Maximum likelihood estimation

Nuno Vasconcelos
UCSD

Bayesian decision theory

- recall that we have
  - Y – state of the world
  - X – observations
  - g(x) – decision function
  - L[g(x),y] – loss of predicting y with g(x)
- Bayes decision rule is the rule that minimizes the risk

$$Risk = E_{X,Y}\left[L(X,Y)\right]$$

- for the "0-1" loss

$$L[g(x),y] = \begin{cases} 1, & g(x) \neq y \\ 0, & g(x) = y \end{cases}$$

- optimal decision rule is the maximum a-posteriori probability rule

MAP rule

- we have shown that it can be implemented in any of the three following ways
  - 1) $$i^*(x) = \arg\max_i P_{Y|X}(i \mid x)$$
  - 2) $$i^*(x) = \arg\max_i \left[P_{X|Y}(x \mid i) P_Y(i) \right]$$
  - 3) $$i^*(x) = \arg\max_i \left[\log P_{X|Y}(x \mid i) + \log P_Y(i) \right]$$
- by introducing a "model" for the class-conditional distributions we can express this as a simple equation
  - e.g. for the multivariate Gaussian

$$P_{X|Y}(x \mid i) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_i \mid}} \exp\left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}$$

The Gaussian classifier

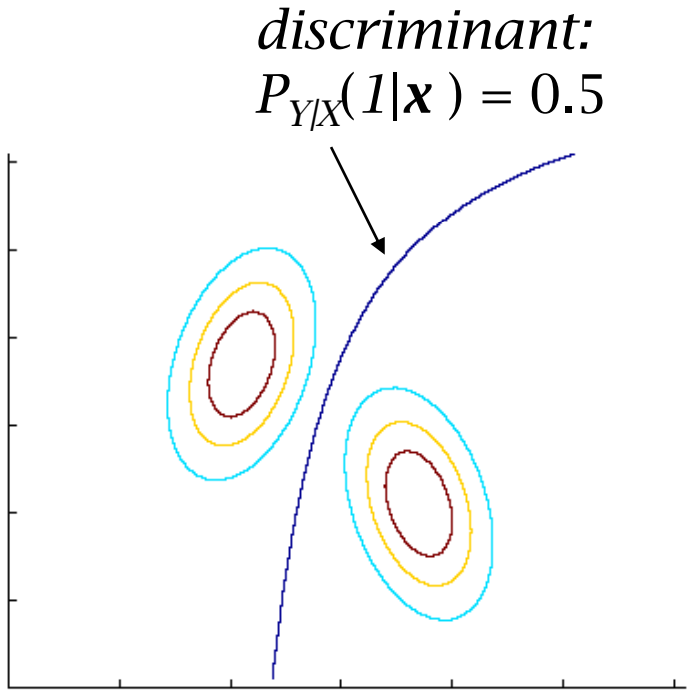
- the solution is

$$i^*(x) = \arg\min_i \left[d_i(x, \mu_i) + \alpha_i \right]$$

with

$$d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y)$$

$$\alpha_i = \log(2\pi)^d |\Sigma_i| - 2 \log P_Y(i)$$



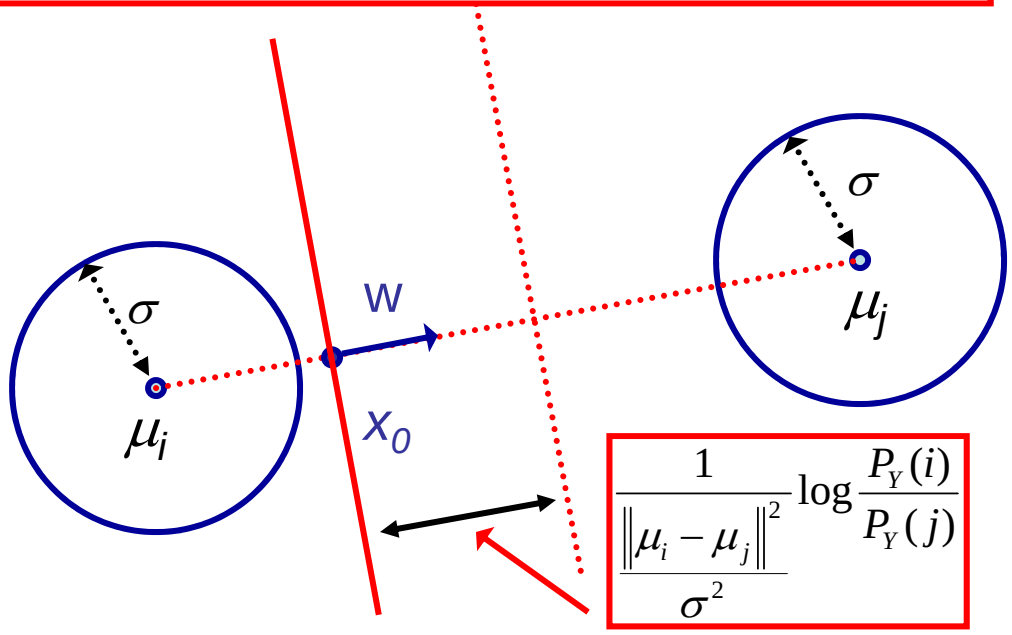
- the optimal rule is to assign x to the closest class
- closest is measured with the Mahalanobis distance $d_i(x,y)$
- can be further simplified in special cases

Geometric interpretation

- for Gaussian classes, equal covariance $\sigma^2 I$

$$w = \frac{\mu_i - \mu_j}{\sigma^2}$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right)$$

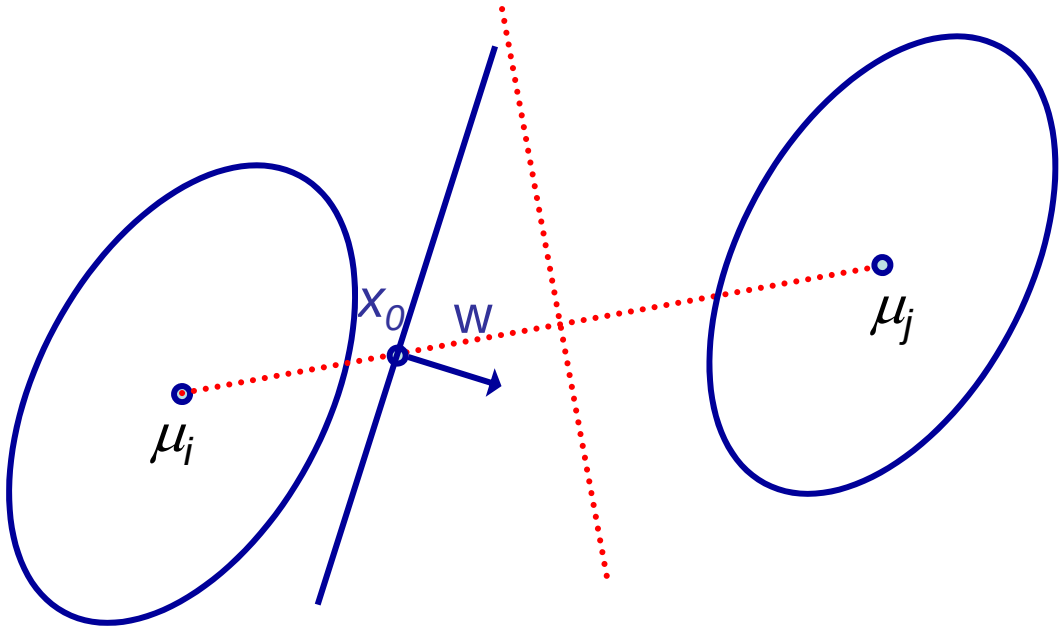


Geometric interpretation

- for Gaussian classes, equal but arbitrary covariance

$$w = \Sigma^{-1}\left(\mu_i - \mu_j\right)$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right)} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right)$$



Bayesian decision theory

- advantages:
  - BDR is optimal and cannot be beaten
  - Bayes keeps you honest
  - models reflect causal interpretation of the problem, this is how we think
  - natural decomposition into "what we knew already" (prior) and "what data tells us" (CCD)
  - no need for heuristics to combine these two sources of info
  - BDR is, almost invariably, intuitive
  - Bayes rule, chain rule, and marginalization enable modularity, and scalability to very complicated models and problems
- problems:
  - BDR is optimal only insofar the models are correct.

Implementation

- we do have an optimal solution

$$w = \Sigma^{-1}\left(\mu_i - \mu_j\right)$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right)} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right)$$

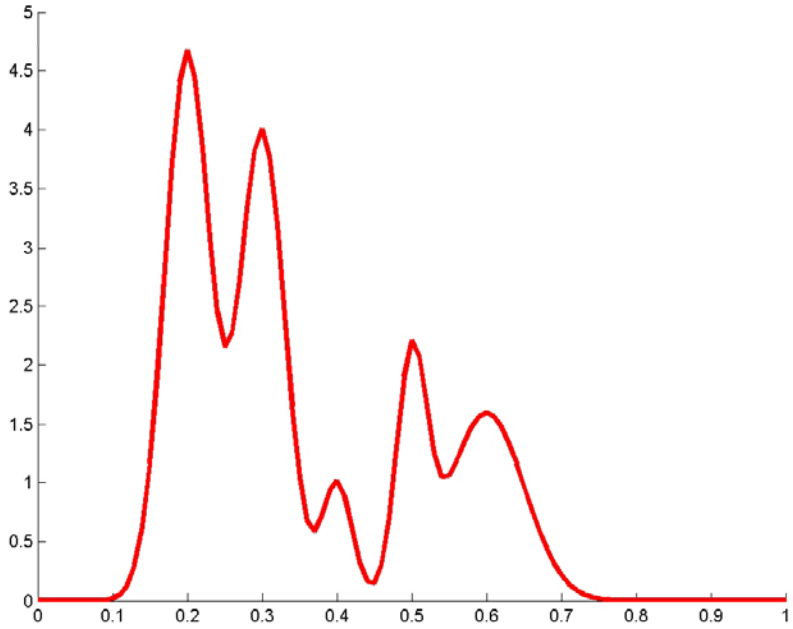
- but in practice we do not know the values of the parameters μ, Σ, $P_Y(1)$
  - we have to somehow estimate these values
  - this is OK, we can come up with an estimate from a training set
  - e.g. use the average value as an estimate for the mean

$$w = \hat{\Sigma}^{-1}\left(\hat{\mu}_i - \hat{\mu}_j\right)$$

$$x_0 = \frac{\hat{\mu}_i + \hat{\mu}_j}{2} - \frac{1}{\left(\hat{\mu}_i - \hat{\mu}_j\right)^T \Sigma^{-1}\left(\hat{\mu}_i - \hat{\mu}_j\right)} \log \frac{\hat{P}_Y(i)}{\hat{P}_Y(j)} \left(\hat{\mu}_i - \hat{\mu}_j\right)$$

Important

- warning: at this point all optimality claims for the BDR cease to be valid!!
- the BDR is guaranteed to achieve the minimum loss when we use the true probabilities
- when we "plug in" the probability estimates, we could be implementing a classifier that is quite distant from the optimal
  - e.g. if the $P_{X|Y}(x|i)$ look like the example above
  - I could never approximate them well by parametric models (e.g. Gaussian)



Maximum likelihood

- this seems pretty serious
  - how should I get these probabilities then?
- we rely on the maximum likelihood (ML) principle
- this has three steps:
  - 1) we choose a parametric model for all probabilities
  - to make this clear we denote the vector of parameters by Θ and the class-conditional distributions by

$$P_{X|Y}(x \mid i; \Theta)$$

  - note that this means that Θ is NOT a random variable (otherwise it would have to show up as subscript)
  - it is simply a parameter, and the probabilities are a function of this parameter

Maximum likelihood

- three steps:
  - 2) we assemble a collection of datasets $\mathcal{D}^{(i)} = \{x_1^{(i)}, ..., x_n^{(i)}\}$ set of examples drawn independently from class i
  - 3) we select the parameters of class i to be the ones that maximize the probability of the data from that class

$$\Theta_i = \arg\max_{\Theta} P_{X|Y}\left(D^{(i)} \mid i;\Theta\right)$$
$$= \arg\max_{\Theta} \log P_{X|Y}\left(D^{(i)} \mid i;\Theta\right)$$

  - like before, it does not really make any difference to maximize probabilities or their logs

Maximum likelihood

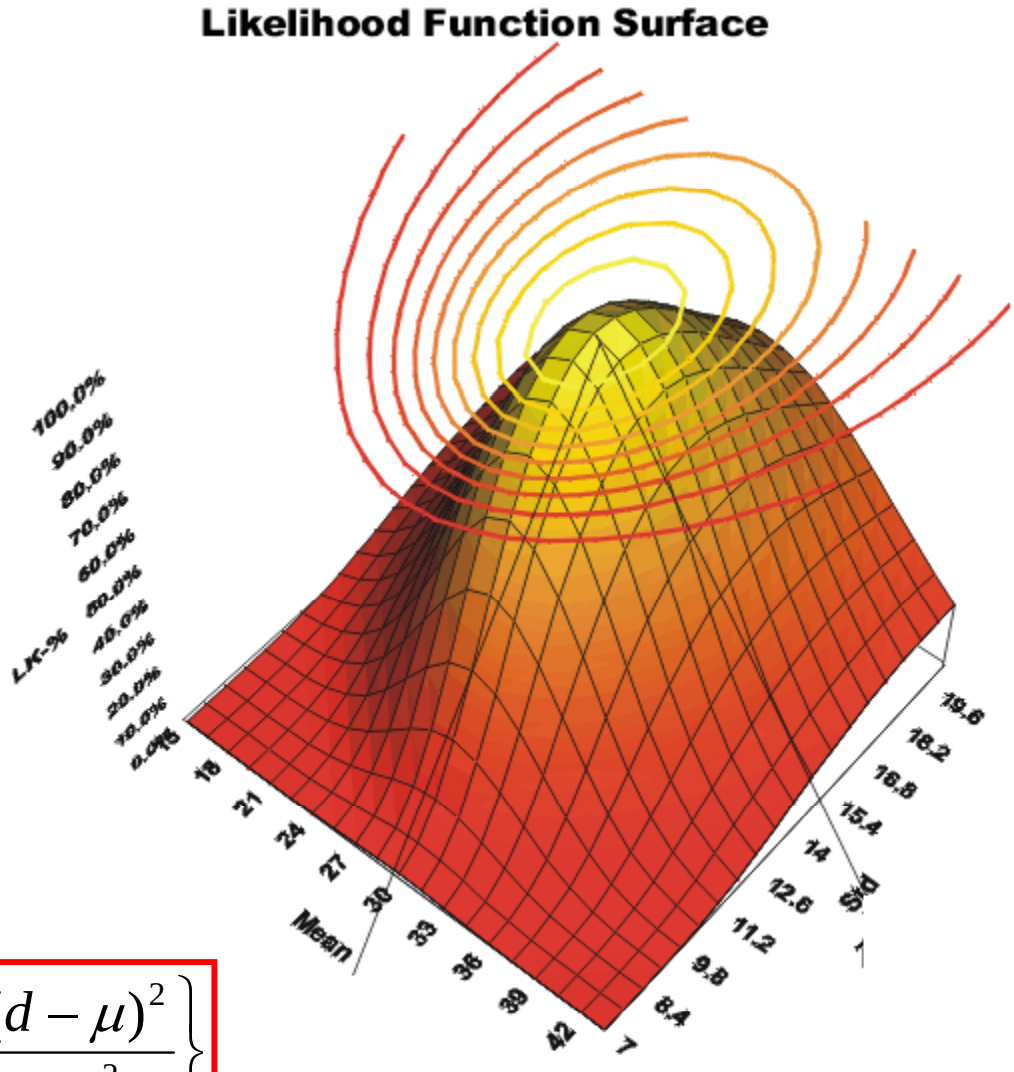
- since
  - each sample $\mathcal{D}^{(i)}$ is considered independently
  - parameter Θ_i estimated only from sample $\mathcal{D}^{(i)}$
- we simply have to repeat the procedure for all classes
- so, from now on we omit the class variable

$$\Theta^* = \arg\max_{\Theta} P_X\left(D;\Theta\right)$$
$$= \arg\max_{\Theta} \log P_X\left(D;\Theta\right)$$

- the function $P_X(\mathcal{D};\Theta)$ is called the likelihood of the parameter Θ with respect to the data
- or simply the likelihood function

Maximum likelihood

- note that the likelihood function is a function of the parameters Θ
- it does not have the same shape as the density itself
- e.g. the likelihood function of a Gaussian is not bell-shaped
- the likelihood is defined only after we have a sample



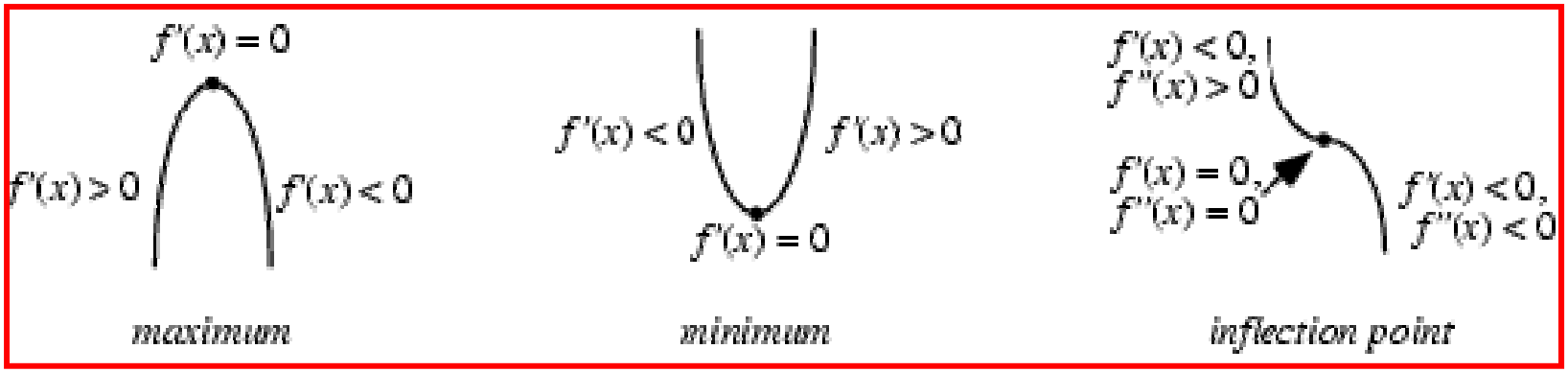
$$P_X(d;\Theta) = \frac{1}{\sqrt{(2\pi)\sigma^2}} \exp\left\{-\frac{(d-\mu)^2}{2\sigma^2}\right\}$$

Maximum likelihood

- given a sample, to obtain ML estimate we need to solve

$$\Theta^* = \arg\max_{\Theta} P_X\left(D;\Theta\right)$$

- when Θ is a scalar this is high-school calculus



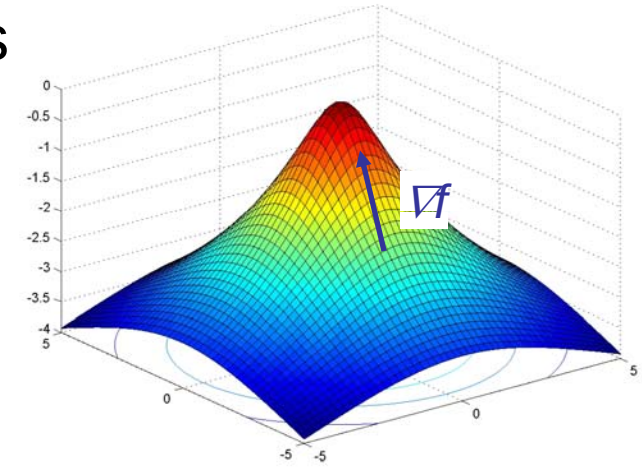
- we have a maximum when
  - first derivative is zero
  - second derivative is negative

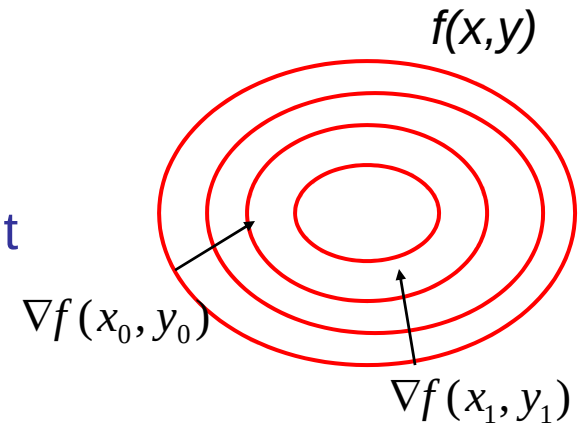
The gradient

- in higher dimensions, the generalization of the derivative is the gradient
- the gradient of a function *f(w)* at *z* is

$$\nabla f(z) = \left(\frac{\partial f}{\partial w_0}(z), \cdots, \frac{\partial f}{\partial w_{n-1}}(z) \right)^T$$

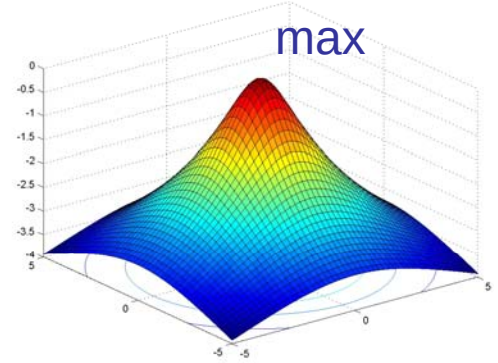
- the gradient has a nice geometric interpretation
  - it points in the direction of maximum growth of the function
  - which makes it perpendicular to the contours where the function is constant

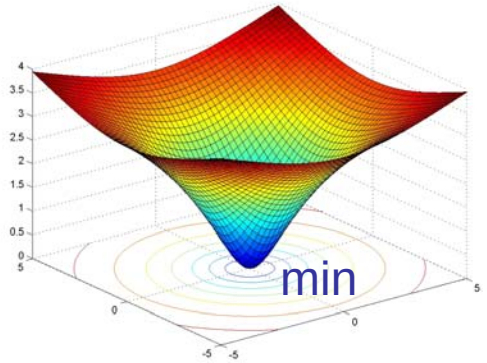


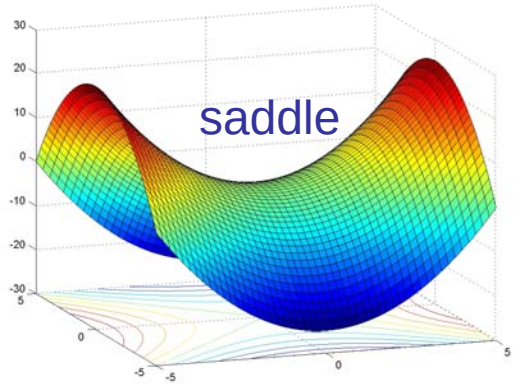


The gradient

- note that if $\nabla f = 0$
  - there is no direction of growth
  - also $-\nabla f = 0$, and there is no direction of decrease
  - we are either at a local minimum or maximum or “saddle” point
- conversely, at local min or max or saddle point
  - no direction of growth or decrease
  - $\nabla f = 0$
- this shows that we have a critical point if and only if $\nabla f = 0$
- to determine which type we need second order conditions







The Hessian

- the extension of the second-order derivative is the Hessian matrix

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(x) \\ \vdots & & \\ \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(x) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(x) \end{bmatrix}$$

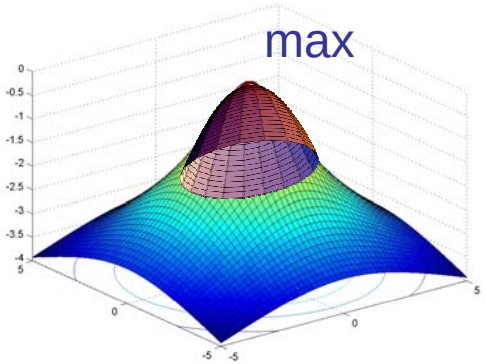
  - at each point x, gives us the quadratic function

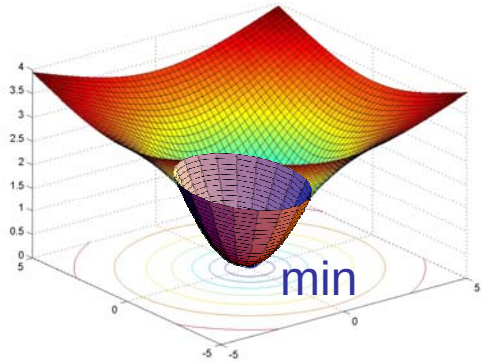
$$x^t \nabla^2 f(x) x$$

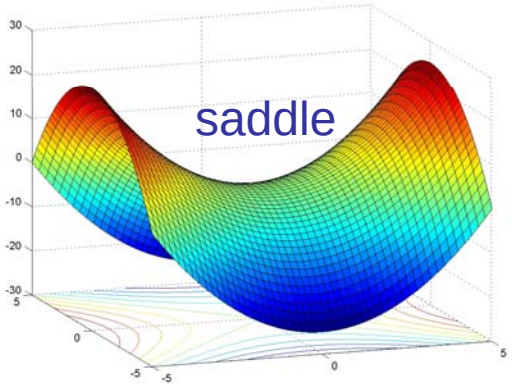
    that best approximates f(x)

The Hessian

- this means that, when gradient is zero at x, we have
  - a maximum when function can be approximated by an "upwards-facing" quadratic
  - a minimum when function can be approximated by a "downwards-facing" quadratic
  - a saddle point otherwise





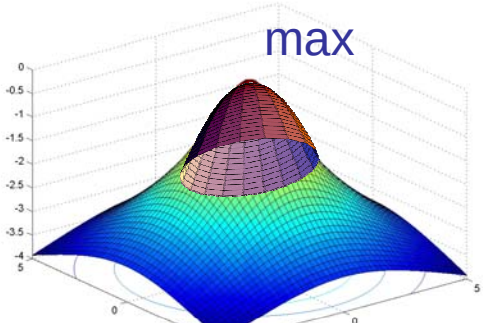


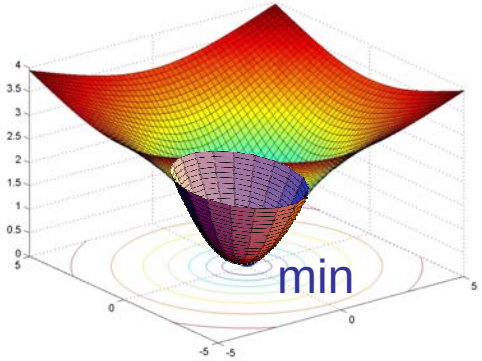
The Hessian

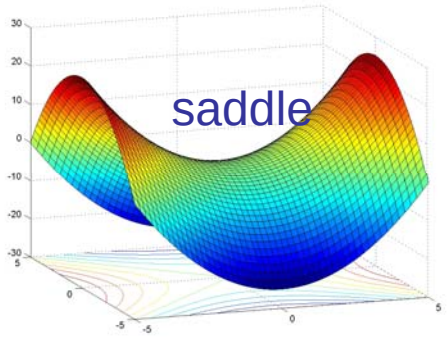
- for any matrix M, the function

$$x^t Mx$$

- is
  - upwards facing quadratic when M is negative definite
  - downwards facing quadratic when M is positive definite
  - saddle otherwise
- hence, all that matters is the positive definiteness of the Hessian
- we have a maximum when Hessian is negative definite







Maximum likelihood

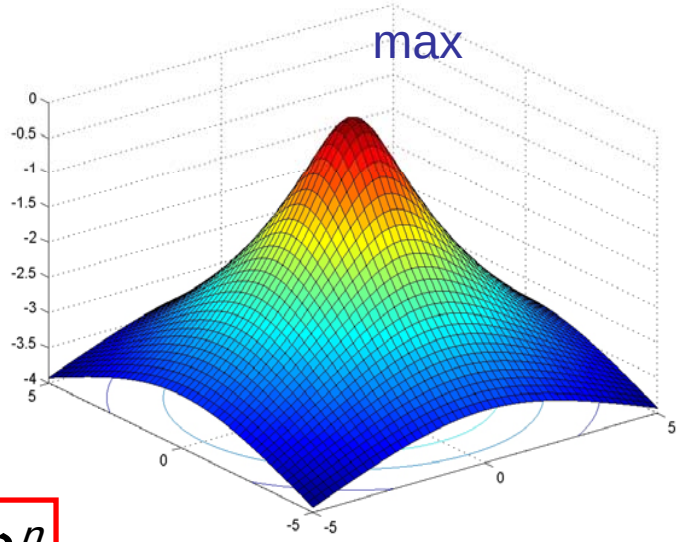
- in summary, given a sample, we need to solve

$$\Theta^* = \arg\max_{\Theta} P_X\left(D;\Theta\right)$$

- the solutions are the parameters such that

$$\nabla_{\Theta} P_X(D;\Theta) = 0$$

$$\theta^t \nabla_{\Theta}{}^2 P_X(D;\theta)\theta \leq 0, \quad \forall \theta \in \Re^n$$



- note that you always have to check the second-order condition!

Maximum likelihood

- let's consider the Gaussian example

$$f(T) = \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{T-\bar{T}}{\sigma_T}\right)^2}$$

- given a sample $\{T_1, \ldots, T_N\}$ of independent points
- the likelihood is

$$L(T_1, T_2, ..., T_N | \bar{T}, \sigma_T) = L = \prod_{i=1}^{N} \left[\frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{T_i-\bar{T}}{\sigma_T}\right)^2} \right]$$

$$L = \frac{1}{(\sigma_T \sqrt{2\pi})^N} e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{T_i-\bar{T}}{\sigma_T}\right)^2}$$

Maximum likelihood

- and the log-likelihood is

$$\Lambda = \ln L = -\frac{N}{2}\ln(2\pi) - N\ln\sigma_T - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{T_i - \bar{T}}{\sigma_T}\right)^2$$

- the derivative with respect to the mean is zero when

$$\frac{\partial(\Lambda)}{\partial\bar{T}} = \frac{1}{\sigma_T^2}\sum_{i=1}^{N}(T_i - \bar{T}) = 0$$

- or

$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N}T_i$$

- note that this is just the sample mean

Maximum likelihood

- and the log-likelihood is

$$\Lambda = \ln L = -\frac{N}{2}\ln(2\pi) - N\ln\sigma_T - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{T_i - \bar{T}}{\sigma_T}\right)^2$$

- the derivative with respect to the variance is zero when

$$\frac{\partial(\Lambda)}{\partial\sigma_T} = -\frac{N}{\sigma_T} + \frac{1}{\sigma_T^3}\sum_{i=1}^{N}(T_i - \bar{T})^2 = 0$$

- or

$$\hat{\sigma}_T^2 = \frac{1}{N}\sum_{i=1}^{N}(T_i - \bar{T})^2$$

- note that this is just the sample variance

Maximum likelihood

- example:
  - if sample is {10,20,30,40,50}

$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$$

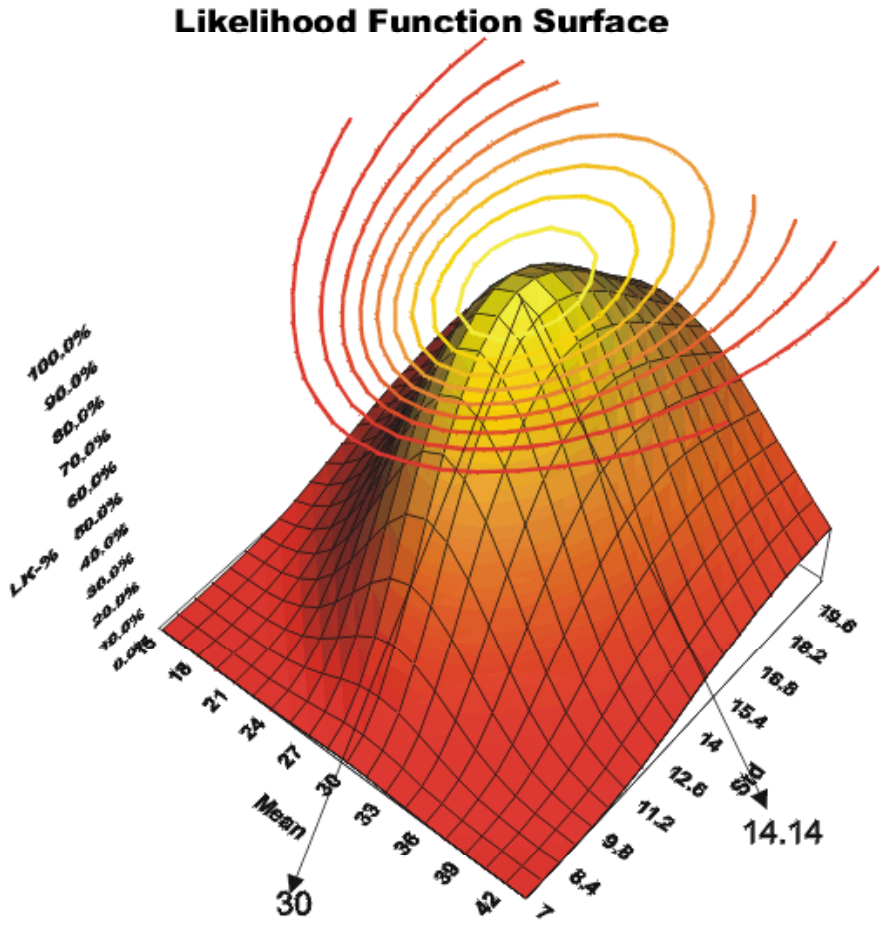
$$= \frac{10+20+30+40+50}{5}$$

$$= 30$$

$$\hat{\sigma}_T = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i-\bar{T})^2}$$

$$= \sqrt{\frac{(10-30)^2+(20-30)^2+(30-30)^2+(40-30)^2+(50-30)^2}{5}}$$

$$= 14.1421$$



Homework

- show that the Hessian is negative definite

$$\theta^t \nabla_\Theta^2 P_X(D;\theta)\theta \le 0, \quad \forall \theta \in \Re^n$$

- show that these formulas can be generalized to the vector case
  - $\mathcal{D}^{(i)} = \{x_1^{(i)}, ..., x_n^{(i)}\}$ set of examples from class i
  - the ML estimates are

$$\mu_i = \frac{1}{n}\sum_j x_j^{(i)} \qquad \Sigma_i = \frac{1}{n}\sum_j (x_j^{(i)} - \mu_i)(x_j^{(i)} - \mu_i)^T$$

- note that the ML solution is usually intuitive

Estimators

- when we talk about estimators, it is important to keep in mind that
  - an estimate is a number
  - an estimator is a random variable

$$\hat{\theta} = f(X_1, \ldots, X_n)$$

- an estimate is the value of the estimator for a given sample.
- if $\mathcal{D} = \{x_1, \ldots, x_n\}$, when we say $\hat{\mu} = \frac{1}{n}\sum_j x_j$

  what we mean is $\hat{\mu} = f(X_1, \ldots, X_n)\big|_{X_1 = x_1, \ldots, X_n = x_n}$ with

$$f(X_1, \ldots, X_n) = \frac{1}{n}\sum_j X_j$$

← the X_i are random variables

Bias and variance

- we know how to produce estimators (by ML)
- how do we evaluate an estimator?
- Q_1: is the expected value equal to the true value?
- this is measured by the bias
  - if $$\hat{\theta} = f(X_1, \ldots, X_n)$$
    then $$Bias(\hat{\theta}) = E_{X_1,\ldots,X_n}\left[f(X_1, \ldots, X_n) - \theta\right]$$
  - an estimator that has bias will usually not converge to the perfect estimate θ, no matter how large the sample is
  - e.g. if θ is negative and the estimator is $f(X_1, \ldots, X_n) = \frac{1}{n}\sum_j X_j^2$ the bias is clearly non-zero

Bias and variance

- the estimators is said to be biased
  - this means that it is not expressive enough to approximate the true value arbitrarily well
  - this will be clearer when we talk about density estimation
- Q_2: assuming that the estimator converges to the true value, how many sample points do we need?
  - this can be measured by the variance

$$Var\left(\hat{\theta}\right) = E_{X_1,\ldots,X_n}\left\{\left(f(X_1,\ldots,X_n) - E_{X_1,\ldots,X_n}\left[f(X_1,\ldots,X_n)\right]\right)^2\right\}$$

  - the variance usually decreases as one collects more training examples

Example

- ML estimator for the mean of a Gaussian $N(\mu, \sigma^2)$

$$Bias(\hat{\mu}) = E_{X_1,\ldots,X_n}[\hat{\mu} - \mu] = E_{X_1,\ldots,X_n}[\hat{\mu}] - \mu$$

$$= E_{X_1,\ldots,X_n}\left[\frac{1}{n}\sum_i X_i\right] - \mu$$

$$= \frac{1}{n}\sum_i E_{X_1,\ldots,X_n}[X_i] - \mu$$

$$= \frac{1}{n}\sum_i E_{X_i}[X_i] - \mu$$

$$= \mu - \mu = 0$$

- the estimator is unbiased

Any questions?