Max Likelihood

independent

$$L(\underbrace{x_1, x_2 \ldots x_n}_{\text{fixed}}; \underset{\substack{\uparrow \\ \text{variate}}}{\theta}) = q(x_1;\theta) \cdot q(x_2;\theta) \cdots q(x_n;\theta)$$

in contrast to PDF / CDF we treat θ as main variate and assume x_i to be known, (given)

Total probability to obtain samples $\{x_i\}$ under some model q with params θ

$$\log L(x_1 \ldots x_n; \theta) = \sum_{i=1}^{n} \log q(x_i;\theta) \quad \leftarrow \text{log-likelihood}$$

Maximised (log-) likelihood shows which θ fits data best

$$\hat{L} = \max_{\theta} L(x_1 \ldots x_n; \theta) \quad \text{OR} \quad \hat{L} = \max_{\theta} \log L(x_1 \ldots x_n; \theta)$$

is a number, can be used alone to evaluate the model

$$\hat{\theta} = \arg\max_{\theta} L(x_1 \ldots; \theta) \quad \leftarrow \text{max. log-like. estimate}$$

Max Likelihood & min D_{KL}

To find maximised (log-) likelihood

- 1st order condition:

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log q(x_i;\theta) = 0 \quad \Rightarrow \hat{\theta}$$

- 2nd order condition

Hessian has to be negative semi-definite at point $\theta = \hat{\theta}$

$$\begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \log L(x;\theta) & \frac{\partial^2}{\partial \theta_1 \partial \theta_m} \log L(x;\theta) \\ \frac{\partial^2}{\partial \theta_m \partial \theta_1} \log L(x;\theta) & \frac{\partial^2}{\partial \theta_m^2} \log L(x;\theta) \end{bmatrix}$$

Minimisation of D_{KL} from a model q to the true (unknown) distr p, from which we draw samples:

$$\min D_{KL}(p \| q) = \min_\theta \left[- \int p(y) \log \frac{q(y;\theta)}{p(y;\theta)} dy \right]$$

$$= \max_\theta \int p(y) \log \frac{p(y)}{q(y;\theta)} dy$$

1st $\Rightarrow$ $\frac{\partial}{\partial \theta} \int p(y) \log p(y) - p(y) \log q(y;\theta) = 0$

$$\int p(y) \frac{\partial}{\partial \theta} \log q(y;\theta) = 0$$

$$E_p\left[\frac{\partial}{\partial \theta} \log q(y;\theta)\right] = 0$$

which can be estimated with samples

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log q(x_i;\theta) = 0 \quad \Rightarrow \hat{\theta}$$

$\boxed{\hat{\theta} \text{ that maximises } L(x,\theta) \text{ also brings min to } D_{KL}(p \| q)}$

Akaike's information criterion

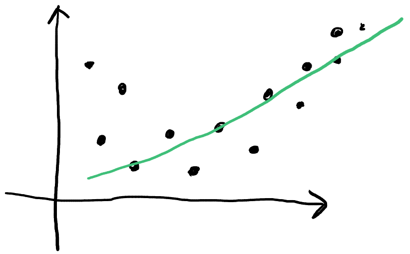
$$AIC = -2\cdot\hat{L} + \underbrace{2d}_{\text{penalty}}$$

Bayesian inf. criterion

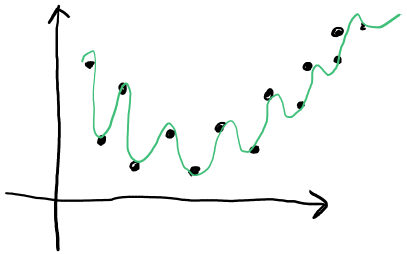
$$BIC = -2\cdot\hat{L} + \underbrace{d \ln n}_{\text{penalty}}$$

d is number of params in model (θ)

n is number of samples



poor fit $\sim \hat{L}$



too many params $\sim$ penalty

- The lower AIC\BIC the better!
- To use AIC\BIC to compare different models (like, linear vs. polynomial) write down $L(x;\theta)$ and do maximisation with resp. to θ $\Rightarrow \hat{L}$

this is usually done numerically with any suitable method.

HW 2. Problem 1

1. Cumulant gen. function $K(k) = \ln \langle e^{ikx} \rangle_x$

$$= \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} \kappa_n$$

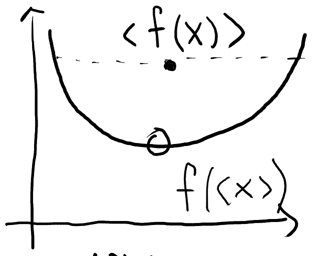
cumulants

$\kappa_1 = \mu_1$

$\kappa_2 = \sigma^2$

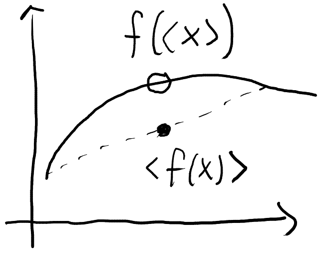
....

2. Jensen's inequality :



convex

$$\langle f(x) \rangle \geq f(\langle x \rangle)$$



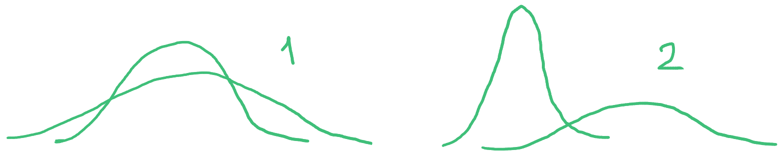
concave

$$\langle f(x) \rangle \leq f(\langle x \rangle)$$

3. $D_{KL}(p \,//\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$, replace with $\int$ for continuous

„from q to p"

$$= -\sum_x p(x) \log \frac{q(x)}{p(x)}$$



$D_{KL_1} < D_{KL_2}$

HW2. Problem 1

Some properties of D_{KL}

1. non-negative

2. $D_{KL}(p \,//\, q) = 0 \iff p = q$ almost everywhere

3. assymetric

   $D_{KL}(p \,//\, q) \neq D_{KL}(q \,//\, p)$

   (that's why D_{KL} is not a distance)

4. $D_{KL}(p \,//\, q) =$ cross-entropy(p, q) − entropy(p)

5. additive for independent p & q prob. distr.

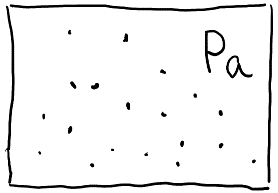
   $D_{KL}(p_x \cdot p_y \,//\, q_x \cdot q_y) = D_{KL}(p_x \,//\, q_x) + D_{KL}(p_y \,//\, q_y)$

example of distance: L_2 norm

$$\|f - g\|_{L_2} = \sqrt{\int (f(x) - g(x))^2 dx}$$

$$\|f - g\|_{L_2} = \|g - f\|_{L_2}$$

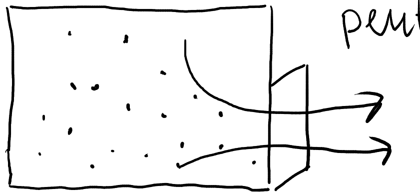
HW2. Problem 1

1. 

room

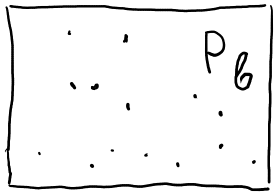
ω – velocities of molecules

r – could be a temperature

2. 

perturbation happens

somebody opens a door

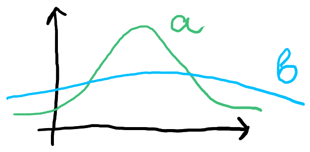
3. 

Hints:

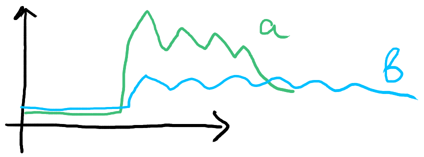
- perturbation process isn't important for this task,

  we just have some P_a & P_b — in this problem they remain unknown!

- „P_a and P_b have the same support" means you can safely divide P_a over P_b



OR



- For steps 1, 2, 3 you need just a few lines, no need to simplify them.

- Use result of a previous step to write the next one

- the final relation includes only D_{KL}, moments (or cumulants) and absolute value of responce