Maximum Likelihood

MLAI Lecture 3

Neil D. Lawrence

Department of Computer Science
Sheffield University

27th September 2012

Outline

Maximum Likelihood

Entropy

- Last time we computed $-\langle \log P(x) \rangle_{P(x)}$.
- This special expectation is known as the entropy of a distribution.
- It is a measure of how much "uncertainty" is in a distribution (learn it!).

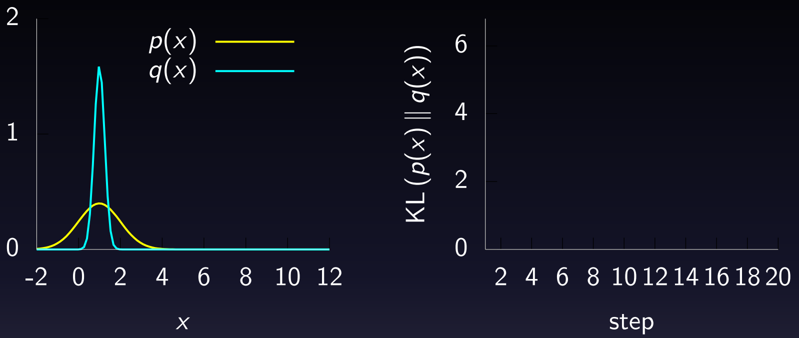
$$\mathcal{H}(x) = -\sum_{x} P(x) \log P(x)$$

Kullback Leibler Divergence

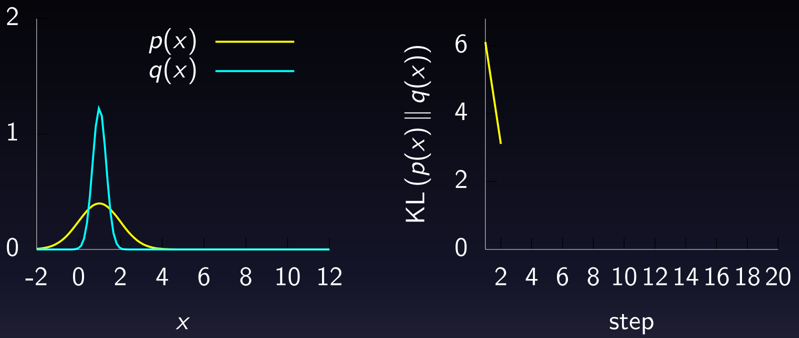
- The Kullback Leibler divergence is another special expectation (learn it!).

$$\mathrm{KL}\left(P(x)\,\|\,Q(x)\right) = \left\langle \log \frac{P(x)}{Q(x)} \right\rangle_{P(x)} = \langle \log P(x) \rangle_{P(x)} - \langle \log Q(x) \rangle_{P(}$$

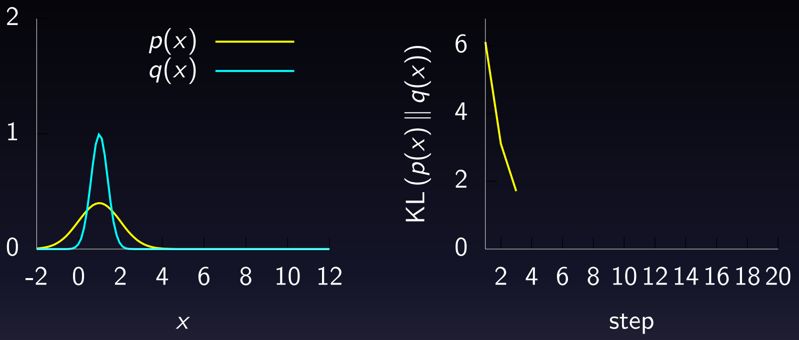
- It is a measure of divergence between two distributions $Q(x)$ and $P(x)$.
- It is zero if they are identical (this is obviously true).
- It is positive if they are different (this is less obvious).



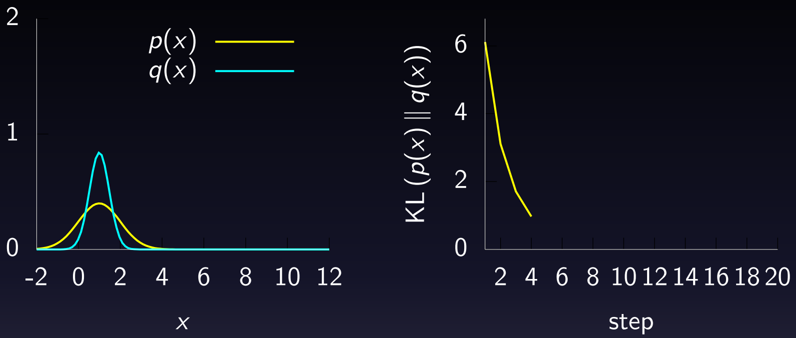
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



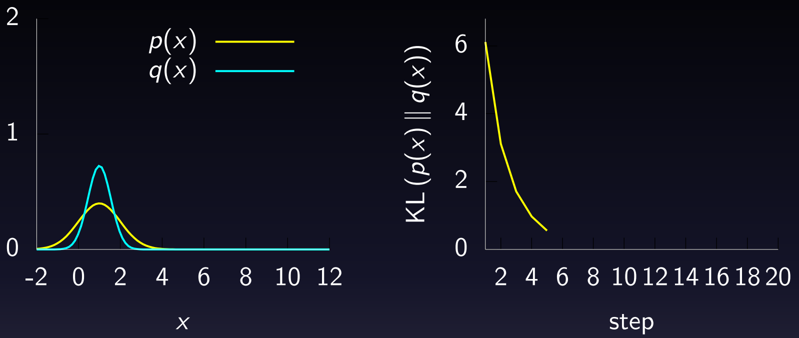
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



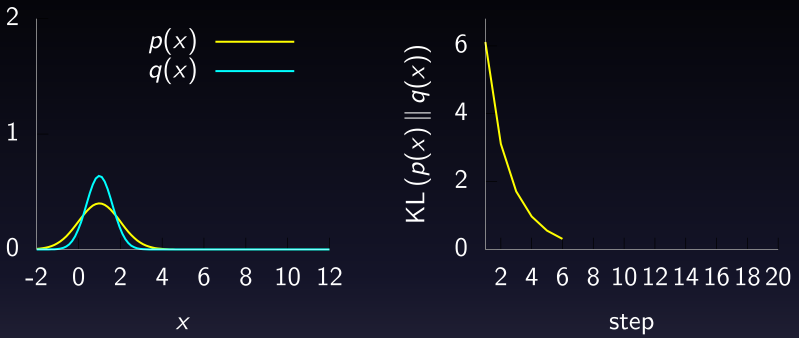
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



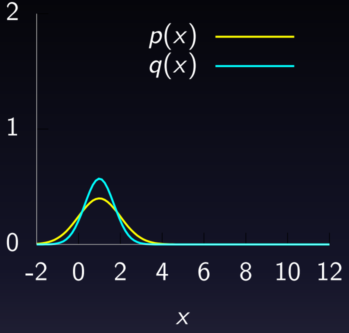
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.

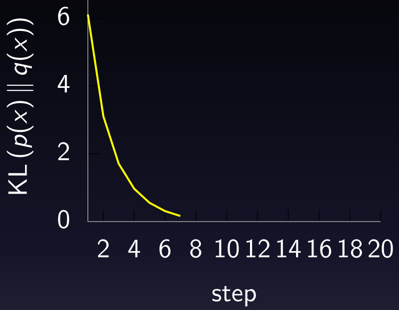


As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.

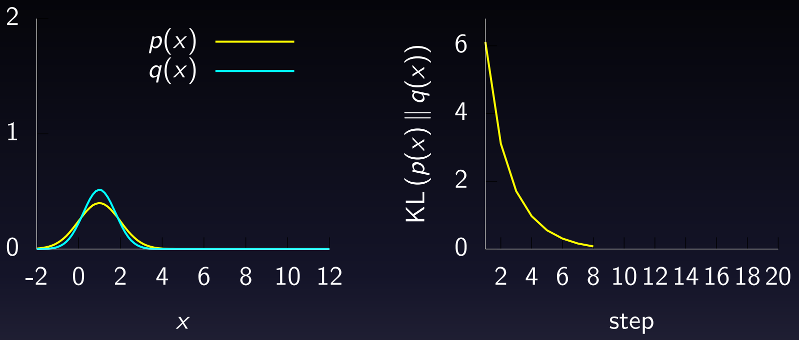


As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

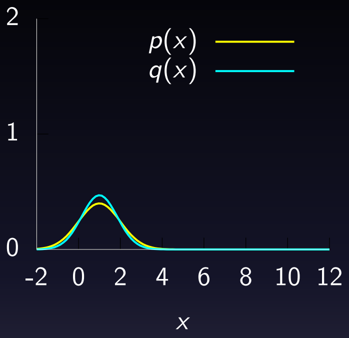


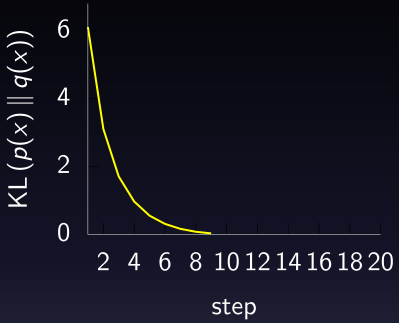


As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

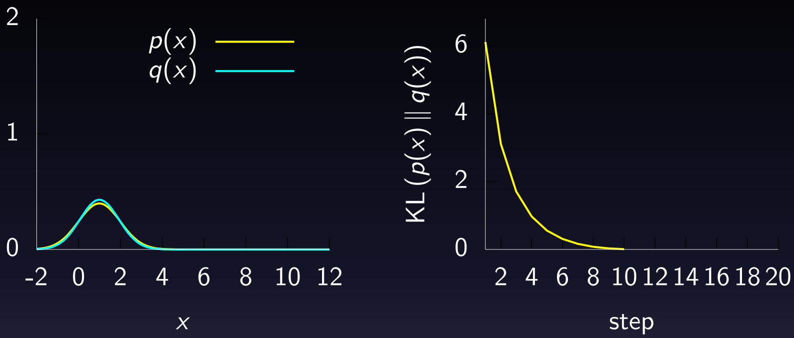


As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.

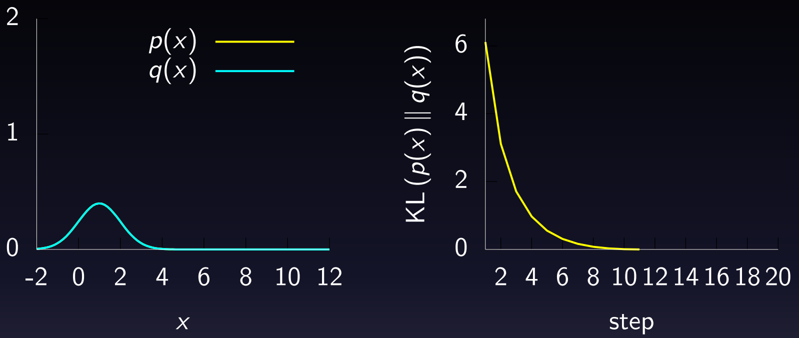




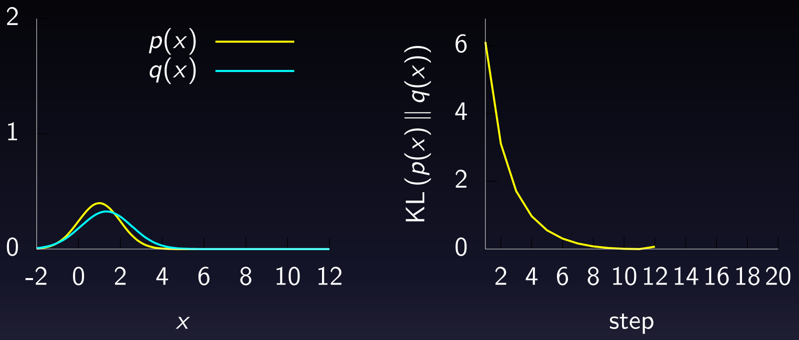
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



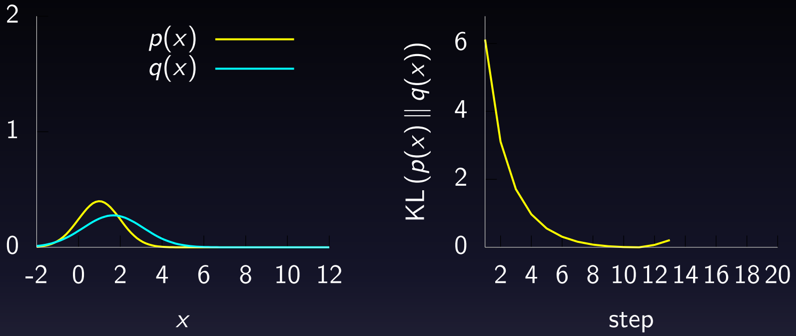
As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



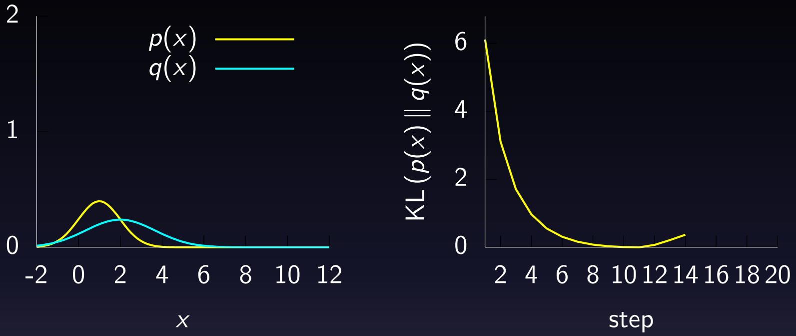
As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



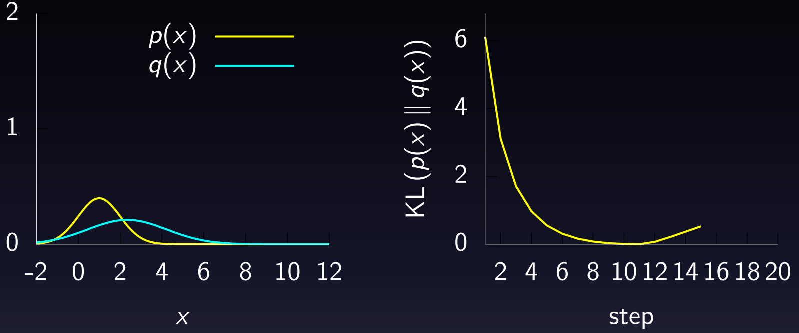
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



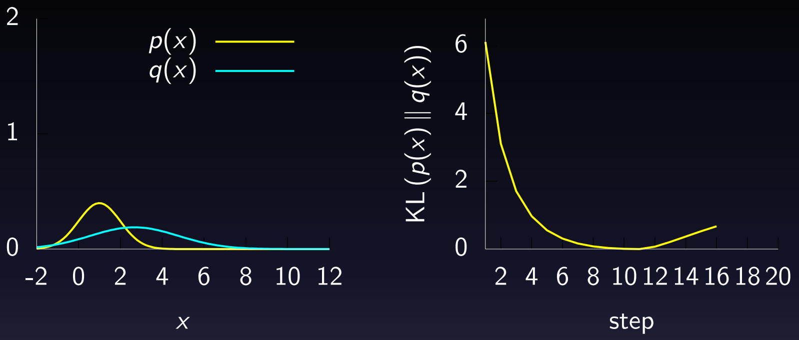
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



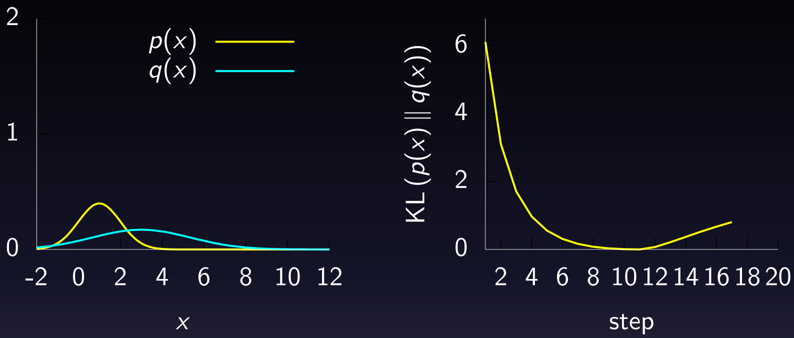
As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



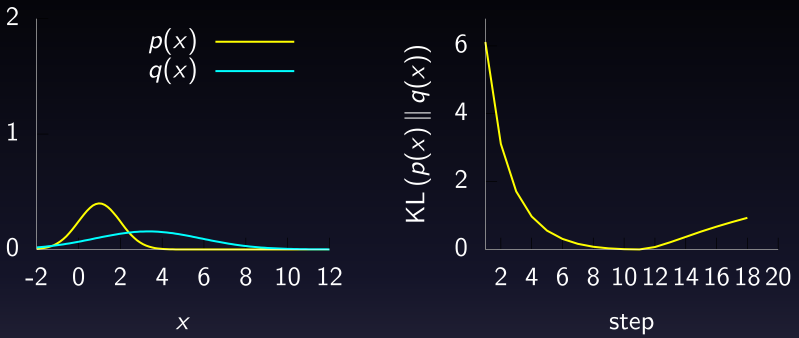
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



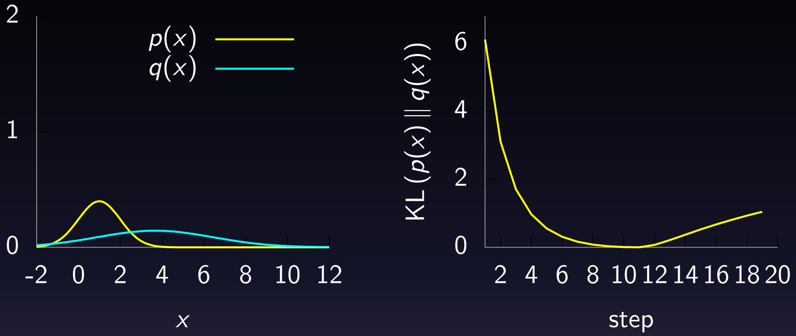
As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



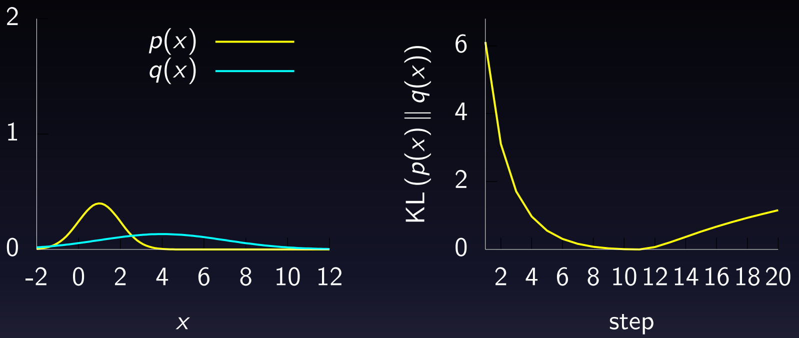
As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



As the cyan Gaussian density $(q(x))$ approaches the yellow Gaussian density $(p(x))$ the KL divergence approaches zero. As they move apart, KL divergence increases again.



As the cyan Gaussian density ($q(x)$) approaches the yellow Gaussian density ($p(x)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

Matching Two Distributions

- To match two distributions $P(x)$ and $Q(x)$ we can *minimize* the KL divergence.
- If we know the form of $Q(x)$ (our approximation) and it has parameters like a and b for the Gamma or mean and variance for Gaussian, we can change these parameters to find the best fit of $Q(x)$ to $P(x)$.
- If we have only got *samples* from $P(x)$ we use a sample based approximation.

Sample Based Approximation to the KL

$$\mathrm{KL}\left(P(x) \,\|\, Q(x)\right) \approx \frac{1}{N}\sum_{i=1}^{N} \log P(x_i) - \frac{1}{N}\sum_{i=1}^{N} \log Q(x_i)$$

- *Can't* compute the first term, but it *doesn't* depend on $Q(x)$ anyway.
- *Can* compute the second term. It is known as the negative log likelihood.

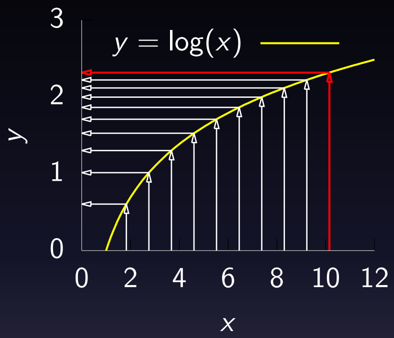
Maximum Likelihood

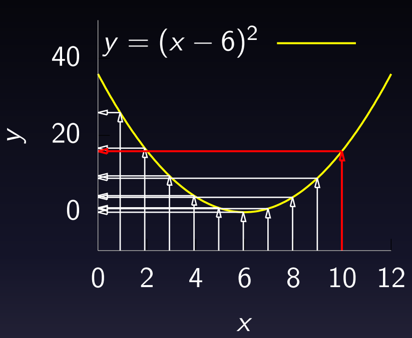
- Minimizing sample based KL divergence is equivalent to maximum likelihood (ML).
- The likelihood is defined as

$$\P(\mathbf{x}|\boldsymbol{\theta})$$

  where $\mathbf{x}$ is a vector containing the data and $\boldsymbol{\theta}$ is a vector of parameters. i.e. this is the probability of the data given the parameters.
- Maximizing log likelihood is equivalent to maximizing likelihood because log is a *monotonic* function.

Monotonicity and Ordering





Monotonic functions preserve the ordering of input points, so the largest x is also the largest y. *Left*: gives an impression of this idea, redarrow is largest in x and correspondingly the largest in y. This transformation is log. *Right*: this quadratic function doesn't preserve the ordering and the largest x (again redarrow) is not the largest y value.

Sample Based Approximation implies i.i.d

- The log likelihood is

$$L(\boldsymbol{\theta}) = \log P(\mathbf{x}|\boldsymbol{\theta})$$

- If the likelihood is independent over the individual data points,

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(x_i|\boldsymbol{\theta})$$

- This is equivalent to the assumption that the data is *independent* and *identically* distributed. This is known as *i.i.d.*.
- Now the log likelihood is

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta})$$

  which matches the sample based KL approximation up to a scaling by $-N$.

Maximum Likelihood Properties

Properties of ML arise due to the relationship with the KL divergence, and law of large numbers.

- As $N \to \infty$ *If* class of distributions considered for $Q(x)$ contains $P(x)$ then we will obtain $Q(x) = P(x)$.
- This is known as the consistency of maximum likelihood.
- In practice
  - We won't have infinite data.
  - We cannot prove that $Q(x)$ will include $P(x)$.

Maximum Likelihood, Minimum Error

- To maximize likelihood we use optimization techniques.
- In the optimization community *minimization* is the convention.
- Define the “error function” to be negative log likelihood.

$$E(\boldsymbol{\theta}) = -\log L(\boldsymbol{\theta})$$

- $E(\cdot)$ can also be thought of as an *energy* function. This is a physics interpretation.

Basic Optimization Overview

- To find a minimum, want to find a point where gradient is zero (this is a stationary point).
- If we can show that curvature is positive, this is a minimum.
- Procedure: differentiate the function, find parameters which set derivative to zero.
- This can sometimes be done by a fixed point equation, other times iterative optimization methods are required.

Example: Maximum Likelihood in the Gaussian

$$P(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- Write down error function.
- Differentiate error function.
- Solve such that the derivatives are zero.

Example: Maximum Likelihood in the Gaussian

$$P(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- Write down error function.
- Differentiate error function.
- Solve such that the derivatives are zero.

Example: Maximum Likelihood in the Gaussian

$$P(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- Write down error function.
- Differentiate error function.
- Solve such that the derivatives are zero.

Reading

- Bishop rest of Section 1.2.4, page 26–28 (don't worry about material on bias).

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .