Large Scale Data Analysis Using Deep Learning

Probability and Information Theory

U Kang
Seoul National University

In This Lecture

- Overview of basic probability theory
- Overview of information theory

Why Probability

- Probability theory: a mathematical framework for representing uncertain statements
  - Provides a means of quantifying uncertainty and axioms for making new uncertain statements
  - A fundamental tool of many disciplines of science and engineering

- Use of probability in AI
  - The laws of probability tell us how AI systems should reason, so we design algorithms to compute or approximate various expressions using probability theory
  - Theoretically analyze the behavior of proposed AI systems

Random Variable

- A random variable is a variable that can take on different values randomly
  - With some probabilities for values

- Random variables may be discrete or continuous

Probability Mass Function (PMF)

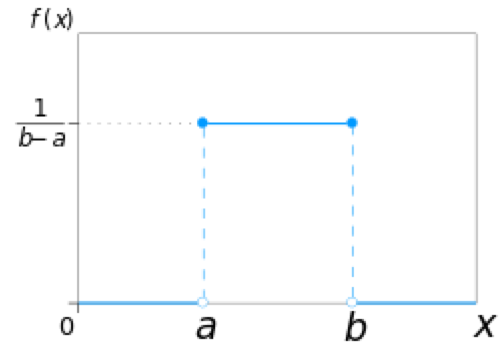
- For discrete random variables
- The domain of P must be the set of all possible states of x
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no 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
- Uniform distribution among k states: $P(x = x_i) = 1/k$

Probability Density Function (PDF)

- For continuous random variables
- 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 $p(x) \leq 1$
- $\int p(x)dx = 1$

- Uniform distribution u(x; a,b) = 1/(b-a)

parameterized by



Computing Marginal Probability with the Sum Rule

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

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

Conditional Probability

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

Chain Rule of Probability

- $P\left(x^{(1)}, \ldots, x^{(n)}\right) = P\left(x^{(1)}\right) \prod_{i=2}^{n} P\left(x^{(i)} | x^{(1)}, \ldots, x^{(i-1)}\right)$

- E.g., $P(a, b, c) = P(a \mid b, c)\, P(b \mid c)\, P(c)$

Independence

- Independence
  - $\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$
  - Notation: $x \perp y$

- Conditional independence
  - $\forall x \in \mathrm{x},\ y \in \mathrm{y},\ \mathrm{z} \in z,$
  - $p(\mathrm{x} = x, \mathrm{y} = y | \mathrm{z} = z) = p(\mathrm{x} = x | \mathrm{z} = z)p(\mathrm{y} = y | \mathrm{z} = z)$
  - Equivalently, $p(x|y, z) = p(x|z)$
    - Pf?

  - Notation: $x \perp y \mid z$

Expectation

- Discrete variable: $E_{x \sim P}[f(x)] = \sum_x P(x) f(x)$
- Continuous variable: $E_{x \sim p}[f(x)] = \int p(x) f(x) dx$
- Linearity of expectations:
  - $E_x[\alpha f(x) + \beta g(x)] = \alpha E_x[f(x)] + \beta E_x[g(x)]$
  - This always holds, even when f(x) and g(x) are dependent

Variance and Covariance

- $\mathrm{Var}\big(\mathrm{f(x)}\big) = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - (E[f(x)])^2$
- Standard deviation: square root of Var

- $\mathrm{Cov(f(x)}, g(y)) = E[(f(x) - E[f(x)])(g(y) - E[g(x)])]$
- Intuition:
  - Positive covariance
  - Negative covariance
- Covariance matrix: $Cov(x)_{i,j} = Cov(x_i, x_j\ )$
  - Diagonal elements $Cov(x)_{i,i} = Var(x_i)$

Bernoulli Distribution

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

- $E[x] = \phi$
- $Var[x] = \phi\,(1\text{-}\,\phi)$
  - Pf?

Multinoulli Distribution

- Categorical Distribution
- A distribution over a single discrete variable with k different states
- Parameterized by a vector $p \in [0, 1]^{k-1}$
- The final, k-th state's probability is given by $1 - \mathbf{1}^{\mathrm{T}} p$

Gaussian Distribution

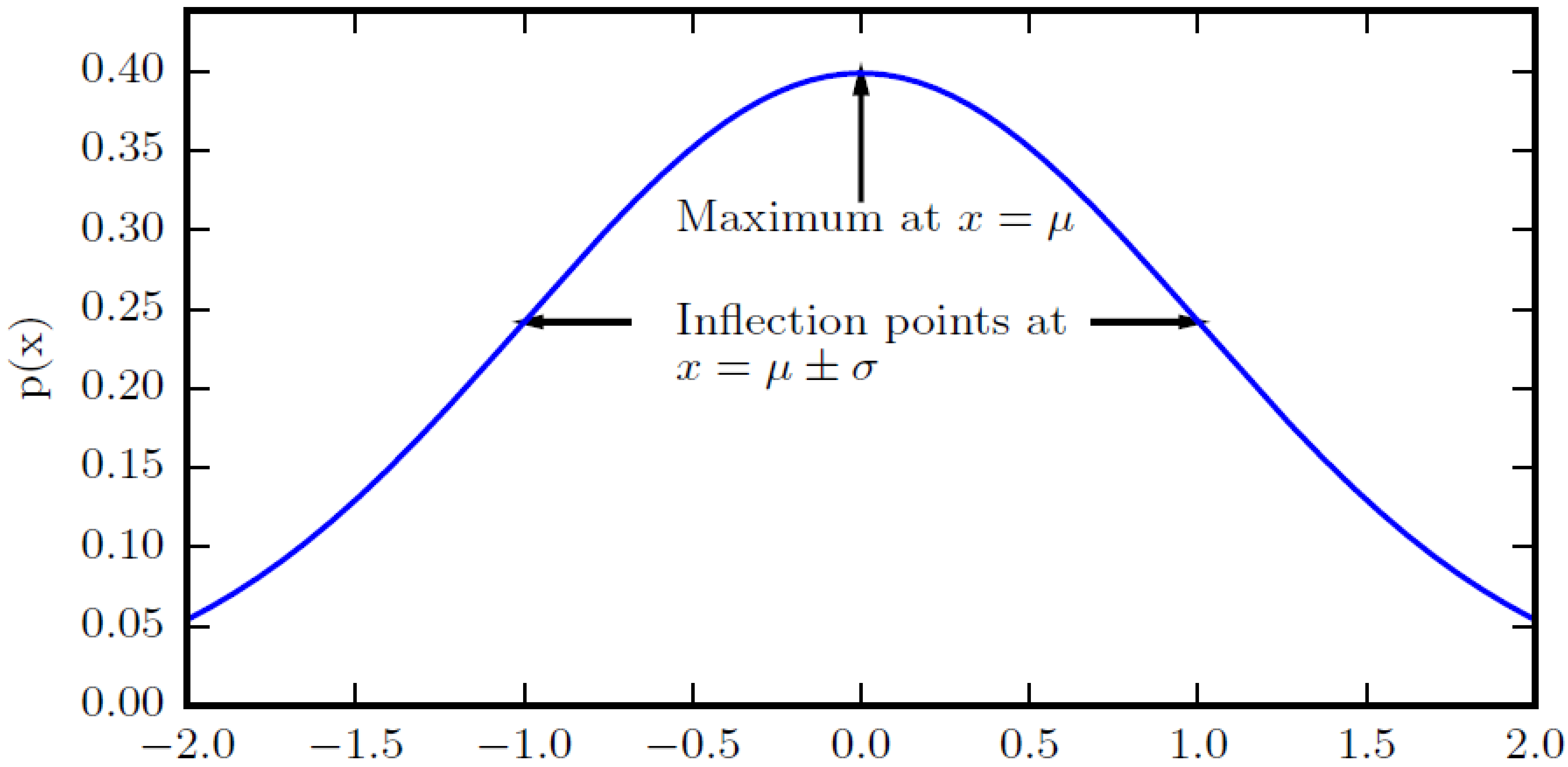
- Parameterized by variance:
  - $\mathrm{E[x]} = \mu,\ \mathrm{Var[x]} = \sigma^2$

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

- Parameterized by precision:

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

Gaussian Distribution



Gaussian Distribution

- Central limit theorem: the sum of many independent random variables is approximately normally distributed
  - $\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \rightarrow N(0,1)$ as $n \rightarrow \infty$

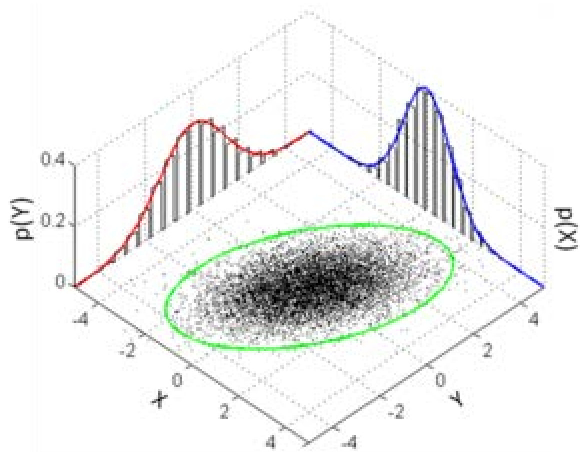
- Law of large numbers: the sample average converges to the expectation as the sample size goes to infinity
  - $\overline{X_n} \rightarrow \mu$ as $n \rightarrow \infty$, where $\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

Multivariate Gaussian

- Parameterized by covariance matrix:

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

- μ is a vector
- Σ is a covariance matrix



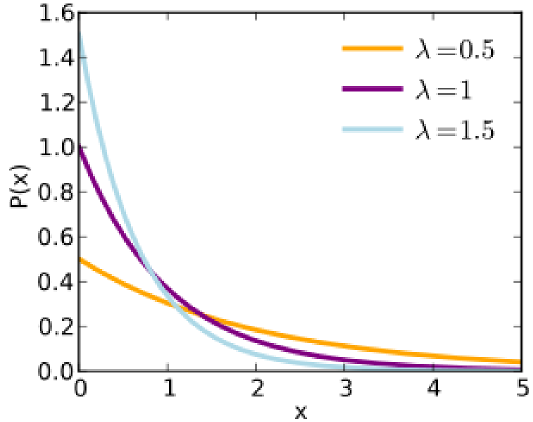
Multivariate Gaussian

- Parameterized by precision matrix:

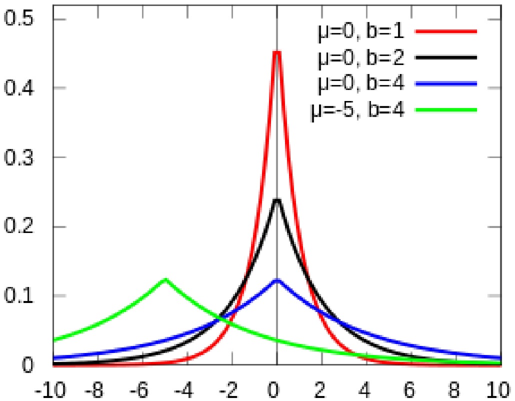
$$\mathrm{N}(x;\mu,\beta^{-1}) = \sqrt{\frac{\det(\beta)}{(2\pi)^n}}\exp(-\frac{1}{2}(x-\mu)^T\beta(x-\mu))$$

More Distributions

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



- Laplace: $p(x;\ \mu, b) = \ \frac{1}{2b}\ \exp(-\frac{|x-\mu|}{b})$



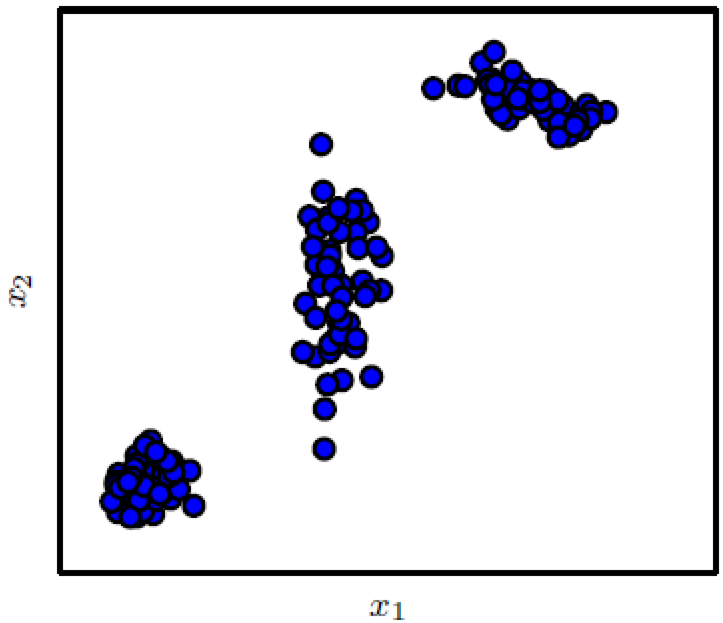
More Distributions

- Dirac Delta: $\mathrm{p(x)} = \delta(x - \mu)$
  - It is zero-valued everywhere except at μ, yet integrates to 1

- Empirical Distribution
  - $\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m} \delta(x - x^{(i)})$

Mixture Distribution

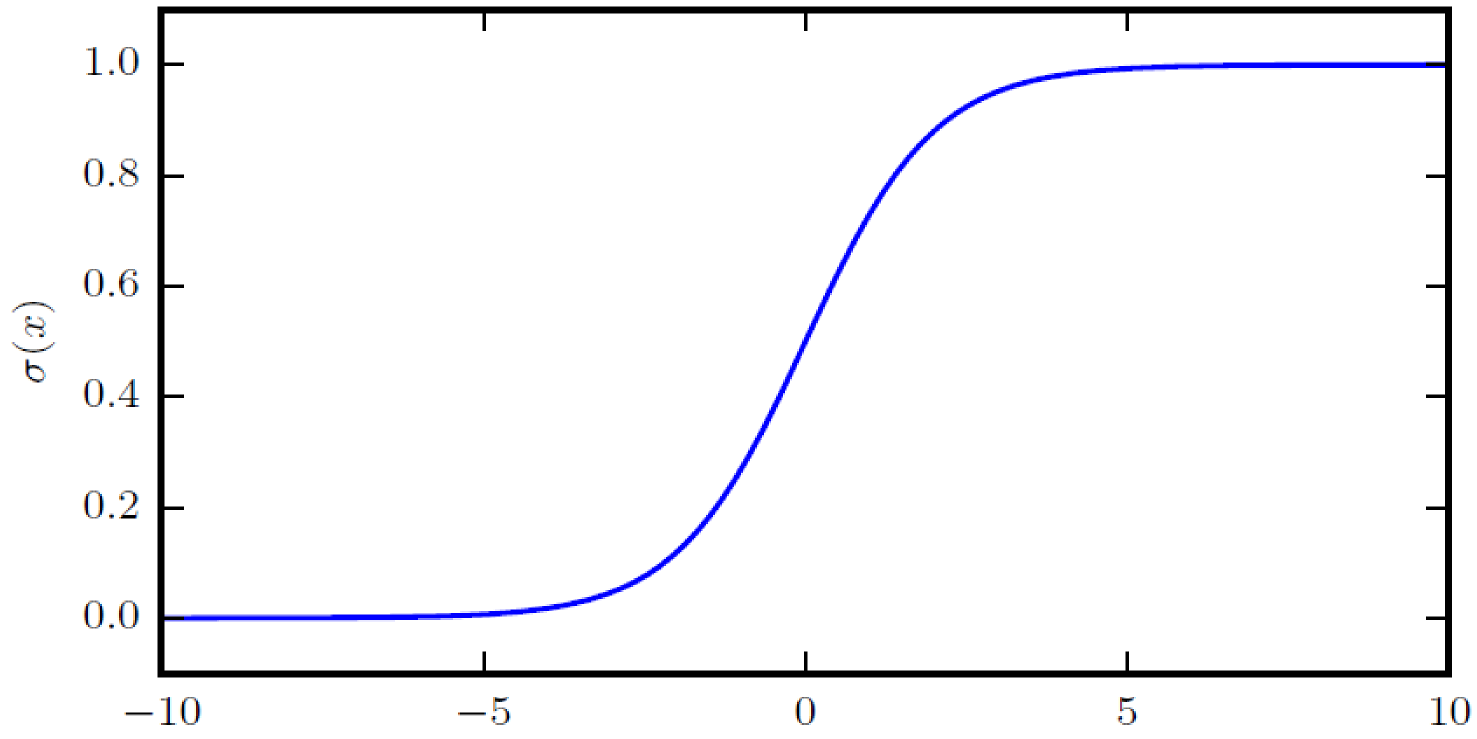
- $P(x) = \sum_i P(c = i)P(x \mid c = i)$
- Gaussian mixture: $P(x \mid c = i)$ is Gaussian



Gaussian mixture with three components

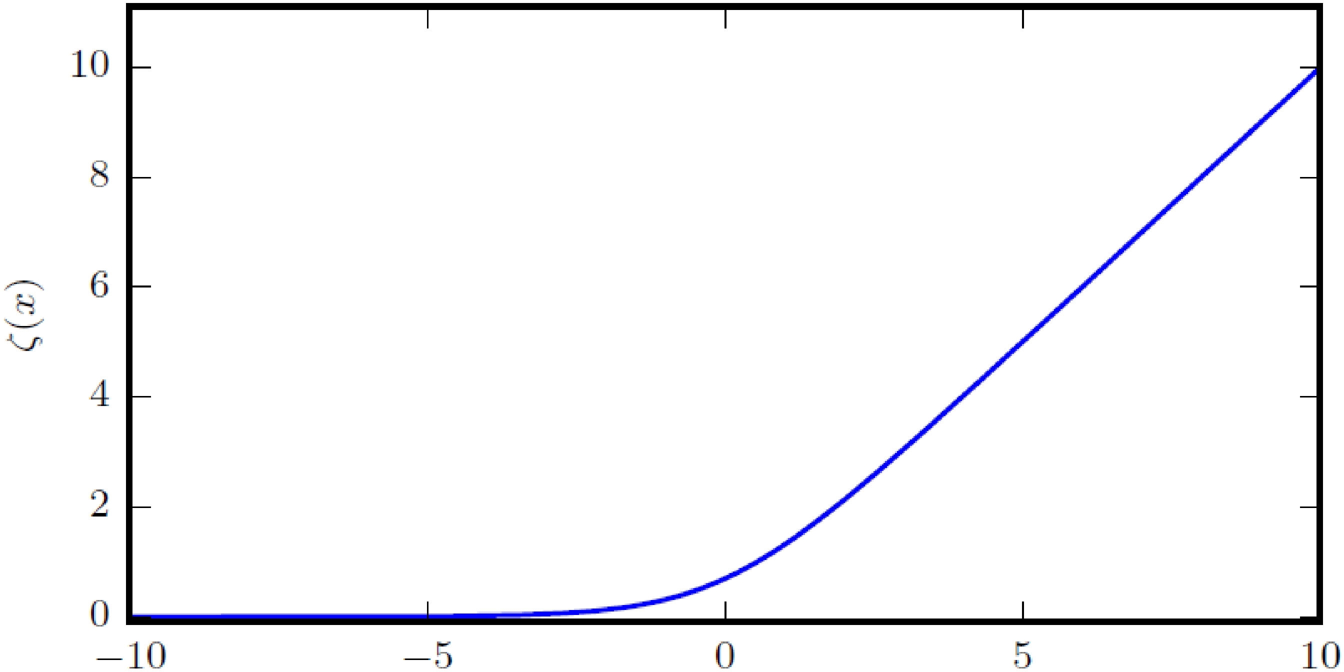
Logistic Sigmoid

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



Softplus Function

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



- "softened" version of $x^+ = \max(0, x)$

"rectified linear unit"

Properties of sigmoid and softplus

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

Bayes Rule

- $P(x \mid y) = \dfrac{P(x)\, P(y \mid x)}{P(y)} = \dfrac{P(x,y)}{\sum_y P(x,y)}$

Change of Variables

- Assume two r.v. x and y such that y = g(x) where g is an invertible, continuous, and differentiable function
- $p_y(y) = p_x(g^{-1}(y))$?
- Example: y = x/2, and x ~ U(0,1)
  - If we use the rule $p_y(y) = p_x(2y)$, p_y will be 0 everywhere except in [0,1/2] where it has 1
  - It means $\int p_y(y)\, dy = 1/2$!

Change of Variables

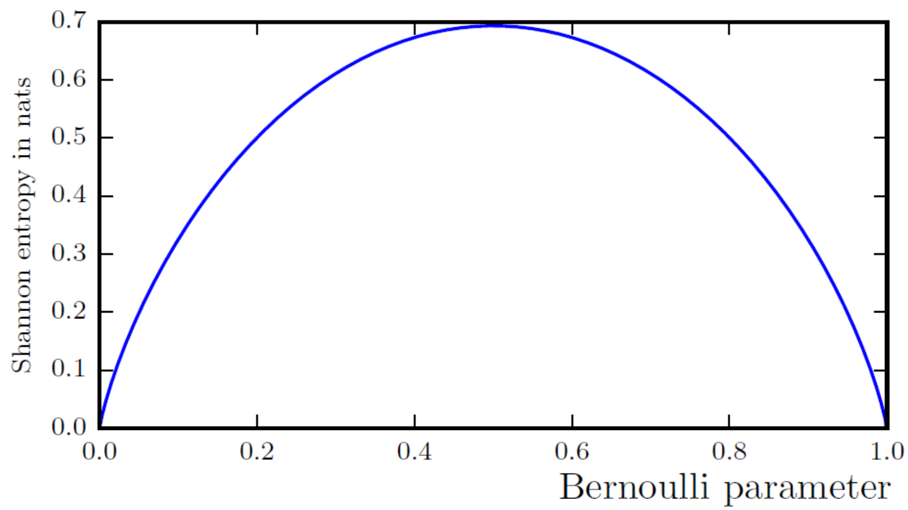
- Assume two r.v. x and y such that y = g(x) where g is an invertible, continuous, and differentiable function
- $p_y(y) = p_x(g^{-1}(y))\frac{dx}{dy}$
  - (pf) $p_y(y)dy = p_x(x)dx$
- Example: y = x/2, and x ~ U(0,1)
  - $p_y(y) = p_x(2y)2 = 2$ (for 0<y<1/2)

Information Theory

- Information theory: quantifying how much information is present in a signal
- Learning that an unlikely event has occurred is more informative than learning that a likely event has occurred
- Self-Information of x
  - $I(x) = -\log P(x)$
  - Intuition: minimum # of bits to express (encode) an event with probability P(x)
  - Rare event has a large information content

Information Theory

- Entropy: expectation of self-information
  - $H(x) = E_{x\sim P}[I(x)] = -E_{x\sim P}[\log P(x)]$
  - Minimum expected # of bits to express a distribution
  - For Bernoulli variable,
    - $H(x) = -p \log p - (1-p)\log(1-p)$



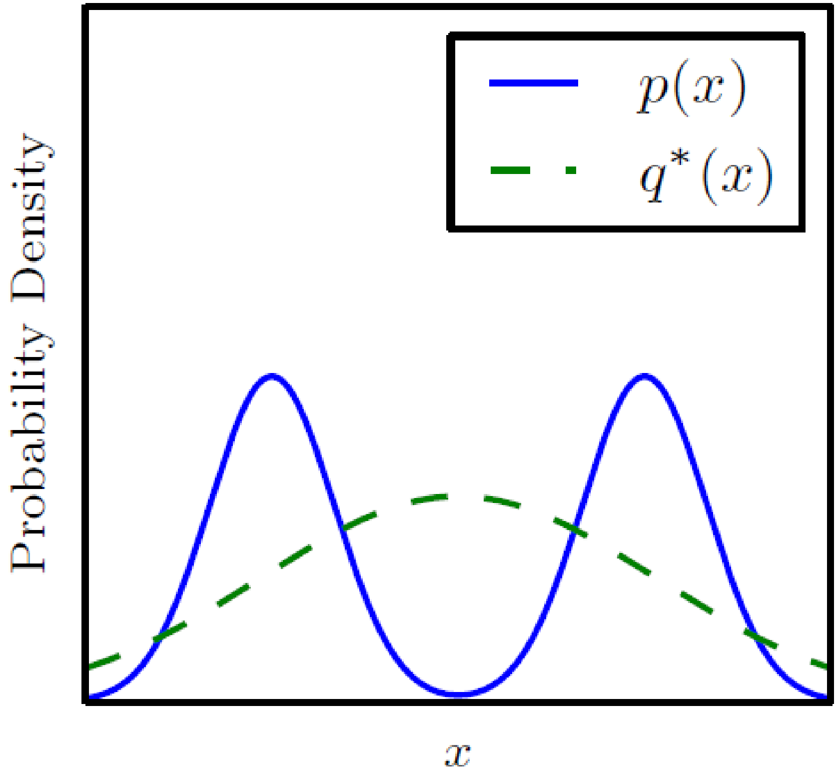
0 log 0 is treated as 0

KL Divergence

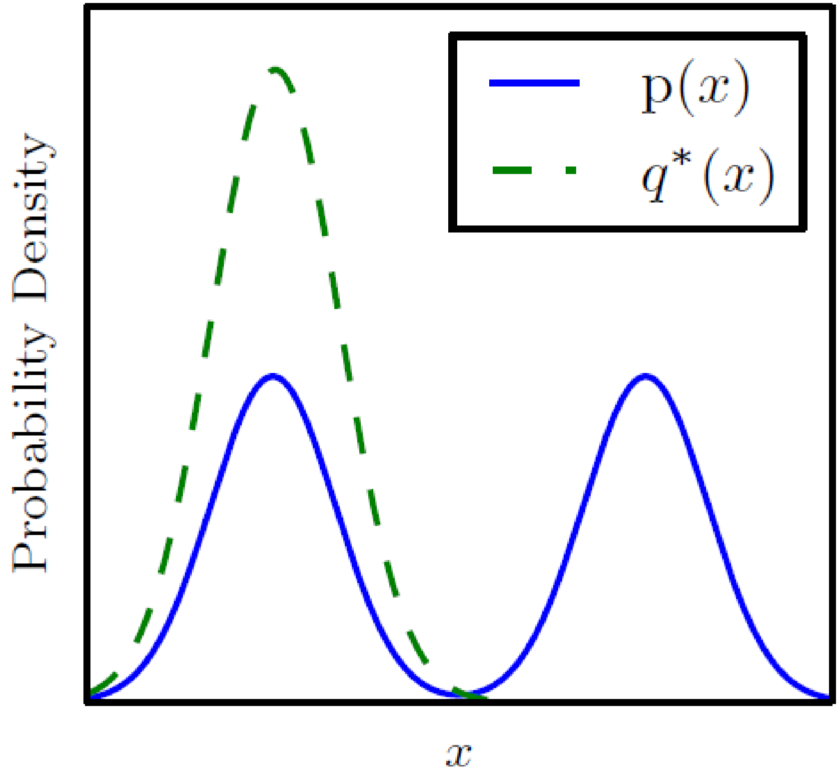
- Measure the difference of two distributions P(x) and Q(x)
- $$D_{KL}(P||Q) = E_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right]$$
  $$= E_{x\sim P}[\log P(x) - \log Q(x)]$$
- Properties
  - Always nonnegative: 0 if and only if P and Q are the same
    - Intuition: If x ~ P, the best (minimal) encoding is given by assigning log P(x) bits for each x
  - Not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

KL Divergence is Asymmetric

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



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



Cross-entropy

- Average # of bits needed to identify an event from the true distribution P, if we use a coding scheme optimized for unnatural distribution Q
- $H(P,Q) = H(P) + \ D_{KL}(P||Q) = -E_{x\sim P} \log Q(x)$
- Minimizing the cross-entropy w.r.t. Q is equivalent to minimizing the KL divergence

What you need to know

- Probability theory concepts
  - PDF and PMF
  - Conditional probability and chain rule
  - Distribution: Bernoulli, Gaussian, …
  - Sigmoid and softplus functions
  - Bayes rule

- Information theory concepts
  - Entropy, KL divergence, and cross-entropy

Questions?