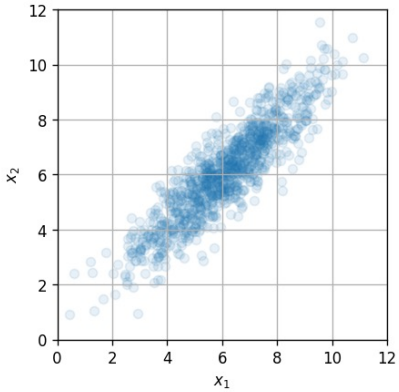
KL Divergence, Fano

- Kullback-Leibler divergence
- Fano's inequality for classifiers

Big Picture



- *unsupervised* machine learning is about finding interesting patterns in data

$\mathcal{D} = \{(x_i)\}_{i=1}^n$ • learn $p(\boldsymbol{x})$

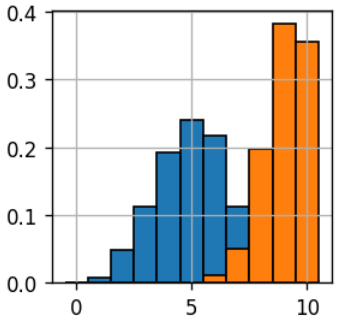
- how well can we approximate $p(\boldsymbol{x})$ or $p(\boldsymbol{x}, y)$?

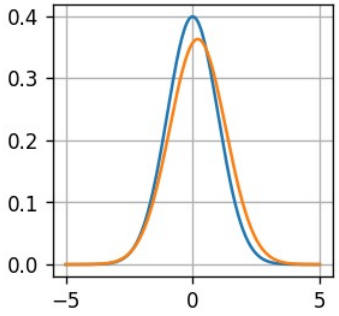
  - simpler question - how well can we estimate $E[\boldsymbol{x}]$?

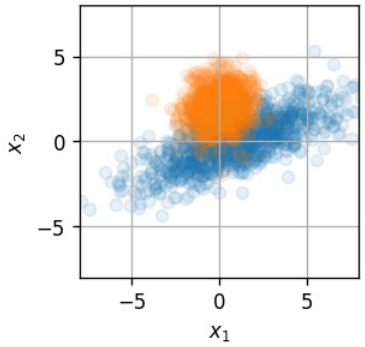
  - intuition: more samples, better job learning $p(\boldsymbol{x})$.

Kullback-Leibler Divergence

- basic task - measuring how different/similar two distributions are:







- many choices: ℓ_1, ℓ_2, KL divergence, total variation, Bhattacharyya distance, Hellinger, ...

- the *Kullback-Leibler Divergence* between to distributions p and q is:

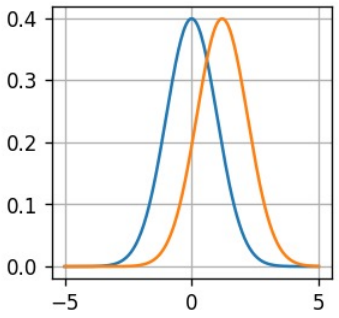
If $p(x) > 0$ implies $q(x) > 0$, then

$$D\left(p||q\right) = \sum p(x) \log \frac{p(x)}{q(x)} \qquad D\left(p||q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

where the sum or integral is over x with $p(x) > 0$. Otherwise, $D(p\|q) = \infty$.

Example

- example: compute $D(p||q)$ when $p \sim \mathcal{N}(0,1)$ and $q \sim \mathcal{N}(\mu,1)$



$$D\left(p||q\right) = E_p\left[\log \frac{p(x)}{q(x)}\right]$$

$$D\left(p||q\right) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \log \frac{e^{-x^2/2}}{e^{-(x-\mu)^2/2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(\frac{\mu^2}{2} - x\mu\right) dx$$

$$= \frac{\mu^2}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx + \mu \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \qquad = \frac{\mu^2}{2}$$

Kullback-Leibler Divergence

- the *Kullback-Leibler Divergence* between to distributions p and q is:

$$D\left(p||q\right) = E_p\left[\log \frac{p(x)}{q(x)}\right]$$

- why is $D(p||q)$ important?
  - imagine a classification task between $X_i \overset{i.i.d.}{\sim} p$ vs. $X_i \overset{i.i.d.}{\sim} q$, and the maximum likelihood rule:

$$\log \frac{p(x)}{q(x)} \lessgtr 0 \qquad\qquad \sum_{i=1}^{n} \log \frac{p(x_i)}{q(x_i)} \lessgtr 0$$

  - $D(p||q)$ is the expected value of the log likelihood ratio (when p is true class)

Chernoff-Stein Lemma:

Consider a simple hypothesis in which you observe n i.i.d. samples from either p or q, and any decision rule that has $\mathbb{P}(\text{ decide } q|p \text{ true }) \leq \epsilon$, then

$$\mathbb{P}\left(\text{decide } p|q \text{ true}\right) \gtrapprox e^{-nD(p||q)}$$

- fix one type of error, and the other error decays exponentially in n and $D(p||q)$.

Kullback-Leibler Divergence

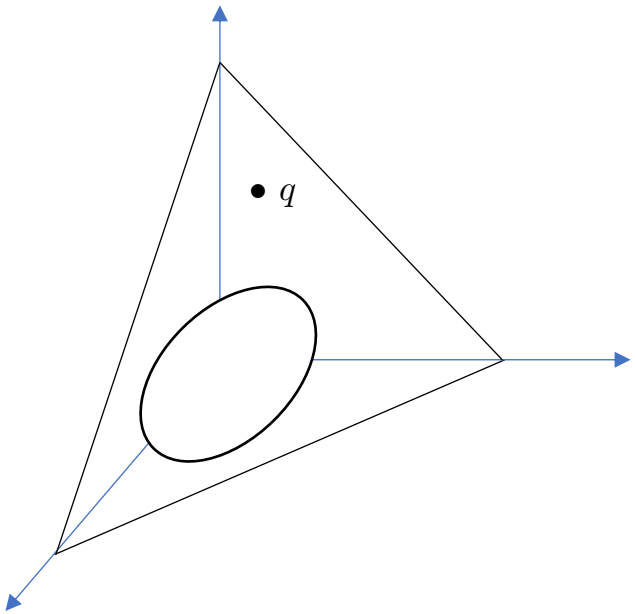
- the *Kullback-Leibler Divergence* between to distributions p and q is:

$$D\left(p||q\right) = \sum p(x) \log \frac{p(x)}{q(x)}$$

- why is $D(p||q)$ important?
  - can be used to find distribution *closest* to prior assumption and observations
    1) assume a prior distribution q
    2) make observations that constrain problem

$\min_p D(p||q)$

subject to p agrees with constraints



- $D(p||q)$ is convex in both p and q

- we'll do this when we look at variational autoencoders

Properties of KL Divergence

- the *Kullback-Leibler Divergence* between to distributions p and q is:

$$D\left(p||q\right) = E_p\left[\log \frac{p(x)}{q(x)}\right]$$

- KL divergence is non-negative: $D(p||q) \geq 0$ and $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x with $p(x) > 0$

sketch of proof:

$$\begin{aligned} -D(p||q) &= -\sum p(x) \log\left(\frac{p(x)}{q(x)}\right) \\ &= \sum p(x) \log\left(\frac{q(x)}{p(x)}\right) \\ &\leq \log\left(\sum p(x)\frac{q(x)}{p(x)}\right) \\ &= 0 \end{aligned}$$

Jensen: $E[f(X)] \leq f(E[X])$ for $f(x)$ concave

Jensen's Inequality

Jensen's inequality: for a convex function $f(\cdot)$ and a random variable X

$$E[f(X)] \geq f(E[X])$$

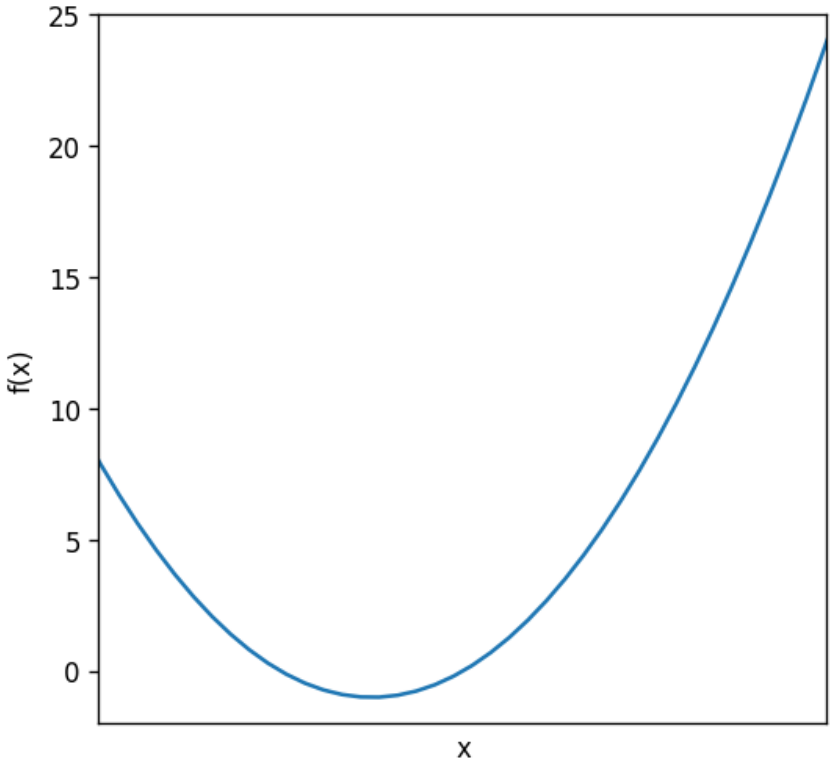
$$\sum p(x)f(x) \geq f(\sum xp(x))$$

sketch of proof:

1) consider binary X

$$p(x_1)f(x_1) + p(x_2)f(x_2) \geq f(x_1p(x_1) + x_2p(x_2))$$

2) by induction



Alternative Derivation of Jensen's Inequality

Recall that f is convex if for $0 \leq \lambda \leq 1$,

$$f\big(u + \lambda(v-u)\big) \leq f(u) + \lambda[f(v) - f(u)].$$

For $0 < \lambda \leq 1$, this is equivalent to

$$\frac{f\big(u + \lambda(v-u)\big) - f(u)}{\lambda} \leq f(v) - f(u).$$

Letting $\lambda \to 0$ implies

$$\nabla f(u)^T (v-u) \leq f(v) - f(u),$$

or

$$f(u) + \nabla f(u)^T (v-u) \leq f(v).$$

Now suppose X is a random vector with mean μ. Put $u = \mu$ and $v = X$ to get

$$f(\mu) + \nabla f(\mu)^T (X-\mu) \leq f(X).$$

Taking expectations yields $f(\mu) \leq \mathbb{E}[f(X)]$, or

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Entropy, Mutual Information, and KL Divergence

- definition - entropy:

$$H(X) = E\left[\log_2\left(\frac{1}{p(x)}\right)\right]$$

- amount of uncertainty in a random variable

- relating conditional entropy to joint (chain rule):

$$H(X,Y) = H(X) + H(Y|X)$$

- mutual information: reduction in uncertainty of Y due to knowledge of X

$$I(Y;X) = H(Y) - H(Y|X)$$

$$I(Y;X) = E\left[\log\left(\frac{p(x,y)}{p(y)p(x)}\right)\right]$$

- KL divergence:

$$D\left(p||q\right) = E_p\left[\log\frac{p(x)}{q(x)}\right]$$

- mutual information is equal to the KL divergence between the joint and the product of marginals

$$I(X;Y) = D\left(p(x,y)||p(x)p(y)\right)$$

Properties of Entropy

- definition - entropy: $H(X) = E\left[\log_2\left(\frac{1}{p(x)}\right)\right]$

- properties of entropy:

$$H(X) \geq 0$$

$$H(X) \leq \log_2(|\mathcal{X}|)$$

proof:

$$D\left(p||q\right) = \sum p(x) \log \frac{p(x)}{q(x)} \qquad \text{set } q(x) = \tfrac{1}{|\mathcal{X}|}, \text{ let } X \sim p(X)$$

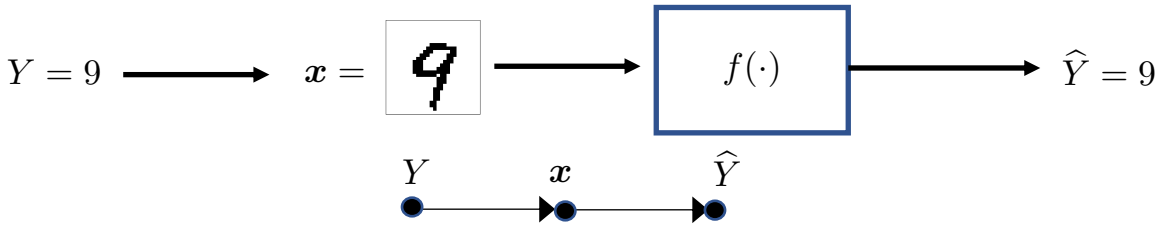
$$= \sum p(x)(\log|\mathcal{X}| + \log p(x))$$

$$= \log|\mathcal{X}| - H(X)$$

$$\geq 0$$

$$H(X) \leq \log|\mathcal{X}| \quad \text{with equality when } p(x) = \tfrac{1}{|\mathcal{X}|}$$

Fano's Inequality



- for any $\widehat{Y} = f(\boldsymbol{x})$ we have $Y \to \boldsymbol{x} \to \widehat{Y}$ and

$$\mathbb{P}(Y \neq \widehat{Y}) \geq \frac{H(Y|\boldsymbol{x}) - 1}{\log(|\mathcal{Y}|)}$$

where $|\mathcal{Y}|$ is the number of classes

- proof: define $E = \mathbb{I}_{\{\widehat{Y} \neq Y\}}$

$$\begin{aligned} H(E, Y|\widehat{Y}) &= H(Y|\widehat{Y}) + H(E|Y, \widehat{Y}) \\ &= H(Y|\widehat{Y}) \\ &\geq H(Y|\boldsymbol{x}) \end{aligned}$$

$$\begin{aligned} H(E, Y|\widehat{Y}) &= H(E|\widehat{Y}) + H(Y|E, \widehat{Y}) \\ &\leq 1 + H(Y|E, \widehat{Y}) \\ &\leq 1 + \mathbb{P}(E = 1)H(Y|E = 1, \widehat{Y}) + \mathbb{P}(E = 0)H(Y|E = 0, \widehat{Y}) \\ &\leq 1 + \mathbb{P}(\widehat{Y} \neq Y)H(Y|E = 1, \widehat{Y}) \\ &\leq 1 + \mathbb{P}(\widehat{Y} \neq Y) \log |\mathcal{Y}| \end{aligned}$$