

## Part III: Optimization

III.1: Minimum Problems:

Convexity and Newton's Method

III.2: Lagrange Multipliers

= Derivatives of the Cost

III.3: Linear Programming, Game Theory,  
and Duality

III.4: Gradient Descent Toward the Minimum

III.5: Stochastic Gradient Descent and ADAM

---

The Expression "argmin"

Consider  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$f(x) = (x-1)^2.$$

Then  $\min_{x \in \mathbb{R}} f(x) = 0.$

That is, the minimum value of  $f(x)$   
is 0.

Any  $\bar{x} \in \mathbb{R}$  such that  $f(\bar{x}) = \min_{x \in \mathbb{R}} f(x)$   
is called an optimal solution.

The set of optimal solutions is  
denoted  $\operatorname{argmin}_{x \in \mathbb{R}} f(x)$ .

Since  $\bar{x}=1$  is the only optimal solution  
in this example, we have

$$\operatorname{argmin}_{x \in \mathbb{R}} f(x) = \{1\}.$$

---

## Multivariable Calculus

If  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is twice differentiable,  
then

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

Where  $\nabla f(x)$  is the gradient vector of  $f$   
at  $x$  and  $\nabla^2 f(x)$  is the Hessian matrix  
of  $f$  at  $x$ .

If  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is differentiable, then

$$f(x + \Delta x) \approx f(x) + J(x)\Delta x$$

where  $J(x)$  is the Jacobian matrix of  $f$  at  $x$ .

Here,  $f(x)$  is a vector,

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix},$$

and  $J(x)$  is the matrix of all first order partial derivatives,

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}.$$

MXN  
matrix

For example, let  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  be given by

$$f(x) = \begin{bmatrix} x_1 x_2 \\ \sin x_1 + \cos x_2 \\ x_1^2 e^{x_2} \end{bmatrix}.$$

Then

$$J(x) = \begin{bmatrix} x_2 & x_1 \\ \cos x_1 & -\sin x_2 \\ 2x_1 e^{x_2} & x_1^2 e^{x_2} \end{bmatrix}.$$

---

## VI. I] Minimum Problems:

### Convexity and Newton's Method

Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  and  $K \subseteq \mathbb{R}^n$ .

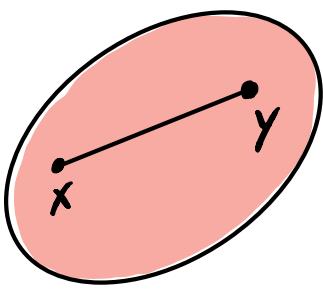
The constrained optimization problem

$\min f(x)$  subject to  $x \in K$

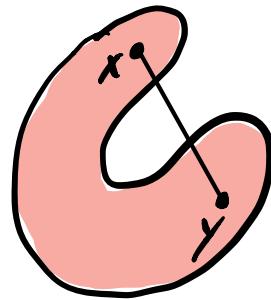
is convex if  $f$  is a convex function and  $K$  is a convex set.

---

A set  $K \subseteq \mathbb{R}^n$  is convex if the line segment between any two points  $x, y \in K$  lies entirely in  $K$ .



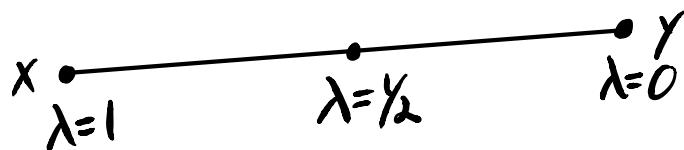
convex



not convex

The line segment between  $x$  and  $y$  can be written as

$$\{\lambda x + (1-\lambda)y : \lambda \in [0, 1]\}$$



Thus, a set  $K \subseteq \mathbb{R}^n$  is convex if  $\lambda x + (1-\lambda)y \in K$  for all  $x, y \in K$  and  $\lambda \in [0, 1]$ .

Examples:

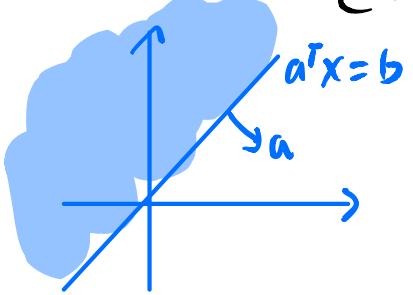
$$(1) \quad K = \{x \in \mathbb{R}^n : Ax = b\}. \quad \begin{matrix} A & m \times n \\ b & m \times 1 \end{matrix}$$

$x, y \in K, \lambda \in [0, 1] \Rightarrow \lambda x + (1-\lambda)y \in K ?$

$$A(\lambda x + (1-\lambda)y) = \lambda Ax + (1-\lambda)Ay = \lambda b + (1-\lambda)b = b$$

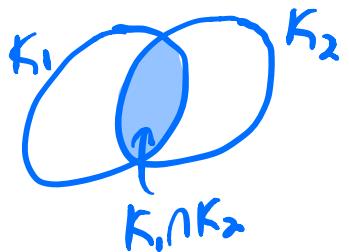
$$\therefore \lambda x + (1-\lambda)y \in K. \therefore K \text{ is convex.}$$

(2)  $K = \{x \in \mathbb{R}^n : a^T x \leq b\}$ . half-space



$$\begin{aligned} a^T(\lambda x + (1-\lambda)y) \\ = \lambda a^T x + (1-\lambda)a^T y \\ \leq \lambda b + (1-\lambda)b = b. \\ \therefore K \text{ is convex.} \end{aligned}$$

(3)  $K = K_1 \cap K_2$ ,  $K_1$  and  $K_2$  convex.

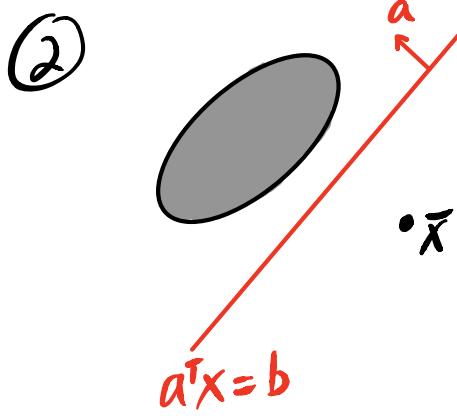
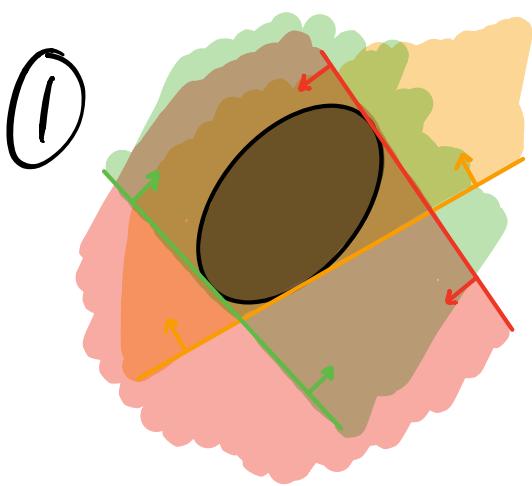


$$\begin{aligned} x, y \in K, \lambda \in [0, 1] \\ x, y \in K_1 \text{ and } x, y \in K_2 \\ \lambda x + (1-\lambda)y \in K_1 \text{ and } \lambda x + (1-\lambda)y \in K_2 \\ \therefore \lambda x + (1-\lambda)y \in K_1 \cap K_2 = K. \\ \therefore K \text{ is convex.} \end{aligned}$$

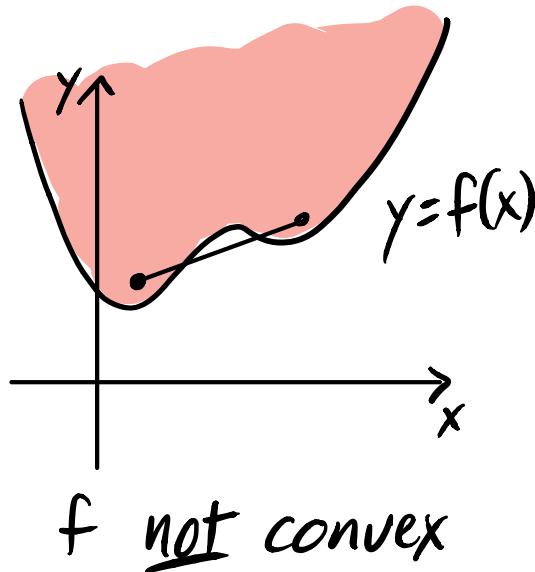
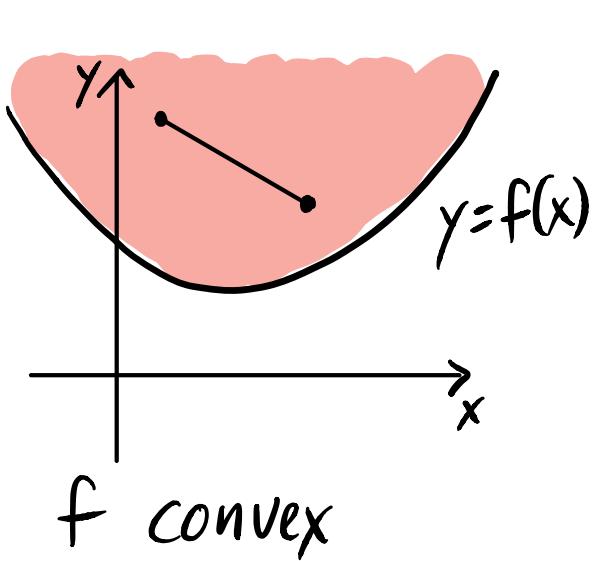
---

Two important properties:

- ① Every closed convex set  $S \subseteq \mathbb{R}^n$  is equal to the intersection of all the half-spaces containing  $S$ .
- ② Let  $S \subseteq \mathbb{R}^n$  be closed and convex and  $\bar{x} \in \mathbb{R}^n$ . If  $\bar{x} \notin S$ , then  $\bar{x}$  can be separated from  $S$  by a hyperplane.



A function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is convex if the set of points on or above the graph of  $f$  is a convex set.

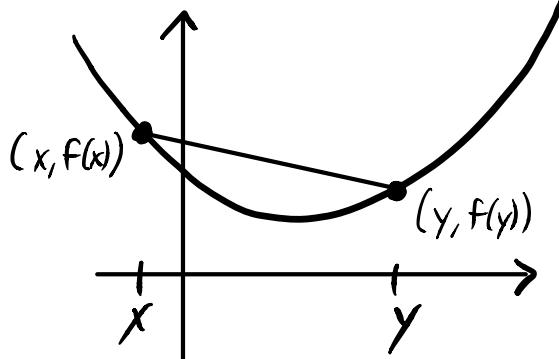


The epigraph of  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is

$$\text{epi}(f) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq y\}.$$

Thus,  $f$  is a convex function iff  $\text{epi}(f)$  is a convex set.

Also, a function  $f$  is convex if the line segment between  $(x, f(x))$  and  $(y, f(y))$  lies on or above the graph of  $f$ .



That line segment is

$$\{(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda)f(y)) : \lambda \in [0,1]\}.$$

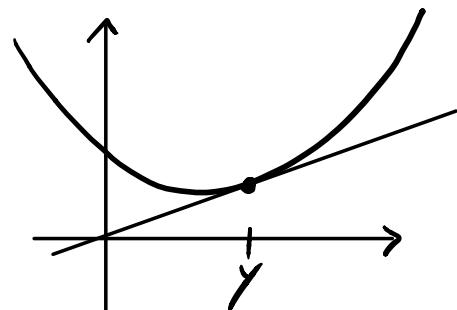
Thus, a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is convex if  $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$  for all  $x, y \in \mathbb{R}^n$  and  $\lambda \in [0,1]$ .

The function is strictly convex if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

for all  $x, y \in \mathbb{R}^n$ , with  $x \neq y$ , and all  $\lambda \in (0,1)$ .

A differentiable convex function must be on or above all of its tangent lines.



Thus,  $f(x) \geq f(y) + \nabla f(y)^T(x-y)$  for all  $x, y \in \mathbb{R}^n$ .

A twice differentiable function is convex iff its Hessian  $\nabla^2 f(x)$  is positive semidefinite for all  $x$ .

If  $\nabla^2 f(x)$  is always positive definite, then  $f$  is strictly convex.

Examples:

$$(1) \quad f(x) = c^T x$$

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= c^T (\lambda x + (1-\lambda)y) \\ &= \lambda c^T x + (1-\lambda)c^T y = \lambda f(x) + (1-\lambda)f(y) \end{aligned}$$

$$(2) \quad f(x) = \frac{1}{2} x^T S x, \quad S \text{ sym. pos. def.}$$

$$\nabla f(x) = Sx$$

$\nabla^2 f(x) = S$  is pos. def.  $\Rightarrow f$  is strictly convex

$$(3) \quad f(x) = x_1^2 + 2x_1x_2 + x_2^2$$

$$\nabla f(x) = \begin{bmatrix} 2x_1 + 2x_2 \\ 2x_1 + 2x_2 \end{bmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$\therefore f$  is convex is pos. semidef.

$$(4) \quad f(x) = x_1^2 - x_2^2$$

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ -2x_2 \end{bmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

$\therefore f$  is not convex is not pos. semidef.

$$(5) \quad f(x) = \|x\| \quad (\text{any vector norm}).$$

$$f(\lambda x + (1-\lambda)y) = \|\lambda x + (1-\lambda)y\|$$

$$\leq \|\lambda x\| + \|(1-\lambda)y\|$$

$$= \lambda \|x\| + (1-\lambda) \|y\|$$

$$= \lambda f(x) + (1-\lambda) f(y).$$

$\therefore f$  is convex

All optimal sol'n's are globally optimal

Suppose  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and  $K \subseteq \mathbb{R}^n$  is convex. If  $\bar{x} \in K$  is a locally optimal sol'n of the convex optimization

$$\min f(x) \text{ s.t. } x \in K$$

then there is no better sol'n in  $K$ .

Pf: Suppose  $\bar{x} \in K$  is a local minimizer of  $f$ . Then all nearby points  $x \in K$  have  $f(x) \leq f(\bar{x})$ . Suppose that  $\exists y \in K$  such that  $f(y) < f(\bar{x})$ . Then

$$\lambda \bar{x} + (1-\lambda)y \in K, \quad \forall \lambda \in [0,1].$$

Also,  $f(\lambda \bar{x} + (1-\lambda)y) \leq \lambda f(\bar{x}) + (1-\lambda)f(y)$

$\therefore \bar{x}$  is not a local min., contradicting our assumption.

$$\begin{aligned} &< \lambda f(\bar{x}) + (1-\lambda)f(\bar{x}) \\ &= f(\bar{x}), \quad \forall \lambda \in [0,1]. \end{aligned}$$

## Newton's Method

Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  be twice continuously differentiable,  $x^* \in \mathbb{R}^n$  be a local minimizer of  $f$  (ie  $\nabla f(x^*) = 0$  and  $\nabla^2 f(x^*)$  pos. def.) and  $x_0 \in \mathbb{R}^n$  be our nearby initial guess. We want to find a point  $x \in \mathbb{R}^n$  that is closer to  $x^*$ .

Two ideas:

- ① Replace  $f(x)$  with its second order Taylor series approximation at the point  $x_0$  and minimize that instead:

$$\left[ \begin{array}{l} \min_x f(x_0) + \nabla f(x_0)^T (x - x_0) \\ \quad + \frac{1}{2} (x - x_0)^T \nabla^2 f(x_0) (x - x_0). \end{array} \right]$$

Let's call that  $g(x)$ . Then

$$\nabla g(x) = \nabla f(x_0) + \nabla^2 f(x_0)(x - x_0).$$

Solving  $\nabla g(x) = 0$ , we obtain:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$x = x_0 - \nabla^2 f(x_0)^{-1} \nabla f(x_0).$$

---

② Linearize the nonlinear system

$$\nabla f(x) = 0$$

by replacing  $\nabla f(x)$  with its first order Taylor series approximation at the point  $x_0$ :

$$\nabla f(x_0) + \nabla^2 f(x_0)(x - x_0) = 0.$$

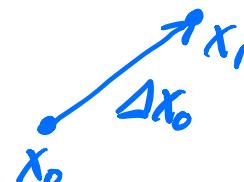
Solving this linear system for  $x$ , we obtain

$$x = x_0 - \nabla^2 f(x_0)^{-1} \nabla f(x_0).$$

---

Thus, we let  $x_1 = x_0 + \Delta x_0$ , where  $\Delta x_0$  is the solution of the linear system:

$$\nabla^2 f(x_0) \Delta x_0 = -\nabla f(x_0).$$



In general, we define  $x_{k+1}$  as

$$x_{k+1} = x_k + \Delta x_k, \quad \nabla^2 f(x_k) \Delta x_k = -\nabla f(x_k).$$

This is Newton's Method for minimizing a function of  $n$  variables. Since it involves the first and second derivatives of  $f$ , it is called a second-order method.

If  $f$  is twice continuously differentiable, and Newton's method is converging to  $x^*$  where  $\nabla f(x^*) = 0$  and  $\nabla^2 f(x^*)$  is pos. def., then Newton's method converges quadratically to  $x^*$ :

$$\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2.$$

## Examples:

① Rosenbrock's banana function:

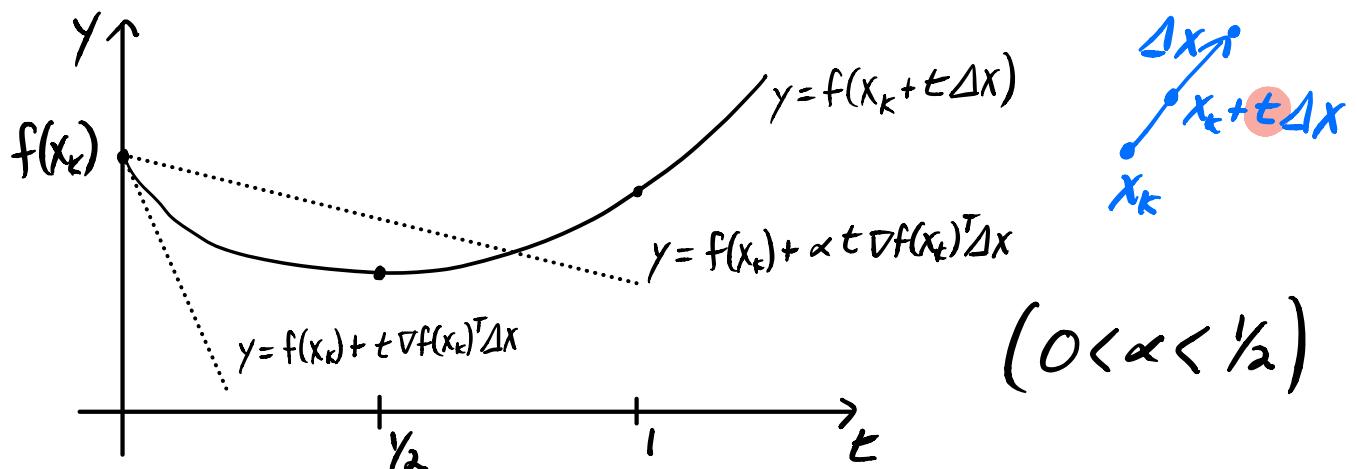
$$f(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$$

② Himmelblau's function:

$$f(x) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

## A practical Newton's Method

To help Newton's method behave better in practice, it is important to include a line search for the step size  $t$ .



Starting with  $t=1$ , we check if the Armijo-Goldstein sufficient decrease condition

$$f(x_k + t\Delta x) \leq f(x_k) + \alpha t \nabla f(x_k)^T \Delta x$$

is satisfied. This is one of the two Wolfe conditions, the second being a slope condition.

If the Armijo-Goldstein condition is not satisfied, we reduce  $t$  by a factor of  $0 < \beta < 1$ , and check the Armijo-Goldstein condition again. We keep doing this until the Armijo-Goldstein condition is satisfied. This is a backtracking line search.

In order to use the backtracking line search, we need to make sure that  $\Delta x$  is a descent direction:

$$\nabla f(x_k)^T \Delta x < 0.$$

If  $\nabla^2 f(x_k)$  is pos. def., then

$$\nabla f(x_k)^T \Delta x = -\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) < 0.$$

However, if  $\nabla^2 f(x_k)$  is not pos. def., then  $\nabla f(x_k)^T \Delta x \geq 0$  is possible. In this case we use the direction of steepest descent:

$$\Delta x = -\nabla f(x_k).$$

If  $\nabla f(x_k) \neq 0$ , then

$$\nabla f(x_k)^T \Delta x = -\|\nabla f(x_k)\|^2 < 0$$

so we are guaranteed to have a descent direction and we can use the backtracking line search.

---

# Levenberg - Marguardt for Nonlinear Least Squares

Suppose we have a nonlinear fitting function  $\hat{y}: \mathbb{R}^n \rightarrow \mathbb{R}^m$  that depends on  $n$  parameters  $P = (P_1, \dots, P_n)$ .

We want to find the parameters  $P$  so that  $\hat{y}(P)$  is as close as possible to the data points  $y = (y_1, \dots, y_m)$ , where the distance/error is measured by the sum of squares loss function:

$$E(P) = \frac{1}{2} \sum_{i=1}^m (y_i - \hat{y}_i(P))^2 = \frac{1}{2} \|y - \hat{y}(P)\|_2^2.$$

We want to solve the nonlinear least squares problem:

$$\min_{P \in \mathbb{R}^n} \frac{1}{2} \|y - \hat{y}(P)\|_2^2.$$

Instead of using Newton's method directly, we approx. the problem with a linear least squares problem by replacing  $\hat{y}(p)$  with its first order Taylor series approximation:

$$\hat{y}(p_k + \Delta p) \approx \hat{y}(p_k) + J(p_k) \Delta p.$$

We find  $\Delta p$  by solving

$$\min_{\Delta p \in \mathbb{R}^n} \frac{1}{2} \| y - \hat{y}(p_k) - J(p_k) \Delta p \|_2^2.$$

The gradient of this objective is

$$J(p_k)^T (y - \hat{y}(p_k) - J(p_k) \Delta p).$$

We want this gradient to be zero, so we solve the linear system

$$J(p_k)^T J(p_k) \Delta p = J(p_k)^T (y - \hat{y}(p_k)).$$

Then  $P_{k+1} = P_k + \Delta p$ . This is the Gauss-Newton method for nonlinear least squares.

When  $P_k$  is far from an optimal  $P^*$ , it is better to use gradient descent:

$$\Delta p = -\nabla E(P_k) = J(P_k)^T(y - \hat{y}(P_k)).$$

The Levenberg - Marquardt method combines Gauss-Newton and gradient descent:

$$(J(P_k)^T J(P_k) + \lambda I) \Delta p = J(P_k)^T(y - \hat{y}(P_k)).$$

Starting with  $\lambda$  large, the direction  $\Delta p$  is more like gradient descent. We increase  $\lambda$  until  $E(p)$  is reduced.

When  $E(p_{k+1})$  is smaller than  $E(p_k)$ , we reduce  $\lambda$ . As  $\lambda$  approaches zero, we obtain the fast convergence of the Gauss-Newton method.

The Levenberg-Marguardt method is equivalent to adding  $\frac{\lambda}{2} \|\Delta p\|_2^2$  to the objective function of the subproblem:

$$\min_{\Delta p \in \mathbb{R}^n} \frac{1}{2} \|y - \hat{y}(p_k) - J(p_k) \Delta p\|_2^2 + \frac{\lambda}{2} \|\Delta p\|_2^2.$$

The gradient of this objective is:

$$-J(p_k)^T (y - \hat{y}(p_k) - J(p_k) \Delta p) + \lambda \Delta p.$$

Setting this to zero and rearranging, we obtain:

$$(J(p_k)^T J(p_k) + \lambda I) \Delta p = J(p_k)^T (y - \hat{y}(p_k)).$$

This is equivalent to solving

$$\begin{aligned} \min \frac{1}{2} \|y - \hat{y}(p_k) - J(p_k) \Delta p\|_2^2 \\ \text{s.t. } \|\Delta p\|_2 \leq \delta \end{aligned}$$

for some  $\delta > 0$ . See also trust region methods, Tikhonov regularization, and Ridge regression.

---

Example:

$$y = a \cos bx + b \sin ax, \quad a=3, \quad b=4$$

$$\hat{y}: \mathbb{R}^2 \rightarrow \mathbb{R}^m \quad (x_1, y_1), \dots, (x_m, y_m)$$

$$\hat{y}(a, b) = \begin{bmatrix} a \cos bx_1 + b \sin ax_1 \\ \vdots \\ a \cos bx_m + b \sin ax_m \end{bmatrix}$$