Entropy

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 15

entropy

- Consider a discrete random variable x. If we observe a specific value, how much information is received?
- If we have two independent events x and y, then the information should satisfy

$$h(\mathrm{x},\mathrm{y}) = h(\mathrm{x}) + h(\mathrm{y})$$

- Since $p(\mathrm{x},\mathrm{y}) = p(\mathrm{x})p(\mathrm{y})$, we can take

$$h(\mathrm{x}) = -\log_2 p(\mathrm{x})$$

entropy

- Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking the expectation with respect to the distribution $p_{\mathrm{x}}(x)$:

$$\mathrm{H}[\mathrm{x}] = -\sum_{x} p_{\mathrm{x}}(x)\ \log_2 p_{\mathrm{x}}(x)$$

- We call $\mathrm{H}[\mathrm{x}]$ the **entropy** of the random variable x.
- Correspondingly, We call $2^{\mathrm{H}[\mathrm{x}]}$ the perplexity of x.

- Note: here we understand $p(x) \log_2 p(x) = 0$ if $p(x) = 0$.

entropy

- Consider a random variable x having 8 possible states, each of which is equally likely. In order to communicate the value of x to a receiver, we would need to transmit a message of length 3 bits.
- Note that the entropy is given by

$$\mathrm{H}[\mathrm{x}] = -\sum_{\text{8 possible states}} \frac{1}{8} \log_2 \frac{1}{8}$$

$$= -8\,\frac{1}{8}\log_2\frac{1}{8} = 3 \text{ bits}$$

entropy

- Now consider an example of a variable x having 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

- We can calculate the entropy

$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2\text{ bits}$$

- Why do we have a smaller number of bits?

entropy

- Consider using the following code strings: 0, 10, 110, 1110, 111100, 111101, 111110, 111111 to encode $\{a, b, c, d, e, f, g, h\}$ whose probabilities are given by

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

- Then, on average,

$$\text{average code length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits}$$

- Shannon’s theorem: entropy is a lower bound on the number of bits needed to transmit the state of a random variable.

entropy

- Up to a scaler of $\ln 2$, we can use the natural logarithms "$\ln$" in defining entropy instead of "$\log_2$" That is, we measure the entropy in units of "nats" instead of "bits".
- A view from physics: considering a set of N identical objects that are to be divided amongst a set of bins, such that there are n_i objects in the i-th bin. The number of ways to do this is (called the (called the multiplicity):

$$W = \frac{N!}{\prod_i n_i!}$$

- The entropy is defined as the logarithm of the multiplicity scaled by $1/N$:

$$\mathrm{H} = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_i \ln n_i!$$

entropy

$$\mathrm{H} = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i!$$

- Consider large number n_i's. By Stirling's formula, $n! \simeq \left(\frac{n}{e}\right)^n$. Therefore, $\ln(n!) \simeq n\ln n - n$.

- We will have

$$\mathrm{H} = -\lim_{N\to\infty}\sum_i \left(\frac{n_i}{N}\right)\ln\left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

entropy

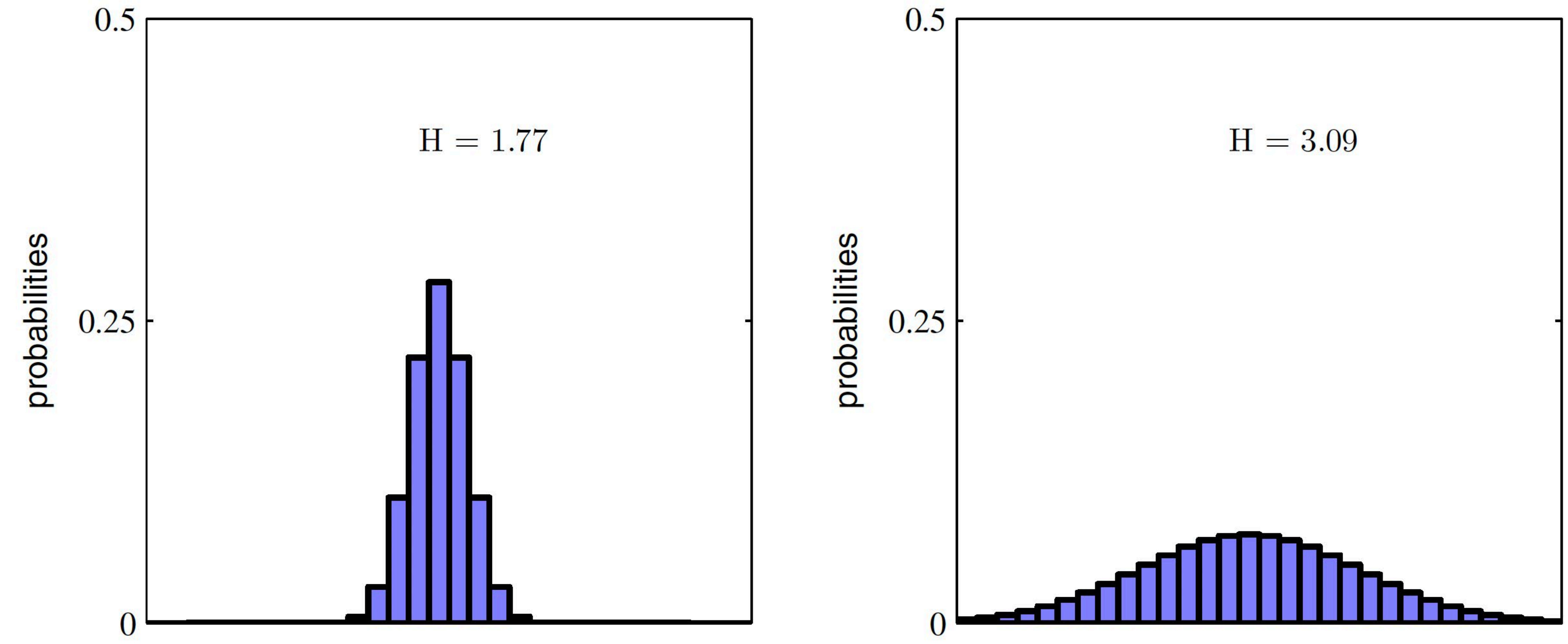


Figure 1.30 Histograms of two probability distributions over 30 bins illustrating the higher value of the entropy H for the broader distribution. The largest entropy would arise from a uniform distribution that would give $\mathrm{H} = -\ln(1/30) = 3.40$.

image taken from [Bi]

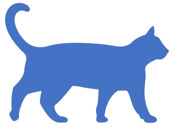
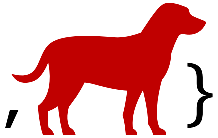
entropy

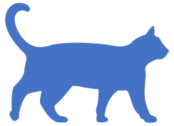
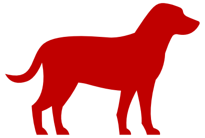
- We can interpret the bins as the states x_i of a discrete random variable x, where $p(\mathrm{x} = x_i) = p_i$
- The entropy of the random variable x is then

$$\mathrm{H}[p] = -\sum_i p(x_i) \ln p(x_i)$$

- The entropy H always satisfies $H \geq 0$. It is minimized at zero: $H = 0$ when one of $p_i = 1$ and all the other $p_{j \neq i} = 0$.

application of entropy

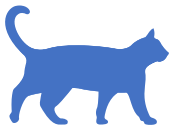
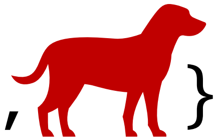
Suppose a random variable X takes two values {, }.
What are

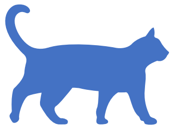
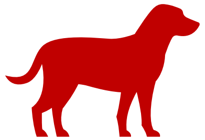
$P(X =$  $)$ and $P(X =$  $)$?

application of entropy

- **The Principle of Maximum Entropy**: the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy.

application of entropy

Suppose a random variable X takes two values {, }.
What are

$P(X =$  $)$ and $P(X =$  $)$?

- Back to this problem, we have:

application of entropy

- Suppose the states are given by $\{x_i\}_{i\in\mathcal{J}}$ and no prior information is given. Following this principle, we need to

$$\text{maximize } -\sum_{i\in\mathcal{J}} p(x_i)\ln p(x_i) \quad \text{subject to } \sum_{i\in\mathcal{J}} p(x_i) = 1.$$

- Using Lagrange multiplier, we need to maximize

$$\widetilde{\mathrm{H}} = -\sum_i p(x_i)\ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right)$$

- Suppose the number of states is $|\mathcal{J}| = M$. Maximizing $\widetilde{\mathrm{H}}$ yields $p(x_i) = \frac{1}{M}$ and the maximal entropy is $H = \ln M$.

application of entropy

- In the continuous case, the **differential entropy** is defined by

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

application of entropy* (optional)

- In the continuous case, the **differential entropy** is defined by

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

- Suppose we have constraints on the first and second moments of $p(\mathbf{x})$. Then maximum entropy principle implies we maximize $H[\mathbf{x}]$ subject to:

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x)\, \mathrm{d}x &= 1 \\
\int_{-\infty}^{\infty} x p(x)\, \mathrm{d}x &= \mu \\
\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, \mathrm{d}x &= \sigma^2
\end{aligned}$$

application of entropy* (optional)

- The Lagrange multiplier method requires maximizing

$$-\int_{-\infty}^{\infty} p(x) \ln p(x)\, \mathrm{d}x + \lambda_1 \left(\int_{-\infty}^{\infty} p(x)\, \mathrm{d}x - 1 \right)$$
$$+\lambda_2 \left(\int_{-\infty}^{\infty} x p(x)\, \mathrm{d}x - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, \mathrm{d}x - \sigma^2 \right)$$

- The form of $p(x)$ is $p(x) = \exp\left\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right\}$

- The Gaussian is, of course, $p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

entropy

- Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be a sequence of random variables drawn i.i.d. according to $p(\mathbf{x})$. By the Law of Large Numbers (lol#), we have

$$-\frac{1}{N}\ln p(\mathbf{x}_1, \cdots, \mathbf{x}_N) \to \mathbb{E}[-\ln p(\mathbf{x})] = H[p]$$

- For $\epsilon > 0$ and any N. The set

$$A_\epsilon^{(N)} = \left\{(\mathbf{x}_1, \cdots, \mathbf{x}_N): \left|-\frac{1}{N}\ln p(\mathbf{x}_1, \cdots, \mathbf{x}_N) - H[p]\right| \le \epsilon\right\}$$

is said to be a typical set.

conditional entropy

- Suppose we have a joint density $p(\mathbf{x}, \mathbf{y})$.
- If a value of $\mathbf{x}$ is already known, then the additional information needed to specify the corresponding value of $\mathbf{y}$ is given by $-\ln p(\mathbf{y}|\mathbf{x})$.
- The average additional information, called the **conditional entropy**, is

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

conditional entropy

- Fact: $\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$

Kullback-Leibler (KL) divergence

- Consider some unknown distribution $p(\mathbf{x})$. Suppose we model this using an approximating distribution $q(\mathbf{x})$.
- The additional information required to specify the value of $\mathbf{x}$ as a result of using q instead of p is called the **relative entropy**, or Kullback-Leibler (**KL) divergence**, given by

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

Kullback-Leibler (KL) divergence

- Fact: $\mathrm{KL}(p\|q) \geqslant 0$

Kullback-Leibler (KL) divergence

- Suppose that data is being generated from an unknown distribution $p(\mathbf{x})$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(\mathbf{x}|\theta)$.
- One way to determine θ is to minimize the KL divergence from $p(\mathbf{x})$ to $q(\mathbf{x}|\theta)$ with respect to θ.
- We cannot do this directly because we don't know $p(\mathbf{x})$. Suppose, however, that we have observed a finite set of training points $\mathbf{x}_n$,for $n = 1, \cdots, N$, drawn from $p(\mathbf{x})$. Then

$$\mathrm{KL}(p\|q) \simeq \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

- What are we doing if we minimize this KL divergence?

mutual information

- If $\mathbf{x}$ and $\mathbf{y}$ are independent, then $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$.
- For a general $p(\mathbf{x}, \mathbf{y})$, how close is it to being independent? We can use KL to measure

$$
\begin{aligned}
\mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}
\end{aligned}
$$

- This is called the **mutual information** between $\mathbf{x}$ and $\mathbf{y}$

mutual information

- Fact: $\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$

information does not hurt

- Fact: $\mathrm{H}[\mathbf{y}|\mathbf{x}] \leq \mathrm{H}[\mathbf{y}]$

independence bound on entropy

- Fact: Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be drawn according to $p(\mathbf{x}_1, \cdots, \mathbf{x}_N)$. Then

$$H[\mathbf{x}_1, \cdots, \mathbf{x}_N] \leq \sum_{n=1}^{N} H[\mathbf{x}_n]$$

KL divergence is convex

- KL divergence $\mathrm{KL}(p \parallel q)$ is convex in (p, q) :

  For any $(p_1, q_1), (p_2, q_2), 0 \leq \lambda \leq 1$,

$$\mathrm{KL}(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2) \leq \lambda\, \mathrm{KL}(p_1 \parallel q_1) + (1-\lambda) \mathrm{KL}(p_2 \parallel q_2)$$

KL divergence is convex

To prove the convexity of KL, we need the Log-Sum Inequality:

For nonnegative numbers $\{a_n\}_{n=1}^{N}$,$\{b_n\}_{n=1}^{N}$,

$$\sum_{n=1}^{N} a_n \ln\left(\frac{a_n}{b_n}\right) \geq \left(\sum_{n=1}^{N} a_n\right) \ln\left(\frac{\sum_{n=1}^{N} a_n}{\sum_{n=1}^{N} b_n}\right)$$

KL divergence is convex

Proof: In order to show $\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda \mathrm{KL}(p_1 \parallel q_1) + (1-\lambda)\mathrm{KL}(p_2 \parallel q_2)$,we only need to show that

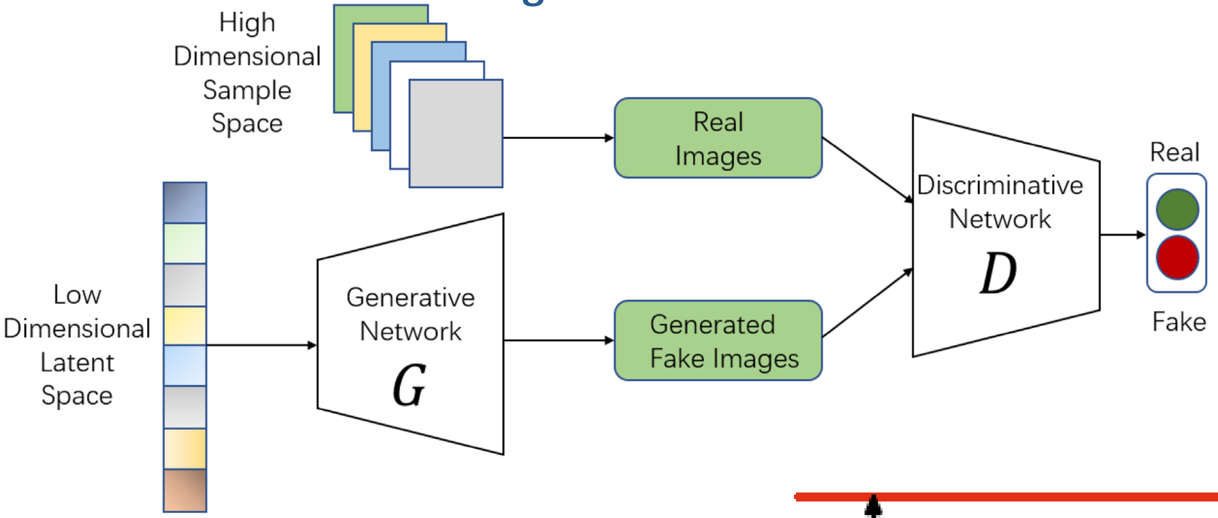
$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \ln \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}$$

$$\leq \lambda p_1(x) \ln \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \ln \frac{p_2(x)}{q_2(x)}$$

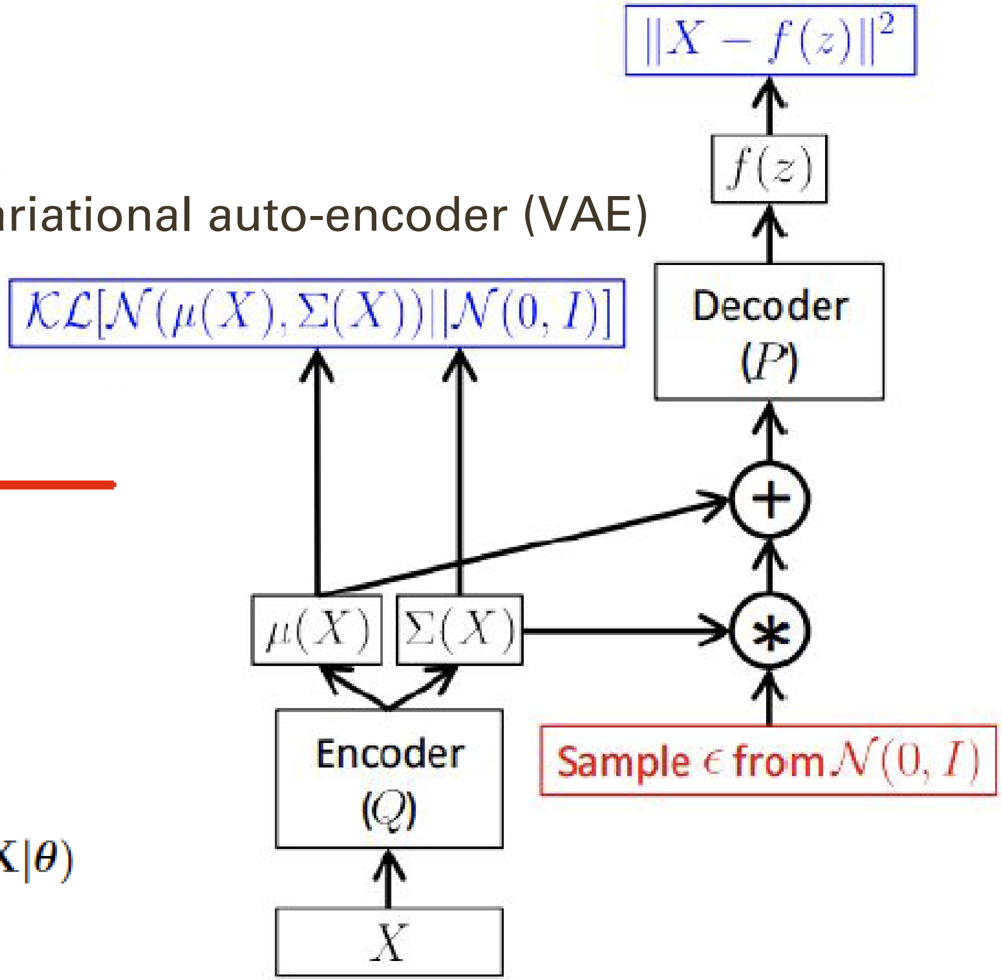
But this immediately follows the Log-Sum Inequality.

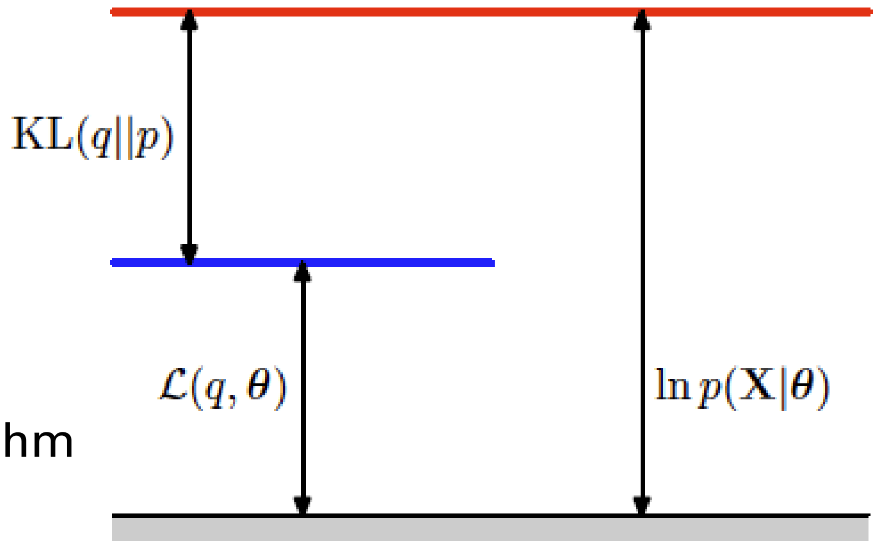
applications of KL divergence

generative adversarial network (GAN)



variational auto-encoder (VAE)





EM algorithm

example: KL of Gaussian

- Let $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^D$, $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$, and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$. Find $\mathrm{KL}(p \parallel q)$.

f-divergence: generalization of KL

- In general, if f is a differentiable convex function satisfying $f(1) = 0$, then we can define a "divergence", called f-divergence, by

$$D_f(p \parallel q) = \int p(\mathbf{x}) f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}$$

- For instance, take $f(u) = \frac{1}{2}(u-1)^2$, then

$$D_f(p \parallel q) = \frac{1}{2} \int \frac{(p(\mathbf{x}) - q(\mathbf{x}))^2}{p(\mathbf{x})} d\mathbf{x}.$$

Questions?

Reference

- *Information theory:*
  - *[Bi] Ch.1.6*