# S&DS 242/542: Theory of Statistics

## Lecture 12: Maximum likelihood estimation and optimization

# Maximum likelihood estimation

The joint PDF or PMF of the data, viewed as a function of the model parameters $\theta$, is called the **likelihood function**.

For data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ from a parametric model $f(x \mid \theta)$, this is given by

$$\text{lik}(\theta) = f(X_1 \mid \theta) \times \ldots \times f(X_n \mid \theta)$$

The **maximum likelihood estimator (MLE)** is the value of $\theta$ in the parameter space of the model that maximizes $\text{lik}(\theta)$.

Intuitively, it is the value of $\theta$ that makes the observed data "most probable" or "most likely".

# Connection to the likelihood ratio statistic

In our discussion of the Neyman-Pearson lemma, recall that for testing

$$H_0 : \mathbf{X} \sim f_0 \qquad \text{vs.} \qquad H_1 : \mathbf{X} \sim f_1$$

the most powerful test rejects $H_0$ for large values of the likelihood ratio statistic

$$L(\mathbf{X}) = f_1(\mathbf{X})/f_0(\mathbf{X}).$$

In the context of a parametric model, we may consider testing $H_0 : X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta_0)$ versus $H_1 : X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta_1)$ for two different parameter values. Then

$$f_0(X_1, \ldots, X_n) = f(X_1 \mid \theta_0) \times \ldots \times f(X_n \mid \theta_0),$$
$$f_1(X_1, \ldots, X_n) = f(X_1 \mid \theta_1) \times \ldots \times f(X_n \mid \theta_1),$$

so the likelihood ratio statistic is exactly $L(\mathbf{X}) = \mathrm{lik}(\theta_1)/\mathrm{lik}(\theta_0)$.

# Connection to empirical risk minimization

Maximizing $\mathrm{lik}(\theta)$ is equivalent to maximizing its logarithm, the **log-likelihood function**. For data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, this is

$$\ell_n(\theta) = \log\left(\mathrm{lik}(\theta)\right) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

It is usually easier to work with the log-likelihood instead of the likelihood itself, because this involves a sum rather than a product.

The MLE maximizes $\ell_n(\theta)$, or equivalently, minimizes

$$\frac{1}{n} \sum_{i=1}^{n} -\log f(X_i \mid \theta)$$

This is an example of *empirical risk minimization*, minimizing the average of the loss function $-\log f(x \mid \theta)$ across the observed data.

# Maximum likelihood in the Poisson model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$.

$$f(x|\lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \Rightarrow \log f(x|\lambda) = -\lambda + x \log \lambda - \log(x!)$$

$$\ell_n(\lambda) = \sum_{i=1}^{n} (-\lambda + x_i \log \lambda - \log(x_i!))$$

$$= -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \log \lambda - \sum_{i=1}^{n} \log(x_i!)$$

Solve $0 = \ell_n'(\hat{\lambda}) = -n + \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} x_i$

$$\Rightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Check: $\ell_n'(\lambda) > 0$ for $\lambda < \hat{\lambda}$

$\ell_n'(\lambda) < 0$ for $\lambda > \hat{\lambda}$

$\Rightarrow \hat{\lambda} = \bar{x}$ is the MLE.

# Maximum likelihood in the Pareto model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Pareto}(\alpha, 1)$.

$$f(x|\alpha) = \frac{\alpha}{x^{\alpha+1}} \quad (\text{for } x > 1) \Rightarrow \log f(x|\alpha) = \log \alpha - (\alpha+1)\log x$$

$$\ell_n(\alpha) = \sum_{i=1}^{n} (\log \alpha - (\alpha+1)\log x_i)$$

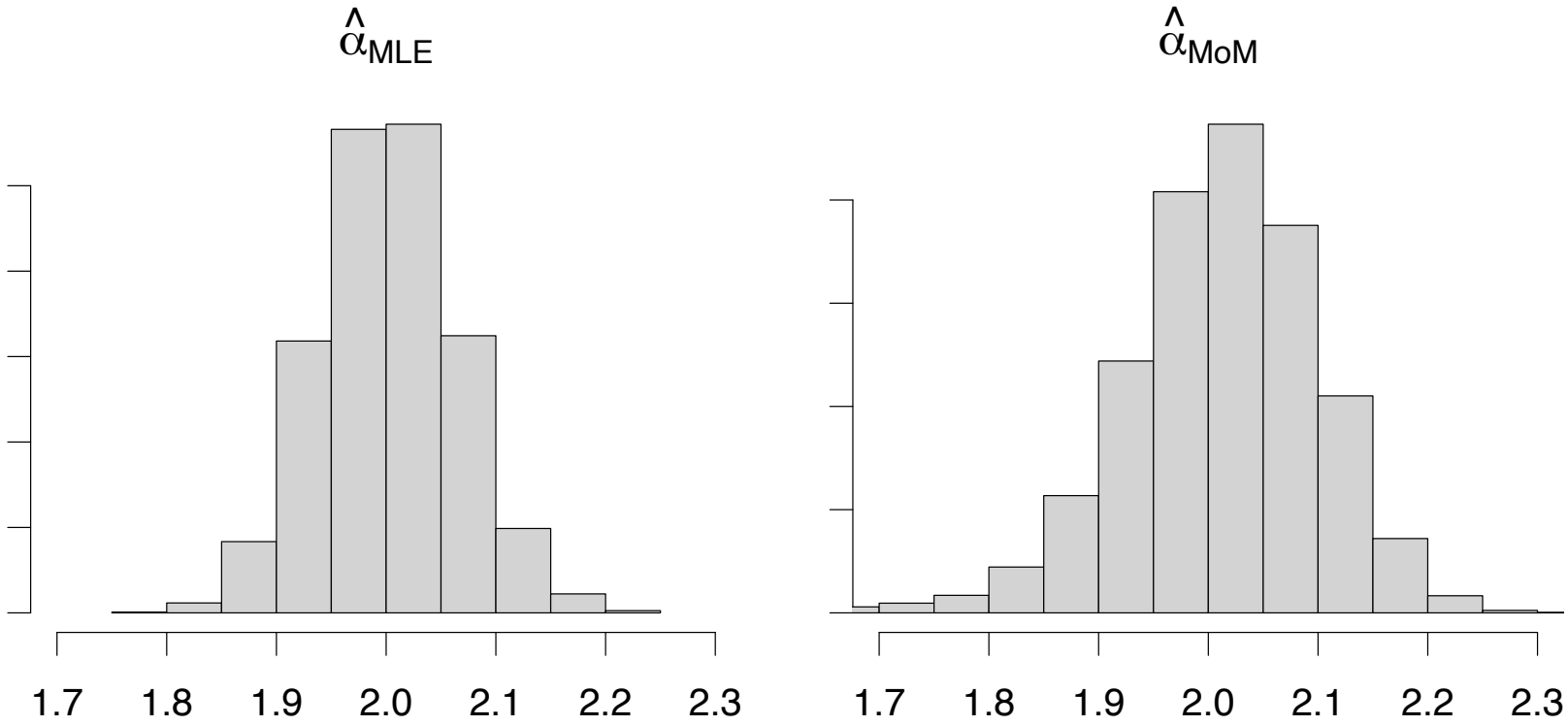
$$= n \log \alpha - (\alpha+1) \sum_{i=1}^{n} \log x_i$$

Solve $0 = \ell_n'(\hat{\alpha}) = \frac{n}{\hat{\alpha}} - \sum_{i=1}^{n} \log x_i$

$$\Rightarrow \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log x_i}$$

Check: $\ell_n'(\alpha) > 0$ for $\alpha < \hat{\alpha}$
$\ell_n'(\alpha) < 0$ for $\alpha > \hat{\alpha}$
$\Rightarrow \hat{\alpha}$ is the MLE

# Maximum likelihood in the Pareto model




True parameter $\alpha = 2$

This estimate $\hat{\alpha}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} \log X_i}$ coincides with a generalized method of moments estimator from last lecture, and is *different* from usual the method of moments estimator $\hat{\alpha}_{\text{MoM}} = \frac{\bar{X}}{\bar{X}-1}$. The variability of $\hat{\alpha}_{\text{MLE}}$ is smaller than that of $\hat{\alpha}_{\text{MoM}}$.

# Maximum likelihood in the normal model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow \log f(x|\mu,\sigma^2) = -\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2$$

$$\ell_n(\mu,\sigma^2) = \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2\right)$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi$$

Solve $$0 = \frac{\partial \ell_n}{\partial \mu}(\hat{\mu},\hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = \frac{1}{\hat{\sigma}^2}\left(\left(\sum_{i=1}^{n}x_i\right) - n\hat{\mu}\right)$$

$$0 = \frac{\partial \ell_n}{\partial \sigma^2}(\hat{\mu},\hat{\sigma}^2) = \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 - \frac{n}{2\hat{\sigma}^2}$$

# Maximum likelihood in the normal model

$$0 = \frac{1}{\hat{\sigma}^2}\left(\left(\sum_{i=1}^n X_i\right) - n\hat{\mu}\right) \qquad 0 = \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^n (X_i - \hat{\mu})^2 - \frac{n}{2\hat{\sigma}^2}$$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X} \qquad \Rightarrow \frac{n}{2\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^n (X_i - \bar{X})^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

- For each fixed $\sigma^2 > 0$, $\hat{\mu} = \bar{X}$ maximizes $\ell_n(\mu, \sigma^2)$
- At $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2$ maximizes $\ell_n(\bar{X}, \sigma^2)$

$\Rightarrow (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)$ is the MLE.

# Maximum likelihood in the Gamma model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$.

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$\Rightarrow \log f(x \mid \alpha, \beta) = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha-1) \log x - \beta x$$

$$\ell_n(\alpha, \beta) = \sum_{i=1}^n (\alpha \log \beta - \log \Gamma(\alpha) + (\alpha-1) \log X_i - \beta X_i)$$

$$= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha-1) \sum_{i=1}^n \log X_i - \beta \sum_{i=1}^n X_i$$

Solve $$0 = \frac{\partial \ell_n}{\partial \alpha}(\hat{\alpha}, \hat{\beta}) = n \log \hat{\beta} - n \cdot \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log X_i$$

$$0 = \frac{\partial \ell_n}{\partial \beta}(\hat{\alpha}, \hat{\beta}) = \frac{n\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n X_i$$

# Maximum likelihood in the Gamma model

$$0 = n \log \hat{\beta} - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^{n} \log X_i \qquad 0 = \frac{n\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^{n} X_i$$

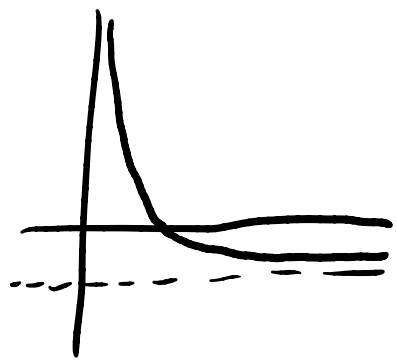
$$\Rightarrow 0 = n \log \frac{\hat{\alpha}}{\bar{X}} - n \cdot \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^{n} \log X_i \qquad \Rightarrow \hat{\beta} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} X_i} = \frac{\hat{\alpha}}{\bar{X}}$$

$$\Rightarrow 0 = \log \hat{\alpha} - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \log \bar{X} + \frac{1}{n} \sum_{i=1}^{n} \log X_i$$

# Maximum likelihood in the Gamma model

One may check that the function

$$f(\alpha) = \log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \bar{X} + \frac{1}{n}\sum_{i=1}^{n} \log X_i$$



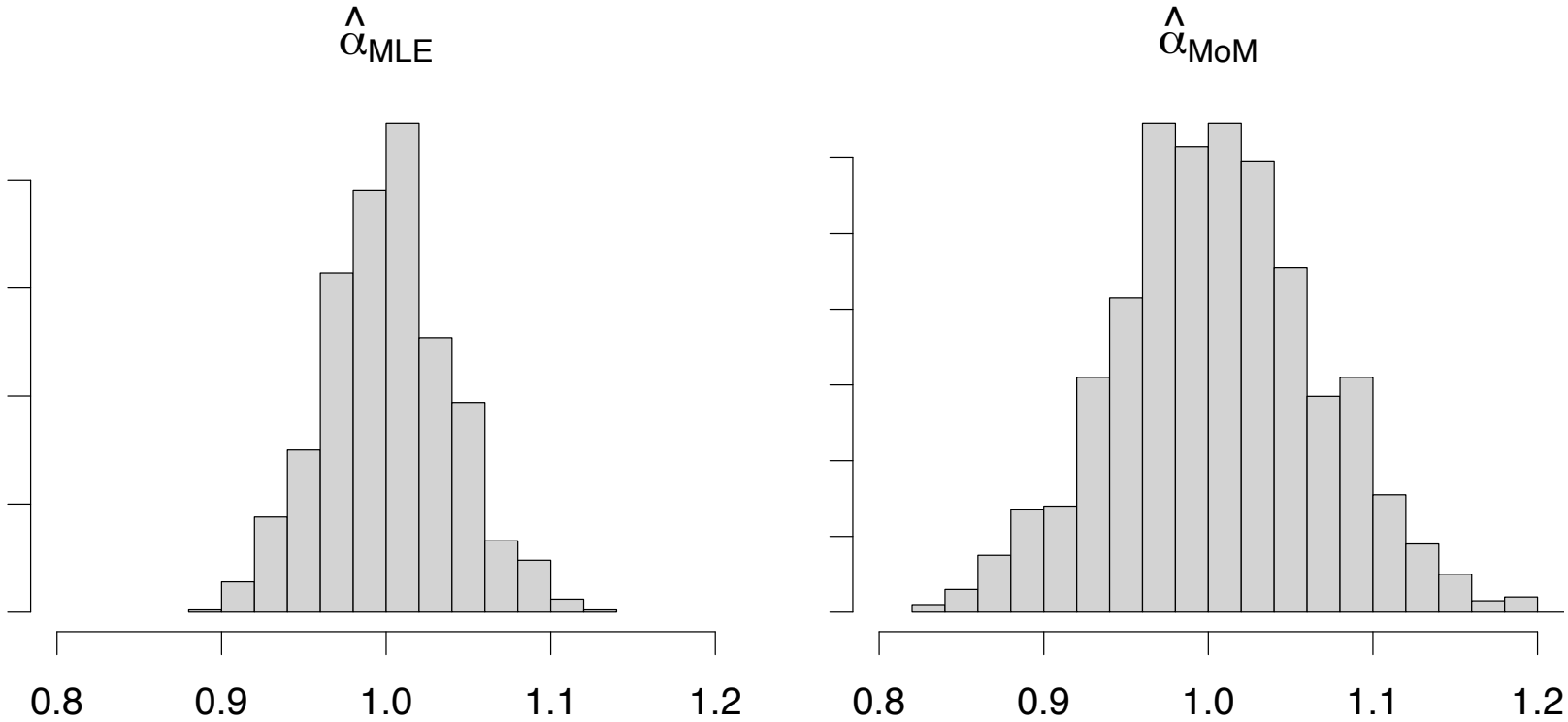
is decreasing over $\alpha \in (0, \infty)$, and

- $\lim_{\alpha \to 0} f(\alpha) = \infty$
- $\lim_{\alpha \to \infty} f(\alpha) = -\log \bar{X} + \frac{1}{n}\sum_{i=1}^{n} \log X_i$

By Jensen's inequality, $\frac{1}{n}\sum_{i=1}^{n} \log X_i < \log \bar{X}$, so $f(\alpha) < 0$ for all large $\alpha$. Thus $0 = f(\alpha)$ has a unique solution $\hat{\alpha}$. [There is no explicit form, and $\hat{\alpha}$ is usually computed numerically.]

The maximum likelihood estimators are $(\hat{\alpha}, \hat{\beta}) = (\hat{\alpha}, \hat{\alpha}/\bar{X})$.

# Maximum likelihood in the Gamma model




True parameters $\alpha = 1$, $\beta = 1$

The maximum likelihood estimates are again different from the method of moments estimates, and have smaller variability.

# Maximum likelihood in the Multinomial model

Let $(X_1, \ldots, X_k) \sim \text{Multinomial}(n, (p_1, \ldots, p_k))$. Here $X_1, \ldots, X_k$ are not IID, but instead are integer counts summing to $n$.

$$\ell_n(p_1, \ldots, p_k) = \log\left[\binom{n}{x_1, x_2, \ldots, x_k} \cdot p_1^{x_1} \cdots p_k^{x_k}\right]$$

$$= \log\binom{n}{x_1, \ldots, x_k} + x_1 \log p_1 + \cdots + x_k \log p_k$$

Lagrange multiplier method:

$$L_n(p_1, \ldots, p_k, \lambda) = \ell_n(p_1, \ldots, p_k) + \lambda(p_1 + \cdots + p_k - 1)$$

$$= \log\binom{n}{x_1, \ldots, x_k} + x_1 \log p_1 + \cdots + x_k \log p_k + \lambda(p_1 + \cdots + p_k - 1)$$

Solve $0 = \frac{\partial L_n}{\partial p_i}(\hat{p}_1, \ldots, \hat{p}_k, \hat{\lambda})$ for each $i = 1, \ldots, k$

$0 = \frac{\partial L_n}{\partial \lambda}(\hat{p}_1, \ldots, \hat{p}_k, \hat{\lambda})$

# Maximum likelihood in the Multinomial model

$0 = \frac{x_i}{\hat{p}_i} + \hat{\lambda}$ for each $i = 1, \ldots, k$ $\quad 0 = \hat{p}_1 + \ldots + \hat{p}_k - 1$

$\Rightarrow \hat{p}_i = -\frac{x_i}{\hat{\lambda}}$ for $i = 1, \ldots, k$ $\quad \Rightarrow 0 = -\frac{x_1 + \ldots + x_k}{\hat{\lambda}} - 1$

$= -\frac{n}{\hat{\lambda}} - 1$

$\Rightarrow \hat{\lambda} = -n$

$\Rightarrow \hat{p}_i = \frac{x_i}{n}$

I.e. $(\hat{p}_1, \ldots, \hat{p}_k) = \left(\frac{x_1}{n}, \ldots, \frac{x_k}{n}\right)$

# Maximum likelihood in the Multinomial model

The formal reasoning behind this Lagrange multiplier method is:

- Fixing any $\lambda \in \mathbb{R}$, maximizing $\ell_n(p_1, \ldots, p_k)$ subject to the constraint $p_1 + \ldots + p_k = 1$ is the same as maximizing $L_n(p_1, \ldots, p_k, \lambda)$ subject to this constraint, because the additional term in $L_n(p_1, \ldots, p_k, \lambda)$ is 0.
- If we instead ignore the constraint $p_1 + \ldots + p_k = 1$, then the unconstrained maximizer of $L_n(p_1, \ldots, p_k, \lambda)$ for each $\lambda$ is $(p_1, \ldots, p_k) = -(1/\lambda)(X_1, \ldots, X_k)$ as we computed.
- Using the specific choice $\lambda = -n$, this unconstrained maximizer satisfies the constraint $p_1 + \ldots + p_k = 1$, so it must also be the constrained maximizer of $L_n(p_1, \ldots, p_k, \lambda)$.

Combining these three statements, $(p_1, \ldots, p_k) = (X_1, \ldots, X_k)/n$ is the constrained maximizer of $\ell_n(p_1, \ldots, p_k)$.

# The Hardy-Weinberg model

In genetics, it is often assumed that the genotypes AA, Aa, and aa at a single locus satisfy *Hardy-Weinberg equilibrium* — they occur with probabilities $(1-p)^2$, $2p(1-p)$, and $p^2$, where $p \in [0,1]$ represents the frequency of the minor allele. Thus the counts of these three genotypes in $n$ samples may be modeled as

$$(X_1, X_2, X_3) \sim \text{Multinomial}\left(n, ((1-p)^2, 2p(1-p), p^2)\right)$$

$$\ell_n(p) = \log\left[\binom{n}{X_1 X_2 X_3} \cdot ((1-p)^2)^{X_1} (2p(1-p))^{X_2} (p^2)^{X_3}\right]$$

$$= \log\binom{n}{X_1 X_2 X_3} + (2X_1 + X_2)\log(1-p) + (2X_3 + X_2)\log p + X_2 \log 2$$

$$\Rightarrow 0 = \ell_n'(\hat{p}) = \frac{2X_1 + X_2}{1-\hat{p}} + \frac{2X_3 + X_2}{\hat{p}} \Rightarrow \hat{p} = \frac{2X_3 + X_2}{2n}$$

# Gradient ascent

Computing the MLE is an optimization problem: For $\theta \in \mathbb{R}$, we wish to maximize $\ell_n(\theta)$, or to solve the *score equation* $0 = \ell_n'(\theta)$.

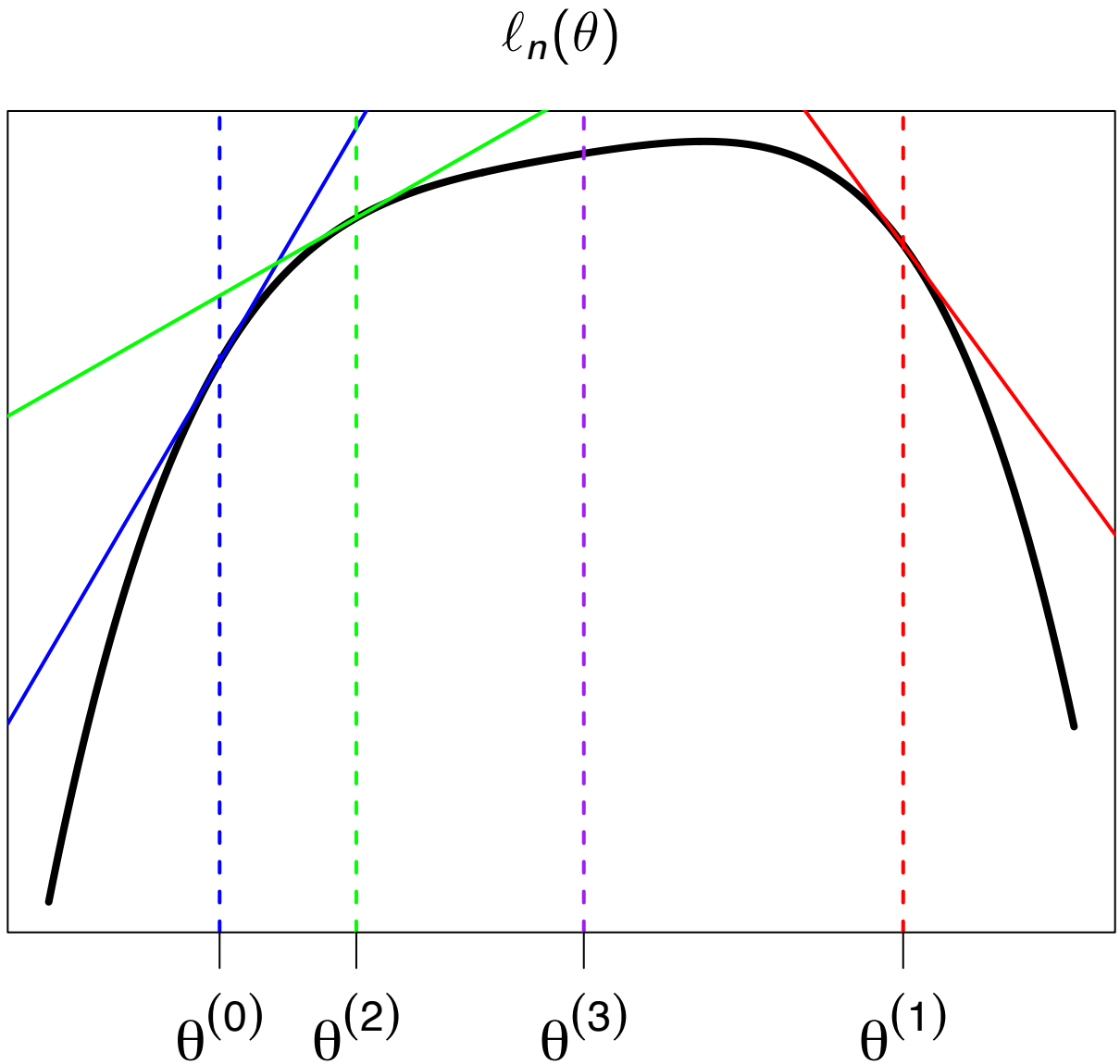
When there is no explicit solution, we oftentimes use numerical optimization procedures. The simplest procedure is **gradient ascent**: Starting with an initial guess $\theta^{(0)}$, iterate

$$\theta^{(t+1)} = \theta^{(t)} + \eta\, \ell_n'(\theta^{(t)})$$

where $\eta > 0$ is a learning rate parameter.

When $0 = \ell_n'(\theta^{(t)})$, this gives $\theta^{(t+1)} = \theta^{(t)}$.

# Gradient ascent



# Newton's method

A second-order procedure is **Newton's method**: Given $\theta^{(t)}$, approximate the solution to $0 = \ell_n'(\theta)$ by a Taylor expansion

$$0 = \ell_n'(\theta) \approx \ell_n'(\theta^{(t)}) + \ell_n''(\theta^{(t)})(\theta - \theta^{(t)}).$$

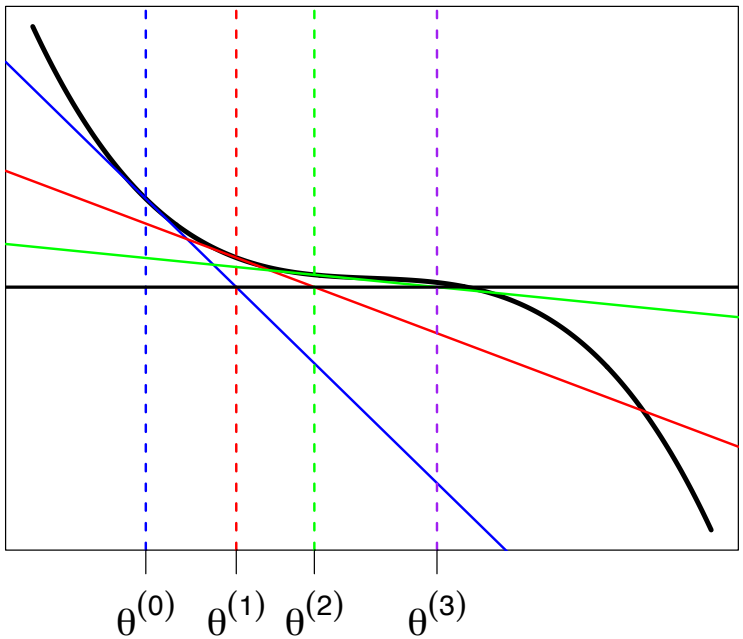
Solve this equation in $\theta$ and set this as $\theta^{(t+1)}$:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell_n'(\theta^{(t)})}{\ell_n''(\theta^{(t)})}$$

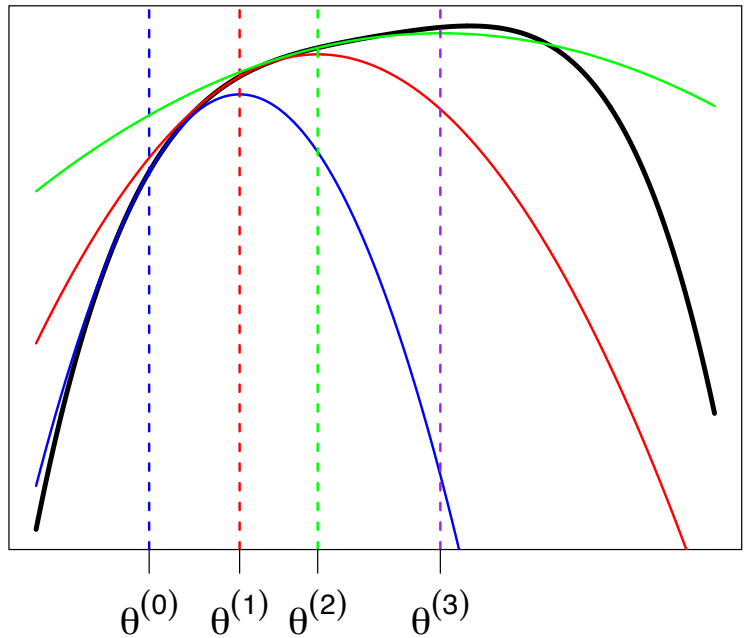
When $0 = \ell_n'(\theta^{(t)})$, again this gives $\theta^{(t+1)} = \theta^{(t)}$.

# Newton's method

$\ell'_n(\theta)$




$\ell_n(\theta)$



# Optimization in higher dimensions

For $\theta \in \mathbb{R}^k$, gradient ascent is the procedure

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla \ell_n(\theta^{(t)})$$

where $\nabla \ell_n(\theta) \in \mathbb{R}^k$ is the *gradient* of $\ell_n(\theta)$, i.e. the vector of its 1st-order partial derivatives.

Newton's method is the procedure

$$\theta^{(t+1)} = \theta^{(t)} - [\nabla^2 \ell_n(\theta^{(t)})]^{-1} \nabla \ell_n(\theta^{(t)})$$

where $\nabla^2 \ell_n(\theta) \in \mathbb{R}^{k \times k}$ is the *Hessian* of $\ell_n(\theta)$, i.e. the matrix of its 2nd-order partial derivatives.