Recap: Global and Local Minimizers

Let $\Omega \subseteq \mathbb{R}^n$ be a nonempty set and let $f : \Omega \to \mathbb{R}$ be given. We define $B_\epsilon(y) := \{x \in \mathbb{R}^n : \|x - y\| < \epsilon\}$ to be the open ball in $\mathbb{R}^n$ with center y and radius $\epsilon > 0$.

The point $x^* \in \mathbb{R}^n$ is said to be a:

- local minimizer, if $x^* \in \Omega$ and there exists $\epsilon > 0$ such that $f(x) \geq f(x^*)$ for all $x \in \Omega \cap B_\epsilon(x^*)$.
- strict local minimizer, if $x^* \in \Omega$ and there is $\epsilon > 0$ with $f(x) > f(x^*)$ for all $x \in (\Omega \cap B_\epsilon(x^*)) \backslash \{x^*\}$.
- global minimizer, if $x^* \in \Omega$ and we have $f(x) \geq f(x^*)$ for all $x \in \Omega$.
- strict global minimizer, if $x^* \in \Omega$ and it holds that $f(x) > f(x^*)$ for all $x \in \Omega \backslash \{x^*\}$.
- **Remark**: global minimizer $\equiv$ global solution $\equiv$ optimal solution.
- The def. for maximizer is identical, changing: $\geq / > \to \leq / <$.

Review: Gradient, Hessian Matrix and Taylor Expansion

- Assume $f(x) = f(x_1, x_2, ..., x_n)$ is continuously differentiable. Then we denote the gradient of f by (an $n \times 1$ vector):

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}; \frac{\partial f}{\partial x_2}; ...; \frac{\partial f}{\partial x_n}\right)$$

The first-order Taylor expansion yields:

$$f(x + td) = f(x) + t\nabla f(x)^\top d + o(t), \quad t \to 0.$$

- If f is twice continuously differentiable, then the Hessian of f (an $n \times n$ matrix) is given by:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j}$$

By a second-order Taylor expansion, we obtain:

$$f(x + td) = f(x) + t\nabla f(x)^\top d + \frac{1}{2}t^2 d^\top \nabla^2 f(x) d + o(t^2), \quad t \to 0.$$

Optimality Conditions for Unconstrained Problems

FONC: First-Order Necessary Conditions

If x^* is a local minimizer of the unconstrained problem $\min_{x \in \mathbb{R}^n} f(x)$, then we must have $\nabla f(x^*) = 0$.

SONC: Second-Order Necessary Conditions

If x^* is a local minimizer of f, then it holds that:

1. $\nabla f(x^*) = 0$;
2. For all $d \in \mathbb{R}^n$: $d^\top \nabla^2 f(x^*) d \geq 0$.

SOSC: Second-Order Sufficient Conditions

Let f be twice continuously differentiable. If x^* satisfies:

1. $\nabla f(x^*) = 0$;
2. For all $d \in \mathbb{R}^n \backslash \{0\}$: $d^\top \nabla^2 f(x^*) d > 0$;

then x^* is a strict local minimizer of f.

- For non-strict local minimizers and saddle points with PSD Hessian (consider x^3 at $x = 0$), SONC is satisfied in both cases. So far we do not have sufficient conditions to identify them.

Definition: Stationary Points and Saddle Points

- A point x satisfying $\nabla f(x) = 0$ is called critical point or stationary point.
- A stationary point is called saddle point if it is neither a local minimizer nor a local maximizer.

Corollary: Saddle Points

Suppose that x^* is a stationary point ($\nabla f(x^*) = 0$) and that the Hessian $\nabla^2 f(x^*)$ is indefinite, then x^* is a saddle point.

Note that the above corollary is just a sufficient condition for saddle point. On the contrary, a saddle point does not necessarily need to have indefinite Hessian as it can have PSD Hessian (called degenerate saddle point), e.g., x^3 at $x = 0$.

Facts:

- If $f(x) = c^\top x$, then $\nabla f(x) = c$.
- If $f(x) = x^\top M x$ (M is symmetric), then $\nabla f(x) = 2Mx$.
- If $f(x) = x^\top M x$ (M is symmetric), then $\nabla^2 f(x) = 2M$.

Least Squares:

$$\underset{\beta}{\text{minimize}} \quad ||X\beta - y||^2 = f(\beta).$$

- Gradient: $\nabla f(\beta) = X^\top(X\beta - y) \implies$ FONC: $X^\top X\beta - X^\top y = 0$.
- Hessian: $\nabla^2 f(\beta) = 2X^\top X \implies$ SONC is always satisfied, SOSC is satisfied if $X^\top X$ is PD.

Weierstrass (Extreme Value) Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and let $\Omega \subset \mathbb{R}^n$ be a bounded, closed, and nonempty set. Then, f attains a global maximum and global minimum on the set Ω.

closed: contains the boundry

- The set Ω is closed if for every convergent sequence $\{x^k\}_k$ with $x^k \in \Omega$ for all k and $\lim_{k\to\infty} x^k = x$, it holds that $x \in \Omega$.
- Ω is bounded if there is $B > 0$ with $\|x\| \leq B$ for all $x \in \Omega$.
- A closed and bounded set is also called compact.

xample:

$$\min \ f(x) = x_1^2 - x_2^2 \quad \text{s.t.} \quad h(x) = x_1^2 + x_2^2 - 4 = 0.$$

- The feasible set $\Omega = \{x : h(x) = 0\} = \{x \in \mathbb{R}^2 : \|x\| = 2\}$ is closed and bounded and f is continuous (on $\mathbb{R}^2$).

→ By Weierstrass: f attains a global max. and min. on Ω.

Definition: Coercivity

A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be coercive if

$$\lim_{\|x\| \to \infty} f(x) = +\infty.$$

- Geometrically, coercivity means that $f(x)$ increases as x moves away from the origin in any possible direction.
- Mathematically, coercivity means:
- The functions $f(x) = x$, $f(x) = x^3$, $f(x) = e^x$, and $f(x) = 1$ are not coercive.

Coercivity is often established by estimating the function and by finding a suitable lower bound for sufficiently large x.

Theorem: Coercivity and Existence of Solutions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous and coercive function. Then, for all $\alpha > 0$, the level set

$$L_{\leq \alpha} := \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$$

is compact and f has at least one global minimizer.

Constrained Problems

Definition: Feasible Direction

Given $x \in \Omega$, we call d a feasible direction at x if there exists $\bar{t} > 0$ such that $x + td \in \Omega$ for all $0 \leq t \leq \bar{t}$.

- If $\Omega = \{x : Ax = b\}$, then all feasible directions at x are given by $\{d : Ad = 0\}$.
- If $\Omega = \{x : Ax \geq b\}$, then the feasible directions at x are given by $\{d : a_i^\top d \geq 0 \text{ if } a_i^\top x = b_i\}$.

Theorem: FONC for Constrained Problems

Let x^* be a local minimum of $\min_{x \in \Omega} f(x)$. Then for any feasible direction d at x^*, we must have $\nabla f(x^*)^\top d \geq 0$.

Definition: Descent Direction

Let f be continuously differentiable. Then d is called a descent direction at x if and only if $\nabla f(x)^\top d < 0$.

If we denote the set of feasible directions at x by $S_\Omega(x)$ and the set of descent directions at x by $S_D(x)$, then the first order necessary condition can be written as: $S_\Omega(x^*) \cap S_D(x^*) = \emptyset$

There are no feasible descent directions at local min x^*.

At a point $x \in \Omega$, the set $\mathcal{A}(x) := \{i : g_i(x) = 0\}$ denotes the set of active constraints. The set of inactive constraints is given by $\mathcal{I}(x) := \{i : g_i(x) <$

Theorem: FONC for Linearly Constrained Problems

If x^* is a local minimum of (1), then there exists some $\mathbb{R}^m \ni y \geq 0$ with

$$\nabla f(x^*) - A^\top y = 0$$
$$y_i \cdot (a_i^\top x^* - b_i) = 0 \quad \forall\, i,$$

where $a_i^\top$ is the ith row of A.

General Nonlinear Optimization Problem:

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad \forall\, i = 1, ..., m, \\ & h_j(x) = 0, \quad \forall\, j = 1, ..., p. \end{aligned}$$

Theorem: Fritz-John Conditions

Let x^* be a local minimum of (3). Then, there exists $\lambda_0, \lambda_1, ..., \lambda_m \geq 0$, which are not all zeros, such that

$$\lambda_0 \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) = 0,$$
$$\lambda_i \cdot g_i(x^*) = 0, \quad \forall\, i = 1, ..., m.$$

Major Drawback: The choice $\lambda_0 = 0$ is allowed.

In this case, the Fritz-John conditions just impose linear dependence of the vectors $\{\nabla g_i(x^*)\}_{i \in \mathcal{A}(x^*)}$.

We want to extend them to more useful cases — KKT conditions, which impose further assumptions to eliminate the possibility $\lambda_0 = 0$.

The first step is to construct the Lagrangian of this problem defined as follows:

1. We associate each constraint with a Lagrangian multiplier (indeed dual variables):

$$\begin{aligned} g_i(x) \leq 0 \quad &\to \quad \lambda_i, \quad i = 1, \ldots, m. \\ h_j(x) = 0 \quad &\to \quad \mu_j, \quad j = 1, \ldots, p. \end{aligned}$$

2. We define the Lagrangian of this problem by:

$$L(x, \lambda, \mu) := f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x).$$

If x is a local minimizer and if a constraint qualification holds, then there exist λ and μ such that:

1. Main Condition

$$\nabla_x L(x, \lambda, \mu) = \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) + \sum_{j=1}^p \mu_j \nabla h_j(x) = 0$$

2. Dual Feasibility

$$\lambda_i \geq 0, \quad i = 1, ..., m.$$

3. Complementarity

$$\lambda_i \cdot g_i(x) = 0, \quad \forall\, i = 1, ..., m.$$

We often add primal feasibility as part of the KKT conditions:

4. Primal Feasibility

$$g_i(x) \leq 0, \quad h_j(x) = 0, \quad \forall\, i, \quad \forall\, j.$$

As before, we require the collection of gradients

$$\{\nabla g_i(x^*) : i \in \mathcal{A}(x^*)\} \cup \{\nabla h_j(x^*) : j = 1, ..., p\} \qquad (\star)$$

to be linearly independent.

- This condition is one candidate for the constraint qualification (CQ) in the theorem and is called Linear Independence Constraint Qualification (LICQ).
- A feasible point x satisfying the LICQ is called regular.

Further Remarks:

- KKT conditions is first order necessary conditions (FONC) for general constrained optimization problems.
- KKT conditions unify all formerly studied FONC.
- A (feasible) point satisfying the KKT conditions is called a KKT point regardless whether it satisfies CQ or not.
- KKT points are candidates for local optimal solutions – just like stationary points.

General Strategy:

- Check LICQ (if required).
- Derive KKT conditions.
- Discuss different easy cases via the complementarity conditions (set multiplier or constraints to 0) to find all KKT points.

Additional Information:

- Check if f is coercive or if Ω is bounded, then the problem has global solutions (which must be KKT points if CQ holds).
- If the LICQ holds, then λ and μ are always unique.
- If CQ holds, then the global optimizer must be a KKT point. If future there is a unique KKT point, then this point must be global minimizer.

Unconstrained	Constrained
FONC: x^* local minimum (+ CQ)	
▶ $\nabla f(x^*) = 0$.	▶ KKT conditions.
SONC: x^* local minimum (+ CQ)	
▶ $\nabla f(x^*) = 0$	▶ KKT conditions
▶ $\nabla^2 f(x^*)$ is positive semi-definite (on $\mathbb{R}^n$).	▶ $\nabla^2_{xx} L(x^*, \lambda^*, \mu)$ is positive semidefinite on $\mathcal{C}(x^*)$.
SOSC:	
▶ $\nabla f(x^*) = 0$	▶ KKT conditions
▶ $\nabla^2 f(x^*)$ is positive definite (on $\mathbb{R}^n$ except zero).	▶ $\nabla^2_{xx} L(x^*, \lambda^*, \mu)$ is positive definite on $\mathcal{C}(x^*)$ except zero.

Convex Optimization

Definition: Convex Set

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

Convex Combination

For any $x_1, ..., x_n$ and $\lambda_1, ..., \lambda_n \geq 0$ satisfying $\lambda_1 + \cdots + \lambda_n = 1$, we call $\sum_{i=1}^n \lambda_i x_i$ a convex combination of $x_1, ..., x_n$.

Definition: Convex Function

A function f on a convex set Ω is said to be convex if for every $x_1, x_2 \in \Omega$ and any $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

Definition: Concave Function

We call f a concave function if and only if $-f$ is convex, i.e., for any x_1, x_2 and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda) x_2) \geq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

Some examples of convex functions:

- $f(x) = x$, $f(x) = x^2$, $f(x) = e^x$, $f(x) = |x|$.

Some examples of concave functions:

- $f(x) = x$, $f(x) = \sqrt{x}$, $f(x) = \log x$

Theorem: Convexity via Hessian

Let f be twice continuously differentiable. Then f is convex (on $\mathbb{R}^n$) if and only if its Hessian matrix is positive semidefinite (PSD), i.e.,

$$d^\top \nabla^2 f(x) d \geq 0 \quad \forall\, d \in \mathbb{R}^n, \quad \forall\, x \in \mathbb{R}^n.$$

In $\mathbb{R}$, this means that the second-order derivative is non-negative.

Taking second-order derivatives is one common way to test convexity.

Different from SOSC, this PSD property is required for all $x \in \mathbb{R}^n$.

Theorem: Concavity via Hessian

Let f be twice continuous differentiable. Then f is concave (on $\mathbb{R}^n$) if and only if its Hessian matrix is negative semidefinite, i.e.,

$$d^\top \nabla^2 f(x) d \leq 0 \quad \forall\, d \in \mathbb{R}^n, \quad \forall\, x \in \mathbb{R}^n.$$

Lemma: Sum Rule

If $a_1, ..., a_m \geq 0$, and $f_1, ..., f_m$ are convex (concave) functions, then $f = a_1 f_1 + \cdots + a_m f_m$ is a convex (concave) function.

Lemma: Composition with Linear Functions

If $g : \mathbb{R}^m \to \mathbb{R}$ is convex (concave) and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ are given, then $g : \mathbb{R}^n \to \mathbb{R}$, $f(x) := g(Ax + b)$, is convex (concave).

Lemma: Taking Maximum

If $f_1, ..., f_m$ are convex functions, then $f(x) = \max\{f_1(x), ..., f_m(x)\}$ is a convex function (this can be extended to uncountably many).

Lemma: Taking Minimum

If $f_1, ..., f_m$ are concave function, then $f(x) = \min\{f_1(x), ..., f_m(x)\}$ is a concave function (this can be extended to uncountably many).

Example: Least Squares $\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \ \|X\beta - y\|^2 = f(\beta)$.

Convexity via Hessian: Recall that $\nabla^2 f(\beta) = 2X^\top X$, which is always PSD for all $\beta \in \mathbb{R}^n$ and hence f is convex.

Convexity via calculus rule: f is a composition of the convex function $g = \|\cdot\|^2$ with a linear function $X\beta - y$ and hence f is convex.

Example: Robust Regression $\underset{\beta}{\text{minimize}} \ \sum_{i=1}^m |y_i - x_i^\top \beta| = \|X\beta - y\|_1 := f(\beta)$.

Yes. We can use convexity via calculus rule. f is a composition of the convex function $g = \|\cdot\|_1$ with a linear function $X\beta - y$ and hence f is convex.

Example: Logistic Regression $f(w) = \frac{1}{m} \sum_{i=1}^m \log\left(1 + \exp\left(-y_i \cdot w^\top x_i\right)\right)$

- f consists of non-negative sum of a set of functions.
- $\log\left(1 + \exp\left(-y_i \cdot w^\top x_i\right)\right)$ is composition of $g(z) = \log(1 + \exp(z))$ with a linear function $z = -y_i \cdot w^\top x_i$ (linear in w).
- By sum and composition rules, f is convex once $g(z)$ is convex for all $z \in \mathbb{R}$.

$$g''(z) = \frac{\exp(z)}{(1 + \exp(z))^2} \geq 0, \quad \forall z \in \mathbb{R}$$

hence g is convex for all $z \in \mathbb{R}$.

- For $x \in \Omega$, the set $\mathcal{A}(x) := \{i : g_i(x) = 0\}$ denotes the set of active constraints.
- The set of inactive constraints is $\mathcal{I}(x) := \{i : g_i(x) < 0\}$.

Definition: Convex Optimization

- Minimize a convex function over a convex feasible region.
- Maximize a concave function over a convex feasible region.

Theorem: Convexity Guarantees Local to be Global Solutions

Let $f:\Omega\to\mathbb{R}$ be a convex function and $\Omega\subset\mathbb{R}^n$ be a convex set. Then any local minimizer of the problem:

$$\begin{aligned}\underset{x}{\text{minimize}}\quad & f(x)\\ \text{subject to}\quad & x\in\Omega\end{aligned}$$

is a global minimizer.

Proof: By contradiction. Assume the local min x^* is not a global min, then there exists $\bar{x}\in\Omega$ such that $f(\bar{x})<f(x^*)$. Then, using convexity, we have

$$f(\lambda\bar{x}+(1-\lambda)x^*)\le\lambda f(\bar{x})+(1-\lambda)f(x^*)<f(x^*)$$

for any $0<\lambda<1$. If $\lambda\to 0$, this is a contradiction to: x^* is a local min. □

Theorem: Stationarity & Global Optimality — Unconstrained Case

Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is convex on $\mathbb{R}^n$ and continuously differentiabl at x^*. Then, x^* is a global minimum if and only if (iff) $\nabla f(x^*)=0$.

- Recall the least squares problem $\min_\beta\|X\beta-y\|^2$, we have its FONC is $X^\top X\beta=X^\top y$. All such β are also global minimizers.
- Similar results apply to the logistic regression problem.
- For robust regression problem, we need a generalized notion of gradient, i.e., subgradient, which is not covered by this course.

Proof **Theorem: First-Order Characterization of Convexity**

Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is differentiable. f is convex ($\mathbb{R}^n$) iff

$$f(y)\ge f(x)+\nabla f(x)^\top(y-x),\quad\forall x,y\in\mathbb{R}^n.$$

If $\nabla f(x^*)=0$, we immediately have

$$f(y)\ge f(x^*),\quad\forall y\in\mathbb{R}^n.$$

The other side is shown by FONC, which completes the proof

Theorem: Stationarity & Global Optimality — Constrained Case

Consider the constrained convex optimization problem

$$\begin{aligned}\underset{x}{\text{minimize}}\quad & f(x)\\ \text{subject to}\quad & x\in\Omega\end{aligned}$$

where f is convex and $\Omega:=\{x:g(x)\le 0,h(x)=0\}$ with g being conve and h being linear. Then, any KKT points for this problem are globa minimizers.

Remarks:

- In a Nutshell: If f and Ω are convex, then KKT points (FONC) are global minimizers.
- Recall that in order to let the necessary direction (KKT conditions) to be true, we need CQ to hold true.
- If f is concave and Ω is convex, then any KKT points of the problem $\max_{x\in\Omega}f(x)$ are global maximizers of f.

Proof KKT point $(\bar{x},\bar{\lambda},\bar{\mu})$ is a solution to the KKT conditions, i.e.,

1. Main Condition
$$\nabla f(\bar{x})+\sum_{i=1}^m\bar{\lambda}_i\nabla g_i(\bar{x})+\sum_{j=1}^p\bar{\mu}_j\nabla h_j(\bar{x})=0.$$
2. Dual Feasibility
$$\bar{\lambda}_i\ge 0\quad i=1,\dots,m.$$
3. Complementarity
$$\bar{\lambda}_i\cdot g_i(\bar{x})=0\quad\forall\, i=1,\dots,m.$$
4. Primal Feasibility
$$g_i(\bar{x})\le 0,\quad h_j(\bar{x})=0\quad\forall\, i,\quad\forall\, j.$$

Since the Lagrangian $x\mapsto L(x,\bar{\lambda},\bar{\mu})$ is convex, the main condition in KK implies $\bar{x}$ is a global minimum of $x\mapsto L(x,\bar{\lambda},\bar{\mu})$ (by the theorem for unconstrained case). Thus,

$$\begin{aligned}f(\bar{x})&=f(\bar{x})+\sum_{i=1}^m\bar{\lambda}_ig_i(\bar{x})+\sum_{j=1}^p\bar{\mu}_jh_j(\bar{x})\quad\text{(complementarity)}\\&=\min_x\left\{f(x)+\sum_{i=1}^m\bar{\lambda}_ig_i(x)+\sum_{j=1}^p\bar{\mu}_jh_j(x)\right\}\\&\le\min_{g_i(x)\le 0,h_j(x)=0}\left\{f(x)+\sum_{i=1}^m\bar{\lambda}_ig_i(x)+\sum_{j=1}^p\bar{\mu}_jh_j(x)\right\}\\&\le\min_{g_i(x)\le 0,h_j(x)=0}f(x)\quad\text{(sign)}.\end{aligned}$$

Lemma: Convex Level Sets

Let g be a convex (concave) function. Then, for any c, the level set $L_{\le c}=\{x:g(x)\le c\}$ ($L_{\ge c}=\{x:g(x)\ge c\}$) is a convex set.

If we have constraints of the form $g(x)\le 0$ and g is convex, then this is a convex constraint.

If we have constraints of the form $g(x)\ge 0$ and g is concave, then this is a convex constraint.

Sometimes, even if a constraint does not appear to be in the above form, it still could be a convex constraint (this lemma is just sufficient but not necessary).

Let $y^1,y^2,\dots,y^k\in\mathbb{R}^2$ be k different given points (generate randomly).

We want to find a circle in $\mathbb{R}^2$ with minimum radius that contains all of these points:

$$\begin{aligned}\text{minimize}_{y\in\mathbb{R}^2,r\in\mathbb{R}}\quad & r\\ \text{subject to}\quad & \|y-y^1\|\le r,\ \|y-y^2\|\le r,\ \dots,\|y-y^k\|\le r\\ & r\ge 0.\end{aligned}$$

Consider the optimization problem:

$$\begin{aligned}\underset{x,y,z}{\text{maximize}}\quad & xyz\\ \text{s.t.}\quad & x+2y+3z\le 3\\ & x,y,z\ge 0\end{aligned}$$

In order for a maximization problem to be a convex optimization problem we need the objective function to be concave.

- However, xyz is not a concave function in x,y,z.

But we can transform this into maximizing $\log(xyz)$ (a common trick in machine learning). The problem becomes:

$$\begin{aligned}\text{maximize}\quad & \log x+\log y+\log z\\ \text{s.t.}\quad & x+2y+3z\le 3\\ & x,y,z\ge 0\end{aligned}$$

which is a convex optimization problem.

nonconvex Algorithms for Unconstrained Problems

Typically, optimization algorithms are iterative procedures:

- Starting from some point x^0, we generate a sequence of iterates $\{x^k\}$.
- The sequence terminates when either no progress can be made or when we know that the current step is already satisfactory.
- Typically, we want to have $f(x^{k+1})<f(x^k)$, i.e., each step we can improve the objective value.
- And hopefully, the sequence $\{x^k\}$ converges to a local minimizer x^* (or even global minimizer).

Definition: Convergence

Let $\{x^k\}$ be a sequence of real vectors. Then $\{x^k\}$ converges to x^* if and only if for every $\epsilon>0$, there exists a positive integer K such that $\|x^k-x^*\|<\epsilon$ for all $k\ge K$.

$\|x\|=\sqrt{x^\top x}=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$

Examples of convergent sequences:

- $x^k:=1/k$ for all k; then $x^k\to 0$.
- $x^k:=(1/2)^k$ for all k; then $x^k\to 0$.

Problems in $\mathbb{R}$

Assume we can find x_ℓ and x_r such that $g(x_\ell)<0$ and $g(x_r)>0$.

By the intermediate value theorem, if g is continuous, there must exist a root of g ($g(x)=0$) in $[x_\ell,x_r]$.

Bisection Method

1. Define $x_m=\frac{x_\ell+x_r}{2}$.
2. If $g(x_m)=0$, then output x_m.
3. Otherwise:
   - If $g(x_m)>0$, then let $x_r=x_m$.
   - If $g(x_m)<0$, then let $x_\ell=x_m$.
4. If $|x_r-x_\ell|<\epsilon$: stop and output $\frac{x_\ell+x_r}{2}$, otherwise go back to step 1.

One can also set the stopping criterion based on $|g(x)|<\epsilon$.

In the bisection method, each iteration will divide the search interval to half.

Therefore, to find an ϵ approximation of x^*, we need at most $\log_2\frac{x_r-x_\ell}{\epsilon}$ many iterations.

Applying the bisection method to f', we can find an approximate stationary point.

- If f is convex, this is an (approximate) global minimizer of f.
- Although simple, the bisection method is very useful in practice because it is easy to implement.

Example: Use bisection method to minimize:

$$f(x)=-\frac{xe^{-x}}{1+e^{-x}}\quad\rightsquigarrow\quad f'(x)=-\frac{e^{-x}(1-x+e^{-x})}{(1+e^{-x})^2}$$

- Sometimes, f' is not available. For example, f sometimes is only a black box, which does not admit an analytical form (thus, the derivative is hard to compute)

However, if we know that f has a unique local minimum x^* in the range $[x_\ell,x_r]$, then we still have a very efficient way to find x^*:

- We call f unimodal if it only has one single stationary point (on $\mathbb{R}$).
- Unimodal functions have the property that the local minimum is already global. (Similarly, if the stationary point is a local maximum).

Golden Section Method

Assume we start with $[x_\ell,x_r]$. Assume $0<\phi<0.5$.

1. Set $x'_\ell=\phi x_r+(1-\phi)x_\ell$ and $x'_r=(1-\phi)x_r+\phi x_\ell$.
2. If $f(x'_\ell)<f(x'_r)$, then the minimizer must lie in $[x_\ell,x'_r]$, so set $x_r=x'_r$.
3. Otherwise, the minimizer must lie in $[x'_\ell,x_r]$, so set $x_\ell=x'_\ell$.
4. If $x_r-x_\ell<\epsilon$, output $\frac{x_\ell+x_r}{2}$, otherwise go back to step 1.

- This is true when

$$\phi=\frac{3-\sqrt{5}}{2}\quad\text{and}\quad 1-\phi=\frac{\sqrt{5}-1}{2}=0.618\ \text{(golden section)}.$$

Higher Dimensional Problems

Solution and General Idea:

- Each time, we first find a search direction.
- Then, we search for a good next step along that direction.

Starting from the initial point x^0, we generate a sequence of points:

$$x^{k+1}=x^k+\alpha_kd^k,\quad k=0,1,\dots$$

We call d^k the search direction (a vector) and α_k the step size (a scalar).

- The key is to choose a proper direction d^k at each iteration.
- d^k typically depends on x^k.
- The step size α_k may be chosen in accordance with some line search rules.

Review: Descent Direction

Let f be continuously differentiable. Then d is called a descent direction at x if and only if $\nabla f(x)^\top d<0$.

Schematic Descent Directions Method

1. Initialization: Select an initial point $x^0\in\mathbb{R}^n$.

For $k=0,1,2,\dots$:

2. Pick a descent direction d^k.
3. Find a stepsize α_k satisfying $f(x^k+\alpha_kd^k)<f(x^k)$.
4. Set $x^{k+1}=x^k+\alpha_kd^k$.
5. If a stopping criterion is satisfied, then STOP and x^{k+1} is the output.

- One simple and possible descent direction is $d^k=-\nabla f(x^k)$. This direction satisfies:

$$\nabla f(x^k)^\top d^k=-\|\nabla f(x^k)\|^2<0$$

as long as $\nabla f(x^k)\ne 0$.

- Choosing $d^k=-\nabla f(x^k)$, the abstract descent method becomes the gradient descent method.

Stopping Criterion:

- A popular stopping criterion is: $\|\nabla f(x^{k+1})\|\le\epsilon$ with tolerance $\epsilon>0$.
  - We stop if x^{k+1} is an approximate stationary point.

Step Sizes

Constant Step Size:

- Choose $\alpha_k=\bar{\alpha}$ for all k.

Exact Line Search:

- An intuitive idea is to choose α_k to achieve the largest descent.

That is, choose α_k such that:

$$\alpha_k=\underset{\alpha\ge 0}{\operatorname{argmin}}\ f(x^k+\alpha d^k).\qquad(1)$$

- If we get the exact α_k in (1), we say we used an exact line search method to find the step size.
- This is a one dimensional problem. We may use the bisection / golden section method to perform the exact line search.
- In some situations, we can even find the exact α analytically.

Backtracking / Armijo Line Search

Assume we have found a descent direction d^k and we want to choose step size α_k.

Let $\sigma,\gamma\in(0,1)$ be given. Choose α_k as the largest element in $\{1,\sigma,\sigma^2,\sigma^3,\dots\}$ such that

$$f(x^k+\alpha_kd^k)-f(x^k)\le\gamma\alpha_k\cdot\nabla f(x^k)^\top d^k.$$

- This condition is called Armijo condition.
- α_k can be determined after finitely many steps if d^k is a descent direction.

Procedure:

1. Start with $\alpha=1$.
2. If $f(x^k+\alpha d^k)\le f(x^k)+\gamma\alpha\cdot\nabla f(x^k)^\top d^k$, choose $\alpha_k=\alpha$. Otherwise, set $\alpha=\sigma\alpha$ and repeat this step.

Why does this work?

- By Taylor expansion, if α is sufficiently small, we have

$$f(x^k+\alpha d^k)\approx f(x^k)+\alpha\nabla f(x^k)^\top d^k<f(x^k)+\gamma\alpha\cdot\nabla f(x^k)^\top d^k.$$

Therefore, as long as α is small enough, the Armijo condition must be satisfied (recall $\nabla f(x^k)^\top d^k=-\|\nabla f(x^k)\|^2<0$).

Illustration:

- Define $\phi_k(\alpha):=f(x^k+\alpha d^k)-f(x^k)$. Then, we have

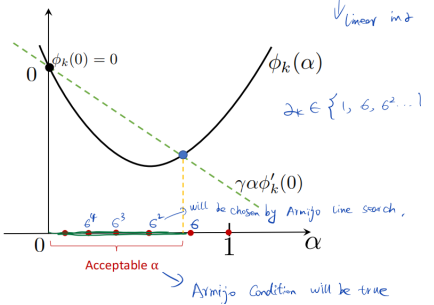
$$\phi'_k(\alpha)=\nabla f(x^k+\alpha d^k)^\top d^k,\quad\phi'_k(0)=\nabla f(x^k)^\top d^k.$$

- The Armijo condition is then equivalent to:

$$\text{find } \alpha \text{ with } \phi_k(\alpha)\le\gamma\alpha\cdot\phi'_k(0).$$

- Notice that $\phi'_k(0)<0$ (since d^k is a descent direction).

Armijo Line Search: Visualization



Example: Exact Line Search

Consider

$$f(x) = b^\top x + \frac{1}{2}x^\top Ax \quad (A \text{ positive definite})$$

At x^k, the gradient descent method will choose:

$$d^k = -\nabla f(x^k) = -(b + Ax^k).$$

To choose the step size, notice that we can explicitly compute

$$f(x^k + \alpha d^k) = b^\top(x^k + \alpha d^k) + \frac{1}{2}(x^k + \alpha d^k)^\top A(x^k + \alpha d^k)$$

$$= \frac{1}{2}\alpha^2 (d^k)^\top A d^k + \alpha(b^\top d^k + (x^k)^\top A d^k) + f(x^k)$$

This is a quadratic function of α with positive second-order term. We can find the optimal $\alpha \geq 0$ by minimizing $\phi(\alpha) = f(x^k + \alpha d^k)$:

$$\alpha_k = \frac{(d^k)^\top d^k}{(d^k)^\top A d^k}.$$

The Gradient Descent Method

Gradient Descent Method

1. Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.

For $k = 0, 1, \ldots$:

2. Pick a stepsize α^k by a line search procedure (exact line search or backtracking) on the function

$$\phi(\alpha) = f(x^k - \alpha \nabla f(x^k)).$$

3. Set

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k).$$

4. If $\|\nabla f(x^{k+1})\| \leq \epsilon$, then STOP and x^{k+1} is the output.

- In step 2, we may also use a constant stepsize by setting $\alpha_k = \bar{\alpha}$ for all $k \geq 0$.
- This constant stepsize is useful when performing line search is too time consuming (i.e., when evaluating the function value is very expensive).

Global Convergence:

- We show that the gradient method can find stationary points independent of the chosen initial point.
- We call such a property global convergence.

Local Convergence and Rate of Convergence:

- Under appropriate assumptions a rate (speed) of convergence can be established.

⇝ Guaranteed and quantifiable progress in each iteration.

Definition: Accumulation Point

A point x is an accumulation point of $\{x^k\}_k$ if for every $\epsilon > 0$, there are infinitely many numbers k with $x^k \in B_\epsilon(x)$.

Several remarks:

- If x is an accumulation point of $\{x^k\}_k$ then there exists a subsequence $\{x^{k_\ell}\}_\ell$ that converges to x.
- If $\{x^k\}_k$ converges to some $x \in \mathbb{R}^n$, then x is the unique accumulation point of $\{x^k\}_k$.
- Bolzano Weierstrass theorem: A bounded sequence always possesses at least one accumulation point.

Examples:

- The sequence $\{a_k\}_k$ with $a_k = (-1)^k$ has the two accumulation points $a = +1$ and $a = -1$.
- The sequence

$$a_k := \begin{cases} k & k \text{ is odd}, \\ 0 & k \text{ is even}, \end{cases}$$

is not bounded. However, it has the accumulation point $a = 0$.

Theorem: Global Convergence

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\{x^k\}_k$ be generated by the gradient method for solving

$$\min_{x \in \mathbb{R}^n} f(x)$$

with one of the following step size strategies:

- exact line search,
- Armijo line search (backtracking) with $\sigma, \gamma \in (0, 1)$.

Then, $\{f(x^k)\}_k$ is nonincreasing and every accumulation point of $\{x^k\}_k$ is a stationary point of f.

- If f is a polynomial function of the variables $x_1, x_2, \ldots, x_n$ and $\{x^k\}_k$ is bounded, the whole sequence $\{x^k\}_k$ converges to a stationary point x^* of f.
- Let x^* be an accumulation point of $\{x^k\}_k$ and suppose that the second order sufficient optimality conditions hold at x^*:

⇝ The sequence $\{x^k\}_k$ converges to the strict local minimizer x^*.

⇝ We require some additional properties to derive rates.

We need to assume that ∇f is Lipschitz continuous over $\mathbb{R}^n$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,$$

where $L \geqslant 0$ is the Lipschitz constant. The class of functions with Lipschitz gradient with constant L is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or $C_L^{1,1}$.

Theorem: Lipschitz Continuity via Hessians

Let f be a twice continuously differentiable function. Then, the following two conditions are equivalent:

- ∇f is Lipschitz continuous with parameter L.
- $\|\nabla^2 f(x)\|_{\text{op}} \leq L$ for any $x \in \mathbb{R}^n$.

The largest eigenvalue of Hessian Matrix

Descent Lemma

Suppose ∇f is Lipschitz continuous with parameter L. Then, we have

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

- This lemma shows that GD with a constant stepsize has descent property. Note that $x^{k+1} - x^k = -\bar{\alpha}\nabla f(x^k)$ by GD update, we have

$$f(x^{k+1}) \leq f(x^k) - \bar{\alpha}\left(1 - \frac{L\bar{\alpha}}{2}\right)\|\nabla f(x^k)\|^2.$$

- Thus, once we choose $\bar{\alpha} \in (0, \frac{2}{L})$, GD has descent property.

Definition: Linear Convergence

We say that $\{x^k\}_k$ converges linearly (linear convergence) with rate $\eta \in (0, 1)$ to $x^* \in \mathbb{R}^n$ if there is $\ell \geq 0$ such that

$$\|x^{k+1} - x^*\| \leq \eta \cdot \|x^k - x^*\|, \quad \forall\, k \geq \ell.$$

Theorem: Rates for (Strongly) Convex Problems

Let f have Lipschitz continuous gradient with parameter L and suppose there exists $\mu > 0$ such that

$$\mu\|d\|^2 \;\leq\; d^\top \nabla^2 f(x) d \;(\leq\; L\|d\|^2) \quad \forall\, d, \forall\, x.$$

Let $\{x^k\}_k$ be generated by the gradient method and let x^* be the unique solution of $\min_x f(x)$. Then:

$$\{x^k\}_k \text{ converges linearly to } x^*.$$

Setting $\eta = 1 - 2M\mu \in (0, 1)$ (see next slide for M), it further follows

$$f(x^{k+1}) - f(x^*) \leq \eta \cdot (f(x^k) - f(x^*)), \quad \forall k \geq 0.$$

Convergence Rate: Remarks

Remarks:

- In the theorem a stronger notion of convexity is required – the so-called strong convexity, i.e., the Hessian is PD for all x.
- The constant M depends on the chosen line search procedure:

$$M = \begin{cases} \bar{\alpha}(1 - \frac{L\bar{\alpha}}{2}) & \text{constant step size: } \bar{\alpha} \in (0, \frac{2}{L}), \\ \frac{1}{2L} & \text{exact line search,} \\ \gamma \min\{1, \frac{2\sigma(1-\gamma)}{L}\} & \text{Armijo line search.} \end{cases}$$

- The optimal rate can be achieved by constant stepsize with $\bar{\alpha} = \frac{1}{L}$ or exact line search. In this case, we have

$$\eta = 1 - \frac{\mu}{L} = 1 - \frac{1}{\kappa} \quad \text{where} \quad \kappa = \frac{L}{\mu}.$$

The rate $\bar{\eta}$ for $\{\|x^k - x^*\|\}_k$ typically has the form $\bar{\eta} = \frac{\kappa - 1}{\kappa + 1} \in (0, 1)$

We have seen that when using exact line search, the directions between consecutive steps are perpendicular, i.e.,

$$(d^{k+1})^\top d^k = 0$$

In fact, this is always true when using exact line search.

Why?

- If α_k is the minimizer of $\phi(\alpha) = f(x^k + \alpha d^k)$. Then, $\phi'(\alpha_k) = 0$ (FONC), which means:

$$0 = \phi'(\alpha_k) = \nabla f(x^k + \alpha_k d^k)^\top d^k = -(d^{k+1})^\top d^k.$$

Newton's Method One Dimension

We want to minimize $f : \mathbb{R} \to \mathbb{R}$:

- The FONC is $g(x) = f'(x) = 0$. We first try to find such stationary points.

Newton's method is an iterative method by approximately solving $g(x) = 0$ at each iteration.

- At each point x^k, we first approximate g using first-order Taylor expansion at x^k:

$$g(x) \approx g(x^k) + g'(x^k)(x - x^k)$$

- We set the right-hand side to be 0 and solve it:

$$x = x^k - \frac{g(x^k)}{g'(x^k)} \quad \text{and} \quad x^{k+1} = x.$$

- Here we assume $g'(x^k) \neq 0$ at each step.

Theorem: Convergence of Newton's Method

If g is twice continuously differentiable and x^* is a root of g at which $g'(x^*) \neq 0$, then provided that $|x^0 - x^*|$ is sufficiently small, the sequence generated by Newton's method:

$$x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)}$$

will satisfy

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

with $C = \frac{1}{2}\sup_{x \in \mathbb{I}} |g''(x)| \cdot \sup_{x \in \mathbb{I}} \left(\frac{1}{|g'(x)|}\right)$, where $\mathbb{I}$ is the interval $\{x : |x - x^*| \leq |x^0 - x^*|\}$.

- We call this convergence speed quadratic convergence.

Remember the gradient descent method can have linear convergence rate (strongly convex case):

$$|x^{k+1} - x^*| \leq \eta |x^k - x^*|.$$

Now, Newton's method has quadratic convergence rate:

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

One Newton step is given by:

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}$$

A gradient descent step is given by:

$$x^{k+1} = x^k - \alpha_k f'(x^k)$$

Observation:

- In the 1-D case, Newton's method simply specifies a specific stepsize in the gradient method (rather than performing line search).
- In the high-dimensional case, however, Newton's method will also change the search direction.

Newton's Method in High Dimensional Case

Newton direction: $d^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$.

Recall that a vector d is a descent direction if $\nabla f(x)^\top d < 0$.

In Newton's method, we have

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

Then, it holds that:

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

- If f is convex, then $\nabla^2 f(x)$ is positive semidefinite and we obtain $\nabla f(x)^\top d \leq 0$.
- If $\nabla^2 f(x)$ is positive definite (f is strongly convex), then $\nabla f(x)^\top d < 0$.

⇝ In this case, Newton's direction is a descent direction.

Step Length

- As we said earlier, Newton's method may not converge unless the starting point is close ⇝ sensitive to initial point.
- One possible way to ensure convergence is to again use a stepsize parameter α_k in

$$x^{k+1} = x^k + \alpha_k d^k$$

where $d^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$ is Newton's direction.

- We can use backtracking line search to determine α_k (which requires that the Newton's direction d^k is a descent direction...).

Complete Procedure of Newton's Method

The Newton Method

1. Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.

For $k = 0, 1, \ldots$:

2. Compute the Newton direction d^k which is the solution of the Newton system

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k).$$

3. Choose a step size α_k by backtracking line search and calculate $x^{k+1} = x^k + \alpha_k d^k$.
4. If $\|\nabla f(x^{k+1})\| \leq \epsilon$, then STOP and x^{k+1} is the output.

- We can also check whether d^k is a descent direction:

$$-\nabla f(x^k)^\top d^k \geq \gamma_1 \min\{1, \|d^k\|^{\gamma_2}\}\|d^k\|^2, \quad \gamma_1, \gamma_2 \in (0, 1).$$

Otherwise, we apply gradient descent step: $d^k = -\nabla f(x^k)$.

- Such a procedure is to globalize Newton's method, which ensures descent of Newton's method and make it insensitive of initial point.

Theorem: Convergence of Newton's Method

Let f be twice continuously differentiable and let x^* be a local minimizer of f. For some given $\epsilon > 0$ assume that:

- There exists $\mu > 0$ with $\nabla^2 f(x) \succeq \mu I$ for any $x \in B_\epsilon(x^*)$.
- There exists $L > \mu$ with $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ for all $x, y \in B_\epsilon(x^*)$.

Let $\{x^k\}_k$ be generated by Newton's method. Then for $k = 0, 1, \ldots$

$$\|x^{k+1} - x^*\| \leq \frac{L}{2\mu}\|x^k - x^*\|^2$$

and in addition, if $\|x^0 - x^*\| \leq \frac{\mu \min\{1, \epsilon\}}{L}$, then

$$\|x^k - x^*\| \leq \frac{2\mu}{L}\left(\frac{1}{2}\right)^{2^k}, \quad k = 0, 1, 2, \ldots$$

Constrained Optimization

We consider the following constrained optimization problem:

$$\begin{aligned} &\underset{x}{\text{minimize}} \; f(x) \\ &\text{subject to } x \in \Omega. \end{aligned} \tag{1}$$

where $\Omega \subset \mathbb{R}^n$ is a convex and closed set.

Under Additional Constraints:

- x^{k+1} can become infeasible.

Project the gradient step back onto the feasible set:

Definition: Euclidean Projection / Orthogonal Projection

Let $\Omega \subset \mathbb{R}^n$ be a nonempty, closed, convex set. The (Euclidean / orthogonal) projection of y onto Ω is defined as the unique minimizer y^* of the constrained optimization problem:

$$\min_x \frac{1}{2}\|x - y\|^2 \quad \text{s.t.} \quad x \in \Omega$$

and we write $y^* = \mathcal{P}_\Omega(y)$.

Observation:

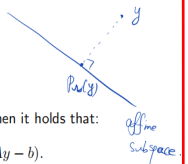
- The projection $y^* = \mathcal{P}_\Omega(y)$ is the point in Ω that has the minimum distance to y.

Example I: Linear Constraints

We first consider the simple case where Ω consists of linear equality constraints:

$$\Omega = \{x : Ax = b\},$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given.



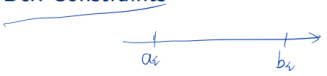
Euclidean Projection:

- Suppose that A has full row rank ($m \le n$), then it holds that:

$$\mathcal{P}_\Omega(y) = y - A^\top(AA^\top)^{-1}(Ay - b).$$

- This yields a special case when $\Omega = \{x : a^\top x = b\}$, i.e., projection onto subspace. We have $\mathcal{P}_\Omega(y) = y - \frac{a^\top y - b}{\|a\|_2^2}a$.

Example II: Box Constraints



Suppose that Ω is given by box constraints:

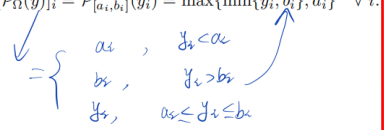
$$\Omega = \{x \in \mathbb{R}^n : x_i \in [a_i, b_i],\ \forall\ i\} = [a, b],$$

where $a, b \in \mathbb{R}^n$, $a \le b$, are given.

Euclidean Projection:

- The projection onto Ω can be computed as follows:

$$[\mathcal{P}_\Omega(y)]_i = \mathcal{P}_{[a_i,b_i]}(y_i) = \max\{\min\{y_i, b_i\}, a_i\} \quad \forall\ i.$$



Example III: Ball Constraints

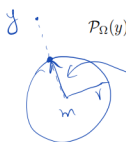
Suppose that Ω is a Euclidean ball with radius $r > 0$ and center $m \in \mathbb{R}^n$, i.e.:

$$\Omega = \{x \in \mathbb{R}^n : \|x - m\| \le r\}.$$

Euclidean Projection:

- The projection onto Ω can be computed as follows:

$$\mathcal{P}_\Omega(y) = \begin{cases} y & \text{if } \|y - m\| \le r, \\ m + \frac{r}{\|y-m\|}(y - m) & \text{if } \|y - m\| \ge r. \end{cases}$$



Observation:

- For many sets, explicit formulae for the projections can be derived.
- Main Tool: KKT-conditions and convexity.
- Many more interesting projections can be expressed efficiently:

$$\mathcal{P}_{\Delta_n}(y), \quad \mathcal{P}_{\mathbb{S}^n_+}(Y), \quad \ldots$$

where $\Delta_n := \{x : \mathbb{1}^\top x = 1,\ x \ge 0\}$ is the n-simplex and $\mathbb{S}^n_+$ is the set of positive semidefinite matrices.

- In general: An optimization problem needs to be solved to obtain $\mathcal{P}_\Omega(y)$.

Optimization Problems with Convex Constraints

Recall the constrained optimization problem (1) as following:

$$\min_x\ f(x) \quad \text{s.t.} \quad x \in \Omega,$$

where $\Omega \subset \mathbb{R}^n$ is a convex and closed set.

- We can derive the following optimality condition:

Theorem: FONC for Problems with Convex Constraints

Let f be continuously differentiable on an open set that contains the convex and closed set $\Omega \subset \mathbb{R}^n$. Let $x^* \in \Omega$ be a local minimizer of (1), then:

$$\nabla f(x^*)^\top(x - x^*) \ge 0, \quad \forall\ x \in \Omega. \tag{2}$$

- If f is convex, then $x^* \in \Omega$ is global minimizer of (1) iff the FONC is satisfied.
- A point satisfying (2) is called a stationary point of (1).

Proof (read it)

If x^* is a local minimizer, then it must satisfy the following first-order optimality condition (abstract FONC):

$$\nabla f(x^*)^\top d \ge 0, \quad \forall d \in S_\Omega(x^*).$$

Here, $S_\Omega(x^*)$ is the set of feasible directions: $d \in S_\Omega(x^*)$ is equivalent to $\exists \bar{t} > 0$, $x^* + td \in \Omega$ for all $t \in (0, \bar{t})$.

Take $x \in \Omega$ arbitrarily, then by convexity of Ω,

$$tx^* + (1-t)x = x^* + (1-t)(x - x^*) \in \Omega, \quad \forall t \in [0, 1].$$

Hence, $x - x^* \in S_\Omega(x^*)$ for any $x \in \Omega$. This, together with the abstract FONC, implies

$$\nabla f(x^*)^\top(x - x^*) \ge 0.$$

Projection Theorem

⇝ Projections are a special case with

$$\min_x \left\{ f(x) = \frac{1}{2}\|x - y\|^2 \right\}, \quad \text{s.t.} \quad x \in \Omega$$

Projection Theorem

Let Ω be a nonempty, closed, and convex set. Then:

- A point y^* is the projection of y onto Ω, i.e., $y^* = \mathcal{P}_\Omega(y)$, if and only if

$$(y^* - y)^\top(x - y^*) \ge 0, \quad \forall\ x \in \Omega.$$

- The mapping $\mathcal{P}_\Omega : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $L = 1$.

$$\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\| \le 1 \cdot \|x - y\|$$

Proof Cauchy-Schwarz inequality: $a^\top b \le \|a\| \cdot \|b\|$

The first iff condition is a direction consequence of the former FONC and (strong) convexity.

We now prove the second Lipschitz condition. Consider $\mathcal{P}_\Omega(x)$ and $\mathcal{P}_\Omega(y)$, by the first iff condition, we have

$$(\mathcal{P}_\Omega(x) - x)^\top(z - \mathcal{P}_\Omega(x)) \ge 0, \quad \forall z \in \Omega \quad (\text{choose } z = \mathcal{P}_\Omega(y))$$
$$(\mathcal{P}_\Omega(y) - y)^\top(z - \mathcal{P}_\Omega(y)) \ge 0, \quad \forall z \in \Omega \quad (\text{choose } z = \mathcal{P}_\Omega(x))$$

Sum these two inequalities

Use the second inequality minus the first one, we have

$$\begin{aligned} 0 &\le (\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y))^\top(\mathcal{P}_\Omega(y) - \mathcal{P}_\Omega(x) + x - y) \\ &= -\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|^2 + (\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y))^\top(x - y) \end{aligned}$$

This gives

$$\begin{aligned} \|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|^2 &\le (\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y))^\top(x - y) \\ &\le \|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\| \cdot \|x - y\|, \end{aligned}$$

where we have used Cauchy-Schwarz inequality.

Rewritten FONC for Problem with Convex Constraints

Recall again problem (1):

$$\min_x\ f(x) \quad \text{s.t.} \quad x \in \Omega,$$

Through projection theorem, its FONC can be rewritten.

Theorem: FONC for Problem with Convex Constraints

The vector x^* is a stationary point of (1) if and only if

$$x^* - \mathcal{P}_\Omega(x^* - \lambda\nabla f(x^*)) = 0 \quad \text{for any } \lambda > 0.$$

Proof: $x^* = \mathcal{P}_\Omega(x^* - \lambda\nabla f(x^*))$ Which is equivalent to

$$(x^* - (x^* - \lambda\nabla f(x^*))^\top(x - x^*) \ge 0, \quad \forall x \in \Omega,$$

which is further equivalent to

$$\lambda\nabla f(x^*)^\top(x - x^*) \ge 0, \quad \forall x \in \Omega.$$

Since $\lambda > 0$, it is equivalent to x^* is a stationary point of (1).

⇝ Next we will study projected gradient method formally.

Motivation & Strategy:

- Setting $x^{k+1} = x^k - \lambda_k\nabla f(x^k)$ might likely generate infeasible iterates: $x^{k+1} \notin \Omega$.
- Idea: We project the step $x^k - \lambda_k\nabla f(x^k)$ back onto Ω:

$$x^{k+1} = \mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)).$$
$$x^{k+1} = \mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)) = x^k + [\mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)) - x^k].$$

- This is close to our usual update form: $x^{k+1} = x^k + \alpha_k d^k$.
- Setting $d^k = \mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)) - x^k$, we can consider:

$$x^{k+1} = x^k + \alpha_k d^k = (1 - \alpha_k)x^k + \alpha_k\mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)).$$

⇝ If $\alpha_k \in [0, 1]$, then the convexity of Ω implies that x^{k+1} will be feasible if $x^k \in \Omega$. often the case

Lemma: Descent Direction

Let $x \in \Omega$ and $\lambda > 0$ be given. If x is not a stationary point of (1), then the direction $d := \mathcal{P}_\Omega(x - \lambda\nabla f(x)) - x$ is a descent direction and it holds that

$$\nabla f(x)^\top d \le -\frac{1}{\lambda}\|d\|^2 < 0.$$

generalized gradient

- We can now reuse similar techniques like backtracking to generate step sizes.
- We can use $\|d^k\| = \|\mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)) - x^k\| \le \epsilon$ as stopping criterion, inspired by formerly re-written FONC.

⇝ Projected gradient method mimics the gradient descent method by changing $d^k = -\nabla f(x^k)$ to $d^k = \mathcal{P}_\Omega(x^k - \lambda_k\nabla f(x^k)) - x^k$, which is justified by FONC and the above lemma.

$$\begin{aligned} \nabla f(x)^\top d &= \nabla f(x)^\top(\mathcal{P}_\Omega(x - \lambda\nabla f(x)) - x) \\ &= \nabla f(x)^\top(\mathcal{P}_\Omega(x - \lambda\nabla f(x)) - \mathcal{P}_\Omega(x)) \\ &= -\frac{1}{\lambda}((x - \lambda\nabla f(x)) - x)^\top(\mathcal{P}_\Omega(x - \lambda\nabla f(x)) - \mathcal{P}_\Omega(x)) \\ &\le -\frac{1}{\lambda}\|\mathcal{P}_\Omega(x - \lambda\nabla f(x)) - \mathcal{P}_\Omega(x)\|^2 \\ &\le -\frac{1}{\lambda}\|\mathcal{P}_\Omega(x - \lambda\nabla f(x)) - x\|^2 \\ &= -\frac{1}{\lambda}\|d\|^2. \end{aligned}$$

use the proof of projection theorem

Theorem: Convergence Projected Gradient Method

Assumptions: f is continuously differentiable, Ω is nonempty, convex, and closed and the step sizes $\{\lambda_k\}_k$ are bounded: $0 < \underline{\lambda} \le \lambda_k \le \lambda \quad \forall\ k$.

Every accumulation point of $\{x^k\}$ is a stationary point.

If ∇f is additionally Lipschitz continuous, we obtain:

- If f is convex, we converge to global solutions of the problem.
- If $\lambda_k \in (0, \frac{2}{L})$, then $\alpha_k = 1$ will be accepted as step size in the backtracking line search procedure.
- If f is (strongly) convex, then $\{x^k\}_k$ converges linearly to a global solution $x^* \in \Omega$.

An integer linear program (IP) is a linear program with the additional constraint that all variables must be integers:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & c^\top x \\ \text{subject to} \quad & Ax = b \\ & x \ge 0 \\ & x \in \mathbb{Z}^n \end{aligned}$$

Here, we use $\mathbb{Z}$ to denote the set of integers.

- One may also encounter mixed integer programs (MIP), in which one set of variables must be integer and the rest are allowed to be continuous.
- A special case is binary integer constraint $x_i \in \{0, 1\}, \forall i$.

Example: Knapsack Problem

John is planning a trip. There are n items he would like to bring.

- The ith item has value v_i.
- The weight of ith item is a_i.
- His bag has a maximum allowable weight C.
- He wants to bring as much value as possible.

Decision Variables:

- x_i: whether to bring ith item or not: $x_i \in \{0, 1\}$.

Optimization Problem:

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & \textstyle\sum_{i=1}^n v_i x_i \\ \text{subject to} \quad & \textstyle\sum_{i=1}^n a_i x_i \le C \\ & x_i \in \{0, 1\} \quad \forall\ i. \end{aligned}$$

Example: Matching Problem

Setup: Assume there are n girls and n boys' information in a dating website.

- Each girl i has rated boy j with a score v_{ij}.
- The website wants to match one girl with one boy so as to maximize

Decision Variables:

- x_{ij}: whether girl i is matched with boy j or not: $x_{ij} \in \{0, 1\}$.

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^n\sum_{j=1}^n v_{ij}x_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^n x_{ij} = 1 \quad \forall\ j, \quad \sum_{j=1}^n x_{ij} = 1 \quad \forall\ i \\ & x_{ij} \in \{0, 1\} \end{aligned}$$

Facility Location Problem:

Decision Variables:

- y_i: whether to open warehouse i or not: $y_i \in \{0, 1\}$.
- x_{ij}: how many units to ship from warehouse i to customer j.

$$\begin{aligned} \underset{x,y}{\text{minimize}} \quad & \textstyle\sum_{i=1}^n b_i y_i + \sum_{i,j} c_{ij}x_{ij} \\ \text{subject to} \quad & \textstyle\sum_{i=1}^n x_{ij} \ge d_j, && \text{for all } j \\ & x_{ij} \le d_j y_i, && \text{for all } i, j \\ & x_{ij} \ge 0 \\ & y_i \in \{0, 1\} \end{aligned}$$

Theorem: LP Relaxation as a Bound for IP

1. For a maximization problem, the optimal value of the LP relaxation provides an upper bound for the optimal value of the IP.
2. For a minimization problem, the optimal value of the LP relaxation provides a lower bound for the optimal value of the IP.

The difference between the optimal value of the LP and the IP is called the integrality gap.

- For maximization problems, the integrality gap is $v^{LP} - v^{IP}$.
- For minimization problems, the integrality gap is $v^{IP} - v^{LP}$.

Theorem

If the optimal solution to the LP relaxation is integer, then the solution must be optimal to the IP problem.

Definition: Total Unimodulartiy (TU)

A matrix A is said to be totally unimodular if the determinant of each submatrix of A is either 0, 1, or -1.

Theorem: Total Unimodularity and Integer Solutions

If the constraint matrix A is totally unimodular and b is an integer vector, then all the BFS are integers and the LP relaxation must have an optimal solution that is an integer solution.

Branching Procedures:

1. Solve the LP relaxation.
   - If the optimal solution is integral, then it is optimal to IP.
   - Otherwise go to step 2.
2. If the optimal solution to the LP relaxation is x^* and x_i^* is fractional, then branch the problem into the following two:
   - One with an added constraint that $x_i \le \lfloor x_i^* \rfloor$.
   - One with an added constraint that $x_i \ge \lceil x_i^* \rceil$.
3. For each of the two problems, use the same method to solve them, and get optimal sol. y_1^* and y_2^* with optimal value v_1^* and v_2^*.
   - Compare to obtain the optimal solution.

Bounding procedures (for maximization):

- Any LP relaxation solution can provide an upper bound for each node in the branching process.
- Any feasible point to the IP can provide a lower bound for the entire problem.

When the optimal value of the LP relaxation of this branch is less than the current lower bound, then we can discard this branch.

- No better solution can be obtained from further exploring this branch.

Bounding is very important for branch-and-bound, it is the key to make it efficient (and practical).