Introduction to AI for postgraduate students

Lecture Note 2-2
Probability Theory

Hyun Jong Yang (hyunyang@postech.ac.kr)

Probability in AI

Probability theory: mathematical framework for representing uncertain statements

- The laws of probability tell us how AI systems should reason
  - We design our algorithms to compute or approximate various expressions derived using probability theory
- We can use probability and statistics to theoretically analyze the behavior of proposed AI systems

This lecture will deal with **only a brief review** of probability theory. You are strongly suggested to read other materials related to probability theory for in-depth understanding.

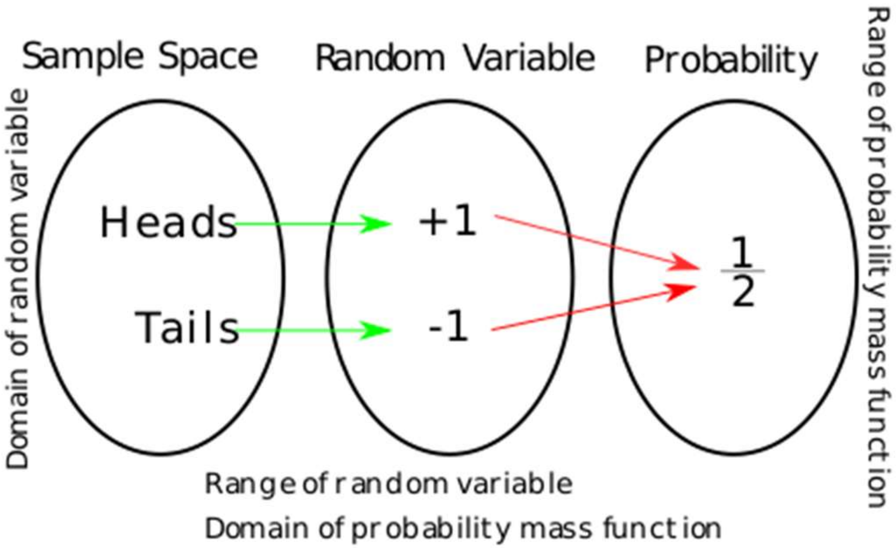
Why Probability?

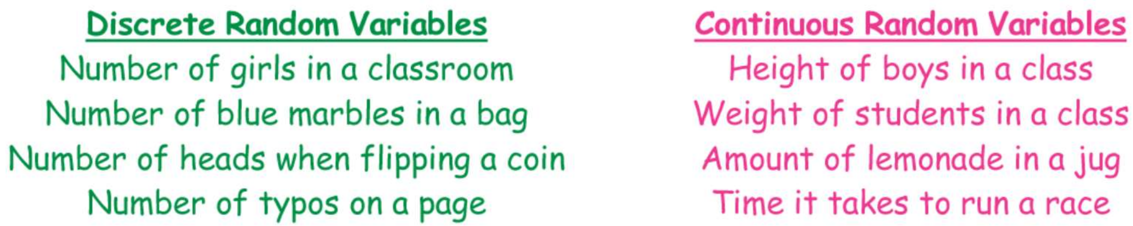
Possible sources of uncertainty

- Inherent stochasticity in the system being modeled
  - E.g., quantum mechanics
- Incomplete observability
  - Even deterministic systems can appear stochastic when we cannot observe all the variables that drive the behavior of the system
- 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 models' prediction

Random Variables

- Variable that can take on different values randomly
- There can be a vector-valued variable, typically denoted as a boldface letter, e.g., $\mathbf{x}$
- May be discrete or continuous





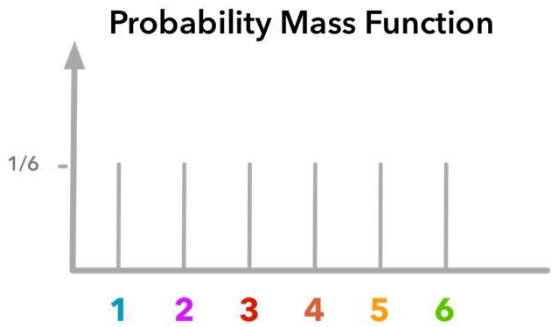
Discrete Variables & PMF

Probability mass function (PMF)

- Probability distribution over discrete variables
- Satisfies the following properties

> - 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.

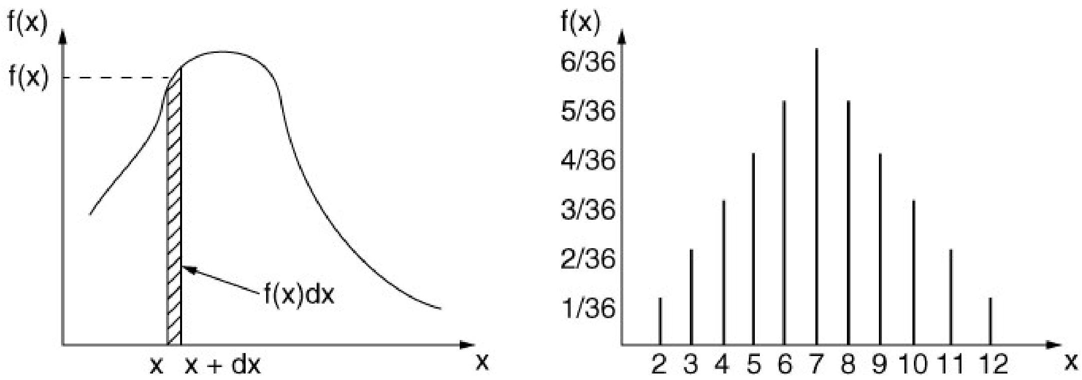
- Example: Uniform distribution



Continuous Variables & PDF

Probability density function (PDF)

- Probability distribution of a continuous random variable
- Satisfies the following properties
  - 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 $p(x) \leq 1$.
  - $\int p(x)dx = 1$.
- Probability that x lies in the interval $[a, b]$: $\int_{[a,b]} p(x)dx$

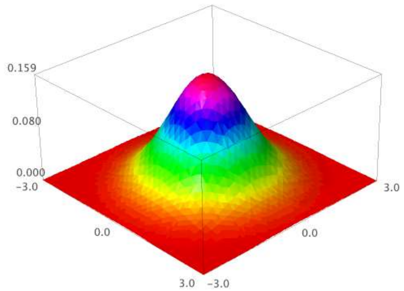


Continuous probability density function   Probability mass function

Joint & Marginal Probability

Joint probability distribution

		Y			
		0	1	2	3
X	1	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$
	2	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{10}$
	3	$\frac{1}{30}$	$\frac{1}{30}$	0	$\frac{1}{10}$



Joint PMF: $P(X = x, Y = y)$

Joint PDF: $p(X = x, Y = y)$

Probability that $X \in [a_1, a_2], Y \in [b_1, b_2]$: $\int_{a_1}^{a_2} \int_{b_1}^{b_2} p(x, y) dxdy$

Marginal probability

Marginal PMF: $\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y)$

Marginal PDF: $p(x) = \int p(x, y) dy$

Conditional Probability

Conditional probability that $Y = y$ given that $X = x$ as

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

Conditional probability is defined only when $P(X = x) > 0$.

Chain rule of conditional probabilities

$$P(A \cap B) = P(B \mid A) \cdot P(A) \Rightarrow \begin{aligned} \mathrm{P}(A_1 \cap A_2 \cap A_3 \cap A_4) &= \mathrm{P}(A_4 \mid A_3 \cap A_2 \cap A_1) \cdot \mathrm{P}(A_3 \cap A_2 \cap A_1) \\ &= \mathrm{P}(A_4 \mid A_3 \cap A_2 \cap A_1) \cdot \mathrm{P}(A_3 \mid A_2 \cap A_1) \cdot \mathrm{P}(A_2 \cap A_1) \\ &= \mathrm{P}(A_4 \mid A_3 \cap A_2 \cap A_1) \cdot \mathrm{P}(A_3 \mid A_2 \cap A_1) \cdot \mathrm{P}(A_2 \mid A_1) \cdot \mathrm{P}(A_1) \end{aligned}$$

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

Independence & Conditional Independence

Two random variables X and Y are independent if

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

$$\Downarrow$$

$$\boxed{\mathrm{x} \perp \mathrm{y}}$$

Two random variables X and Y are conditionally independent given a random variable Z if

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

$$\Downarrow$$

$$\boxed{\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}}$$

Expectation, Variance and Covariance

Expectation:

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

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

Variance:

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

Covariance:

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

Covariance matrix of a random vector $\boldsymbol{x} \in \mathbb{R}^n$

$$\boxed{\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j)} = \mathrm{E}\begin{bmatrix} (X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1]) & \cdots & (X_1 - \mathrm{E}[X_1])(X_K - \mathrm{E}[X_K]) \\ \vdots & \ddots & \vdots \\ (X_K - \mathrm{E}[X_K])(X_1 - \mathrm{E}[X_1]) & \cdots & (X_K - \mathrm{E}[X_K])(X_K - \mathrm{E}[X_K]) \end{bmatrix}$$

Bernoulli Distribution

- A distribution over a single binary random variable

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

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

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

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

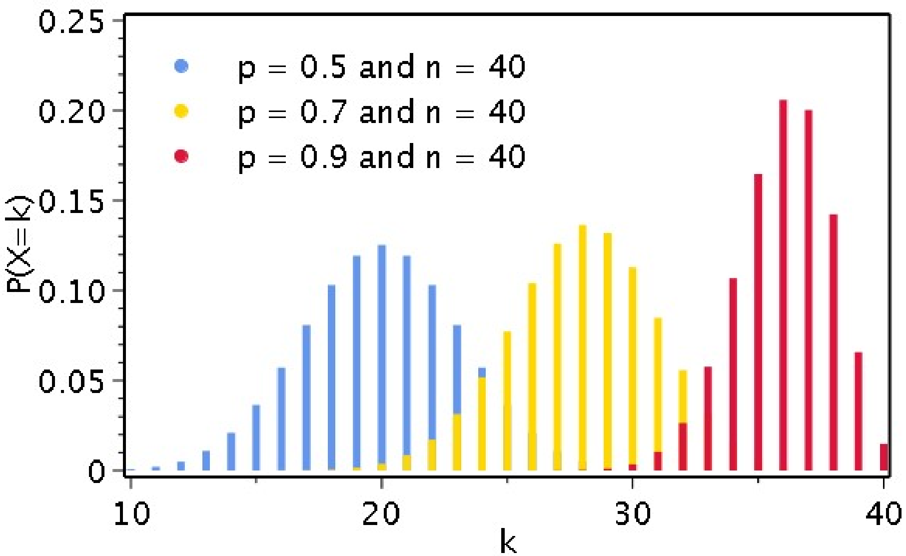
$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

Binomial Distribution

- PMF

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

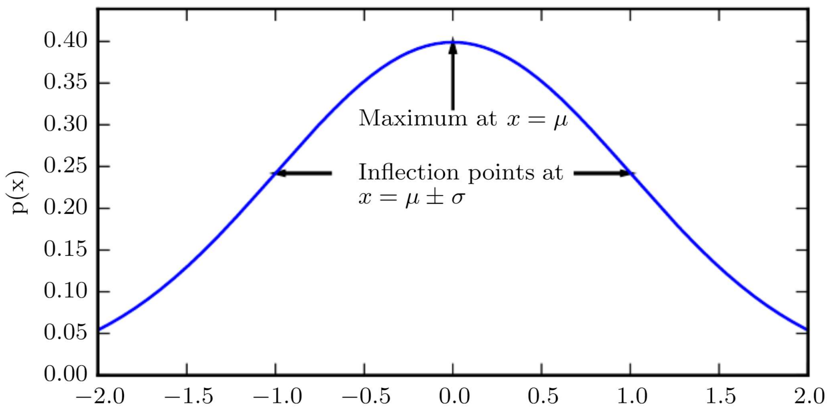


Gaussian Distribution

Gaussian (normal) distribution

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

- $E[x]=\mu$, $\mathrm{Var}(x)=\sigma^2$



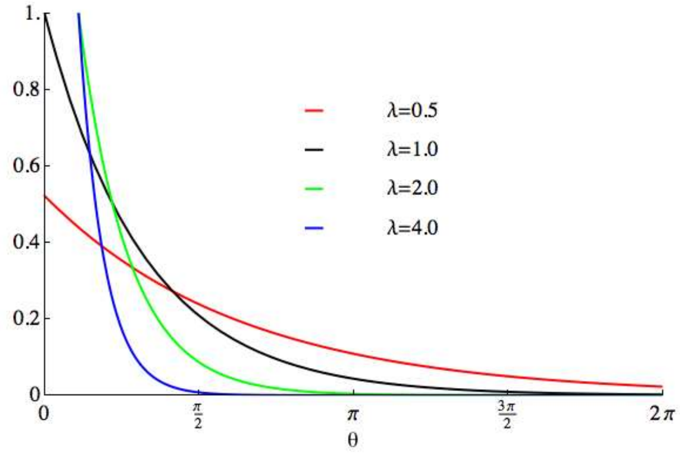
Multivariate normal distribution

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\sqrt{\frac{1}{(2\pi)^n\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\underbrace{\boldsymbol{\Sigma}^{-1}}_{\text{covariance}}(\boldsymbol{x}-\overbrace{\boldsymbol{\mu}}^{\text{mean}})\right)$$

Exponential and Laplace Distributions

Exponential distribution

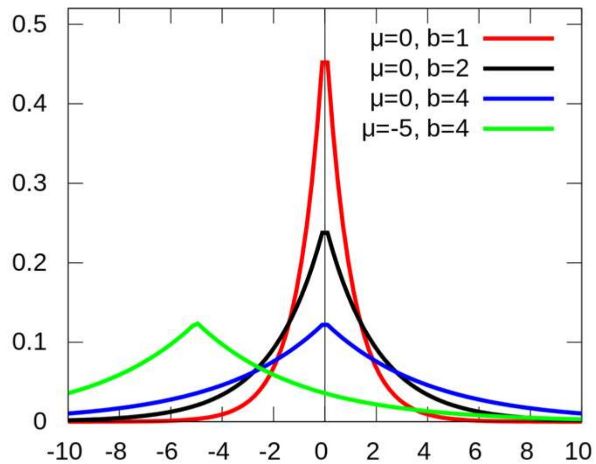
$$p(x;\lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$$



- $\mathbf{1}_{x \geq 0}$: assign probability zero to all negative values of x

Laplace distribution

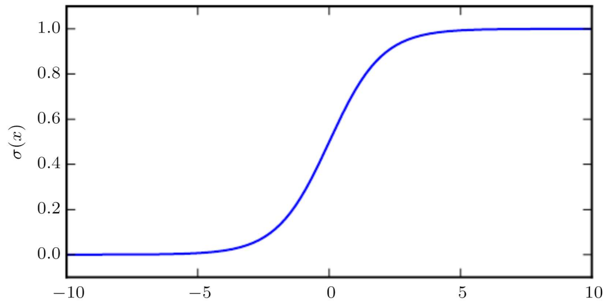
$$\text{Laplace}(x;\mu,\gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$



Useful Functions

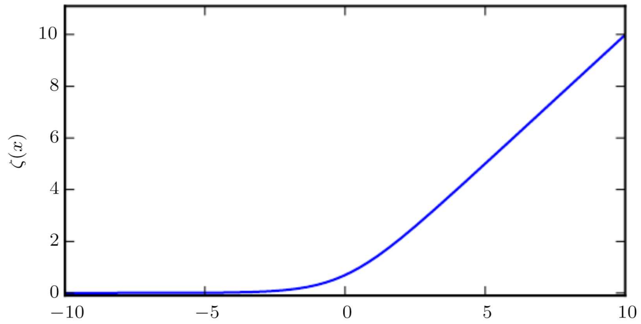
Logistic sigmoid:

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



Softplus:

$$\boxed{\zeta(x) = \log\left(1+\exp(x)\right)}$$



Useful Functions

Useful properties

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

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

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

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

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

logit

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

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

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

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

Bayes' Rule

- Bayes' rule

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

Information Theory

Motivation

- Likely events should have low information content, and in the extreme case, events that are guaranteed to happen should have no information content whatsoever
- Less likely events should have higher information content
- Independent events should have additive information. For example, finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up as heads once

Self-information

- Satisfies all these three properties
- And is defined by

$$\boxed{I(x) = -\log P(x)}$$

Shannon Entropy

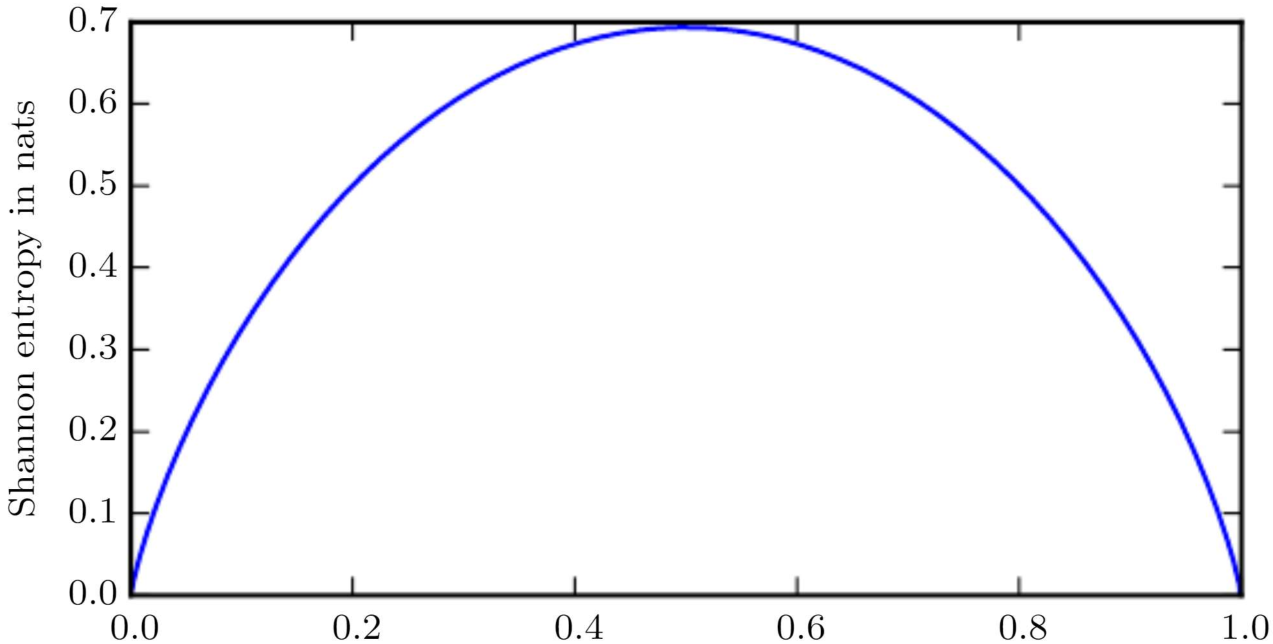
We can quantify the amount of uncertainty in an entire probability distribution using the **Shannon entropy**

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

- When x is continuous, the Shannon entropy is known as the differential entropy

Example of the binary case:

$$H(x) = P \cdot \log\frac{1}{P} + (1-P) \cdot \log\frac{1}{1-P}$$



KL Divergence & Cross-Entropy

If we have two separate probability distributions $P(x)$ and $Q(x)$ over the same random variable x, we can measure how different these two distributions are using the **Kullback-Leibler (KL) divergence**

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

Cross-entropy:

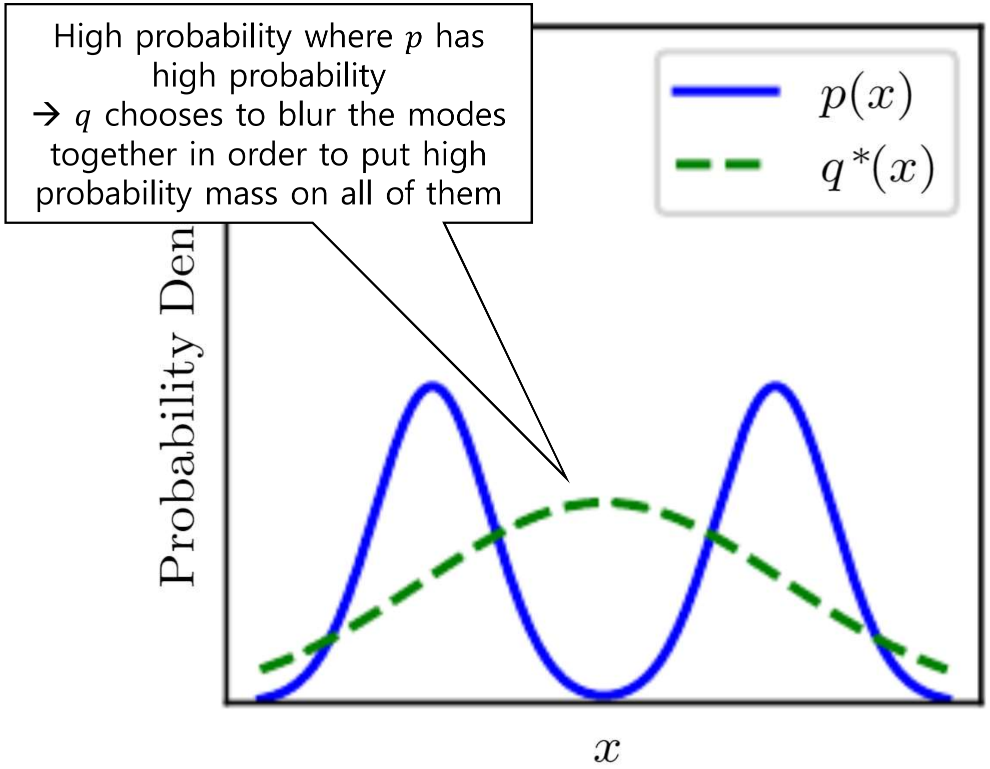
$$\begin{aligned} H(P,Q) &= H(P) + D_{\mathrm{KL}}(P\|Q) \\ &= -\mathbb{E}_{\mathrm{x}\sim P} \log Q(x) \end{aligned}$$

- similar to the KL divergence but lacking the term on the left
- Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term.

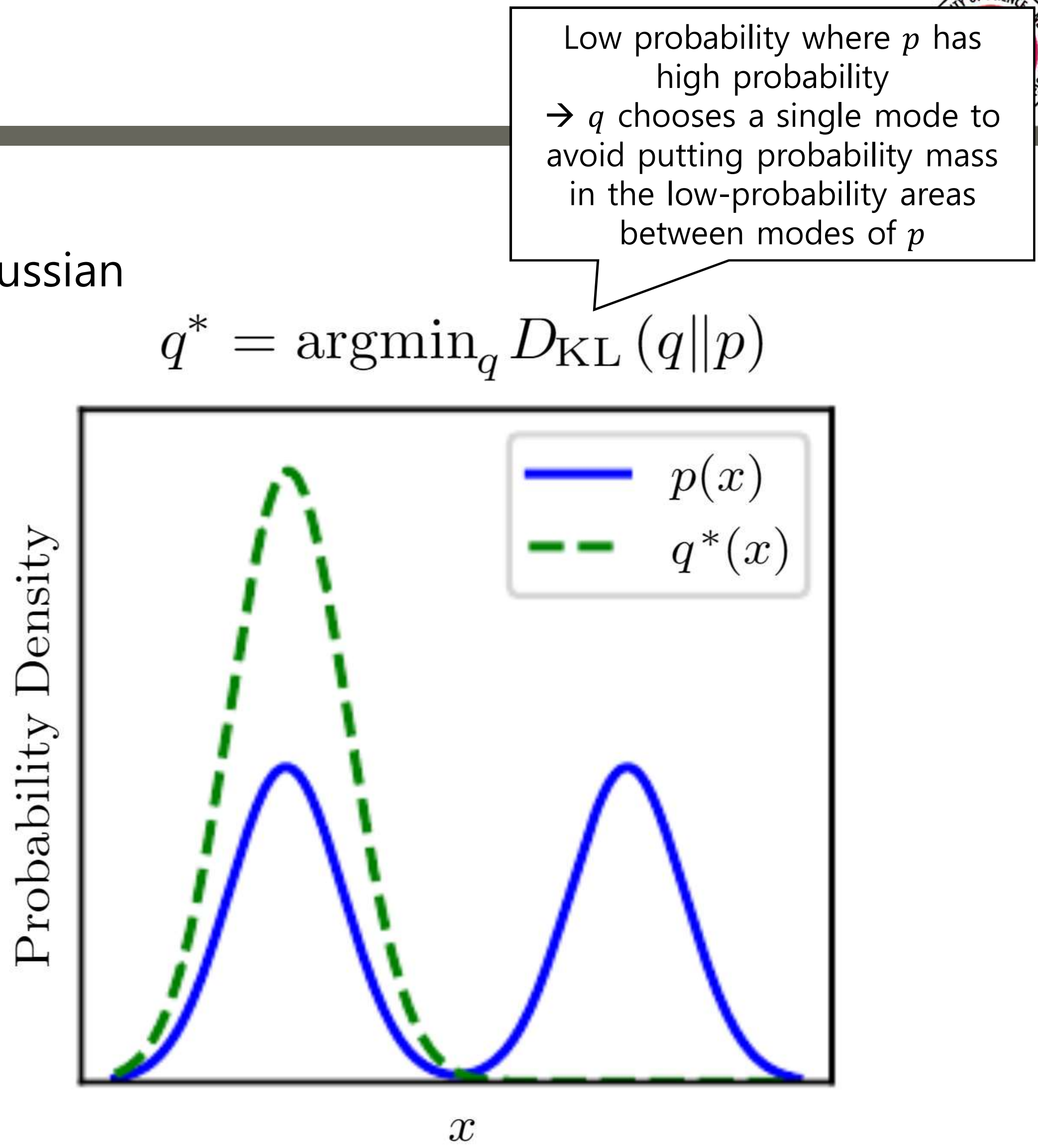
KL Divergence

- We wish to approximate $p(x)$ with $q(x)$
- $p(x)$: mixture of two Gaussians, $q(x)$: a single Gaussian

$$q^* = \text{argmin}_q D_{\text{KL}}(p||q)$$



$$q^* = \text{argmin}_q D_{\text{KL}}(q||p)$$



Structured Probabilistic Models

Probability distribution factorization

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

- Greatly reduce the number of parameters needed to describe the distribution, thereby reducing computational cost

Structured probabilistic model (graphical model)

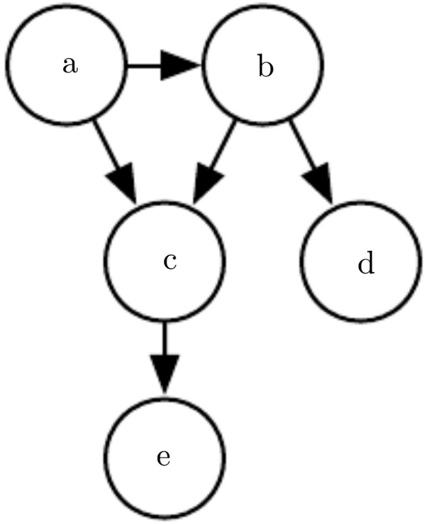
- We can describe these kinds of factorizations using graph theory

Structured Probabilistic Models

Directed graphical models

Parents of x_i

$$p(\mathbf{x}) = \prod_i p\left(\mathrm{x}_i \mid Pa_{\mathcal{G}}\left(\mathrm{x}_i\right)\right)$$

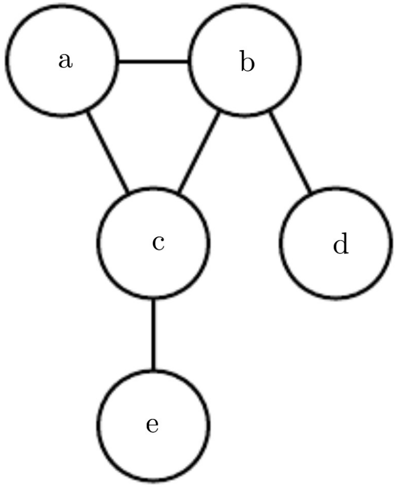


$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c})$$

- We can quickly see some properties of the distribution
  - a and c interact directly
  - a and e interact only indirectly via c

Structured Probabilistic Models

Undirected graphical models



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

Normalization factor

Clique: Set of nodes that are connected to each other

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

- We can quickly see some properties of the distribution
  - a and c interact directly
  - a and e interact only indirectly via c

Any probability distribution may be described in **both** ways, directed and undirected.