Unconstrained Problem and Algorithm I

Jong-June Jeon

October 11, 2023

Department of Statistics, University of Seoul

- Golden Section algorithm
- Gradient descent algorithm
- Newton-Raphson algorithm

Definition 1 (Unconstrained optimization problem)

$$\text{minimize}_{x \in \text{dom}(f)} f(x)$$

Golden section algorithm

Minimization method for a continuous function f on $\mathbb{R}$

(1) Set an interval $[a_0, b_0]$.

(2) Set two points $c_1 < c_2$ in the interval.

(3) Evaluate $f(c_1)$ and $f(c_2)$

(4) If $f(c_1) < f(c_2)$ then drop interval $(c_2, b_0]$ and denote a_0 and c_2 by a_1 and b_1.

(5) If $f(c_1) \geq f(c_2)$ then drop interval $[a_0, c_1)$ and denote c_1 and b_0 by a_1 and b_1.

(6) repeat (2)-(5) until the length of intervals becomes less than the predetermined precision level.

Idea of Golden section algorithm

- First, choose c_1 as an approximation of the minimizer in $[a_0, b_0]$.
- Second, choose c_2 in $[a_0, b_0]$. Suppose that $c_1 < c_2$.
- If $f(c_1) < f(c_2)$, then c_2 becomes a new right limit of the range containing a minimizer.
- If $f(c_1) > f(c_2)$, then c_2 becomes a new approximation of the minimizer. In addition, c_1 becomes a new left limit of the range containing a minimizer.

Here, we ignore the optimal selection of c_1 and c_2.

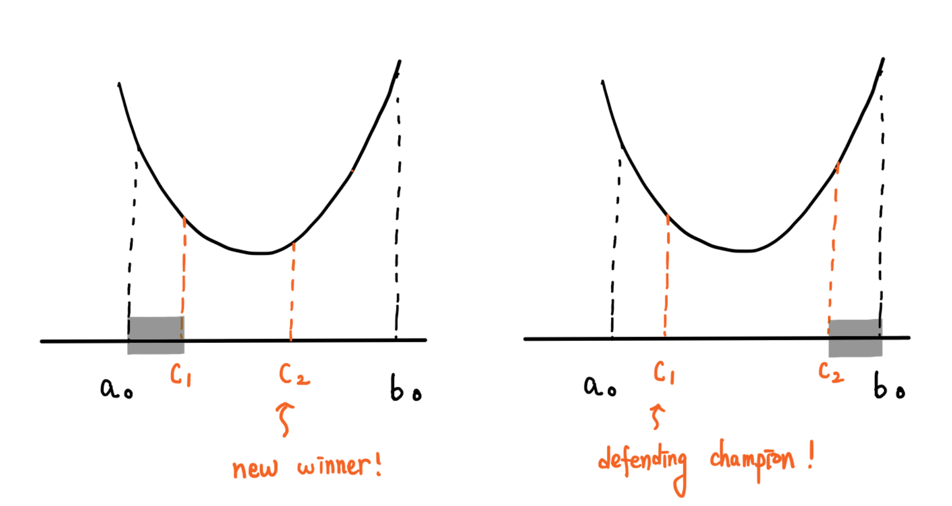


Figure 1: (Golden) search algorithm

Definition 2 (Descent Method)

Consider the update rule:

- Set a current solution $x \in \mathbb{R}^p$
- Set an updating direction $u \in \mathbb{R}^p$ and update the next solution by

$$x^+ = x + \eta u$$

  for a positive learning rate η.

If there exist $\eta > 0$ and $u \in \mathbb{R}$ such that $f(x^+) - f(x) < 0$, then we say that the algorithm is a descent method for minimizing f.

Proposition 1 (Descent method for convex functions)

Suppose that f is differentiable and convex. If an algorithm is a descent method, then it is necessary that $u^\top \nabla f(x) < 0$.

(proof) By convexity and differentiability of f

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

for all x and y. Replacing y with x^+ and write the inequality in terms of η and u, then

$$\frac{f(x + \eta u) - f(x)}{\eta} \geq \nabla f(x)^\top u$$

Assume that for some u, the left side is strictly less than 0, then necessarily $\nabla f(x)^\top u < 0$.

Gradient Descent Algorithm

(1) Set $t = 0$ and an initial value $x^{(t)}$.

(2) Obtain $\nabla f(x^{(t)})$ and set

$$x^{(t+1)} = x^{(t)} - \eta_t \nabla f(x^{(t)})$$

for $\eta_t > 0$

(3) $t \to t + 1$ and repeat (2) until the solution converges.

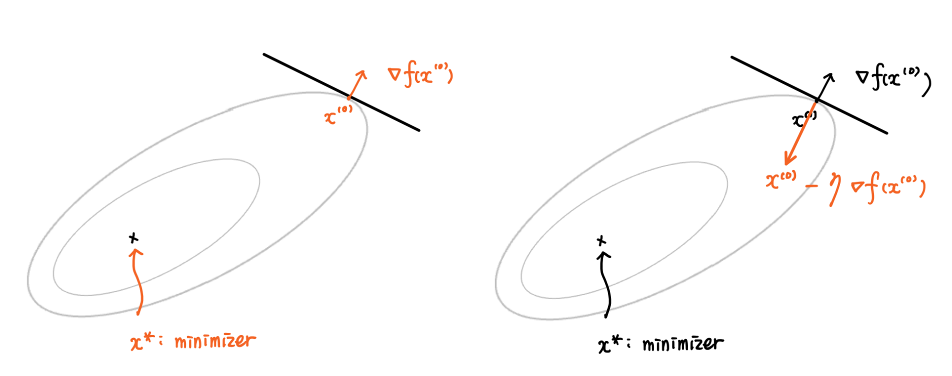


Figure 2: Gradient descent algorithm

Gradient Descent Method: the first order approximation

$$f(x) \simeq f(x^{(t)}) + \nabla f(x^{(t)})^\top (x - x^{(t)}),$$

which is a locally approximated function. The GD updates the current solution with the direction of decreasing the value of the approximated linear function.

Coordinate descent algorithm

(1) Set $k = 0$ and let an initial $x^{(k)} \in \mathbb{R}$.

(2) Find the direction $j = \text{argmax}_k \left| \frac{\partial f(x)}{\partial x_k} \right|$

(3) Obtain the solution

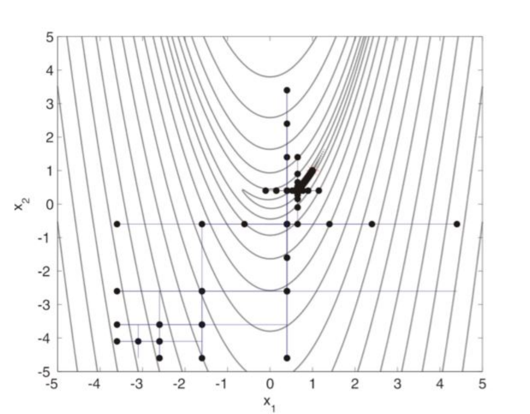
$$\hat{x}_j^{(t+1)} = \text{argmin}_{x \in \mathbb{R}} f(x_1^{(t)}, \cdots, x_{j-1}^{(t)}, x, x_{j+1}^{(t)}, \cdots, x_p^{(t)})$$

and let

$$x^{(t+1)} = (x_1^{(t)}, \cdots, x_{j-1}^{(t)}, x_j^{(t+1)}, x_{j+1}^{(t)}, \cdots, x_p^{(t)})$$

(4) $t \to t + 1$ and repeat (2) until the solution converges.

Coordinate descent algorithm



Optimality function for the strictly convex function

Assume that $f : \mathbb{R} \to \mathbb{R}$ is strictly convex and differentiable. If x^* satisfies $f'(x^*) = 0$, then x^* is the unique minimizer.

Therefore, it is sufficient to solve the equation $\nabla f(x^*) = 0$ for obtaining the minimizer of f.

The Newton-Raphson method is an algorithm to solve the nonlinear equations. We let the estimating equation be $\nabla f(x) = 0$ for $x \in \mathbb{R}$, and assume that ∇f is differentiable. Then, the Newton-Raphson algorithm is following:

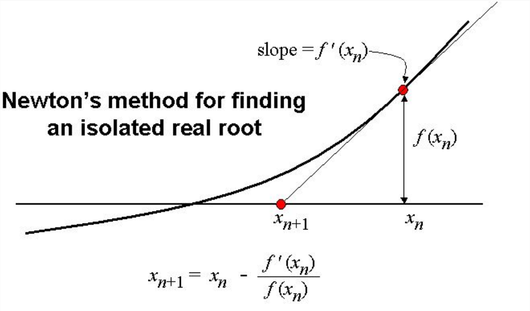
Newton-Raphson method on $\mathbb{R}$

(1) Set $t = 0$ and an initial value $x^{(t)}$.

(2) Obtain $x^{(t+1)}$ which is a solution of $\nabla^2 f(x^{(t)})(x - x^{(t)}) + \nabla f(x^{(t)}) = 0$

(3) $t \to t + 1$ and repeat (2) until the solution converges.

Under some conditions, the convergence of the solution is proved. The Newton-Raphson method is illustrated in figure 1.



Second-order approximation

Let the objective function be $f : \mathbb{R}^p \mapsto \mathbb{R}$. Set an initial solution $x^{(k)}$ for $k = 0$ and consider the second order approximation of $f(x)$ at $x^{(k)}$.

$$
\begin{aligned}
f(x) \simeq Q(x; x^{(k)}) \quad = \quad & f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) \\
& + \frac{1}{2}(x - x^{(k)})^\top \nabla^2 f(x^{(k)})(x - x^{(k)}).
\end{aligned}
$$

The function $Q(x; x^{(k)})$ is a quadratic function. Investigate the minimizer of Q.

First order approximation of $\nabla f(x)$

$$\nabla f(y) \simeq \nabla f(x) + \nabla^2 f(x)(y-x)$$

Thus,

$$\nabla f(x+s) \simeq \nabla f(x) + \nabla^2 f(x)s$$

Second-order approximation: minimizer of $Q(x; x^{(k)})$

$$\frac{\partial Q(x; x^{(k)})}{\partial x} = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)}).$$

Let $\nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)}) = 0$. The minimizer of Q is given by

$$x = x^{(k)} - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}),$$

which is an equal procedure in the Newton-Raphson method.

contour plot

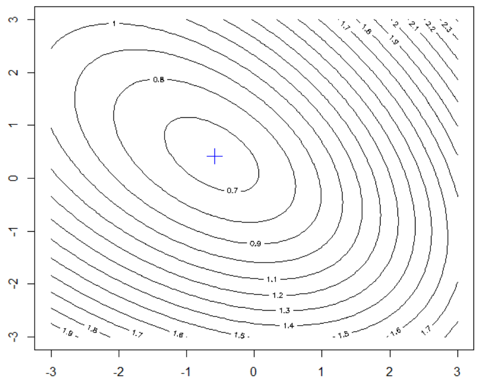


Figure 3: contour plot of an $l(x)$

set an initial

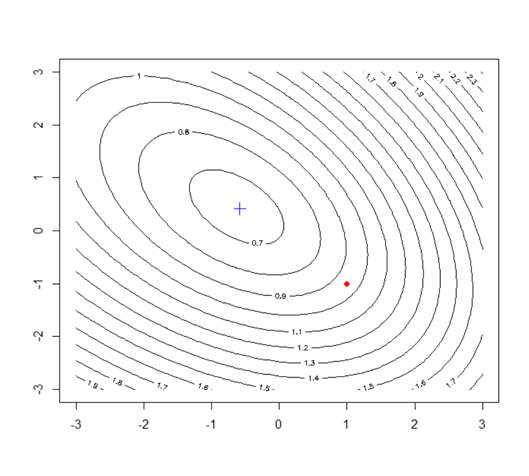


Figure 4: contour plot of an $l(x)$

Quadratic approximation and updating solution

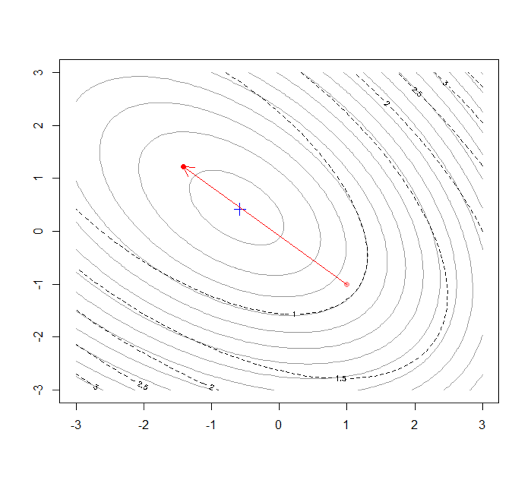


Figure 5: dashed curve is the contour of quadratic function approximated at the initial points

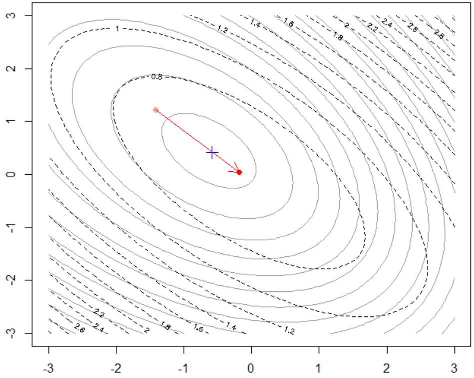


Figure 6: dashed curve is the contour of quadratic function approximated at the updated points

Newton-Raphson algorithm

(1) Set $t = 0$ and an initial value $x^{(t)} \in \mathbb{R}^p$.

(2) compute the gradient $\nabla f(x^{(t)})$ and the Hessian $H = \nabla^2 f(x^{(t)})$.

(3)

$$x^{(t+1)} \leftarrow x^{(t)} - H^{-1} \nabla f(x^{(t)})$$

where H is hessian matrix of F.

(4) $t \to t + 1$ and repeat (2) until the solution converges.

Note

The drawbacks of a second-order approximation method are as follows: 1) The computation of the Hessian matrix requires substantial computational resources, and 2) the second-order approximation may be inaccurate when an initial value is far away from the optimal solution, leading to convergence issues. For these reasons, corrective methods for improving the accuracy of the solution are often employed.

- Trust region method
- Line search method
- Backtracking method

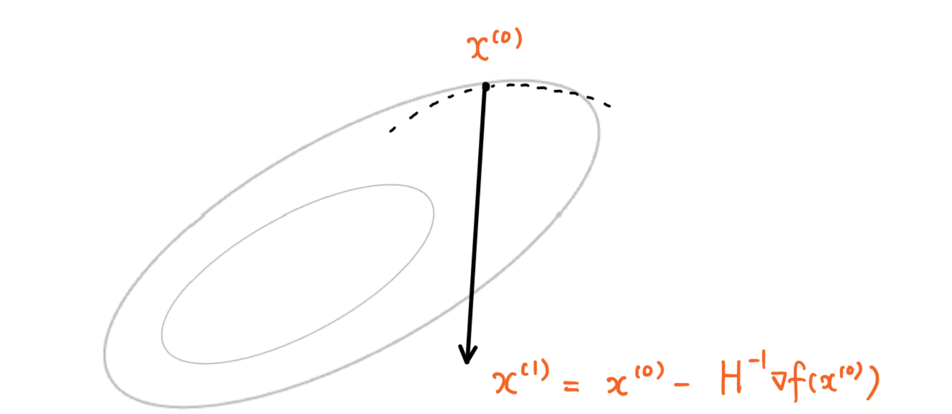


Figure 7: Failed step in Newton- Raphason algorithm

Trust region

Let $x^{(t)}$ be a current solution and let $x^{(t)} + \delta^{(t)}$ be the updated solution. We approximate the value of the objective function at the updated solution by

$$f(x^{(t)} + \delta) \simeq f(x^{(t)}) + \nabla f(x^{(t)})^\top \delta + \frac{1}{2}\delta^\top H \delta \equiv q(\delta)$$

and let $\delta^{(t)} = -H^{-1}\nabla f(x^{(t)})$, the minimizer of $q(\delta)$, then this algorithm becomes the Newton algorithm.

Trust region

If the norm of $\delta^{(t)}$ is too large, we may be concerned with the approximation of $q(\delta)$ to $f(x^{(t)}+\delta)$. We expect $q(\delta)$ to be close to $f(x^{(t)} + \delta)$ (in fact, this is the reason why we minimize $q(\delta)$) but we cannot trust the approximation anymore for such large $\delta^{(t)}$.

Trust region

Instead, we constrain the norm of δ:

$$\begin{aligned} \delta &= \text{argmin}_\delta \; q(\delta) \\ &\quad \text{subject to } \|\delta\|_2^2 \leq \gamma_t^2 \end{aligned}$$

This is the l_2 shrinkage to prevent too large δ, and it is known that the shrinkage is the eigenvalue regularization of H in $q(\delta)$.

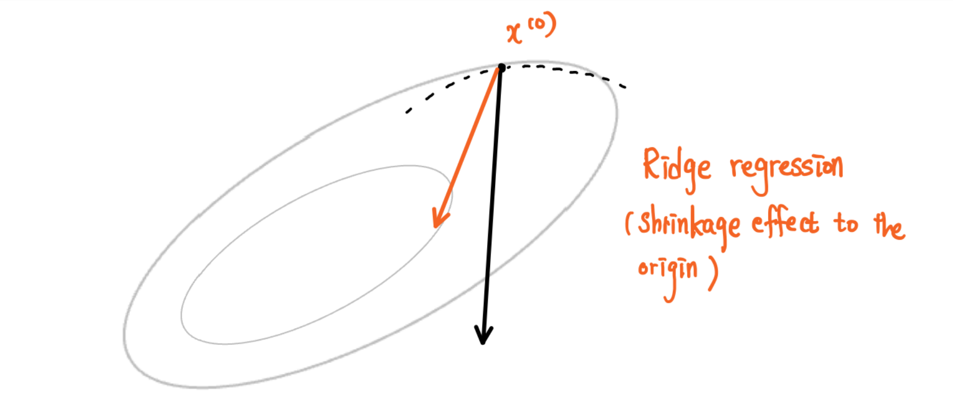


Figure 8: Trust region method

Line Search

Let $\delta^{(t)} = -H^{-1}\nabla f(x^{(t)})$ and find the minimizer

$$\alpha^* = \text{argmin}_\alpha f(x^{(t)} + \alpha\delta^{(t)}).$$

Update the solution by

$$x^{(t+1)} = x^{(t)} + \alpha^*\delta^{(t)}$$

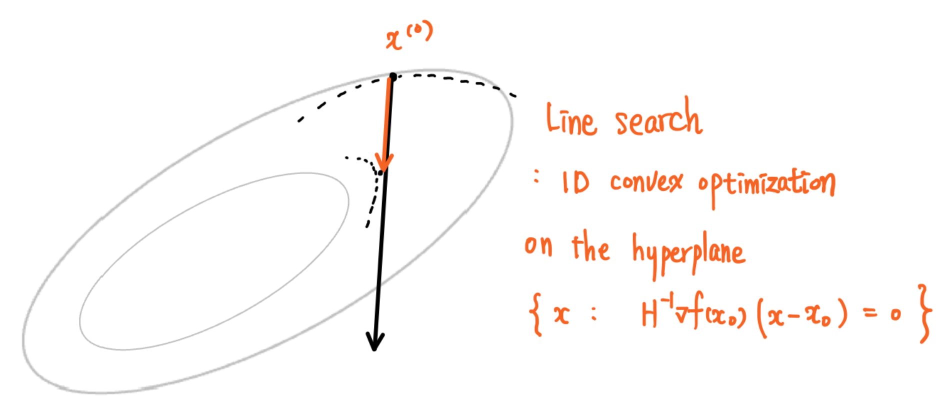


Figure 9: Line search method

Backtracking

A kind of inexact line search

Minimize

$$f(x^{(t)} + \alpha\delta^{(t)})$$

for $\alpha = 1, \tau, \tau^2, \cdots$

Example 3 (Logistic regression model)

- $y \in \{0, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^p$
- $y|\mathbf{x} \sim \text{Bernoulli}(\theta(\mathbf{x}_i; \boldsymbol{\beta}))$ where

$$\theta(\mathbf{x}; \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^\top \boldsymbol{\beta})}.$$

- Let $(y_i, \mathbf{x}_i)$ for $i = 1, \cdots, n$ be independent random samples, then the negative loglikelihood function is given by

$$l(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} [-y_i \mathbf{x}_i^\top \boldsymbol{\beta} + \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}))]$$

The partial derivative of $l(\beta)$ is given by

$$\begin{aligned}\frac{\partial}{\partial \beta_k} l(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big[-y_i x_{ik} + \frac{x_{ik}\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}\Big] \\ &= -\frac{1}{n}\sum_{i} x_{ik}\left(y_i - \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}\right)\end{aligned}$$

for all k.

Since $\hat{y} = \frac{\exp(\mathbf{x}_i^\top \beta)}{1+\exp(\mathbf{x}_i^\top \beta)}$, we can write

$$\frac{\partial}{\partial \beta} l(\beta) = -X^\top (y - \hat{y})/n$$

The hessian matrix H evaluated at β is given by

$$(H)_{jk} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_{ij}x_{ik}\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}))^2}$$

Using matrix notations

$$H = X^\top W X/n$$

where W is the $n \times n$ diagonal matrix whose elements are

$$\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{(1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta}))^2}$$

for $i = 1, \cdots, n$.

Example 4 (Possion regression)

- $y \in \mathbb{Z}_+$ and $\mathbf{x}_i \in \mathbb{R}^p$
- $y|\mathbf{x} \sim$ Poisson$(\theta(\mathbf{x}_i; \boldsymbol{\beta}))$ where

$$\theta(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}^T \boldsymbol{\beta})$$

- Let $(y_i, \mathbf{x}_i)$ for $i = 1, \cdots, n$ be independent random samples, then the negative loglikelihood function is given by

$$l(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] + \log y_i! \right)$$

The estimating equation in the Poisson regression model is given by

$$\frac{\partial}{\partial \beta_k} l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} [y_i x_{ik} - x_{ik} \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] = 0.$$

for all k. Thus $\nabla l(\boldsymbol{\beta}) = -X^\top (Y - \hat{Y})$.

The Hessian matrix H is given by

$$(H)_{jk} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \exp(\mathbf{x}_i^T \boldsymbol{\beta}),$$

and thus $H = X^\top W X / n$, where $W = \text{diag}(\exp(\mathbf{x}_1^\top \boldsymbol{\beta}), \cdots, \exp(\mathbf{x}_n^\top \boldsymbol{\beta}))$

Appendix

Let $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ and denote the image of f by $(f_1(x_1, x_2), f_2(x_1, x_2))$, where $f_j : \mathbb{R}^2 \mapsto \mathbb{R}$

ex) $f(x_1, x_2) = (x_1, x_1^2 + x_2)$

How to write the change of output f according to a small perturbation? $\rightarrow$ Jacobian!

$$J_f(x) = \begin{pmatrix} \frac{\partial f_1(x_1,x_2)}{\partial x_1} & \frac{\partial f_1(x_1,x_2)}{\partial x_2} \\ \frac{\partial f_2(x_1,x_2)}{\partial x_1} & \frac{\partial f_2(x_1,x_2)}{\partial x_2} \end{pmatrix}$$

Appendix

Composition and derivatives

Let $h : \mathbb{R}^p \mapsto \mathbb{R}^q$ and $f : \mathbb{R}^q \mapsto \mathbb{R}$ (f and h are continuously differentiable.)

$$\frac{\partial}{\partial x_1} f(h(x_1, \cdots, x_p))?$$

Let $h(x_1, \cdots, x_p) = (h_1(x_1, \cdots, x_p), \cdots, h_p(x_1, \cdots, x_p))$ where $h_j : \mathbb{R}^p \mapsto \mathbb{R}$. Jacobian of h is given by

$$J_h(x) = \begin{pmatrix} \frac{\partial h_1(x)}{\partial x_1} & \dots & \frac{\partial h_1(x)}{\partial x_p} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_q(x)}{\partial x_1} & \dots & \frac{\partial h_q(x)}{\partial x_p} \end{pmatrix}$$

Composition and derivatives

Let $z = f(u_1, \cdots, u_q)$. By fundamental lemma

$$dz = \frac{\partial z}{u_1}\Delta u_1 + \cdots + \frac{\partial z}{u_q}\Delta u_q.$$

Let $u_j = h_j(x)$ then

$$\begin{aligned}\frac{\partial}{\partial x_1} f(h(x_1, \cdots, x_p)) &= \sum_{j=1}^{q} \frac{\partial f(u)}{\partial u_j} \frac{\partial h_j(x)}{\partial x_1} \\ &= (J_h(x)^\top \nabla f(u))[0, :]\end{aligned}$$

Appendix

Composition and derivatives

$$\frac{\partial f(h)}{\partial x} = J_h(x)^\top \nabla f(u)$$

Jacobian and Hessian

Let $f : x \in \mathbb{R}^p \mapsto \mathbb{R}$ and $\nabla f = (f_1, \cdots, f_p)^\top$, where f_i is the derivative of f. The Hessian matrix is the Jacobian matrix of ∇f.

$$J_{\nabla f}(x) = \frac{\nabla f}{\partial x^\top} = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_p} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_p(x)}{\partial x_1} & \cdots & \frac{\partial f_p(x)}{\partial x_p} \end{pmatrix} = \nabla^2 f$$

The hessian matrix of f represents the change of ∇f according to a small perturbation of x.

Appendix

Directional derivatives

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ and let $v \in \mathbb{R}^p$. Denote $f_j(x) = \frac{\partial f(x)}{\partial x_j}$.

$$\frac{\partial f(x+tv)}{\partial t} = \sum_{j=1}^{p} f_j(x+tv)v_j = \nabla f(x+tv)^\top v$$

Thus, the direction derivatives along to v are equal to $\nabla f(x)^\top v$.

Appendix

Directional derivatives

Conversely $\nabla f(x)^\top v$ is approximated by the direction derivatives along to v:

$$\nabla f(x)^\top v \simeq \frac{f(x+tv)-f(x)}{t}$$

Appendix

Hessian and Directional derivatives

$$(\nabla^2 f)v = J_{\nabla f}(x)v = \begin{pmatrix} \sum_{j=1}^{p} \frac{\partial f_1(x)}{\partial x_j} v_j \\ \vdots \\ \sum_{j=1}^{p} \frac{\partial f_p(x)}{\partial x_j} v_j \end{pmatrix} = \begin{pmatrix} \nabla f_1(x)^\top v \\ \vdots \\ \nabla f_p(x)^\top v \end{pmatrix} \simeq \begin{pmatrix} (f_1(x+tv) - f_1(x))/t \\ \vdots \\ (f_p(x+tv) - f_1(x))/t \end{pmatrix}$$

as $t \to 0$. In summary,

$$(\nabla^2 f)v \simeq t^{-1}(\nabla f(x+tv) - \nabla f(x))$$

as $t \to 0$.