Maximum Likelihood Estimation (MLE)

Nuno Vasconcelos

(Ken Kreutz-Delgado)

UCSD

BDR (under 0/1 Loss)

- For the zero/one loss, the following three decision rules are optimal and equivalent
  - 1) $$\boxed{i^*(x) = \arg\max_i P_{Y|X}(i \mid x)}$$
  - 2) $$\boxed{i^*(x) = \arg\max_i \left[P_{X|Y}(x \mid i) P_Y(i) \right]}$$
  - 3) $$\boxed{i^*(x) = \arg\max_i \left[\log P_{X|Y}(x \mid i) + \log P_Y(i) \right]}$$
  - **Form 1) is usually hard to use, 3) is frequently easier than 2)**

The Gaussian Classifier

- One important case is that of Multivariate Gaussian Classes
  - The pdf of class i is a Gaussian of mean μ_i and covariance Σ_i

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

- The BDR is

$$\boxed{\begin{aligned} i^*(x) = \arg\max_i \Bigg[& -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \\ & -\frac{1}{2}\log(2\pi)^d \left|\Sigma_i\right| + \log P_Y(i) \Bigg] \end{aligned}}$$

Implementation

- To design a Gaussian classifier (e.g. homework)
  - Start from a collection of datasets, where the *i*-th class dataset $\mathcal{D}^{(i)} = \{x_1^{(i)}, ..., x_n^{(i)}\}$ is a set of $n^{(i)}$ examples from class *i*
  - For each class *estimate* the Gaussian parameters :

$$\boxed{\hat{\mu}_i = \frac{1}{n^{(i)}} \sum_j x_j^{(i)}} \quad \boxed{\hat{\Sigma}_i = \frac{1}{n^{(i)}} \sum_j (x_j^{(i)} - \hat{\mu}_i)(x_j^{(i)} - \hat{\mu}_i)^T} \quad \boxed{\hat{P}_Y(i) = \frac{n^{(i)}}{T}}$$

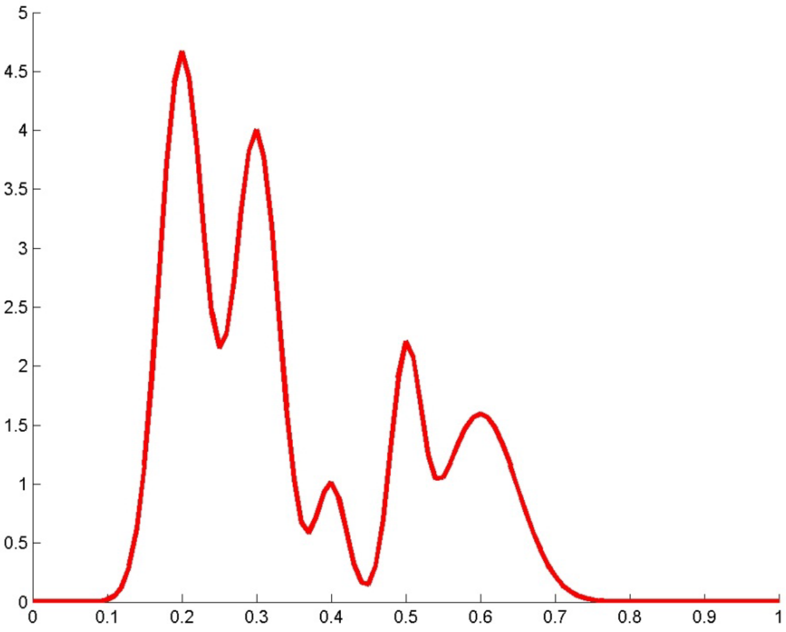
where T is the total number of examples over all c classes

- the BDR is approximated as

$$\boxed{\begin{aligned} i^*(x) = \arg\max_i \Bigg[& -\frac{1}{2}(x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (x - \hat{\mu}_i) \\ & -\frac{1}{2}\log(2\pi)^d \left|\hat{\Sigma}_i\right| + \log \hat{P}_Y(i) \Bigg] \end{aligned}}$$

Important

- Warning: at this point all optimality claims for the BDR cease to be valid!!
- The BDR is guaranteed to achieve the minimum loss *only* when we use the true probabilities
- When we "plug in" *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 one could never approximate it well by using simple parametric models (e.g. a single Gaussian).

Maximum likelihood Estimation (MLE)

- Given a parameterized pdf how should one estimate the parameters which define the pdf?
- There are many techniques of *"parameter estimation."* We shall utilize 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

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

    - Note: This is a classical statistics approach, which means that θ is *NOT* a random variable. It is *a deterministic but unknown parameter*, and the probabilities are a function of this unknown parameter.

Maximum Likelihood Estimation (MLE)

- The three steps continued:

  - 2) Assemble a collection of datasets:

    $\mathcal{D}^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\}$ = set of examples from each class i

  - 3) Select the values of the parameters of class i to be the ones that maximize the probability of the data from that class

$$\boxed{\begin{aligned}\hat{\theta}_i &= \arg\max_{\theta \in \Theta} P_{D^{(i)}|Y}\left(\mathrm{D}^{(i)} \mid i;\theta\right) \\ &= \arg\max_{\theta \in \Theta} \log P_{D^{(i)}|Y}\left(\mathrm{D}^{(i)} \mid i;\theta\right)\end{aligned}}$$

Note that it does not make any difference to maximize probabilities or their logs.

Maximum Likelihood Estimation (MLE)

- Since
  - Each sample $\mathcal{D}^{(i)}$ is considered independently
  - Each parameter vector θ_i is 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 *i* :

$$\boxed{\begin{aligned}\hat{\theta}_{\mathrm{ML}} &= \arg\max_{\theta\in\Theta} P_X(\mathcal{D};\theta)\\ &= \arg\max_{\theta\in\Theta} \log P_X(\mathcal{D};\theta)\end{aligned}}$$

- The function $L(\theta;\mathcal{D}) = P_X(\mathcal{D};\theta)$ is the *likelihood* of the parameter θ given the data $\mathcal{D}$, or simply the *likelihood function*.

The Likelihood Function

- Given a parameterized family of pdf's (aka known as a *statistical model*) for the data $\mathcal{D}$, we define a Likelihood of the parameter vector θ given $\mathcal{D}$:

$$\boxed{L_{\mathrm{D}}(\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathrm{D}) = \boldsymbol{\alpha}(\mathrm{D})P_{\mathbf{D}}(\mathrm{D}\,|\,\boldsymbol{\theta})}$$

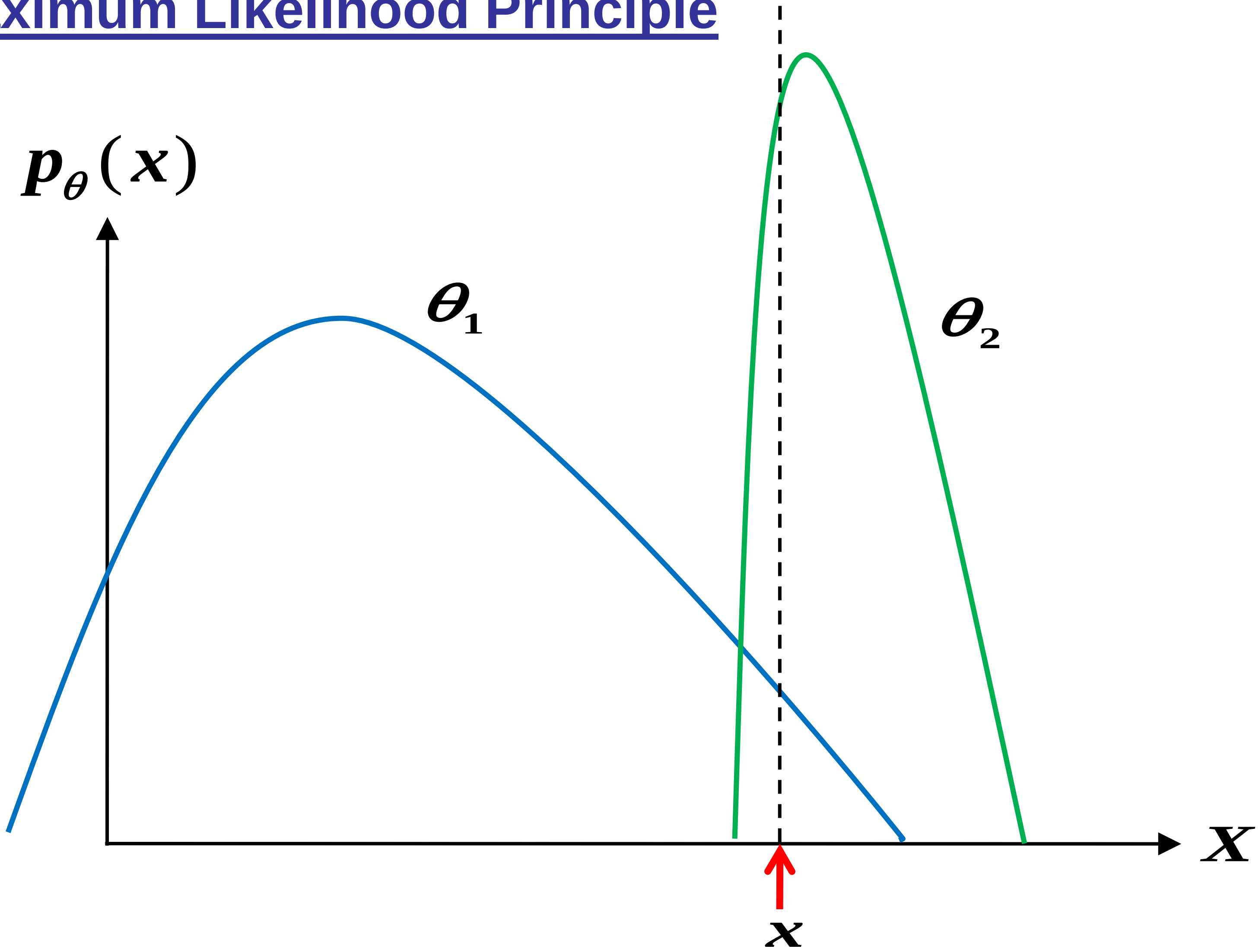
where $\alpha(\mathcal{D}) > 0$ for all $\mathcal{D}$, and $\alpha(\mathcal{D})$ is independent of the parameter θ.

- The choice $\alpha(\mathcal{D}) = 1$ yields the

  Standard Likelihood: $L(\theta\,;\,\mathcal{D}) = P_{\mathrm{D}}(\mathcal{D}\,;\theta)$

  which was shown on the previous slide.

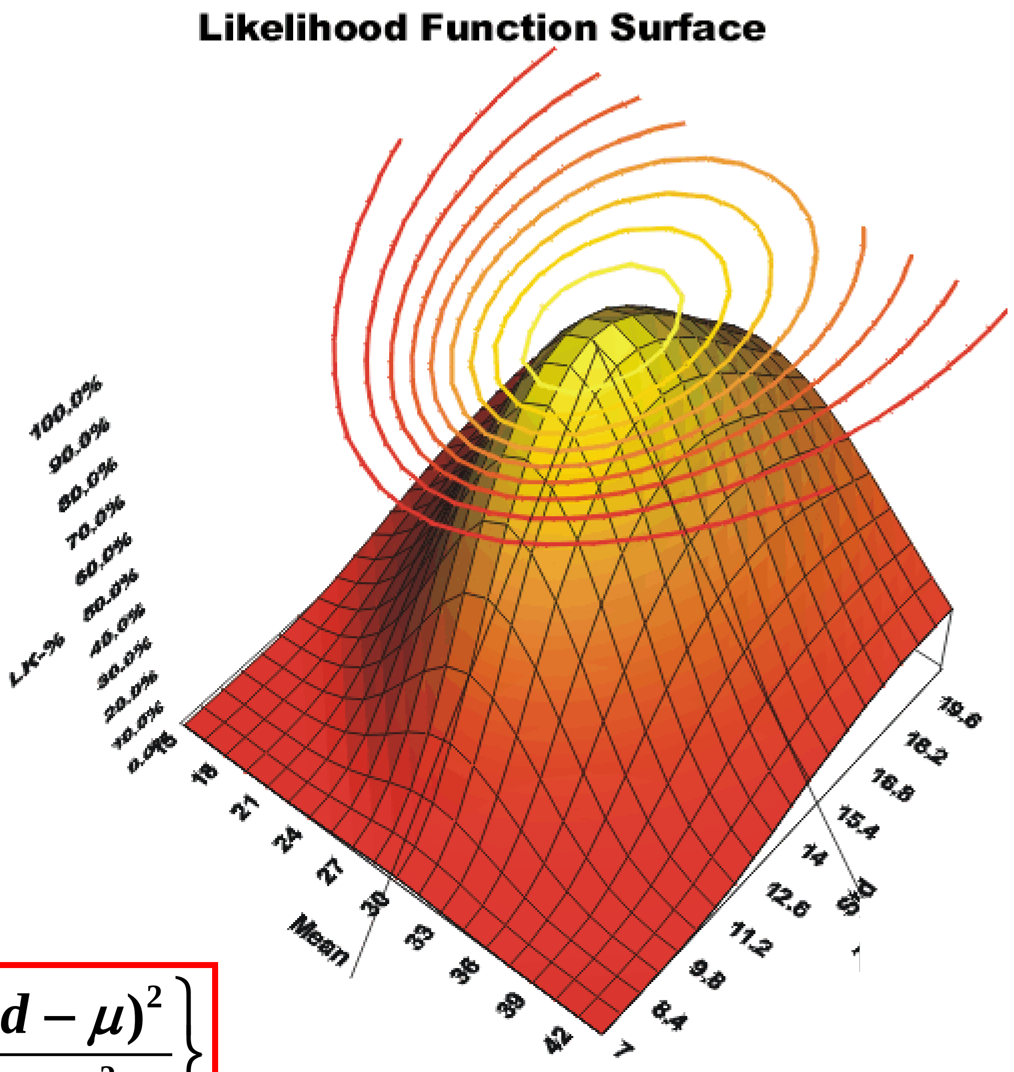
Maximum Likelihood Principle



$$L_x(\theta_2) = P_{\theta_2}(x) > P_{\theta_1}(x) = L_x(\theta_1)$$

The Likelihood Function

- 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 data sample



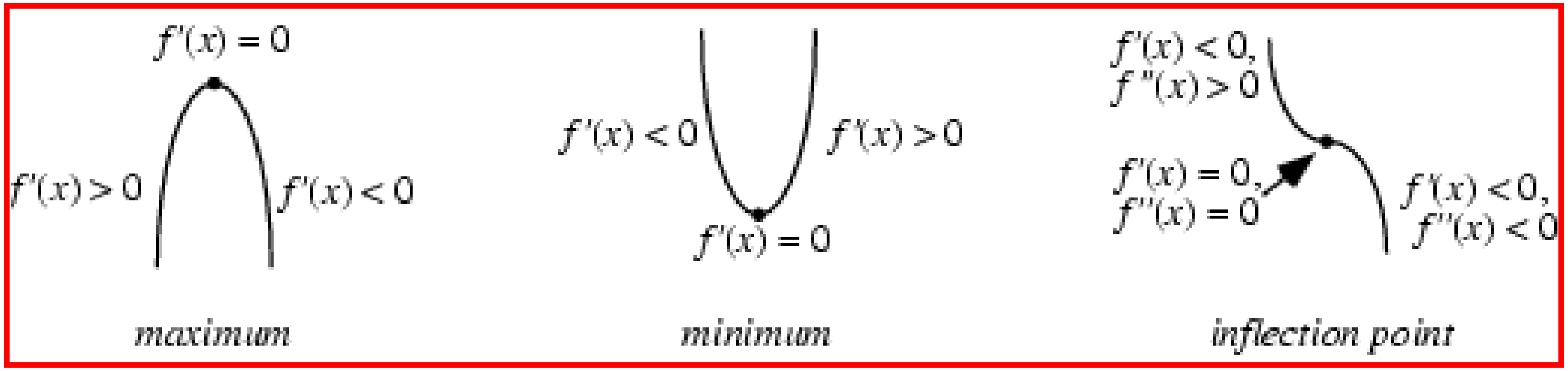
$$P_X(d;\theta) = \frac{1}{\sqrt{(2\pi)\sigma^2}} \exp\left\{-\frac{(d-\mu)^2}{2\sigma^2}\right\}$$

Maximum Likelihood Estimation (MLE)

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

$$\boxed{\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} P_{\mathrm{D}}(\mathrm{D};\theta)}$$

- When θ is a scalar, this is high-school calculus:



- We have a *local maximum of f(x) at a point x* when
  - The first derivative at *x* is zero. (*x* is a stationary point.)
  - The second derivative is negative at *x*.

MLE Example

- *Gaussian with unknown mean & standard deviation*:

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

- Given a data sample $\mathcal{D} = \{T_1, \ldots, T_N\}$ of independent and identically distributed (iid) measurements, the (standard) likelihood is

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

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

MLE Example

- The log-likelihood is

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

- The derivative with respect to the mean is zero when

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

yielding

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

- Note that this is just the sample mean

MLE Example

- The log-likelihood is

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

- The derivative wrt the standard deviation 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

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

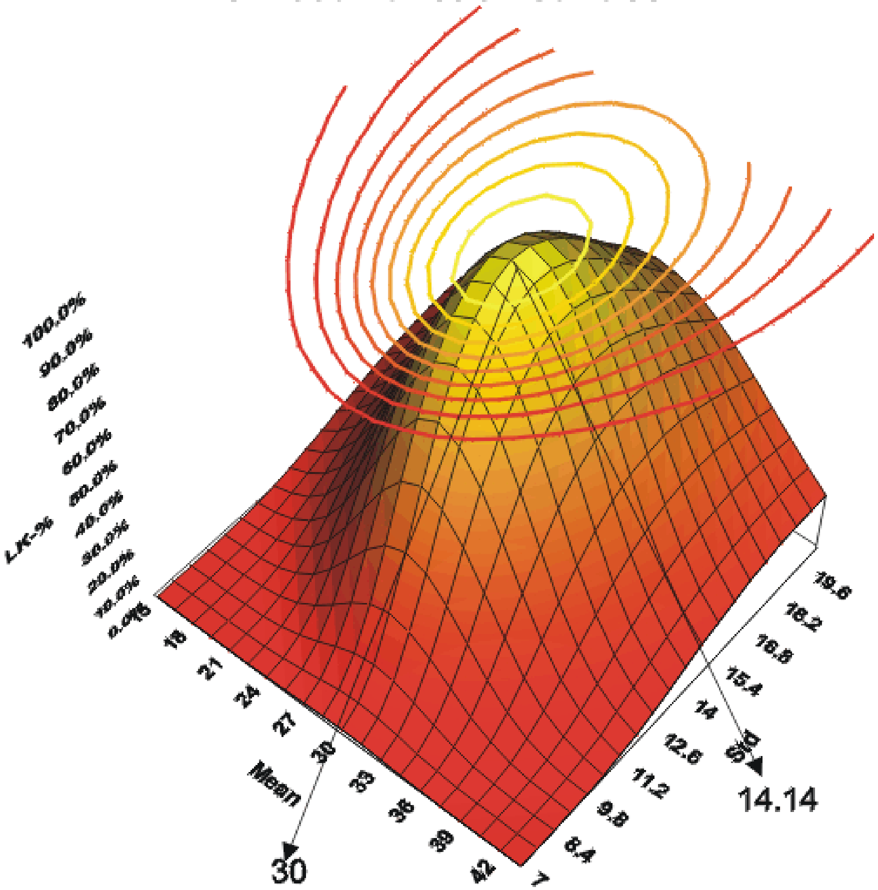
Note that this is just the sample variance.

MLE Example

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

$$
\begin{aligned}
\hat{\bar{T}} &= \frac{1}{N}\sum_{i=1}^{N} T_i \\
&= \frac{10+20+30+40+50}{5} \\
&= 30
\end{aligned}
$$

Likelihood Function Surface



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

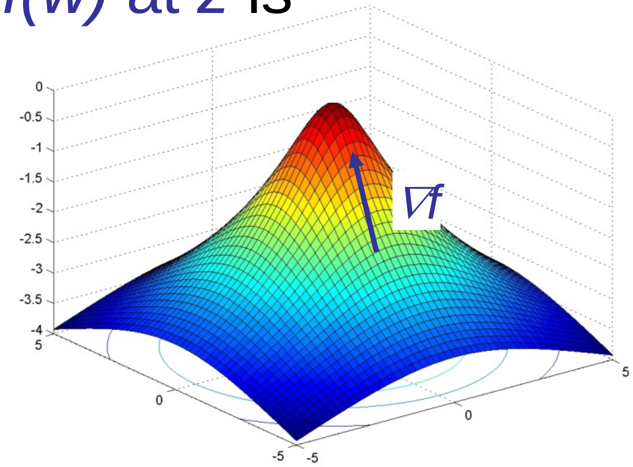
$$
= \sqrt{\frac{(10-30)^2+(20-30)^2+(30-30)^2+(40-30)^2+(50-30)^2}{5}}
$$

$$
= 14.1421
$$

The Gradient

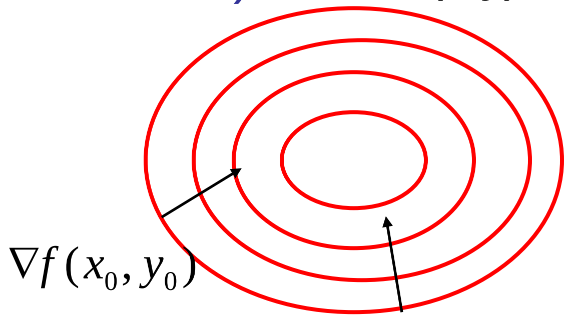
- In higher dimensions, the generalization of the derivative is the gradient
- The (Cartesian) gradient of a function *f(w)* at *z* is

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



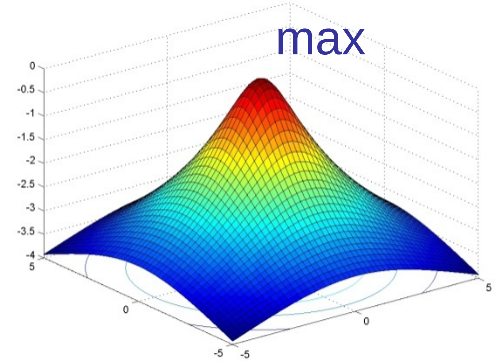
- The gradient has a nice geometric interpretation
  - It points in the direction of maximum growth of the function. (Steepest Ascent Direction.)
  - Which makes it perpendicular to the contours where the function is constant.
  - The above is the gradient for the simple (unweighted) Euclidean Norm (aka the Cartesian Gradient).

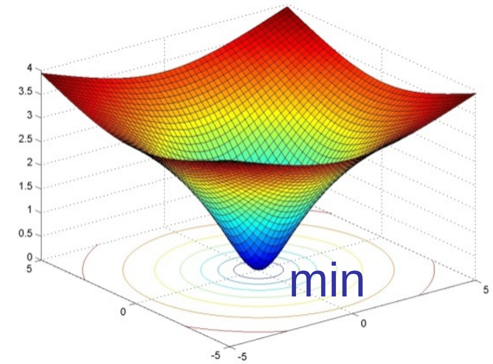
f(x,y)

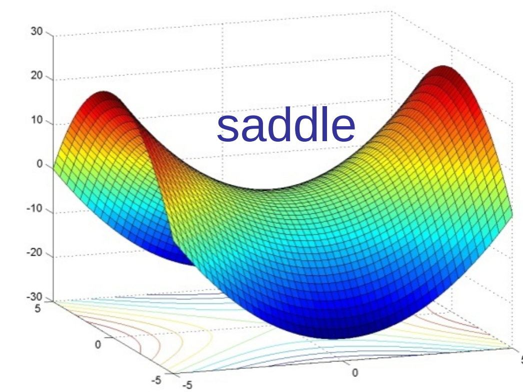


The Gradient

- Note that if $\nabla f(x) = 0$
  - There is no direction of growth at x
  - also $-\nabla f(x) = 0$, and there is no direction of decrease at x
  - We are either at a local minimum or maximum or “saddle” point at x
- Conversely, if there is a local min or max or saddle point at x
  - There is no direction of growth or decrease at x
  - $\nabla f(x) = 0$
- This shows that we have a *stationary point* at x if and only if $\nabla f(x) = 0$
- To determine which type holds we need second order conditions







The Hessian

- The extension of the scalar second-order derivative is the *Hessian matrix of second partial derivatives*:

$$\boxed{\mathrm{H}(x) = \frac{\partial^2 f}{\partial x^2}(x) = \frac{\partial}{\partial x}\left(\frac{\partial f(x)}{\partial x}\right)^T = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(x) \\ & \vdots & \\ \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(x) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(x) \end{bmatrix}}$$

Note that the Hessian is symmetric.

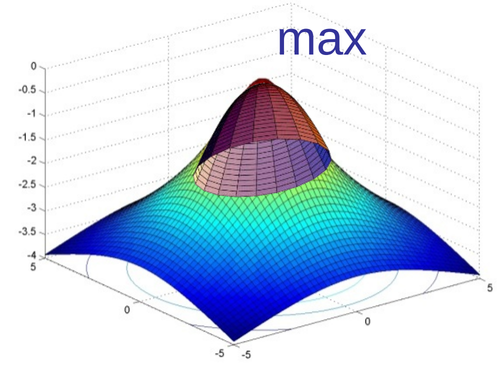
- The Hessian gives us the *quadratic function*

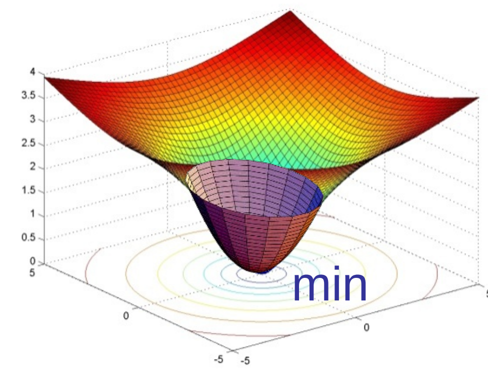
$$\boxed{\tfrac{1}{2}(x - x_0)^T \mathrm{H}(x_0)(x - x_0)}$$

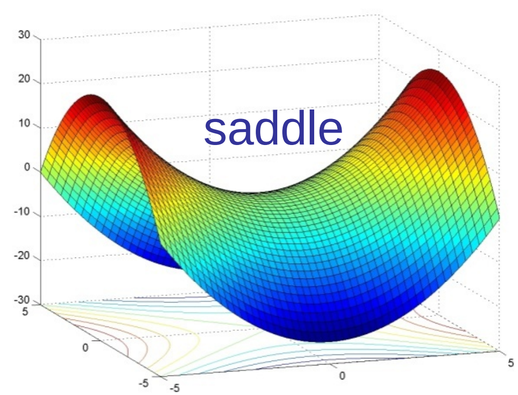
that *best approximates f(x) at a stationary point* x_0.

Hessian as a Quadratic Approximation

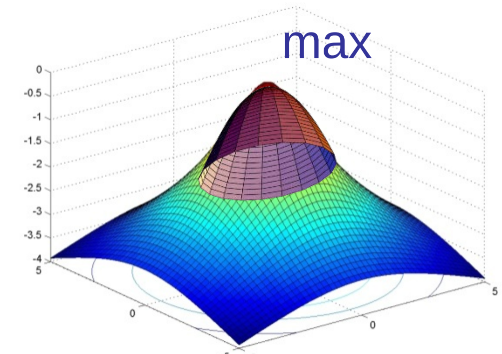
- E.g. this means that *if the gradient is zero at x_0*, we have
  - a maximum when the function *f(x)* can be locally approximated by an "upwards pointing" quadratic bowl ($\mathcal{H}(x_0)$ is neg-def)
  - a minimum when the function can be locally approximated by a "downwards pointing" quadratic bowl ($\mathcal{H}(x_0)$ is pos-def)
  - a saddle point otherwise ($\mathcal{H}(x_0)$ is indefinite)







Hessian Gives Local Behavior

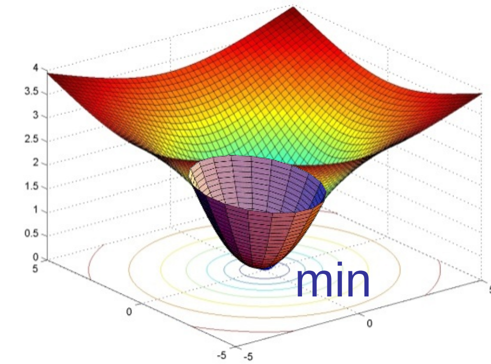


- This is something that we already saw:

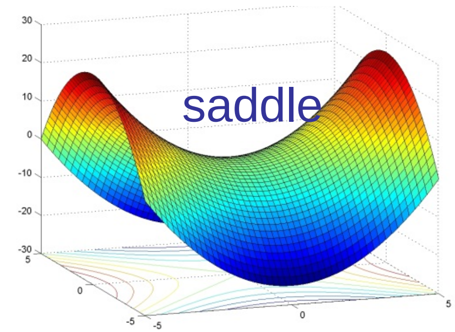
  For any matrix M, the quadratic function

$$\boxed{x^T M x}$$

  - is an upwards pointing quadratic bowl at the point $x = 0$ when M is negative definite
  - is a downwards pointing quadratic bowl at $x = 0$ when M is positive definite
  - is a saddle point at $x = 0$ otherwise



- Hence, similarly, what matters is the definiteness property of the Hessian at a stationary point x_0



- E.g., we have a maximum at a stationary point x_0 when the Hessian is negative definite at x_0

Optimality Conditions

In summary:

- w_0 is a local minimum of *f(w)* if and only if
  - *f* has zero gradient at w_0

$$\boxed{\nabla f(w_0) = 0}$$

and the Hessian of *f* at w_0 is positive definite

$$\boxed{d^T \mathrm{H}(x_0) d > 0, \quad \forall d \in \mathbb{R}^n,\ d \neq 0}$$

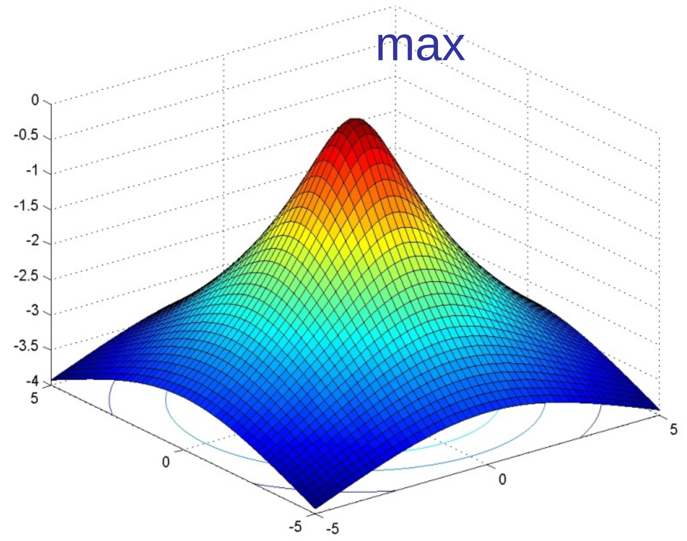
where

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

Maximum Likelihood Estimation (MLE)

- Given a sample, to obtain an MLE we want to solve

$$\boxed{\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} P_D(\mathrm{D};\theta)}$$



- Candidate solutions are the parameter values $\hat{\theta}$ such that

$$\boxed{\frac{\partial}{\partial \theta} P_D(\mathrm{D};\hat{\theta}) = 0}$$

$$\boxed{\theta^T \mathrm{H}(\hat{\theta})\theta < 0, \quad \forall \theta \neq 0}$$

- Note that you *always* have to check the second-order Hessian condition

MLE Example

- Back to our 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 *iid samples* $\{T_1, \ldots, T_N\}$ the *likelihood* is

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

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

MLE Example

- The log-likelihood is

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

- The derivative of Λ with respect to the mean is

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

from which we compute the second-order derivatives

$$\frac{\partial^2\Lambda}{\partial\bar{T}^2} = -\frac{N}{\sigma_T^2} \qquad \frac{\partial^2\Lambda}{\partial\bar{T}\,\partial\sigma_T} = -\frac{2}{\sigma_T^3}\sum_i\left(T_i - \bar{T}\right)$$

MLE Example

- The derivative of Λ with respect to the standard deviation is

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

which yields the second-order derivatives

$$\frac{\partial^2 \Lambda}{\partial(\sigma_T)^2} = \frac{N}{\sigma_T^2} - \frac{3}{\sigma_T^4}\sum_i \left(T_i - \bar{T}\right)^2 \qquad \frac{\partial^2 \Lambda}{\partial \sigma_T \partial \bar{T}} = -\frac{2}{\sigma_T^3}\sum_i \left(T_i - \bar{T}\right)$$

- The stationary parameter values are,

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

MLE Example

- The elements of the Hessian are:

$$\frac{\partial^2 \Lambda}{\partial \overline{T}^2} = -\frac{N}{\sigma_T^2} \qquad \frac{\partial^2 \Lambda}{\partial \overline{T} \partial \sigma_T} = -\frac{2}{\sigma_T^3} \sum_i \left(T_i - \overline{T} \right) = 0$$

$$\frac{\partial^2 \Lambda}{\partial \sigma_T \partial \overline{T}} = -\frac{2}{\sigma_T^3} \sum_i \left(T_i - \overline{T} \right) = 0 \qquad \frac{\partial^2 \Lambda}{\partial \left(\sigma_T \right)^2} = \frac{N}{\sigma_T^2} - \frac{3N}{\sigma_T^4} \sigma_T^2 = -\frac{2N}{\sigma_T^2}$$

- Thus the Hessian is

$$\boxed{\mathrm{H}(\theta) = -\frac{N}{\sigma_T^2} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}}$$

which is clearly negative definite at the stationary point. This we have determined the MLE of the parameters.

Combining the MLE Examples

- For Gaussian Classes all of the above formulas can be generalized to the random vector case as follows:
  - $\mathcal{D}^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\}$ = set of iid vector examples from each class i , i = 1, … , d .
  - The MLE estimates in the vector random data case are:

$$\boxed{\hat{\mu}_i = \frac{1}{n_i}\sum_j x_j^{(i)}}$$

$$\boxed{\hat{P}_Y(i) = \frac{n_i}{N}}$$

$$\boxed{\hat{\Sigma}_i = \frac{1}{n_i}\sum_j (x_j^{(i)} - \hat{\mu}_i)(x_j^{(i)} - \hat{\mu}_i)^T}$$

  - These are the sample estimates given earlier with no justification.
  - The ML solutions are intuitive, which is usually the case.

2nd MLE Example

- To find the MLE's of the two prior class probabilities $P_Y(i)$ note that

$$P_Y(i) = \begin{cases} \pi, & i = 1 \\ 1-\pi, & i = 0 \end{cases}$$

can be written as

$$\boxed{P_Y(x) = \pi^x (1-\pi)^{1-x} \qquad x \in \{0,1\}}$$

where x is the so-called indicator (or 0-1) function.

- Given iid indicator samples $\mathcal{D} = \{x_1, \dots, x_N\}$, we have

$$\boxed{L(\pi; \mathrm{D}) = P_Y(\mathrm{D}; \pi) = \prod_{i=1}^{N} \pi^{x_i} (1-\pi)^{1-x_i}}$$

2nd MLE Example

- Therefore

$$\log P_Y(\mathrm{D};\pi) = \sum_{i=1}^{n}\left\{ x_i \log\pi + (1-x_i)\log(1-\pi) \right\}$$

- Setting the derivative of the log-likelihood with respect to π equal to zero,

$$\frac{\partial \log P_Y(\mathrm{D})}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{N} x_i - \frac{N}{1-\pi} + \frac{1}{1-\pi}\sum_{i=1}^{N} x_i$$

$$= \frac{1}{\pi(1-\pi)}\sum_{i=1}^{N} x_i - \frac{N}{1-\pi} = 0,$$

2nd MLE Example

yields the MLE estimate

$$\boxed{\hat{\pi}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{n_i}{N} \quad where \quad n_i = \sum_{i=1}^{N} x_i}$$

Note that this is just the relative frequency of occurrence of the value “1” in the sample. I.e. the MLE is just the count of the number of 1’s over the total number of points!

Again we see that the MLE yields an intuitively pleasing estimate of the unknown parameters.

2^{nd} MLE Example

- Check that the second derivative is negative:

$$\frac{\partial^2 \log P_Y(D)}{\partial \pi^2} = -\frac{1-2\pi}{\pi^2(1-\pi)^2}\sum_{i=1}^{N} x_i - \frac{N}{(1-\pi)^2}$$

$$= \frac{N}{(1-\pi)^2}\left[-\frac{1-2\pi}{\pi^2}\pi - 1\right]$$

$$= \frac{N}{(1-\pi)^2}\left[1-\frac{1}{\pi}\right] < 0$$

for $\pi < 1$.

Any questions?