Convexity, Global Minima and Global Landscape

Ruoyu Sun

This Lecture

- Convexity and Global Minima
- After this lecture, you will be able to
  - **describe** sufficient conditions for existence of global optima
  - **check** whether a function is convex
  - **understand** the importance of convexity
  - **plot** the optimization landscape of a function
  - **compute** all stationary points and global optima of a quadratic minimization problem

Outline

Recall: Limitation of Optimality Conditions

Recall tentative-method 1: check all stationary points, and among them find x^* with the minimal function value.

However, x^* may or may not be a global-min.

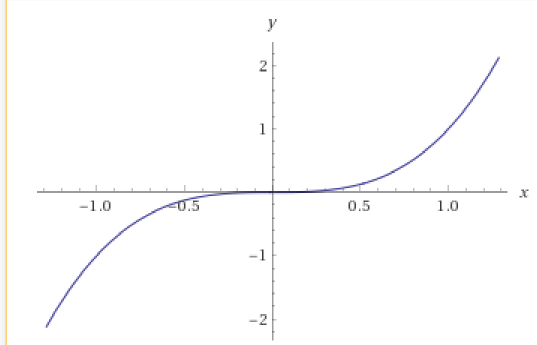
Correction:

- **Positive** side: verify that $f(x) \geq f(x^*), \forall x$ (then x^* is global-min)
  - Eg1: in machine learning applications, the loss $f \geq 0$.
    so if a candidate solution x^* has a value $f(x^*) = 0$, then done.
- **Negative** side: show that $f(x)$ can go to $-\infty$ (then no global-min exists)

Next: Let us analyze the cause, and propose other corrections.

Failure of Optimality Condition

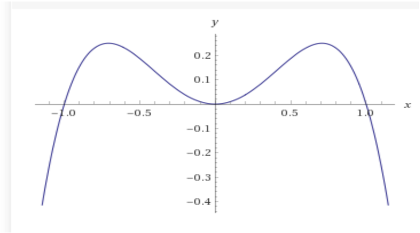
- Example 1: $f(x) = x^3$.
- 1st order optimality condition: $3(x^*)^2 = 0$, i.e. $x^* = 0$.
- 2nd order optimality condition (necessary): $\nabla^2 f(x^*) = 6x^* = 0$.
- $x^* = 0$ is the only “candidate”, but not a global-min:



What if Sufficient Condition Holds?

- In Example 1, only necessary conditions are satisfied.
- What if sufficient conditions are also satisfied? Could you give one counter-example?
- Example 2: $f(x) = x^2 - x^4$.

  $x^* = 0$ is a unique local-min satisfying the sufficient condition, but ...



this unique local-min is NOT a global-min

Is Lower-bounded Enough?

- In the above two examples, the function has no lower bound, so no global-min exists.
- Is "lower bounded" enough for existence of global-min?

Conjecture 1: Consider a differentiable function f. Suppose:

- f has a global lower bound, i.e, $f(x) \geq f_0, \forall x$.
- The set of stationary points is S, and $f(x^*) \leq f(x), \forall x \in S$.

Then x^* is the global minimum of f^*.

Counter-example to Conjecture 1

Example 3:

$$\min_{x\in\mathbb{R}} \exp(-x^2) =? \tag{1}$$

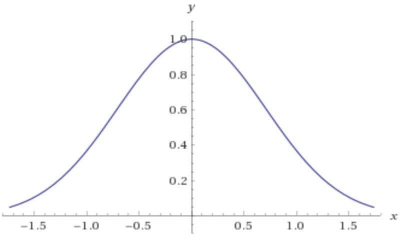
Since $f'(x) = -2x\exp(x^2)$, we have: $f'(x^*) = 0$ iff $x^* = 0$.

- Thus x^* is the unique stationary point.
- In addition, $\exp(-x^2) \geq 0, \forall x$; i.e., has a lower bound 0

The conjecture states: $x^* = 0$ is a global-min.

Counter-example to Conjecture 1

Let's draw the plot of $f(x) = \exp(-x^2)$:



$x^* = 0$ is not a global-min!

In fact, the function has no global-min!

- It has global infimum ∞ and $-\infty$, and infimum value 0

How to Fix

- **Answer**: besides applying Method 1 and 2, only need to ensure existence of global-min

Claim 1: Consider a differentiable function f. Suppose:

- (C1) f has at least one global minimizer;
- (C2) The set of stationary points is S, and $f(x^*) \leq f(x), \forall x \in S$.

Then x^* is a global minimizer of f^*.

Think: how does this extra condition fix the logical gap?

Proof: Existence of Global-min is a Fix (reading)

Claim 1 (repeat): Consider a differentiable function f. Suppose:

- (C1) f has at least one global minimizer;
- (C2) The set of stationary points is S, and $f(x^*) \leq f(x), \forall x \in S$.

Then x^* is a global minimizer of f^*.

Proof: Suppose $\hat{x}$ is a global minimizer of f, i.e.,

$$f(\hat{x}) \leq f(x), \forall x. \tag{2}$$

By the necessary optimality condition, we have $\nabla f(\hat{x}) = 0$, thus $\hat{x} \in S$.
By (C2), we have

$$f(x^*) \leq f(\hat{x}). \tag{3}$$

Combining (2) and (3), we have $f(\hat{x}) \leq f(x^*) \leq f(\hat{x})$, thus $f(\hat{x}) = f(x^*)$.
Plugging into (2), we have $f(x^*) \leq f(x), \forall x$. Thus x^* is a global minimizer of f^*. $\Box$

Local-min and Global-min On a Set

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X \end{array}$$

- **Objective function** $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function
- **Optimization variable** $x \in X$
- **local minimum** of f on X: $\exists \epsilon > 0$ s.t. $f(x) \geq f(\hat{x})$, for all $x \in X$ such that $\|x - \hat{x}\| \leq \epsilon$;
  i.e., x^* is the best in the intersection of a small neighborhood and X
- **Global minimum** of f on X: $f(x) \geq f(x^*)$ for all $x \in X$
- "Strict global minimum", "strict local minimum" "local maximum", "global maximum" of f on X are defined accordingly

Existence of Global-min

- **Bolzano-Weierstrass Theorem** (compact domain)
  Any continuous function f has at least one global minimizer on any compact set X.
  That is, there exists an $x^* \in X$ such that $f(x) \geq f(x^*), \forall x \in X$.

- **Corollary** (bounded level sets): Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. If for a certain c, the level set

$$\{x \mid f(x) \leq c\} \tag{4}$$

  is non-empty and compact, then the global minimizer of f exists, i.e., there exists $x^* \in \mathbb{R}^d$ s.t.

$$f(x^*) = \inf_{x \in \mathbb{R}^d} f(x).$$

- **Corollary** (coercive): Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. If $f(x) \to \infty$ as $\|x\| \to \infty$, then the global minimizer of f over $\mathbb{R}^d$ exists.
  - "$f(x) \to \infty$ as $\|x\| \to \infty$" means: for any sequence $x^k \to \infty$, we have $f(x^k) \to \infty$.

Examples: Checking Existence

- **Example 1**: $f(x) = x^3$. Level sets $\{x \mid x^3 \leq c\}$ is $\{x \mid x \leq c^{1/3}\}$: unbounded.

  If $x \to -\infty$, then $f(x) \to -\infty$. So NOT coercive.

- **Example 2**: $f(x) = x^2$.

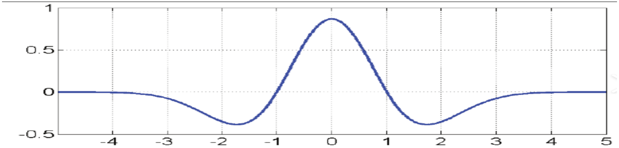
  Level set $\{x \mid x^2 \leq 1\}$ is $\{x \mid -1 \leq x \leq 1\}$: non-empty compact.

  Thus there exists a global minimum.

Coercive implies One Bounded Level Set (reading)

Coercive $\Rightarrow$ one non-empty bounded level set; but not the other way.

- **Eg**: Mexican hat function has a bounded level set $\{x \mid f(x) \leq -0.1\}$, but NOT coercive.



Claim (all level sets bounded $\Leftrightarrow$coercive): Let f be a continuous function, then f is coercive iff $\{x \mid f(x) \leq \alpha\}$ is compact for any α.

Proof:

Proof. We first show that the coercivity of f implies the compactness of the sets $\{x \,|f(x) \leq \alpha\}$. We begin by noting that the continuity of f implies the closedness of the sets $\{x \,|f(x) \leq \alpha\}$. Thus, it remains only to show that any set of the form $\{x \,|f(x) \leq \alpha\}$ is bounded. We show this by contradiction. Suppose to the contrary that there is an $\alpha \in \mathbb{R}^n$ such that the set $S = \{x \,|f(x) \leq \alpha\}$ is unbounded. Then there must exist a sequence $\{x^\nu\} \subset S$ with $\|x^\nu\| \to \infty$. But then, by the coercivity of f, we must also have $f(x^\nu) \to \infty$. This contradicts the fact that $f(x^\nu) \leq \alpha$ for all $\nu = 1, 2, \ldots$. Therefore the set S must be bounded.

Let us now assume that each of the sets $\{x \,|f(x) \leq \alpha\}$ is bounded and let $\{x^\nu\} \subset \mathbb{R}^n$ be such that $\|x^\nu\| \to \infty$. Let us suppose that there exists a subsequence of the integers $J \subset \mathbb{N}$ such that the set $\{f(x^\nu)\}_J$ is bounded above. Then there exists $\alpha \in \mathbb{R}^n$ such that $\{x^\nu\}_J \subset \{x \,|f(x) \leq \alpha\}$. But this cannot be the case since each of the sets $\{x \,|f(x) \leq \alpha\}$ is bounded while every subsequence of the sequence $\{x^\nu\}$ is unbounded by definition. Therefore, the set $\{f(x^\nu)\}_J$ cannot be bounded, and so the sequence $\{f(x^\nu)\}$ contains no bounded subsequence, i.e. $f(x^\nu) \to \infty$. $\square$

Use of Optimality Condition: Finding Optimal Solutions

- How to find a global minimum? (modify Tentative-method-1 $\&$ 2)

  Method of finding-global-min-among-stationary-points (FGMSP):

  Step 0: Verify coercive or bounded level set:
  - **Case 1**: success, go to Step 1.
  - **Case 2**: otherwise, try to show non-existence of global-min. If success, exit and report "no global-min exists".
  - **Case 3**: cannot verify coercive or bounded level set; cannot show non-existence of global-min. Exit and report "cannot decide".

  Step 1: Find all stationary points (candidates) by solving $\nabla f(\mathbf{x}) = 0$;

  Step 2 (optional): Find all candidates s.t. $\nabla^2 f(\mathbf{x}) \succeq 0$.

  Step 3: Among all candidates, find one candidate with the minimal value. Output this candidate, and report "find a global min".

Remarks

Remark 1: The method in the last page is not a “practical algorithm”.

- **Main reason**: finding *all* stationary points can be quite hard.
- Educational-algorithm: find global-min for very simple functions in homework/exam.

Remark 2: “cannot decide” is due to the lack of available tools.

- For any given function, either there exists a global-min, or there does not exist a global-min.
  But we may or may not be able to tell which case it is.

Correct Use of Optimality Condition

- **Example 1:** $\min \ \frac{1}{2}(x-b)^2$
- **Step 0**: Since $f(x) \to \infty$ as $|x| \to \infty$, f is coercive.
  Step 1: Set gradient $x^* - b = 0$, get $x^* = b$. It is the unique global-min.

- **Example 2:** $\min \ x^2 + 2y^2 + 3xy.$
- **Step 0**: Denote $f(x,y) = x^2 + 2y^2 + 3xy = (x + 1.5y)^2 - 0.25y^2$.
  - Let $y = M, x = -1.5M$, then $f(x,y) = -0.25M^2$.
  - As $M \to \infty$, $f(x,y) \to -\infty$, so there is no global minimizer!

Outline

Convexity and Optimal Conditions

- Sufficient condition for global optimality? Difficult to find.
- Most well-known conditions:

  Convexity $+$ first order condition $\Rightarrow$ global optimal.

  For a convex function, any stationary point is a global-min.

Convexity and Optimal Conditions

- **Convex set** C: $x, y \in C$ implies $\lambda x + (1-\lambda)y \in C$, for any $\lambda \in [0,1]$.

- **Convex function** (0-th order): f is convex in a convex set C iff

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \forall x, y \in C, \forall \alpha \in [0,1].$$

- **Property** (1st order) If f is differentiable, then f is convex iff

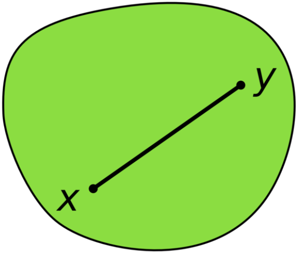
$$f(z) \geq f(x) + (z-x)'\nabla f(x), \quad \forall x, z \in C.$$

- **Property** (2nd order): If f is twice differentiable, then f is convex iff

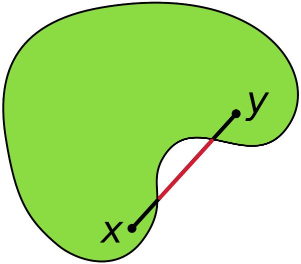
$$\nabla^2 f(x) \succeq 0, \quad \forall x \in C.$$

- Strictly convex: when $\geq$ becomes $>$ in any of the above relations.

Illustration of Convex Sets



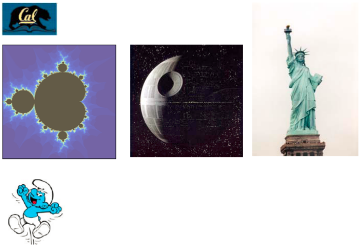
(a) Convex set



(b) Non-convex set

Illustration of Convex Sets

(c) Convex set



(d) Non-convex set

Illustration of Convex Functions

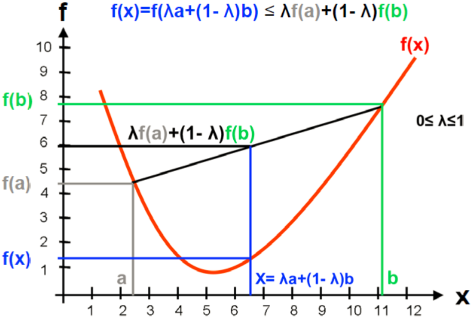
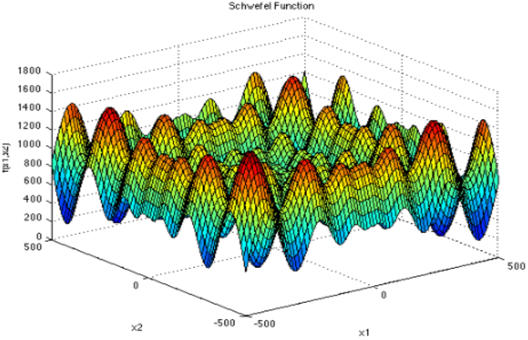


Illustration of Non-Convex Functions



$$f(\mathbf{x}) = 418.9829d - \sum_{i=1}^{d} x_i \sin(\sqrt{|x_i|})$$

Convex and Non-convex Functions

Convex Functions:

- Linear: $a'x + b$
- exponential: e^x, $-\log x$
- (convex) quadratic: x^2, $\|Ax - b\|^2$

Non-convex Functions:

- Bilinear: $(wv - 1)^2$, $\|UV - M\|_F^2$;
- Neural network: $\|Y - U\phi(VX)\|_F^2$; or

$$\min_{v \in \mathbb{R}^{m \times 1}, W \in \mathbb{R}^{m \times d}} \sum_{i=1}^{n} (y_i - v^T \sigma(Wx_i))^2,$$

where $(x_k, y_k), k = 1, 2, \ldots, n$ are the training data.

See more non-convex functions at
https://www.sfu.ca/ ssurjano/optimization.html

Convexity and Optimal Conditions

- **Proposition 1** (Prop. 1.1.2 of textbook): Let $f : X \longmapsto \mathbb{R}$ be a convex function over the convex set X.

  (a) A local-min of f over X is also a global-min over X.

  (b) If X is open (e.g. $\mathbb{R}^n$), then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for x^* to be a global minimum.

- Proof based on a property (Prop. B.3): If f is differentiable over C, then f is convex iff

$$f(z) \geq f(x) + (z - x)' \nabla f(x), \quad \forall x, z \in C.$$

Concave Function and Optimal Conditions

Definition: A function f is a concave function iff $-f$ is a convex function.

Corollary 1 (corollary of Prop 1): Let $f : X \longmapsto \mathbb{R}$ be a **concave** function over the convex set X.

(a) A local-max of f over X is also a global-max over X.

(b) If X is open (e.g. $\mathbb{R}^n$), then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for x^* to be a global maximum.

Some functions are more "convex"

Convex functions may look quite different from each other.

Left to right: more and more "convex".

How to measure the "degree of convexity"?

Strong convexity

Definition: We say $f : C \to \mathbb{R}$ is a μ-strongly convex function in a convex set C if f is differentiable and

$$\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2, \quad \forall w, v \in C. \tag{5}$$

- If f is twice differentiable, then f is μ-strongly convex iff

$$\nabla^2 f(x) \succeq \mu I, \quad \forall x \in C.$$

- Namely, all eigenvalues of the Hessian at any point is at least μ.
- if $f(w)$ is convex, then $f(w) + \frac{\mu}{2}\|w\|^2$ is μ-strongly convex.
  - In machine learning, easy to change a convex function to a strongly convex function: just add a regularizer

Outline

Non-convex functions

Convex optimization is a very important branch of optimization, since they are tractable (e.g. Steven Boyd's book “Convex optimization”.)

Nonconvex problems are much harder, since there may exist sub-optimal local-min (or stationary points).

Naive division: convex = easy; non-convex = hard.

However, some non-convex problems are much easier than others.

One way to get a bit more understanding of an unconstrained problem: **visualization**.

Visualization of Non-convex functions

Two types of visualization:

- **Image** (θ, f), where θ is the argument, $f(\theta)$ is the function value.
- **Contour** (level sets): $\{\theta \mid f(\theta) \leq c\}$, for $c = 0, 0.1, 0.3, 0.5, 1, 1.5$, etc. Can color it .

One example of "nice" non-convex function: $F(v, w) = (vw - 1)^2$.

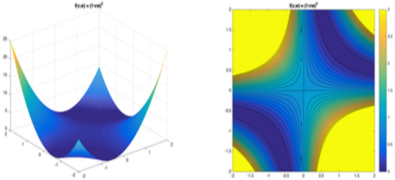


Figure: Visualization of $(vw - 1)^2$. Left: 3D plot. Right: contour.

Coding tips: for matlab, you can search "Creating 3-D plots, Mathworks" at google.

- Check commands "plot", "surf", "contour"

Visualization of Non-convex functions: High-dim functions

Plots only show low-dimension (at most 2 or 3), what about **high-dim**?

Idea: **Projection onto low-dim space**!

Consider visualizing a function $f(\theta)$, where $\theta \in \mathbb{R}^d$.

Method: 1) Pick a point $\hat{\theta}$ that you want to visualize around;

2) Pick two vectors $u, v \in \mathbb{R}^d$ (e.g. random Gaussian vectors);

3) Define a new function $f_{low}(s,t) = f(\hat{\theta} + su + tv)$. Visualize $f_{low}(s,t)$ for $s, t \in [-1, 1]$.

- To visualize it, can draw 3D plot $(s, t, f_{low}(s,t)), s, t \in [-1, 1]$, or draw the contour.

Practical tips: To get a good result, you may need to adjust u, v.

- Eg1: Multiply u, v by constant C (e.g., 0.01, 10, 1000) to see how the plot changes
- Eg2: If $\angle(u, v)$ is too small, then you may re-sample u, v

Visualization of Non-convex functions

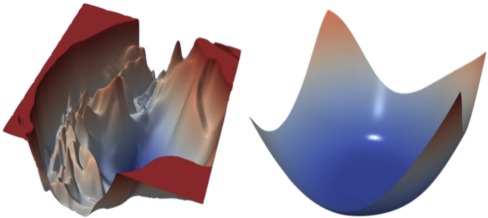


Figure: 3D Visualization of two neural networks. Left: bad; right: good.

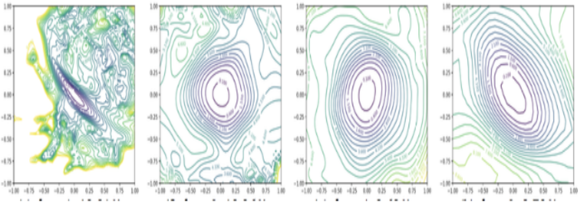


Figure: Contour Visualization of four neural networks.

Outline

Existence of Optimal Solutions

Convexity, Global Optimality Condition

Visualization of Landscape

Case Study: Quadratic Minimization

Unconstrained Quadratic Optimization: Toy Problems

Toy Problem 1:

$$\min_{x,y\in\mathbb{R}} x^2 + y^2 + \alpha xy.$$

Discuss the set of stationary points, global minima and global optimal value for every value of α.

Toy Problem 2:

$$\min_{x,y\in\mathbb{R}} y^2 - x.$$

Solution to Toy Problem 1

Toy Problem 1: $\min_{x,y\in\mathbb{R}} f(x,y) \triangleq x^2 + y^2 + \alpha xy.$

Step 1: First order condition: $2x^* + \alpha y^* = 0, 2y^* + \alpha x^* = 0.$

- We get $4x^* = -2\alpha y^* = \alpha^2 x^*$. So $(4-\alpha^2)x^* = 0.$
- **Case 1**: $\alpha^2 = 4$. If $x^* = -\alpha y^*/2$, then (x^*, y^*) is a stationary point.
- **Case 2**: $\alpha^2 \neq 4$. Then $x^* = 0$; $y^* = -\alpha x^*/2 = 0$. So $(0,0)$ is stat-pt.

Step 2: Check convexity. Hessian $\nabla^2 f(x,y) = \begin{pmatrix} 2 & \alpha \\ \alpha & 2 \end{pmatrix}$.

Eigenvalues λ_1, λ_2 satisfy $(\lambda_i - 2)^2 = \alpha^2,\ i = 1,2.$
Thus $\lambda_{1,2} = 2 \pm |\alpha|$.

- If $|\alpha| \leq 2$, then $\lambda_i \geq 0, \forall i$. Thus f is convex. Any stat-pt is global-min.
- If $|\alpha| > 2$, at least one $\lambda_i < 0$, thus f is not convex.

Step 3: For non-convex case ($|\alpha| > 2$), prove no lower bound.
$f(x,y) = (x + \alpha y/2) + (1 - \alpha^2/4)y^2$. Pick $y = M, x = -\alpha M/2$, then
$f(x,y) = (1-\alpha^2/4)M^2 \to -\infty$ as $M \to \infty$.
Summary: If $|\alpha| > 2$, no global-min, $(0,0)$ is stat-pt;
if $|\alpha| = 2$, any $(-0.5\alpha t, t), t \in \mathbb{R}$ is a stat-pt and global-min;
if $|\alpha| < 2$, $(0,0)$ is the unique stat-pt and global-min.

What is Special About Toy Problem 1?

We have studied a similar problem before ($x^2 + 2y^2 + 3xy$); toy problem 1 considers more general α.

Observation 1: Compare to FGMSP method, here we introduce an extra step of checking convexity

- we do not check coercive or bounded level sets

Observation 2: For cvx case, stat-pts are global-min. For non-convex case, no global-min exists.

- Implication: we can either find a global-min, or decide "no global-min exists"
- There is no case of "cannot decide" (which might happen in FGMSP method)

Next: this property holds for general quadratic problems!

Unconstrained Quadratic Optimization

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T \mathbf{Q}\mathbf{w} - \mathbf{b}^T \mathbf{w} \\ \text{subject to} \quad & \mathbf{w} \in \mathbb{R}^d, \end{aligned}$$

where $\mathbf{Q}$ is a symmetric $d \times d$ matrix. (what if non-symmetric?)

- Necessary condition for (local) optimality

$$\mathbf{Q}\mathbf{w} = \mathbf{b}, \quad \mathbf{Q} \succeq 0 \tag{6}$$

- **Case 1**: $\mathbf{Q}\mathbf{w} = \mathbf{b}$ has no solution, i.e. $\mathbf{b} \notin R(\mathbf{Q})$.
  No stationary point, can achieve $-\infty$ (how?).
- **Case 2**: $\mathbf{Q}$ is not PSD (f is non-convex)
  No local-min. Can achieve $-\infty$ (how?).
- **Case 3**: $\mathbf{Q} \succeq 0$ and $\mathbf{b} \in R(\mathbf{Q})$.

  Any stationary point is a global optimal solution.

Proof (reading)

Claim 1: If $\mathbf{Q}\mathbf{w} = \mathbf{b}$ has no solution, then: (i) there is no stationary point; (ii) $f(\mathbf{w})$ can achieve $-\infty$.

Proof: (i) is because a stationary point must satisfy $\mathbf{Q}\mathbf{w} = \mathbf{b}$.
Now we prove (ii). $\mathbf{Q}$ must be singular (otherwise $\mathbf{Q}\mathbf{w} = \mathbf{b}$ has a solution). We can write $\mathbf{b} = \mathbf{b}_{\parallel} + \mathbf{b}_{\perp}$, and $\mathbf{w} = \mathbf{w}_{\parallel} + \mathbf{w}_{\perp}$, where $\mathbf{b}_{\parallel}, \mathbf{w}_{\parallel} \in R(\mathbf{Q})$ and $\mathbf{b}_{\perp}, \mathbf{w}_{\perp} \perp R(\mathbf{Q})$. By $\mathbf{Q}\mathbf{w} = \mathbf{b}$ has no solution, we have $\mathbf{b}_{\perp} \neq 0$. Then

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{Q}\mathbf{w} - \mathbf{b}^T\mathbf{w} = \frac{1}{2}\mathbf{w}_{\parallel}^T\mathbf{Q}\mathbf{w}_{\parallel} - \mathbf{b}_{\parallel}^T\mathbf{w}_{\parallel} - \mathbf{b}_{\perp}^T\mathbf{w}_{\perp}$$

Pick $\mathbf{w}_{\perp} = M\mathbf{b}_{\perp}$ and $\mathbf{w}_{\parallel} = 0$, we have $f(\mathbf{w}) = -M\|\mathbf{b}_{\perp}\|^2 \to -\infty$ as $M \to -\infty$. Thus $f(\mathbf{w})$ can achieve $-\infty$. $\square$

Claim 2: If $\mathbf{Q}$ is not PSD, then: (i) there is no local-min; (ii) $f(\mathbf{w})$ can achieve $-\infty$.
Proof: (i) is because a local-min must satisfy $\mathbf{Q} \succeq 0$. To prove (ii), we write the eigen-decomposition of $\mathbf{Q}$ as $\mathbf{Q} = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\top}$ where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$. Since $\mathbf{Q}$ is not PSD, λ_1 must be negative. Pick $\mathbf{w} = M\mathbf{v}_1$, then $f(\mathbf{w}) = 0.5M^2\lambda_1 - M\mathbf{v}_1^T\mathbf{b}$. Since $\lambda_1 < 0$, as $M \to \infty$, $f(\mathbf{w}) \to -\infty$. $\square$

Linear Regression (Least Squares)

$$\begin{array}{ll} \text{minimize} & f(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}^T\mathbf{w} - \mathbf{y}\|^2 \\ \text{subject to} & \mathbf{w} \in \mathbb{R}^d, \end{array}$$

where $\mathbf{X} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, $\mathbf{y} \in \mathbb{R}^{n \times 1}$

- n data points, d features
- $\mathbf{X}$ may be wide (under-determined), tall (over-determined), or rank-deficient
- Note that comparing with the previous case, $\mathbf{Q} = \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{d \times d}$, $\mathbf{b} = \mathbf{X}\mathbf{y} \in \mathbb{R}^{d \times 1}$
  - $\mathbf{Q} \succeq 0$; Case 2 never happens!
- First order condition $\mathbf{X}\mathbf{X}^\top \mathbf{w}^* = \mathbf{X}\mathbf{y}$.
  - It always has a solution (why?); Case 1 never happens!

Claim: Linear regression problem is always convex; it has global-min.

Linear Regression (Least Squares)

- First order condition

$$\mathbf{X}\mathbf{X}^\top \mathbf{w}^* = \mathbf{X}\mathbf{y}.$$

which always has a solution.

- If $XX^\top \in \mathbb{R}^{d\times d}$ is invertible (only happen when $n \geq d$), then there is a unique stationary point $x = (A^\top A)^{-1}A^\top b$. It is also a global minimum.
- If $XX^\top \in \mathbb{R}^{d\times d}$ is not invertible, then there can be infinitely many stationary points, which are the solutions to the linear equation.

  All of them are global minima, giving the same function value.

Summary

Two conditions that ensure existence of global minimizers:

- Coercive
- One (non-empty) bounded level set

Convexity ensures every stationary point is global-min.

High-dim function landscape can be visualized by projection onto low-dim space

Minimizing quadratic function $\mathbf{x}^T\mathbf{Q}\mathbf{x} - 2\mathbf{x}^T\mathbf{b}$:

- Case 1: $\mathbf{b} \notin R(\mathbf{Q})$: no stationary point; no lower bound
- Case 2: $\mathbf{Q}$ not PSD: non-convex; no lower bound
- Case 3: $\mathbf{Q}$ PSD; $\mathbf{b} \in R(\mathbf{Q})$: convex; has global-min

Linear regression: always Case 3.