EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT

Deep Neural Networks

Chapter 3: Probability and Information Theory

3.1 Why Probability?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- There are three possible sources of uncertainty:
  1. Inherent stochasticity in the system being modeled.
  2. Incomplete observability.
  3. Incomplete modeling.
- In many cases, it is more practical to use a simple but uncertain rule rather than a complex but certain one
- “Most birds fly” is cheap to develop and is broadly useful, while a rule “All birds fly, except very young ones, very sick or injured ones, ostriches, penguins, kiwis,…” is not.

3.2 Random Variables

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- A random variable is a variable that can take on different values randomly. We write them as x, y, … (non-italics).
- We write values of random variables in the usual way, as x, y, … (in italics).
- $P(\mathrm{x} = x)$ is the probability that the random variable x takes on the value x.
- A random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states is.
- Random variables may be discrete or continuous.
- A discrete random variable has a finite or countably infinite number of states.

3.3 Probability distributions

- A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states.
- A probability distribution over discrete variables may be described using a probability mass function (PMF) P.
- We often simply call this a probability P.
- The probability that $\mathrm{x} = x$ is denoted as $P(x)$
- $P(x) = 1$ means that $\mathrm{x} = x$ is certain
- $P(x) = 0$ means that $\mathrm{x} = x$ is impossible.
- Sometimes we write explicitly: $P(\mathrm{x} = x)$.
- A probability distribution over many variables is known as a joint probability distribution.

3.3.1 Probability Mass Functions

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- $P(\mathrm{x} = x, \mathrm{y} = y)$ denotes the probability that $\mathrm{x} = x$ and $\mathrm{y} = y$. We may also write $P(x, y)$ for brevity.
- To be a probability mass function on a random variable x, a function P must satisfy the following properties:
  1. The domain of P is the set of all possible states of x.
  2. $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$.
     An impossible event has probability 0.
     An event that is guaranteed to happen has probability 1.
     Any event has a probability between these extremes.
  3. $\sum_{x \in \mathrm{x}} P(x) = 1$. All probabilities of a random variable sum up to 1. They are normalized.

Uniform Distribution

- A single discrete random variable x with k different states x_i, which are all equally probable (which have a uniform distribution), has a probability mass function

$$P(\mathrm{x} = x_i) = \frac{1}{k}$$

- We see that

$$\sum_i P(\mathrm{x} = x_i) = \sum_i \frac{1}{k} = \frac{k}{k} = 1$$

so the distribution is properly normalized.

3.3.2 Continuous Variables and PDFs

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- For continuous random variables we describe probability distributions using a probability density function (PDF) rather than a probability mass function.
- To be a probability density function, a function p must satisfy the following properties:
  1. The domain of p is the set of all possible states of x.
  2. $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
  3. $\int p(x)dx = 1$.
- In the univariate example, the probability that x lies in the interval $[a, b]$ is given by $\int_{[a,b]} p(x)dx$.

Uniform distribution

- Consider a uniform distribution on an interval of $\mathbb{R}$. We write this as function $u(x; a, b)$, where a and b are the endpoints of the interval, with $a < b$.
  x is the argument of u, while a and b are parameters.
- To ensure that there is no probability mass outside the interval, we say $u(x; a, b) = 0$ for all $x \notin [a, b]$.
- Within $[a, b]$, $u(x; a, b) = \frac{1}{b - a}$.
- This is nonnegative everywhere, and it integrates to 1.
- We denote that x follows the uniform distribution on $[a, b]$ by writing $x \sim U(a, b)$.

3.4 Marginal Probability

- Sometimes we know the probability distribution over a set of variables and we want to know the probability distribution over just a subset of them.
- The probability distribution over the subset is known as the marginal probability distribution.
- If we have discrete random variables x and y, we can compute $P(x)$ from $P(x, y)$ with the sum rule:

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

- For continuous variables, we need to use integration instead of summation:

$$p(x) = \int p(x, y) dy$$

3.5 Conditional Probability

- A conditional probability is the probability of some event, given that some other event has happened.
- $P(\mathrm{y} = y \mid \mathrm{x} = x)$ is the conditional probability that $\mathrm{y} = y$ given $\mathrm{x} = x$.
- It can be computed as

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

- It is only defined when $P(\mathrm{x} = x) > 0$.

3.6 Chain Rule of Cond. Probabilities

- Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

- This is known as the chain rule or product rule of probability.
- For example, we can compute

$$P(\mathrm{a}, \mathrm{b}, \mathrm{c}) = P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c})$$

$$P(\mathrm{b}, \mathrm{c}) = P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c})$$

$$P(\mathrm{a}, \mathrm{b}, \mathrm{c}) = P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c})$$

- Two random variables x and y are independent if their prob. distribution can be expressed as a product of two factors, one involving only x and one involving only y:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, \quad p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)\, p(\mathrm{y} = y)$$

- Two random variables x and y are conditionally independent given a random variable z if the conditional probability distribution over x and y factorizes in this way for every value of z:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},$$

$$p(\mathrm{x} = x, \mathrm{y} = y \,|\, \mathrm{z} = z) = p(\mathrm{x} = x \,|\, \mathrm{z} = z)\, p(\mathrm{y} = y \,|\, \mathrm{z} = z)$$

- Notation: $x \perp y$ means that x and y are independent,
  $x \perp y \mid z$ means that x and y are conditionally independent given z.

3.8 Expectation, Variance, Covariance

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- The expectation or expected value of some function $f(x)$ with respect to a prob. distr. $P(\mathrm{x})$ is the average or mean value that f takes on when x is drawn from P.
- For discrete variables this is computed by summation:

$$\mathbb{E}_{\mathrm{x}\sim P}\left[f(x)\right] = \sum_{x} P(x) f(x)$$

- For continuous variables it is computed with an integral:

$$\mathbb{E}_{\mathrm{x}\sim P}\left[f(x)\right] = \int P(x) f(x) dx$$

- We often write $\mathbb{E}_{\mathrm{x}}\left[f(x)\right]$ or $\mathbb{E}\left[f(x)\right]$ when x is clear.
- Expectations are linear:

$$\mathbb{E}_{\mathrm{x}}\left[\alpha f(x) + \beta g(x)\right] = \alpha \mathbb{E}_{\mathrm{x}}\left[f(x)\right] + \beta \mathbb{E}_{\mathrm{x}}\left[g(x)\right]$$

Variance, Covariance

- The variance gives a measure of how much the values of a function of a random variable x vary:

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

- The covariance indicates how much two values are linearly related to each other, as well as their scale:

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

- The covariance matrix of a random vector $x \in \mathbb{R}^n$ is an $n \times n$ matrix such that

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

- The diagonal elements of the matrix give the variance

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

3.9.0 Uniform Distribution

see 3.3.2

3.9.1 Bernoulli Distribution

- The Bernoulli distribution is a distribution over a single binary random variable. The parameter $p \in [0,1]$ gives the probability of the random variable being equal to 1.
- Properties: $P(\mathrm{x} = 1) = p$

$$P(\mathrm{x} = 0) = 1 - p$$

$$P(\mathrm{x} = x) = p^{x}(1-p)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = p$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = p(1-p)$$

3.9.1b Binomial Distribution

- The binomial distribution with parameters n and p is the discrete probability distr. of the number k of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p.
- Its probability mass function is

$$P(\mathrm{x} = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

with $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ being the binomial coefficients.

- The formula says: k successes occur with probability p^k and $n - k$ failures occur with probability $(1 - p)^{n-k}$. The k successes can occur anywhere among the n trials, and there are n over k different ways of distributing k successes in a sequence of n trials.

Binomial Distribution

- The expectation of the binomial distribution is

$$\mathbb{E}_x[x] = np$$

- The variance is

$$\mathrm{Var}_x(x) = np(1-p)$$

3.9.2 Multinomial Distribution

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- The multinomial distribution is a generalization of the binomial distribution. It models the probability of counts for rolling a k-sided die n times.
- When $n = 1$ and $k = 2$ (a coin) the multinomial distribution is the Bernoulli distribution.
- When $n > 1$ and $k = 2$ (a coin) it is the binomial distribution.
- When $n = 1$ and $k > 2$ it is the categorical distribution. In the deep learning book this is given the new name multinoulli distribution.
- (Note that k here denotes the number of categories or sides of a die, while in 3.9.1 k denoted the number of successes throwing a die with two categories, i.e. a coin, multiple times)

Multinomial Distribution

- The multinomial distribution has the following PMF:

$$P(\mathrm{x}_1 = x_1, ..., \mathrm{x}_k = x_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

- with $\sum_i p_i = 1$
- Expectation value: For each i the random variable x_i is binomially distributed with parameters n and p_i, thus

$$\mathbb{E}[\mathrm{x}_i] = n \cdot p_i$$

- Variance

$$\mathrm{Var}(\mathrm{x}_i) = n \cdot p_i (1 - p_i)$$

- Covariance

$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j) = -n \cdot p_i p_j$$

3.9.3 Gaussian Distribution

- The normal distribution, also known as the Gaussian distribution is:

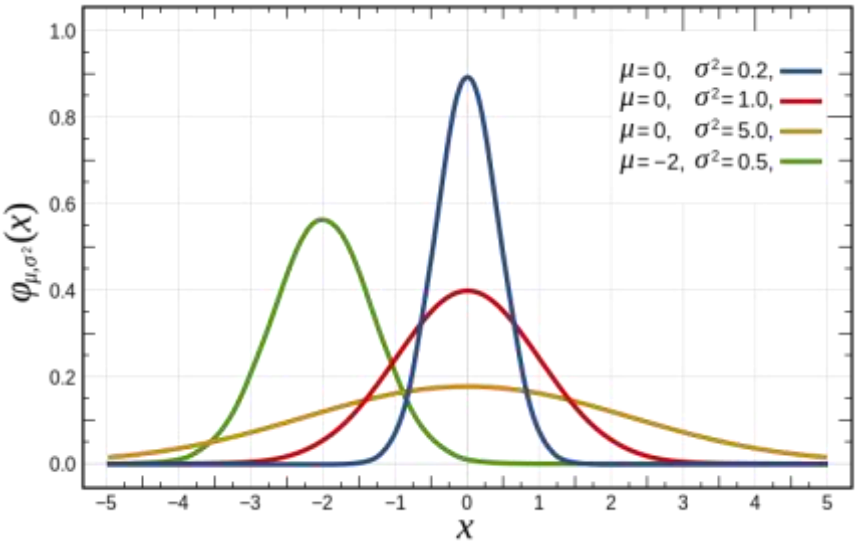
$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

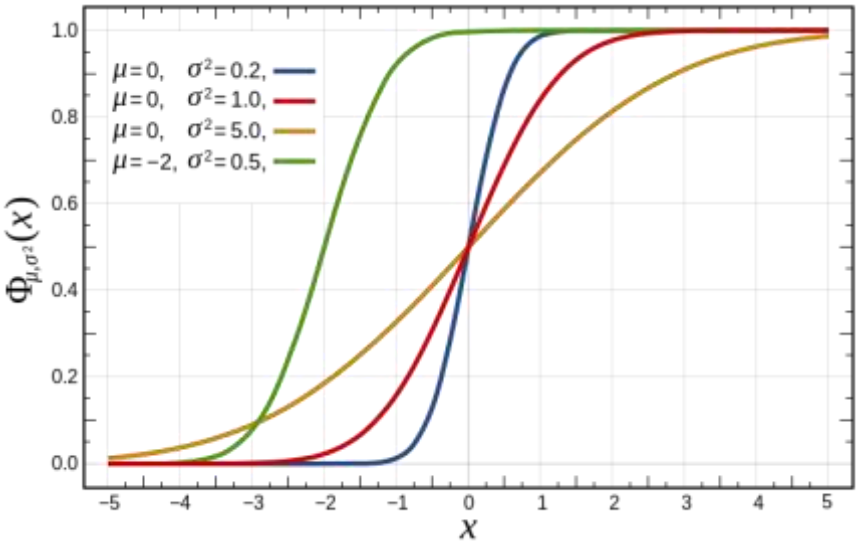
- $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ control the normal distribution. μ gives the coordinate of the central peak
- Mean of the distribution: $\mathbb{E}[\mathrm{x}] = \mu$.
- Standard deviation: σ,
- Variance: $\mathrm{Var}(\mathrm{x}) = \sigma^2$

- On the next page is a plot of the PDF of the normal distribution

Gaussian Distribution

- Probability density function Cumulative distrib. function





- The normal distribution $\mathcal{N}(x;\mu,\sigma^2)$ exhibits a “bell curve” shape, with the x coordinate of its central peak given by μ, and the width of its peak controlled by σ.

Gaussian Distribution

- If we frequently need to evaluate the PDF of the normal distribution it is more efficient to use the parameter $\beta \in (0, \infty)$ to control the precision (inverse variance):

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

- Many distributions we wish to model are truly close to being normal distributions.
- The central limit theorem shows that the sum of many independent random variables is approximately normally distributed.
- Out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers.

Gaussian Distribution

- The normal distribution generalized to $\mathbb{R}^n$, called the multivariate normal distribution

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

- Now x and μ are vectors, the parameter Σ is the covariance matrix of the distribution.
- Again, if we wish to evaluate the PDF many times for different parameter values, we can use a precision matrix β instead of the covariance Matrix Σ.

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

- We often fix the covariance matrix to be a diag. matrix.

3.9.4 Exponential and Laplace Distrib.

- The exponential distribution has a sharp point at $x = 0$:

$$p(x;\lambda) = \lambda \cdot \mathrm{l}_{x \geq 0} \exp(-\lambda x)$$

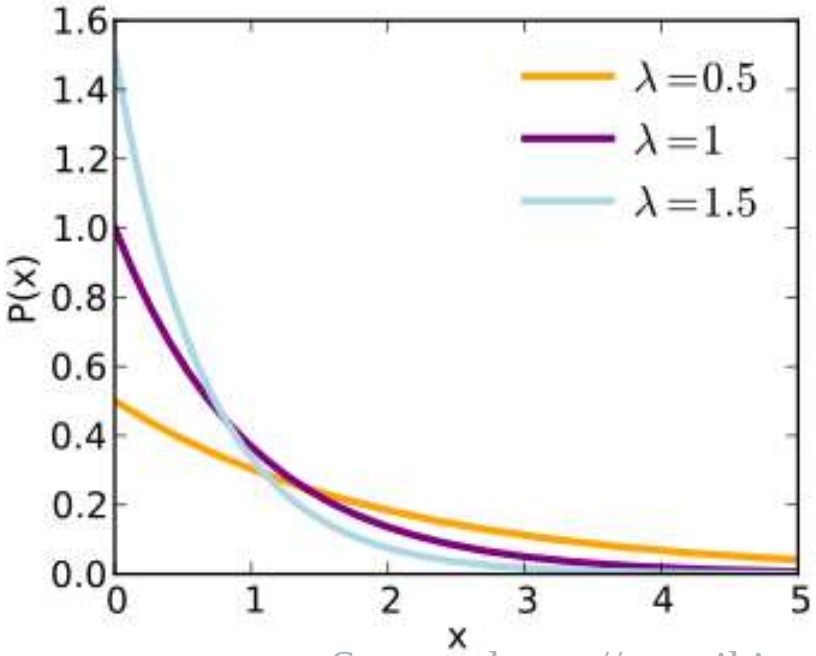
- It uses the indicator function $\mathrm{l}_{x \geq 0}$ to assign probability zero to all negative values of x. One could also write

$$p(x;\lambda) = \begin{cases} \lambda \cdot \exp(-\lambda x) & \text{if } x > 0 \\ 0 & \text{else} \end{cases}$$

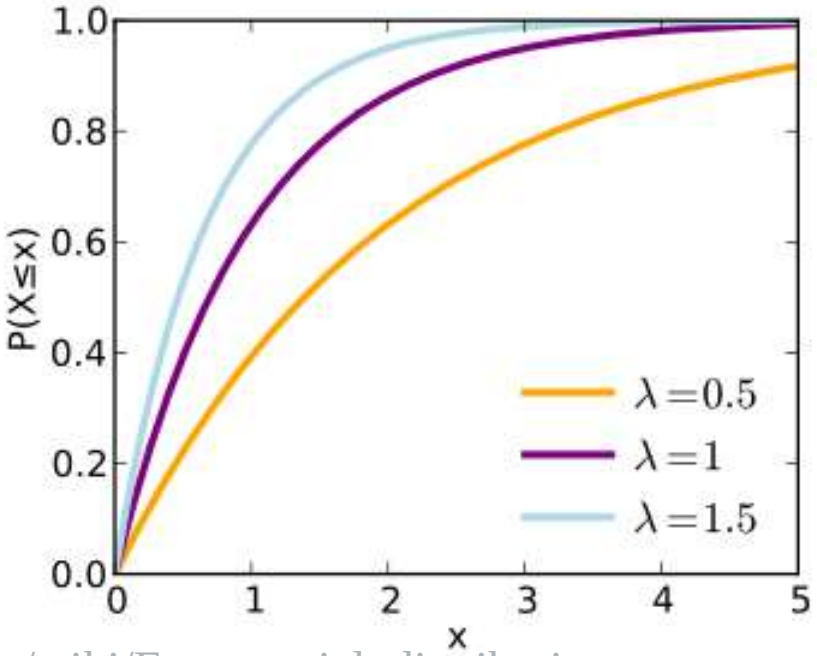
- It has the following properties:
- Mean: $\mathbb{E}[\mathrm{x}] = \lambda^{-1}$
- Variance: $\mathrm{Var}(x) = \lambda^{-2}$

Exponential Distribution

- Probability density funct.



Cumulative distribution funct.



Source: https://en.wikipedia.org/wiki/Exponential_distribution

Laplace Distribution

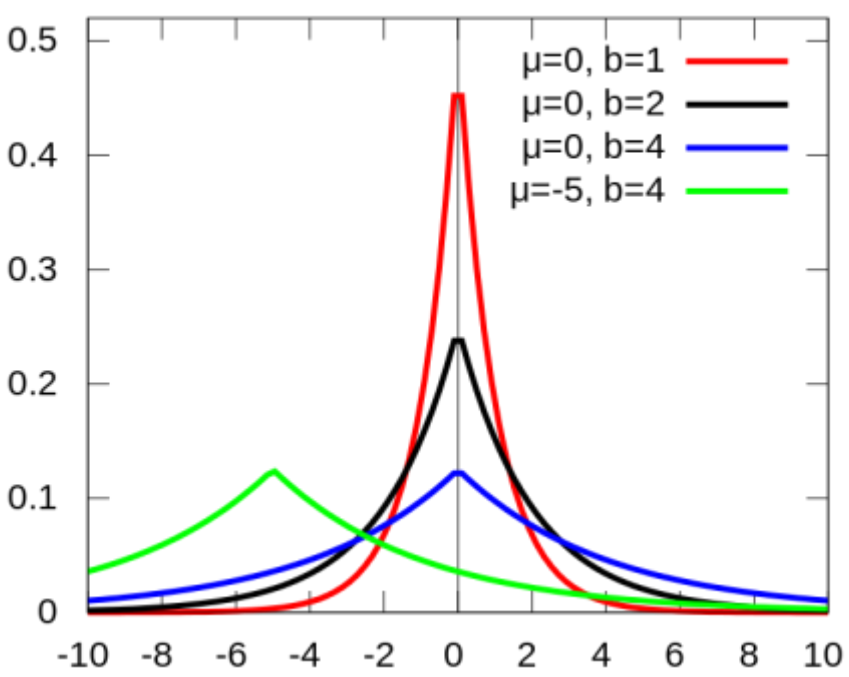
- The Laplace distribution allows to place a sharp peak at an arbitrary location μ.

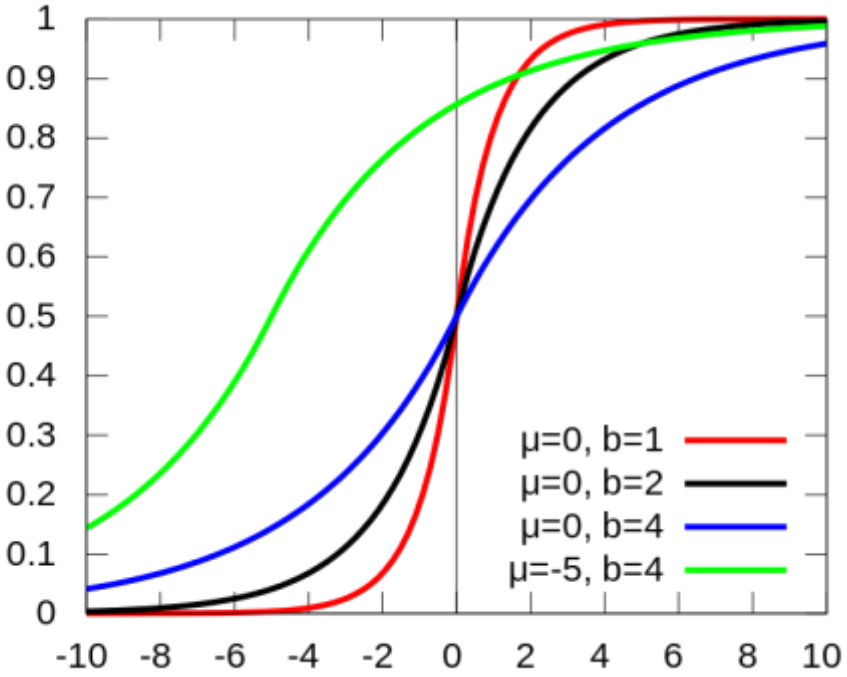
$$\text{Laplace}(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

- Here, μ is a location parameter and $b > 0$, called diversity, is a scale parameter.
- If $\mu = 0$ and $b = 1$, the positive half-line is exactly an exponential distribution scaled by 1/2.
- Whereas the normal distribution is expressed in terms of $|x-\mu|^2$, the Laplace distribution is expressed in terms of $|x-\mu|$. Consequently, the Laplace distribution has fatter tails than the normal distribution.

Laplace Distribution

- Probability density funct. Cumulative distribution funct.





https://en.wikipedia.org/wiki/Laplace_distribution#/media/File:Laplace_cdf_mod.svg

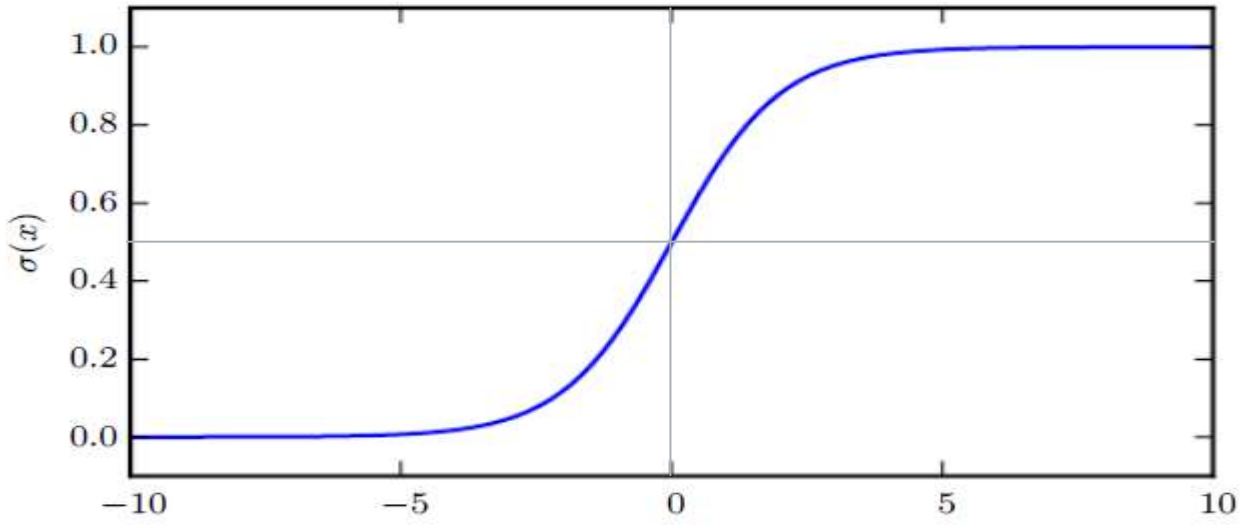
- Mean: $\mathbb{E}[\mathrm{x}] = \mu$, $\mathrm{Var}(x) = 2b^2$

3.10 Useful Properties of Some Functions

- Certain functions arise often while working with probability distributions used in deep learning models.
- The logistic function or logistic sigmoid:

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

It was frequently used as activation function of neurons.



The Softplus Function

- Another commonly encountered function is the softplus function

$$\zeta(x) = \log(1 + \exp(x))$$

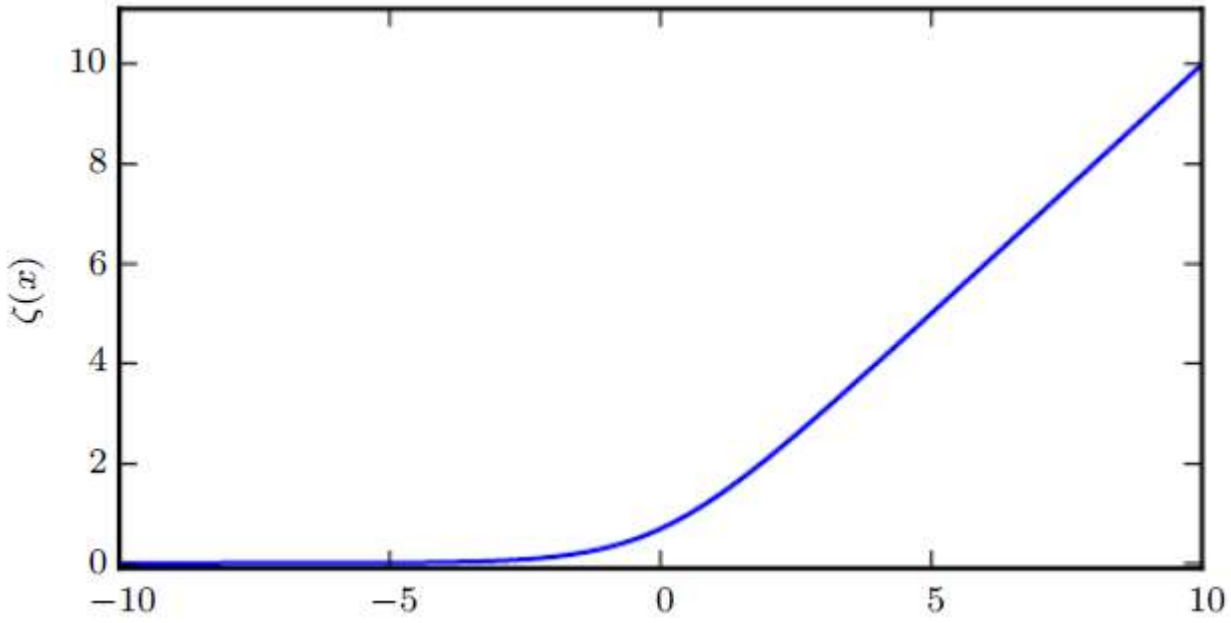
- The softplus function is useful for producing the β or σ parameter of a normal distrib. because its range is $(0,\infty)$.
- The name of the softplus function comes from the fact that it is a smoothed or "softened" version of

$$x^{+} = \max(0, x)$$

- Note that its slope is 1 and thus is much steeper than the slope of the logistic function.

The Softplus Function

- The softplus function



3.10 Useful Properties of Some Functions

- Here are some useful properties of functions:

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

  - The derivative of the logistic function is easy to compute

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

  - The logistic function is point symmetric

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

  - It is somehow related to the softplus function

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

  - The logistic function is actually the derivative of softplus!

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

Logistic and Softplus function

- Further properties of logistic and softplus functions:
  - The inverse of the logistic function is a simple logarithm

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

  - The inverse of softplus also has a simple form

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

  - Softplus is the integral of the logistic function

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y) dy$$

  - Softplus is a smoothed version of the positive part function $x^+ = \max(0, x)$:

$$\zeta(x) - \zeta(-x) = x$$

3.11 Bayes' Rule

- We often find ourselves in a situation where we know $P(\mathrm{y} \mid \mathrm{x})$ and need to know $P(\mathrm{x} \mid \mathrm{y})$. Fortunately, if we also know $P(\mathrm{x})$, we can compute the desired quantity using Bayes' rule:

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

- Note that while $P(\mathrm{y})$ appears in the formula, it is usually feasible to compute $P(\mathrm{y}) = \sum_{\mathrm{x}} P(\mathrm{y} \mid \mathrm{x})P(\mathrm{x})$, so we do not need to begin with knowledge of $P(\mathrm{y})$.
- It is named after the Reverend Thomas Bayes, who first discovered a special case of the formula. The general version was independently discovered by Pierre-Simon Laplace.

3.13 Information Theory

- Information theory is a branch of applied mathematics dealing with quantifying how much information is present in a signal.
- We here use some ideas from information theory to characterize probability distributions or quantify similarity between probability distributions.
- Intuition: learning of an unlikely event is more informative than learning of a likely event.
- We want the following properties:
  - Likely events should have low information content.
  - Less likely events should have higher information content.
  - Independent events should have additive information.

Information Theory

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- In order to satisfy all three of these properties, we define the self-information of an event $x = x$ to be

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

  the logarithm is the natural logarithm with base e.

- Our definition of $I(x)$ is therefore written in units of nats. One nat is the amount of information gained by observing an event of probability $1/e$. If we use base-2 logarithms the units are called bits; information in bits is just a rescaling of information measured in nats.

Shannon Entropy

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

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

- Shannon entropy of a distribution P is the expected amount of information in an event drawn from that distr. It gives a lower bound on the number of bits or nats needed on average to encode symbols drawn from P.
- Distributions that are nearly deterministic have low entropy; distributions that are closer to uniform have high entropy.
- When x is continuous, the Shannon entropy is known as the differential entropy.

Shannon Entropy of Distributions

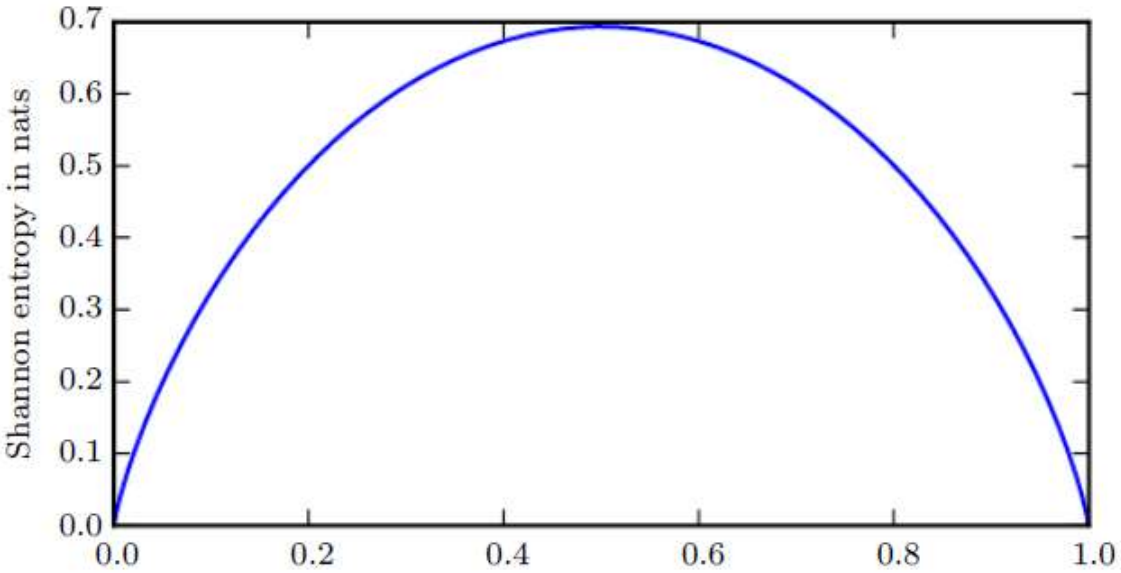


Fig. 3.5: This plot shows how distributions that are closer to deterministic have low Shannon entropy while distributions that are close to uniform have high Shannon entropy. On the horizontal axis, we plot p , the probability of a binary random variable. The entropy is given by $(p-1)\log(1-p) - p \log p$.
When p is near 0 or near 1, the distribution is nearly deterministic.
When $p = 0.5$, the entropy is maximal, because the distribution is uniform over the two outcomes.

Kullback-Leibler Divergence

- The Kullback-Leibler (KL) divergence measures how different two probability distributions $P(x)$ and $Q(x)$ over the same random variable x are:

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

- The KL divergence is non-negative.
- It is 0 iff P and Q are the same distribution (discrete variables) or equal “almost everywhere” (contin. var.)
- The KL divergence is no true distance measure because it is not symmetric:

$$D_{\mathrm{KL}}(P \,\|\, Q) \neq D_{\mathrm{KL}}(Q \,\|\, P) \quad \text{for some } P \text{ and } Q.$$

Kullback-Leibler Divergence

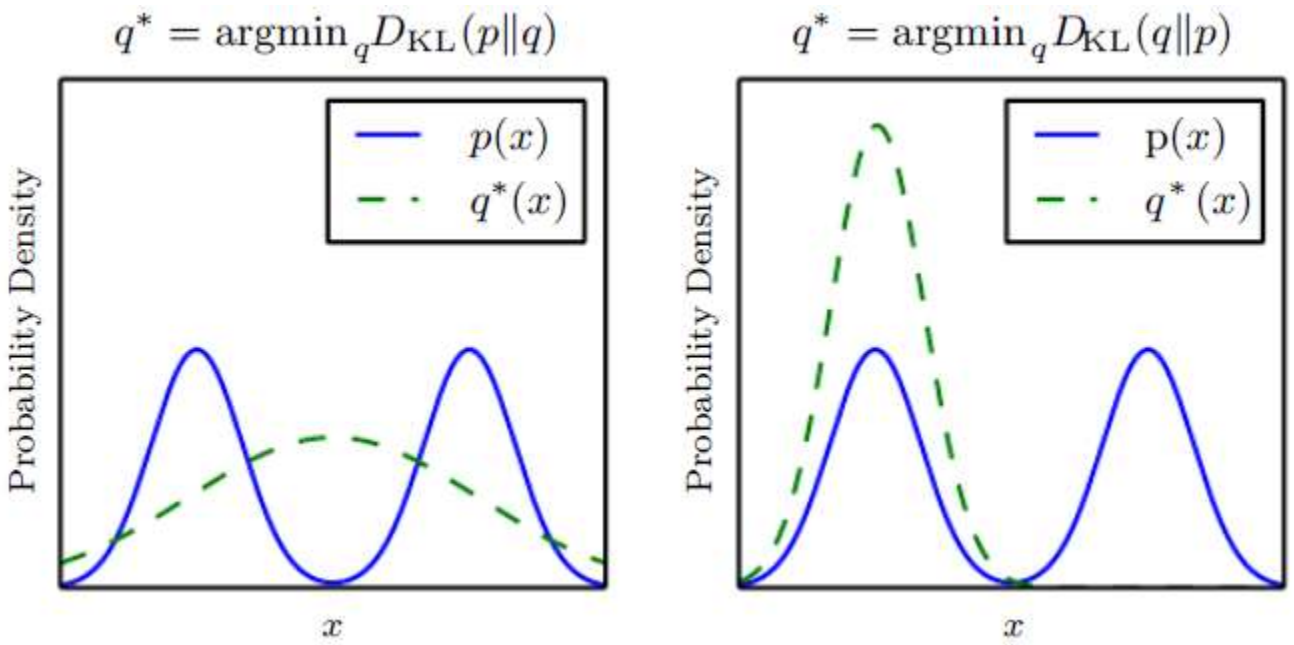


Fig. 3.6: The KL divergence is asymmetric. Suppose we have a distribution p(x) and wish to approximate it with a distribution q(x). We have may minimize either $D_{KL}(p\|q)$ or $D_{KL}(q\|p)$. We use a mixture of 2 Gaussians for p , and one Gaussian for q. *(Left)*The effect of minimizing $D_{KL}(p\|q)$. Here we select a q that has high probability where p has high probability. When p has multiple modes, q blurs the modes together. *(Right)*The effect of minimizing $D_{KL}(q\|p)$. Here we select a q that has low probability where p has low probability.

Cross Entropy

- The cross-entropy of two distributions P and Q is:

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

- It is similar to the KL divergence but lacking the term on the left:

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

- Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term.
- When computing many of these quantities, it is common to encounter expressions of the form $0 \log 0$. By convention, we treat these expressions as $\lim_{x \to 0} x \log x = 0$.

3.14 Structured Probabilistic Models

- ML algorithms often involve probability distributions over a very large number of random variables, but with few interactions between these variables.
- We then can split a probability distribution into many factors that we multiply together.
- Suppose we have three random variables: a, b and c. Suppose that a influences b and b influences c, but that a and c are independent given b. We can write

$$p(\mathrm{a},\mathrm{b},\mathrm{c}) = p(\mathrm{a})\,p(\mathrm{b}\,|\,\mathrm{a})\,p(\mathrm{c}\,|\,\mathrm{b})$$

- These factorizations can greatly reduce the number of parameters needed to describe the distribution.
- We can describe these kinds of factorizations using graphs. We call such a graph a graphical model.

Directed Graphical Models

- Directed graphical models use graphs with directed edges
- A directed model contains one factor for every RV x_i in the distribution, and that factor consists of the conditional distrib. over x_i given the parents of x_i, denoted $Pa_{\mathcal{G}}(x_i)$:

$$p(\mathbf{x}) = \prod_i p(x_i \mid Pa_{\mathcal{G}}(x_i))$$

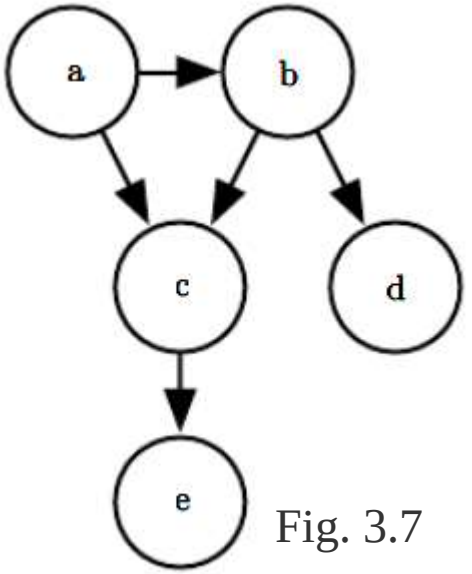


Fig. 3.7

- Example (top right): A directed graphical model over RVs a, b, c, d and e that corresponds to probability distributions that can be factored as

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

Undirected Graphical Models

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Undirected graphical models use graphs with undirected edges, and they represent factorizations into a set of functions; unlike in the directed case, these functions are usually not probability distributions of any kind.
- Any set of nodes that are all connected to each other in $\mathcal{G}$ is called a clique.
- Each clique $\mathcal{C}^{(i)}$ in an undirected model is associated with a factor $\phi^{(i)}(\mathcal{C}^{(i)})$. These factors are just functions, not probability distributions. We therefore normalize the product of the functions with a constant Z, defined to be the sum or integral over all the states of the product:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \phi^{(i)}(\mathcal{C}^{(i)})$$

Undirected Graphical Models

- Example: An undirected graphical model over random variables a, b, c, d and e.
- This graph corresponds to probability distributions that can be factored as

$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a},\mathrm{b},\mathrm{c})\phi^{(2)}(\mathrm{b},\mathrm{d})\phi^{(3)}(\mathrm{c},\mathrm{e})$$

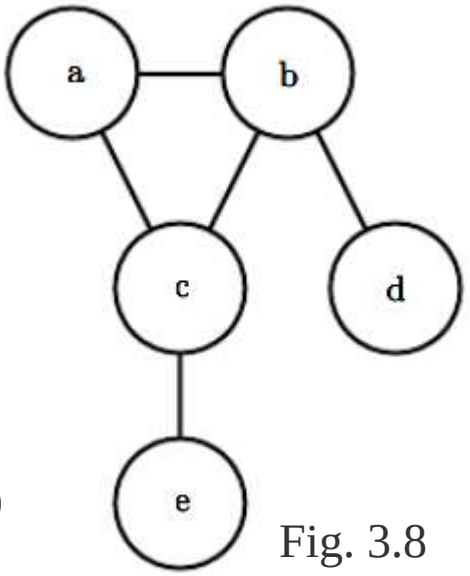


Fig. 3.8

- Being directed or undirected is not a property of a probability distribution; it is a property of a particular description of a probability distribution, but any probability distribution may be described in both ways.