Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

Rappels #5 (optimization)

Calculus

Given a smooth function $f : \mathbb{R}^n \to \mathbb{R}$

Its gradient, $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is

$$\nabla f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \cdots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right)$$

Its Hessian matrix is $H = \nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n\times n}$

$$H(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\mathbf{x}^\top} = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

Classical rules for differentiable $\mathbb{R} \to \mathbb{R}$ functions

- $h(x) = \alpha f(x) + \beta g(x)$, $h'(x) = \alpha f'(x) + \beta g'(x)$,
- $h(x) = f(x)g(x)$, $h'(x) = f'(x)g(x) + f(x)g'(x)$
- $h(x) = f\big(g(x)\big)$, $h'(x) = f'\big(g(x)\big)g'(x)$
- $h = f^{-1}$, $h'(y) = \dfrac{1}{f'(h^{-1}(y))}$
- $h(x) = f(x)^n$, $h'(x) = nf'(x)f(x)^{n-1}$
- $h(x) = \dfrac{f(x)}{g(x)}$, $h'(x) = \dfrac{f'(x)g(x) - f(x)g'(x)}{g(x)}^2$,
- $h(x) = \log[f(x)]$, $h'(x) = \dfrac{f'(x)}{f(x)}$

Let $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^\top \mathbf{x} = \sum_{i=1}^{n} a_i x_i$, $\dfrac{\partial \mathbf{a}^\top \mathbf{x}}{\partial x_i} = a_i$

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial x_1}, \frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial x_2}, \cdots, \frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial x_n}\right) = (a_1, a_2, \cdots, a_n) = \mathbf{a}^\top$$

More generally, for multivariate linear or quadratic functions,

- $\dfrac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$ analogous of: if $f(x) = ax$, $f'(x) = a$
- $\dfrac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}^\top$
- $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top \left(\mathbf{A} + \mathbf{A}^\top\right)$ if $f(x) = ax^2$, $f'(x) = 2ax$
- $\dfrac{\partial^2 \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}\partial \mathbf{x}^\top} = \mathbf{A} + \mathbf{A}^\top$ analogous of: if $f(x) = ax^2$, $f''(x) = 2a$
- $\dfrac{\partial \langle \mathbf{x}, \mathbf{x}\rangle}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = \dfrac{\partial \|\mathbf{x}\|^2}{\partial \mathbf{x}} = 2\mathbf{x}^\top$

Optimisation: continuous (differentiable)

The problem is to solve $\min_{y\in\mathbb{R}}\{f(y)\}$

Note: $\min_{y\in\mathbb{R}}\{f(y)\} = \max_{y\in\mathbb{R}}\{-f(y)\}$

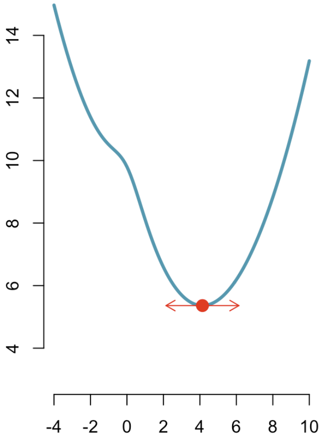
Note: $y^\star \in \underset{y\in\mathbb{R}}{\text{argmin}}\{f(y)\}$

and $\min_{y\in\mathbb{R}}\{f(y)\} = f(y^\star)$.

First order condition

$$f'(y^\star) = \left.\frac{\partial f(y)}{\partial y}\right|_{y=y^\star} = 0$$

(necessary condition)



Optimisation: continuous (differentiable)

First order condition

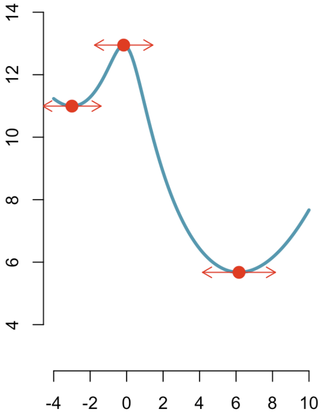
$$f'(y^\star) = \left.\frac{\partial f(y)}{\partial y}\right|_{y=y^\star} = 0$$

might be not sufficient

$$f''(y^\star) = \left.\frac{\partial^2 f}{\partial y^2}\right|_{y=y^\star} > 0: \text{ minimum}$$

$$f''(y^\star) = \left.\frac{\partial^2 f}{\partial y^2}\right|_{y=y^\star} < 0: \text{ maximum}$$

can be a local minimum...



Optimisation: continuous (differentiable)

Example: $\{y_1, \cdots, y_n\}$ in $\mathbb{R}$, let

$$f(y) = \sum_{i=1}^{n} (y_i - y)^2$$

$$\frac{\partial f(y)}{\partial y} = \frac{\partial}{\partial y} \sum_{i=1}^{n} (y_i - y)^2 = \sum_{i=1}^{n} \frac{\partial (y_i - y)^2}{\partial y} = \sum_{i=1}^{n} -2(y_i - y)$$

so

$$\left.\frac{\partial f(y)}{\partial y}\right|_{y=y^\star} = 0 \text{ if and only if } \sum_{i=1}^{n} (y_i - y^\star) = 0 \text{ or } \sum_{i=1}^{n} y_i = n y^\star$$

i.e. $y^\star = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$.

Optimisation: continuous (differentiable)

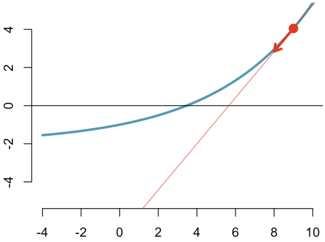
Solving $f'(y^\star) = 0$ numerically
Newton's method: solve $g(y^\star) = 0$

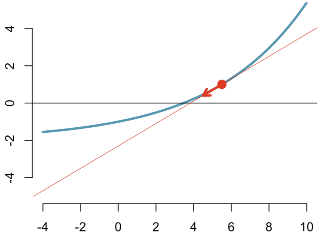
$$g(y) \simeq g(y_0) + g'(y_0)(y - y_0)$$

If $g(y) \simeq 0$, $g(y_0) + g'(y_0)(y - y_0) \simeq 0$

Start from y_0, then

$$y_{k+1} = y_k - \frac{g(y_k)}{g'(y_k)}$$



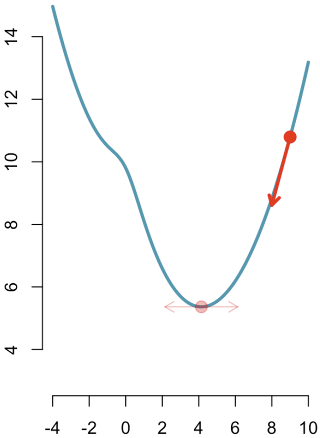


Optimisation: continuous (differentiable)

To solve $f'(y^\star)=0$ numerically
Start from y_0, then

$$y_{k+1}=y_k-\frac{f'(y_k)}{f''(y_k)}$$

$f'(y_k)$ gives the direction
$f''(y_k)$ gives the speed of convergence
(close to the minimum $f''(y_k)>0$)

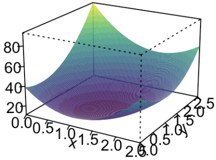


Optimisation: continuous (differentiable)

```
> v = c(0.89367,-1.04729,1.97133,-0.38363,1.65414)
> mean(v)
[1] 0.617644
> f = function(x) sum((v-x)^2)
> optim(0, f)
$par
[1] 0.6175781
$value
[1] 6.757535
```

Optimisation: continuous (differentiable)

The problem is $\min_{\mathbf{y}\in\mathbb{R}^p}\{f(\mathbf{y})\}$



or $\min_{(y_1,\cdots,y_p)\in\mathbb{R}^p}\{f(y_1,\cdots,y_p)\}$

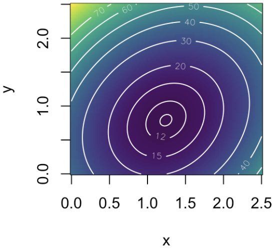
First order conditions: $\nabla f(\mathbf{y}^\star)=\mathbf{0}$,

$$\left.\frac{\partial f(y_1,y_2,\cdots,y_p)}{\partial y_1}\right|_{\mathbf{y}=\mathbf{y}^\star}=0$$

$$\left.\frac{\partial f(y_1,y_2,\cdots,y_p)}{\partial y_2}\right|_{\mathbf{y}=\mathbf{y}^\star}=0$$

$$\vdots$$

$$\left.\frac{\partial f(y_1,y_2,\cdots,y_p)}{\partial y_p}\right|_{\mathbf{y}=\mathbf{y}^\star}=0$$



Optimisation: continuous (differentiable)

Example: $\{(x_1,y_1),\cdots,(x_n,y_n)\}$ in $\mathbb{R}^2$, let

$$f(a,b)=\sum_{i=1}^n (y_i-[a+bx_i])^2$$

$$\frac{\partial f(a,b)}{\partial a}=-2\sum_{i=1}^n (y_i-[a+bx_i])=-2\Big(n\overline{y}-[a+bn\overline{y}]\Big)$$

$$\frac{\partial f(a,b)}{\partial b}=-2\sum_{i=1}^n (y_i-[a+bx_i])x_i$$

$$\left.\frac{\partial f(a,b)}{\partial a}\right|_{(a,b)=(a^\star,b^\star)}=0 \text{ means that } \overline{y}=a^\star+b^\star\overline{x},$$

$$\left.\frac{\partial f(a,b)}{\partial b}\right|_{(a,b)=(a^\star,b^\star)}=0 \text{ means that } \widehat{\boldsymbol{\varepsilon}}\perp\mathbf{x},\ \widehat{\varepsilon}_i=y_i-[a^\star+b^\star x_i],$$

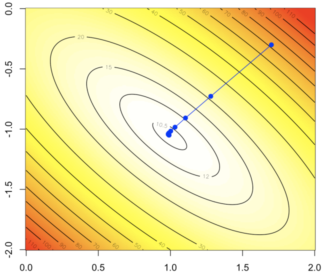
Optimisation: continuous (differentiable)

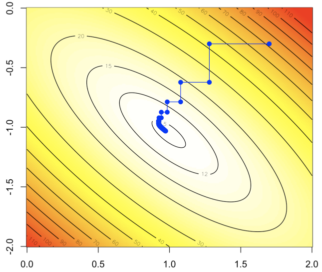
To solve $\nabla f(\mathbf{y}^\star) = \mathbf{0}$ numerically
Start from $\mathbf{y}_0$, then

$$\mathbf{y}_{k+1} = \mathbf{y}_k - \mathbf{H}_k^{-1} \nabla f(\mathbf{y}_k)$$

$\nabla f(\mathbf{y}_k)$ gives the direction
$\mathbf{H}_k$ gives the speed of convergence
$\mathbf{H}_k^{-1}$ is the inverse of the Hessian matrix

One could also consider some numerical tricks, see coordinate descent
where we iterate on the dimension
(univariate optimisation problems)





The problem is $\displaystyle\min_{(x,y)\in\mathbb{R}^2}\{f(x,y)\}$ subject to $g(x,y)\leq 0$,

or $\displaystyle\min_{(x,y)\in\mathbb{R}^2}\{f(x,y)\}$ subject to $g(x,y)=0$.

$f(x,y)$ is the objective function
$g(x,y)$ is the constraint.

The trick is to consider the Lagrangian,

$$\mathcal{L}(x,y,\lambda)=f(x,y)+\lambda g(x,y)$$

Constrained Optimisation: Lagrangian

The optimization problem becomes

$$\min_{(x,y,\lambda)} \{\mathcal{L}(x,y,\lambda)\}$$

The first order conditions are now

$$\frac{\partial \mathcal{L}(x^\star,y^\star,\lambda^\star)}{\partial x} = \frac{\partial f(x^\star,y^\star)}{\partial x} + \lambda^\star \frac{\partial g(x^\star,y^\star)}{\partial x} = 0$$

$$\frac{\partial \mathcal{L}(x^\star,y^\star,\lambda^\star)}{\partial y} = \frac{\partial f(x^\star,y^\star)}{\partial y} + \lambda^\star \frac{\partial g(x^\star,y^\star)}{\partial y} = 0$$

$$\frac{\partial \mathcal{L}(x^\star,y^\star,\lambda^\star)}{\partial \lambda} = g(x^\star,y^\star) = 0$$

Constrained Optimisation: Lagrangian

Interpretation: the ratios of the partial derivatives are all equal, and equal to $-\lambda$,

$$-\lambda = \frac{\partial f(x^\star, y^\star)/\partial x}{\partial g(x^\star, y^\star)/\partial x} = \frac{\partial f(x^\star, y^\star)/\partial y}{\partial g(x^\star, y^\star)/\partial y}$$

(ratios of marginal benefit to marginal cost are all equals)

Note: duality in the optimization problem

Primal problem, $\min_{(x,y)\in\mathbb{R}^2} \{f(x,y)\}$ subject to $g(x,y) = 0$

Dual problem, $\max_{(x,y)\in\mathbb{R}^2} \{g(x,y)\}$ subject to $f(x,y) = f^\star$

Constrained Optimisation: Lagrangian

Example : $\{(x_{1,1}, x_{2,1}, y_1), \cdots, (x_{1,n}, x_{2,n}, y_n)\}$ in $\mathbb{R}^3$, let

$$f(b_1, b_2) = \sum_{i=1}^{n} (y_i - [b_1 x_{1,i} + b_2 x_{2,i}])^2$$

Goal: find $\displaystyle\min_{(b_1,b_2)\in\mathbb{R}^2} \{f(b_1, b_2)\}$ subject to $b_1^2 + b_2^2 \leq s$

or $\displaystyle\min_{\mathbf{b}\in\mathbb{R}^2} \{f(\mathbf{b})\}$ subject to $\|\mathbf{b}\|^2 \leq s$ (see Ridge regression)

The Lagrangian is

$$\mathcal{L}(b_1, b_2, \lambda) = \sum_{i=1}^{n} (y_i - [b_1 x_{1,i} + b_2 x_{2,i}])^2 + \lambda(b_1^2 + b_2^2 - s)$$

$$\frac{\partial \mathcal{L}(b_1, b_2, \lambda)}{\partial b_j} = -2 \sum_{i=1}^{n} x_{j,i}(y_i - [b_1 x_{1,i} + b_2 x_{2,i}]) + 2\lambda b_j$$

$$\frac{\partial \mathcal{L}(b_1, b_2, \lambda)}{\partial \lambda} = (b_1^2 + b_2^2 - s)$$

Constrained Optimisation: Lagrangian

To go further... using matrix notations,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{2,1} \\ \vdots & \vdots \\ x_{1,n} & x_{2,n} \end{pmatrix}$$

The solution of $\min_{\mathbf{b}\in\mathbb{R}^2}\{f(\mathbf{b})\}$ with $f(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^\top(\mathbf{y} - \mathbf{X}\mathbf{b})$
is $\mathbf{b}^\star = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

The solution of $\min_{\mathbf{b}\in\mathbb{R}^2}\{f(\mathbf{b})\}$ subject to $\|\mathbf{b}\|^2 \leq s$
is $\mathbf{b}^\star = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^\top\mathbf{y}$ for some $\lambda > 0$.