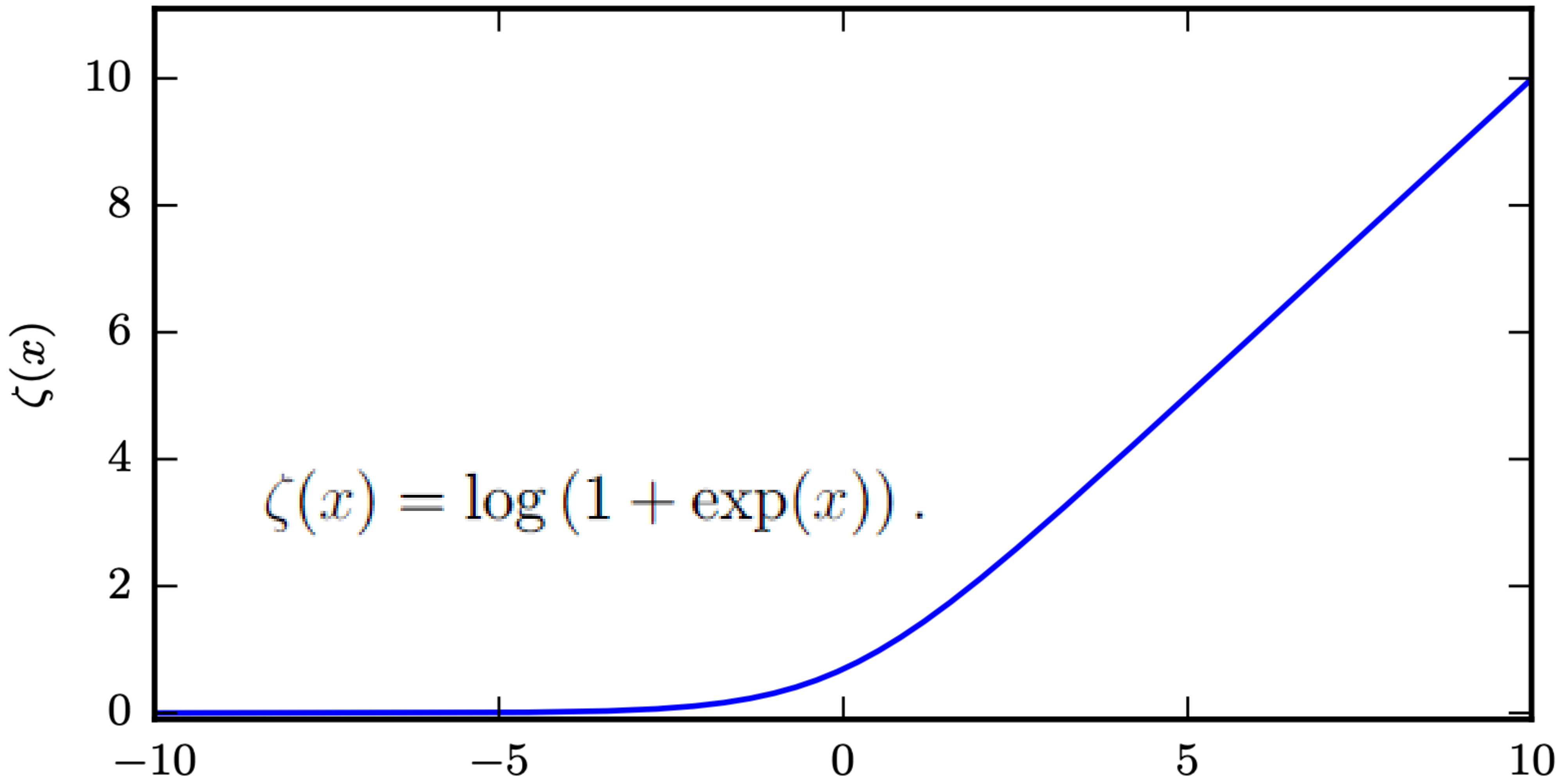


Figure 3.4: The softplus function.

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

(Goodfellow 2016)

Useful Properties

- $\sigma(x) = \frac{\exp(x)}{\exp(x)+\exp(0)}$
- $\frac{d}{dx}\sigma(x) = \sigma(x)\big(1-\sigma(x)\big)$
- $1-\sigma(x) = \sigma(-x)$
- $\log\big(\sigma(x)\big) = -\zeta(-x)$
- $\frac{d}{dx}\zeta(x) = \sigma(x)$
- $\forall x \in (0,1), \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$
- $\forall x > 0, \zeta^{-1}(x) = \log(\exp(x)-1)$
- $\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$
- $\zeta(x) - \zeta(-x) = x$

(Goodfellow 2016)

Bayes' Rule

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}. \qquad (3.42)$$

(Goodfellow 2016)

Bayes Rule

Posterior

Likelihood

Prior

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Marginal likelihood

computed by the total probability rule:

$$P(B) = \sum_{a} P(B|A=a)\,P(A=a)$$

Bag-of-words Naïve Bayes:

- Features: W_i is the word at position i
- Called "bag-of-words" because model is insensitive to word order or reordering:
  - In a bag-of-words model, each position is identically distributed

$$P(Spam|W_1, W_2, \ldots, W_N) \propto P(W_1|Spam)P(W_2|Spam) \ldots P(W_N|Spam)$$

- Start with a bunch of probabilities:
  - **Prior** distribution $P(Spam), P(Ham)$
  - and the **likelihood** probabilities (The CPT tables) $P(W_i|Y)$
- Use standard inference to compute the **posterior** probabilities $P(Y|W_1 \ldots W_n)$
- We can use the normalization trick: $P(Ham|W_1 \ldots W_n)+P(Spam|W_1 \ldots W_n) = 1$
- Computing the **log posterior** (instead of the posterior) prevents numerical errors

Example

$P(Y)$

ham	0.66
spam	0.33

$P(W|\text{spam})$

the	0.0156
to	0.0153
and	0.0115
of	0.0095
you	0.0093
a	0.0086
with	0.0080
from	0.0075
...	

$P(W|\text{ham})$

the	0.0210
to	0.0133
of	0.0119
2002	0.0110
with	0.0108
from	0.0107
and	0.0105
a	0.0100
...	

Word	P(w | spam)	P(w | ham)	Total spam (log)	Total ham (log)
(prior)	0.333	0.666	-1.1	-0.4
The	0.005	0.013	-5.27	-4.27
Year	…	…	…	…
2002				
…				

Bag of words exercise

- Spam messages:
  - Offer is secret
  - Click secret link
  - Secret sports link
- Ham messages:
  - Play sports today
  - Went play sports
  - Secret sports event
  - Sports is today
  - Sports costs money

- Size of vocabulary= ?
- P(SPAM) = ?
- P_{ML}("Secret"|SPAM) = ?
- P_{ML}("Secret"|HAM)=?
- how many parameters to represent the Naïve Bayes Network?
- P(SPAM|"Sports")=?
- P(SPAM| "Secret is secret") = ?
- P(SPAM|"Today is secret")=?

Change of Variables

Example of common mistake:

$y = \frac{x}{2}$ and $x \sim U(0,1)$

$$p_y(y) = p_x\left(\frac{x}{2}\right) \Rightarrow \begin{cases} y = 1 \text{ if } x \in [0, ½] \\ y = 0 \; elsewhere \end{cases}$$

$$\int p(y)dy = \frac{1}{2}\,!!$$

The right thing is: $\left|p_y(g(x))dy\right| = |p_x(x)dx|$

$$p_x(x) = p_y\big(g(x)\big)\left|\frac{\partial y}{\partial x}\right|$$

In higher dimensions:

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x}))\left|\det\left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|.$$

(Goodfellow 2016)

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!

(Goodfellow 2016)

Self-Information

$$I(x) = -\log P(x). \tag{3.48}$$

Log base e => unit is **nats**
Log base 2 => unit is **bits** or **shannons**

(Goodfellow 2016)

Entropy

Entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x} \sim P}[I(x)] = -\mathbb{E}_{\mathrm{x} \sim P}[\log P(x)]. \tag{3.49}$$

- Entropy is a lower bound on the number of bits needded 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

(Goodfellow 2016)

Entropy of a Bernoulli Variable

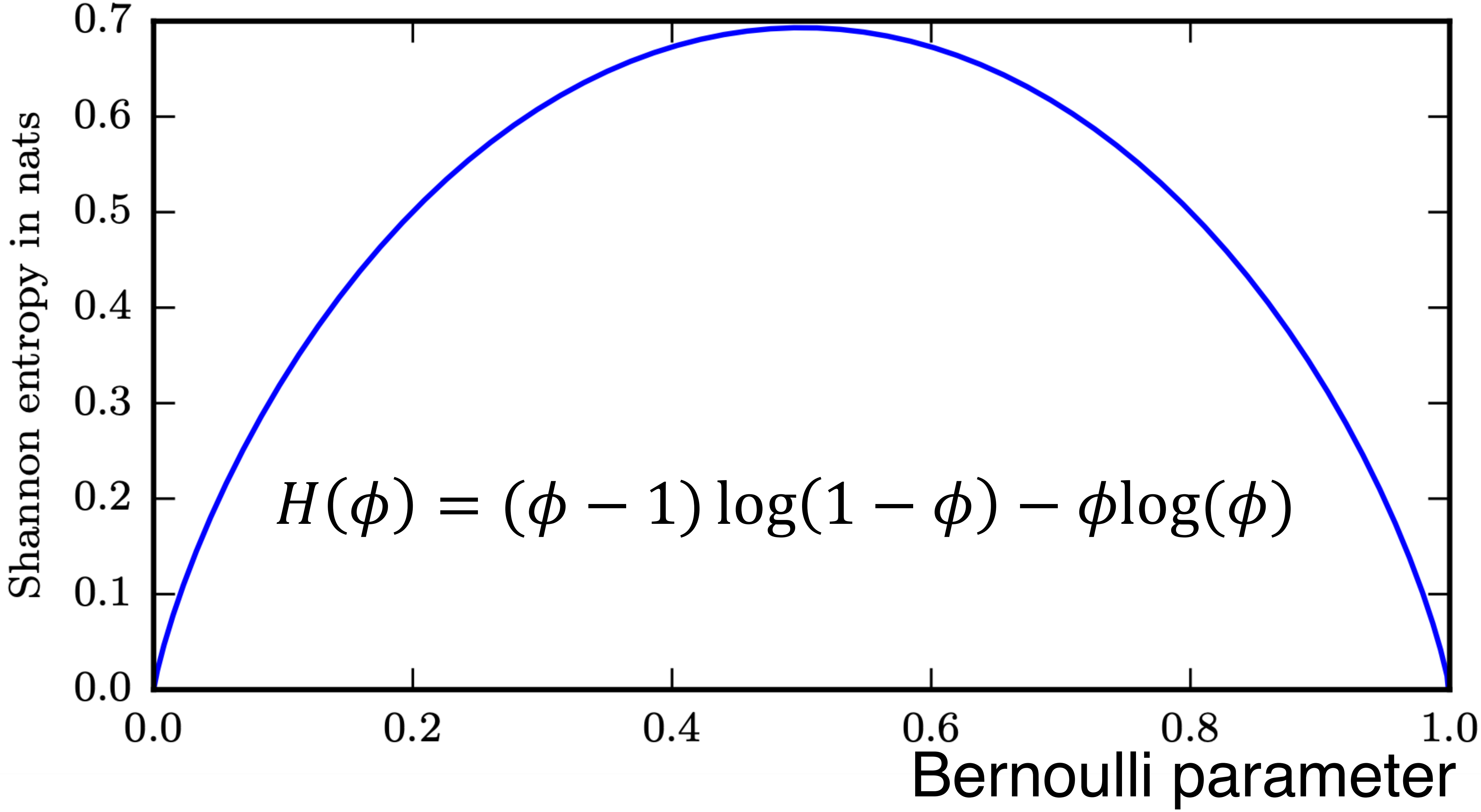


Figure 3.5

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

(Goodfellow 2016)

KL-divergence

- It can be used as a distance measure between distributions
- But it is not a true distance measure since it is not symmetric:
  - $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

(Goodfellow 2016)

The KL Divergence is Asymmetric

Mixture of two Gaussians for P, One Gaussian for Q

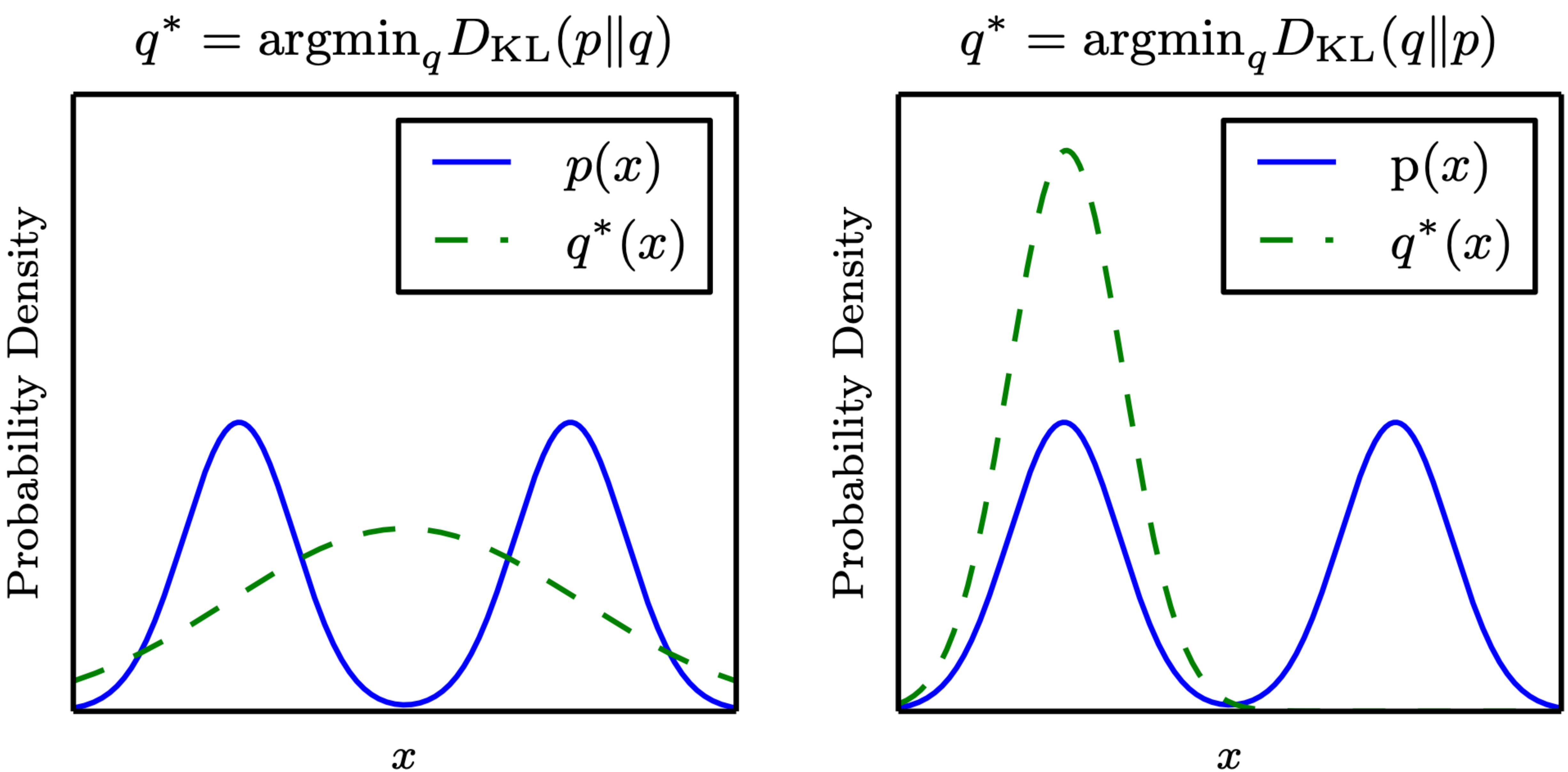


Figure 3.6

(Goodfellow 2016)

Cross-entropy

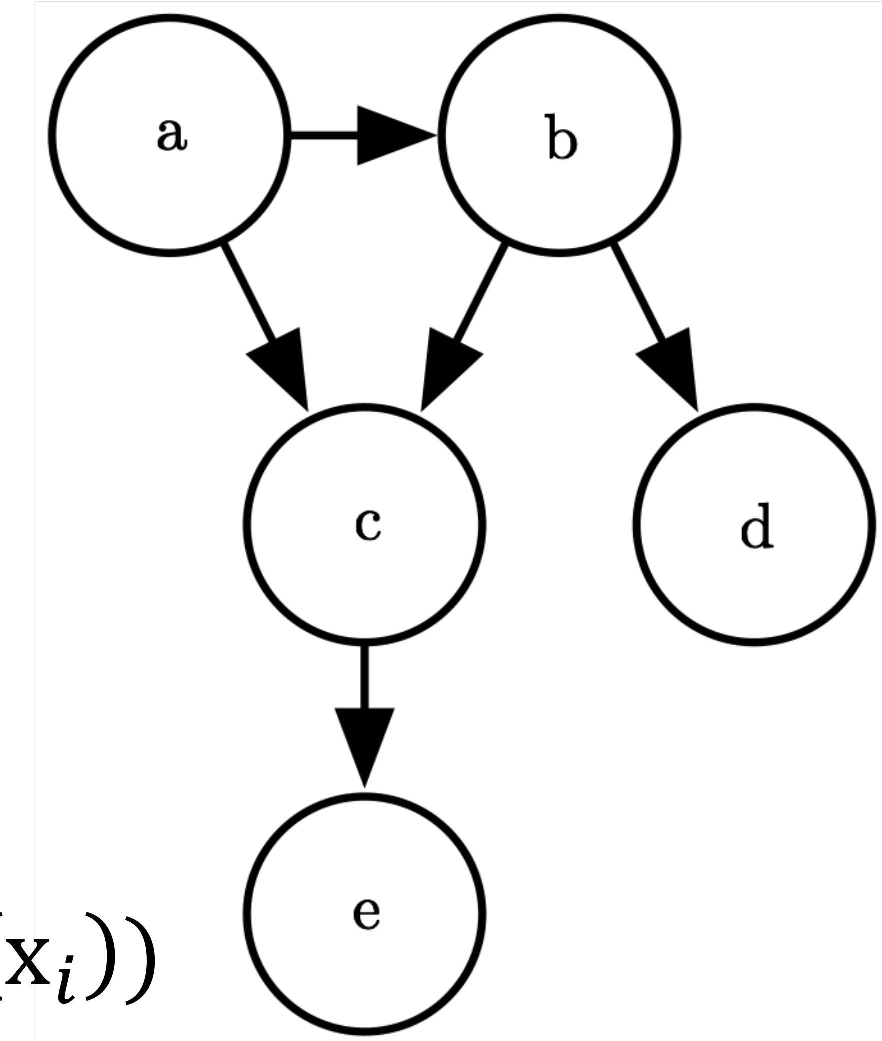
- $H(P, Q) = H(P) + D_{KL}(P||Q)$
  $= -\mathbb{E}_{x\sim P}\,(\log P(x)) + \mathbb{E}_{x\sim P}(\log P(x)) - \mathbb{E}_{x\sim P}(\log Q(x))$
  $= -\mathbb{E}_{x\sim P}(\log Q(x))$
- Minimzing the cross entropy with respect to Q is equivalent to minimize the KL divergence!

- Remark: usually we consider $0 \log 0 = 0$

(Goodfellow 2016)

Directed Model

Figure 3.7



$$p(\mathrm{x}) = \prod_i p(\mathrm{x}_i | Pa_{\mathcal{G}}(\mathrm{x}_i))$$

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}). \quad (3.54)$$

(Goodfellow 2016)