Part VI: Optimization

VI.1: Minimum Problems:
Convexity and Newton's Method

VI.2: Lagrange Multipliers
= Derivatives of the Cost

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

VI.4: Gradient Descent Toward the Minimum

VI.5: Stochastic Gradient Descent and ADAM

The Expression "argmin"

Consider $f: \mathbb{R} \to \mathbb{R}$ 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 <u>optimal solution</u>.

The <u>set of optimal solutions</u> 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 $f: \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then

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

where $\nabla f(x)$ is the <u>gradient vector</u> of f at x and $\nabla^2 f(x)$ is the <u>Hessian matrix</u> of f at x.

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

$$\boxed{f(x+\Delta x) \approx f(x) + J(x)\Delta x}$$

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,

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

$m \times n$ matrix

For example, let $f: \mathbb{R}^2 \to \mathbb{R}^3$ 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.1 Minimum Problems: Convexity and Newton's Method

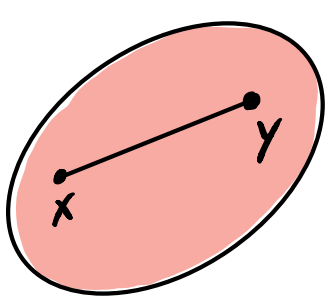
Let $f: \mathbb{R}^n \to \mathbb{R}$ and $K \subseteq \mathbb{R}^n$.

The constrained optimization problem

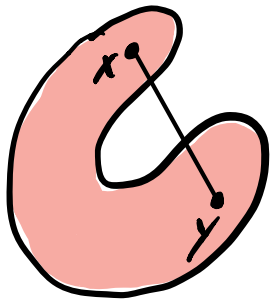
$$\boxed{\min f(x) \text{ subject to } x \in K}$$

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.



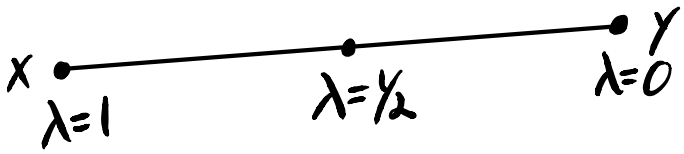
convex



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 <u>convex</u> if $\lambda x + (1-\lambda) y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$.

<u>Examples</u>:

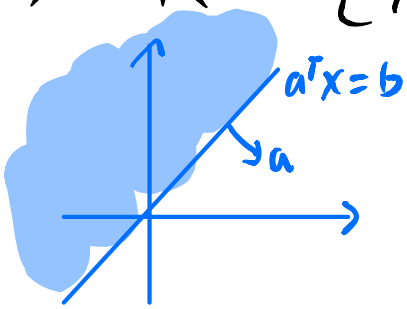
(1) $K = \{x \in \mathbb{R}^n : Ax = b\}$. A $m \times n$, b $m \times 1$

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

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

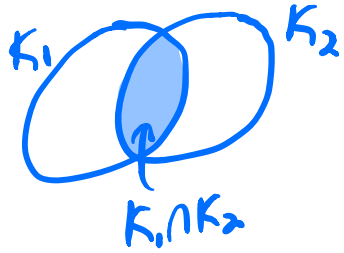
$\therefore \lambda x + (1-\lambda) y \in K$. $\therefore K$ is convex.

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



$a^T(\lambda x + (1-\lambda)y)$
$= \lambda a^T x + (1-\lambda) a^T y$
$\le \lambda b + (1-\lambda) b = b.$
$\therefore$ K is convex.

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



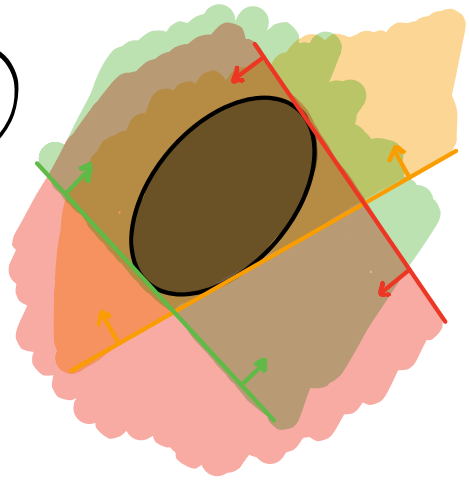
$x, y \in K$, $\lambda \in [0,1]$
$x, y \in K_1$ and $x, y \in K_2$
$\lambda x + (1-\lambda)y \in K_1$ and $\lambda x + (1-\lambda)y \in K_2$
$\therefore \lambda x + (1-\lambda)y \in K_1 \cap K_2 = K.$
$\therefore$ K is convex.

Two important properties:

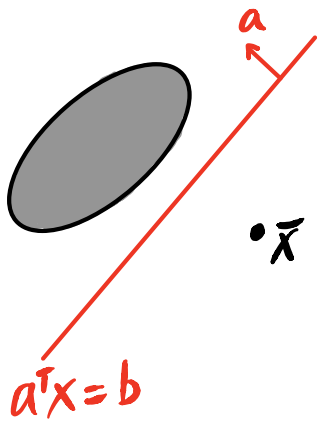
① Every closed convex set $S \subseteq \mathbb{R}^n$ is equal to the intersection of all the half-spaces containing S.

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.

(1)

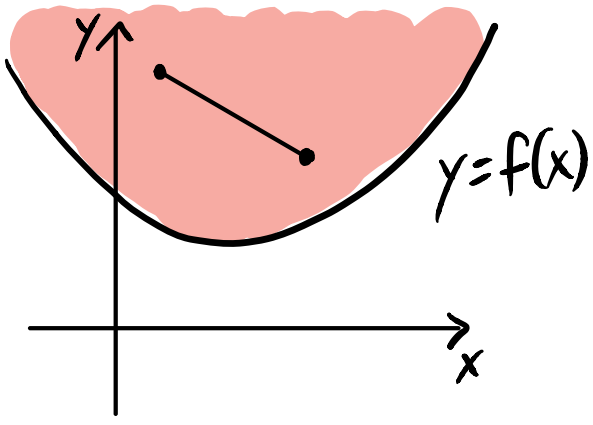


(2)

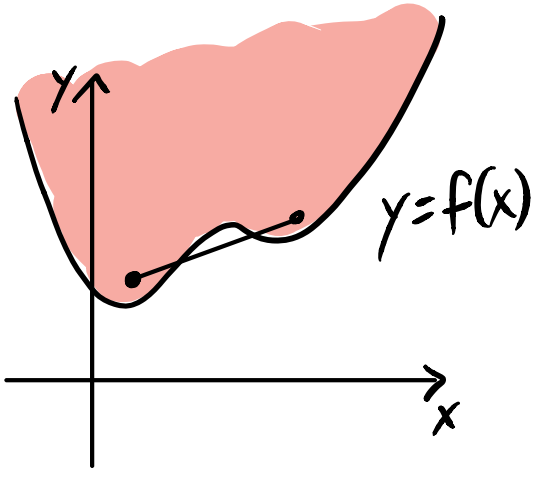


$a^T \bar{x} < b$ and $a^T x \geq b$, $\forall x \in S$

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



f convex



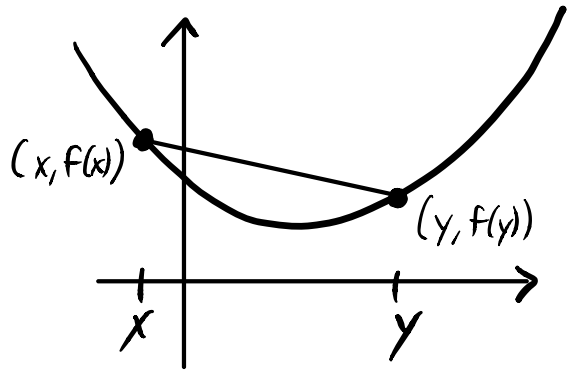
f <u>not</u> convex

The <u>epigraph</u> of $f: \mathbb{R}^n \to \mathbb{R}$ is

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

Thus, f is a convex function <u>iff</u> 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

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

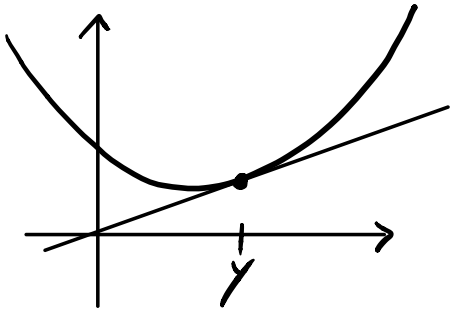
Thus, a function $f: \mathbb{R}^n \to \mathbb{R}$ is <u>convex</u> if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$ for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$.

The function is <u>strictly convex</u> 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) $f(x) = c^T x$

$f(\lambda x + (1-\lambda)y) = c^T(\lambda x + (1-\lambda)y)$

$= \lambda c^T x + (1-\lambda) c^T x = \lambda f(x) + (1-\lambda) f(y)$

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

$\nabla f(x) = Sx$

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

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

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

is pos. semidef.

$\therefore$ f is convex

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

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

is <u>not</u> pos. semidef.

$\therefore$ f is <u>not</u> convex

(5) $f(x) = \|x\|$ (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'ns are globally optimal

Suppose $f:\mathbb{R}^n \to \mathbb{R}$ 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

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

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 $f(\bar{x}) \le f(x)$. 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, $f(\lambda \bar{x} + (1-\lambda) y) \le \lambda f(\bar{x}) + (1-\lambda) f(y)$

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

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

Newton's Method

Let $f: \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, $x^* \in \mathbb{R}^n$ be a <u>local minimizer</u> of f (<u>ie</u> $\nabla f(x^*) = 0$ 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 $x_1 \in \mathbb{R}^n$ that is closer to x^*.

<u>Two ideas</u>:

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

$$\min_x \; f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0).$$

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

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

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

$x = x_0 - f(x_0)/f'(x_0)$

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

---

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:

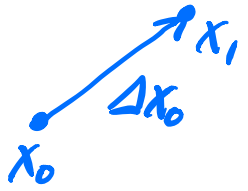
$$\nabla f(x_0) + \nabla^2 f(x_0)(x - x_0) = 0.$$

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 Δx_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:

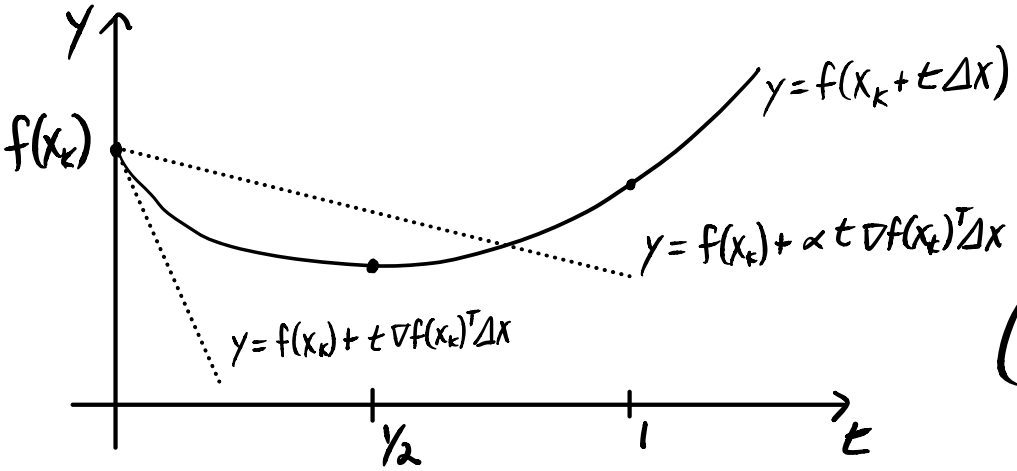
$$f(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$$

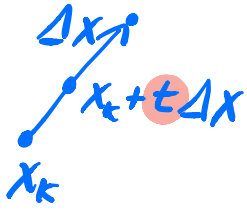
② Himmelblau's function:

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

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.





$(0 < \alpha < 1/2)$

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 Δx 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:

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

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-Marquardt for Nonlinear Least Squares

Suppose we have a <u>nonlinear</u> fitting function $\hat{y}: \mathbb{R}^n \to \mathbb{R}^m$ that depends on n parameters $p = (p_1, \ldots, p_n)$.

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

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

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 change the problem into 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 Δp 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

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

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:

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

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

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

The Levenberg-Marquardt 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

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

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

Example:

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