Optimization

Wenxiao Yang[*]

[*]Department of Mathematics, University of Illinois at Urbana-Champaign

2022

Contents

1 Unconstrained Optimization **5**
1.1 Conditions for Optimality 5
1.2 Global minimizer, Local minimizer 5
1.3 Optimization in $\mathbb{R}$ 5
1.3.1 Theorem: local minimizer $\Rightarrow f'(x^*) = 0$ 5
1.3.2 Theorem: $f'(x^*) = 0, f''(x^*) \geq 0 \Rightarrow$ local minimizer 5
1.4 Optimization in $\mathbb{R}^n$ 5
1.4.1 First Order Necessary Condition: $\nabla f(x^*) = 0$ 5
1.4.2 Stationary Point, Saddle Point 6
1.4.3 Second Order Necessary Condition $\nabla^2 f(x^*) \succeq 0$ 6
1.4.4 Sufficient Conditions for Optimality $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$ 7
1.4.5 Using Optimality Conditions to Find Minimum 7
1.4.6 Fix Conditions for Global Optimality 8
1.5 Optimization in a Set 8
1.5.1 Existence of Global-min 8
1.6 Method of finding-global-min-among-stationary-points (FGMSP) 9

2 Convexity **9**
2.1 Definition 9
2.2 Convex$\Rightarrow$Stationary point is global-min 11
2.3 Unconstrained Quadratic Optimization 11
2.4 Strongly Convexity 12
2.4.1 μ-Strongly Convex: $\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2$ 12
2.4.2 μ-strongly convex $\Leftrightarrow \nabla^2 f(x) \succeq \mu I$ 12
2.4.3 Lemma: Strongly convexity $\Rightarrow$ Strictly convexity 13
2.4.4 Lemma: $\nabla^2 f(x) \succeq mI \Rightarrow f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|^2$ 13

3 Gradient Methods **13**
3.1 Steepest Descent 14
3.2 Methods for Choosing α_k 14
3.3 Optimal(Exact) Line Search 14
3.4 Armijo's Rule 15
3.5 Armijo's Rule for Steepest Descent 15

4 Convergence of GD with Constant Stepsize **16**
4.1 Lipschitz Gradient (L-Smooth) . . . 16
4.1.1 Theorem: $-MI \preceq \nabla^2 f(x) \preceq MI \Rightarrow \nabla f(x)$ is Lipschitz with constant M . . . 16
4.1.2 Descent Lemma: $\nabla f(x)$ is Lipschitz with constant $L \Rightarrow f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$ . . . 17
4.1.3 Co-coercivity Condition: $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$ . . . 18
4.2 Convergence of Steepest Descent with Fixed Stepsize . . . 18
4.2.1 Theorem: f has Lipschitz gradient $\Rightarrow \{x_k\}$ converges to stationary point . . . 18
4.3 Convergence of GD for convex functions . . . 20
4.3.1 Theorem: f is convex and has Lipschitz gradient $\Rightarrow f(x_k)$ converges to global-min value with rate $\frac{1}{k}$ . . . 20
4.4 Convergence of GD for strongly convex functions . . . 21
4.4.1 Theorem: Strongly convex, Lipschitz gradient $\Rightarrow \{x_k\}$ converges to global-min geometrically . . . 21
4.4.2 Example . . . 22
4.5 Convergence of Gradient Descent on Smooth Strongly-Convex Functions . . . 23
4.6 From convergence rate to iteration complexity . . . 25

5 Newton's Method **26**
5.1 Generalization to Optimization . . . 26
5.2 A New Interpretation of Newton's Method . . . 26
5.3 Convergence of Newton's Method . . . 27
5.4 Note: Cons and Pros . . . 28
5.5 Modifications to ensure global convergence . . . 29
5.6 Quasi-Newton Methods . . . 29
5.6.1 BFGS Method . . . 29
5.7 Trust-Region Method . . . 31
5.8 Cubic Regularization . . . 31

6 Neural Networks **31**
6.1 Neuron . . . 31
6.2 Multilayer Neural Network . . . 32
6.3 Back Propagation Algorithm . . . 33
6.4 Other Methods . . . 35

7 Constrained Optimization and Gradient Projection **35**
7.1 Constrained Optimization: Basic . . . 35
7.1.1 Def: Optimality . . . 35
7.1.2 Prop: local-min$\Rightarrow \nabla f(x^*)^T(x-x^*) \geq 0, \forall x \in \&$ $\Leftrightarrow$ global-min in convex . . . 35
7.1.3 Def: Interior Point . . . 36
7.2 Constrained Optimization Example . . . 36
7.3 Projection onto Closed Convex Set . . . 36
7.3.1 Def: Projection $[z]^{\&}$ . . . 36
7.3.2 Prop: unique projection $[z]^{\&}$ on closed convex subset of $\mathbb{R}^n$ . . . 37
7.3.3 Projection Theorem: $x = [z]^{\&}$ is projection on closed conex subset of $\mathbb{R}^n \Leftrightarrow (z-x)^T(y-x) \leq 0, \forall y \in \&$ . . . 37
7.3.4 Prop: Projection is non-expansive $\|[x]^{\&} - [z]^{\&}\| \leq \|x-z\|, \forall x, z \in \mathbb{R}^n$ . . . 38
7.4 Projection on (Linear) Subspaces of $\mathbb{R}^n$ . . . 38
7.4.1 Orthogonality Principle in subspaces of $\mathbb{R}^n$: $(z-y^*)^T x = 0, \forall x \in \&$ . . . 38
7.5 Gradient Projection Method . . . 39

7.5.1 Def: fixed point in fixed step-size steepest descent method, $\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$ 39
7.5.2 Prop: $L-$smooth, $0 < \alpha < \frac{2}{L} \Rightarrow$ limit point is a fixed point (in fixed step-size steepest descent method) . . . 40
7.5.3 Prop: x is minimizer in convex func $\Leftrightarrow$ fixed point (in fixed step-size steepest descent method) . . . 40
7.5.4 Thm: Convergence of Gradient Projection: Convex, $L-$smooth, $0 < \alpha < \frac{2}{L} \Rightarrow$ $f(x_k) \to f(x^*)$ at rate $\frac{1}{k}$ . . . 40
7.5.5 Thm: Strongly convex, Lipschitz gradient $\Rightarrow \{x_k\}$ converges to x^* geometrically 41

8 Optimization with Equality Constraints **41**
8.1 Basic . . . 41
8.2 Lagrange Mutiplier Theorem . . . 41
8.2.1 First-order necessary condition: $\exists \lambda, \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0$ . . . 41
8.2.2 Second-order necessary condition: $z^T \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*)\right) z \geq 0, \forall z \in V(x^*)$ . . . 43
8.2.3 Sufficient Condition: $\exists \lambda$ 1. $\nabla f(x^*)+\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0$ 2. $z^T \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*)\right) z > 0, \forall z \in V(x^*), z \neq 0$ . . . 44
8.2.4 Lagrangian Function . . . 44
8.2.5 Example . . . 44
8.2.6 Sensitivity Analysis $f(x^*(u)) = f(x^*) - \lambda^T u + O(\|u\|)$ . . . 45
8.2.7 Linear Constraints . . . 46

9 Optimization with Inequality Constraints **47**
9.1 Basic . . . 47
9.1.1 Active vs. Inactive Inequality Constraints . . . 47
9.1.2 ICP $\to$ ECP . . . 47
9.1.3 Intuition $\mu_j \geq 0, \forall j \in A(x^*)$ . . . 48
9.1.4 Complementary Slackness . . . 48
9.2 Karush–Kuhn–Tucker (KKT) Necessary Conditions . . . 48
9.3 Karush–Kuhn–Tucker (KKT) Sufficient Conditions . . . 50
9.4 General Sufficiency Condition . . . 51
9.5 Barrier Method . . . 52
9.6 An Exmaple Using KKT or Barrier . . . 53
9.6.1 Solution using KKT conditions . . . 53
9.6.2 Solution using logarithmic barrier . . . 54
9.7 Penalty Method (For ECP) . . . 54

10 Duality **55**
10.1 Weak Duality Theorem: $\max_{(\lambda,\mu)\in G} D(\lambda, \mu) \leq \min_{x \in F} f(x)$ . . . 56
10.2 Strong Duality Theorem: under some conditions, $\max_{(\lambda,\mu)\in G} D(\lambda, \mu) = \min_{x \in F} f(x)$. 56
10.2.1 Slater's sufficient condition for strong duality . . . 58
10.2.2 Example . . . 58
10.3 Dual of Linear Program . . . 58

11 Augmented Lagrangian Method (adjusted penalty method) **59**
11.1 Motivation . . . 59
11.2 Augemented Lagrangian Method . . . 59
11.2.1 Method of Multipliers . . . 60

12 Sub-gradient Methods **61**
12.1 Sub-gradient 62
12.2 Sub-differential 62
12.3 More examples 63
12.4 First-order necessary conditions for optimality in terms of subgradient 64
12.5 Properties of Subgradients 64
12.6 Sub-gradient Descent for Unconstrained Optimization 64

1 Unconstrained Optimization

1.1 Conditions for Optimality

Function: $f : \mathbb{R}^n \to \mathbb{R}^n$, $x \in \&$, $\& \subseteq \mathbb{R}^n$.
Terminology: x^* will always be the optimal input at some function.

1.2 Global minimizer, Local minimizer

Definition 1.
Say x^* is a global minimizer(minimum) of f if $f(x^*) \leq f(x), \forall x \in \&$.
Say x^* is a unique global minimizer(minimum) of f if $f(x^*) < f(x), \forall x \neq x^*$.
Say x^* is a local minimizer(minimum) of f if $\exists r > 0$ so that $f(x^*) \leq f(x)$ when $\|x - x^*\| < r$.

A minimizer is strict if $f(x^*) < f(x)$ for all relevant x.

1.3 Optimization in $\mathbb{R}$

1.3.1 Theorem: local minimizer $\Rightarrow f'(x^*) = 0$

Theorem 1. *If $f(x)$ is differentiable function and x^* is a local minimizer, then $f'(x^*) = 0$.*

Proof.
Def of $f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$
Def of local minimizer: $f(x^*) - f(x) \geq 0, |x^* - x| < r$
when $0 < h < r$, $\frac{f(x+h)-f(x)}{h} \geq 0$; when $-r < h < 0$, $\frac{f(x+h)-f(x)}{h} \leq 0$. Then $f'(x) = 0$. $\square$

1.3.2 Theorem: $f'(x^*) = 0, f''(x^*) \geq 0 \Rightarrow$ local minimizer

Theorem 2. *If $f : \mathbb{R} \to \mathbb{R}$ is a function with a continuous second derivative and x^* is a critical point of f (i.e. $f'(x) = 0$), then:*
(1): If $f''(x) \geq 0,\ \forall x \in \mathbb{R}$, then x^* is a global minimizer on $\mathbb{R}$.
(2): If $f''(x) \geq 0,\ \forall x \in [a, b]$, then x^* is a global minimizer on $[a, b]$.
(3): If we only know $f''(x^*) \geq 0$, x^* is a local minimizer.

Proof.
(1)$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2}f''(\xi)(x - x^*)^2 = f(x^*) + 0 + \textit{something non negative} \geq f(x^*)\ \forall x$
(2) Similar to (1)
(3)$f''(x^*) \geq 0,\ f''$ continuous $\Rightarrow \exists r$ s.t. $f''(x) \geq 0\ \forall x \in [x^* - \frac{r}{2}, x^* + \frac{r}{2}]$, then x is a local minimizer. $\square$

1.4 Optimization in $\mathbb{R}^n$

1.4.1 First Order Necessary Condition: $\nabla f(x^*) = 0$

A base point x, we consider an arbitrary direction u. $\{x + tu | t \in \mathbb{R}\}$
For $\alpha > 0$ sufficiently small:

1. $f(x^*) \leq f(x^* + \alpha u)$
2. $g(\alpha) = f(x^* + tu) - f(x^*) \geq 0$
3. $g(\beta)$ is continuously differentiable for $\beta \in [0, \alpha]$

By chain rule,

$$g'(\beta) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i$$

By Mean Value Theorem,

$$g(\alpha) = g(0) + g'(\beta)\alpha \text{ for some } \beta \in [0, \alpha]$$

Thus

$$g(\alpha) = \alpha \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i \geq 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i \geq 0$$

Letting $\alpha \to 0$ and hence $\beta \to 0$, we get

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*)u_i \geq 0 \text{ for all } u \in \mathbb{R}^n$$

By choosing $u = [1, 0, ..., 0]^T$, $u = [-1, 0, ..., 0]^T$, we get

$$\frac{\partial f(x^*)}{\partial x_1} \geq 0, \ \frac{\partial f(x^*)}{\partial x_1} \leq 0 \Rightarrow \frac{\partial f(x^*)}{\partial x_1} = 0$$

Similarly, we can get

$$\nabla f(x^*) = [\frac{\partial f(x^*)}{\partial x_1}, \frac{\partial f(x^*)}{\partial x_2}, ..., \frac{\partial f(x^*)}{\partial x_n}]^T = 0$$

Theorem 3. *If f is continuously differentiable and x^* is a local extremum. Then $\nabla f(x^*) = 0$.*

1.4.2 Stationary Point, Saddle Point

All points x^* s.t. $\nabla f(x^*) = 0$ are called stationary points.
Thus, all extrema are stationary points.
But not all stationary points have to be extrema.
Saddle points are the stationary points neither local minimum nor local maximum.

Example 1. *$f(x) = x^3$, $x = 0$ is a stationary point but not extrema. (saddle point)*

1.4.3 Second Order Necessary Condition $\nabla^2 f(x^*) \succeq 0$

Definition 2. *The Hessian of f at point x is an $n \times n$ symmetric matrix denoted by $\nabla^2 f(x)$ with $[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$*

Theorem 4. *Suppose f is twice continuously differentiable and x^* in local minimum. Then*

$$\nabla f(x^*) = 0 \text{ and } \nabla^2 f(x^*) \succeq 0$$

Proof.
$\nabla f(x^*) = 0$ already proved before.
Let α be small enough so that $g(\alpha) = f(x^* + \alpha u) - f(x^*) \geq 0$.

By Taylor series expansion,

$$
\begin{aligned}
g(\alpha) &= g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(0) + O(\alpha^2) \\
g'(\alpha) &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u) u_i = \nabla f(x^* + \alpha u)^T u \\
g''(\alpha) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x^* + \beta u) u_i u_j = u^T \nabla^2 f(x^* + \alpha u) u
\end{aligned}
$$

$$
\begin{aligned}
g'(0) = \nabla f(x^*)^T u = 0;\ g''(0) &= u^T \nabla^2 f(x^*) u \\
g(\alpha) &= \frac{\alpha^2}{2} u^T \nabla^2 f(x^*) u + O(\alpha^2) \geq 0 \\
\text{When } \alpha \to 0, \text{ we get } u^T \nabla^2 f(x^*) u &\geq 0,\ \forall u \in \mathbb{R}^n \\
&\Rightarrow \nabla^2 f(x^*) \succeq 0
\end{aligned}
$$

□

1.4.4 Sufficient Conditions for Optimality $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$

Theorem 5. *Suppose f is twice continuously differentiable in a neighborhood of x^* and (1) $\nabla f(x^*) = 0$; (2) $\nabla^2 f(x^*) \succ 0$ ($u^T \nabla^2 f(x^*) u > 0$, $\forall u \in \mathbb{R}^n$). Then x^* is local minimum.*

Proof.
Consider $u \in \mathbb{R}^n$, $\alpha > 0$ and let

$$
\begin{aligned}
g(\alpha) &= f(x^* + \alpha u) - f(x^*) \\
&= \frac{\alpha^2}{2} u^T \nabla^2 f(x^*) u + O(\alpha^2) \geq 0 \\
&= \frac{\alpha^2}{2} [u^T \nabla^2 f(x^*) u + 2\frac{O(\alpha^2)}{\alpha^2}] \\
& u^T \nabla^2 f(x^*) u > 0;\ \frac{O(\alpha^2)}{\alpha^2} \to 0 \\
& \Rightarrow g(\alpha) > 0 \text{ for } \alpha \text{ sufficiently small for all } u \neq 0 \\
& \Rightarrow x^* \text{ is local minimum.}
\end{aligned}
$$

(specially if $\|u\| = 1$, $u^T \nabla^2 f(x^*) u \geq \lambda_{\min}(\nabla^2 f(x^*))$, $\lambda_{\min}(\nabla^2 f(x^*))$ is the minimal eigenvalues of $\nabla^2 f(x^*)$.) □

1.4.5 Using Optimality Conditions to Find Minimum

1. Find all points satisfying necessary condition $\nabla f(x) = 0$ (all stationary points)
2. Filter out points that don't satisfy $\nabla^2 f(x) \geq 0$
3. Points with $\nabla^2 f(x) > 0$ are strict local minimum.
4. Among all points with $\nabla^2 f(x) \geq 0$, declare a global minimum, one with the smallest value of f, assuming that global minimum exists.

Example 2. $f(x) = 2x^2 - x^4$

$$
\begin{aligned}
f'(x) &= 4x - 4x^3 = 0 \\
&\Rightarrow x = 0, x = 1, x = -1 \text{ are stationary points} \\
f''(x) &= 4 - 12x^2 = \begin{cases} 4 & \text{if } x = 0 \\ -8 & \text{if } x = 1, -1 \end{cases} \\
&\Rightarrow x = 0 \text{ is the only local min, and it is strict}
\end{aligned}
$$

But $-f(x) \to \infty$ as $|x| \to \infty \Rightarrow$ no global min, but global max exists. $f(1), f(-1)$ are strict local max and both global max.

1.4.6 Fix Conditions for Global Optimality

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^*.

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

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

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

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

Combining the two inequalities, we have $f(\hat{x}) \leq f(x^*) \leq f(\hat{x})$, thus $f(\hat{x}) = f(x^*)$. Plugging into the second inequality, we have $f(x^*) \leq f(x), \forall x$. Thus x^* is a global minimizer of f^*. □

1.5 Optimization in 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

1.5.1 Existence of Global-min

Theorem 6 (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 1 (bounded level sets). *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. If for a certain c, the level set*

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

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

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

Example 3. $f(x) = x^2$. *Level set* $\{x|x^2 \leq 1\}$ *is* $\{x| -1 \leq x \leq 1\}$*: non-empty compact. Thus there exists a global minimum.*

Corollary 2 (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.*

Proof. Let $\alpha \in \mathbb{R}^d$ be chosen so that the set $S = \{x|f(x) \leq \alpha\}$ is non-empty. By coercivity, this set is compact. □

Coercive $\Rightarrow$ one non-empty bounded level set; but not the other way.
Claim (all level sets bounded $\Leftrightarrow$ coercive): Let f be a continuous function, then f is coercive iff $\{x|f(x) \leq \alpha\}$ is compact for any α.

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

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 ".

2 Convexity

2.1 Definition

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 $f(z) \geq f(x) + (z-x)^T \nabla f(x)$, $\forall x, z \in C$. The inequality is strict for strict convexity.

Proof.

(i) "$\Rightarrow$"

$$\begin{aligned} f(x+\alpha(y-x)) &\leq (1-\alpha)f(x) + \alpha f(y), \forall \alpha \in (0,1) \\ \Rightarrow \frac{f(x+\alpha(y-x)) - f(x)}{\alpha} &\leq f(y) - f(x) \\ \text{Limit as } \alpha \to 0 \Rightarrow (y-x)^T \nabla f(x) &\leq f(y) - f(x) \end{aligned}$$

(ii) "$\Leftarrow$" Let $g = \alpha x + (1-\alpha)y$

$$\begin{aligned} f(g) + (x-g)^T \nabla f(g) &\leq f(x) \\ f(g) + (y-g)^T \nabla f(g) &\leq f(y) \\ \Rightarrow f(g) &\leq \alpha f(x) + (1-\alpha)f(y) \\ f(\alpha x + (1-\alpha)y) &\leq \alpha f(x) + (1-\alpha)f(y) \end{aligned}$$

$\square$

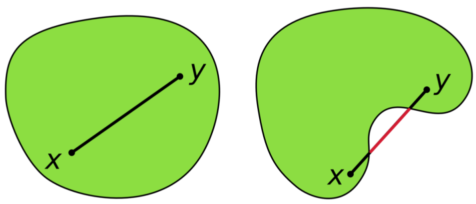
Property (2nd order): If f is twice differentiable, then f is convex iff

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

Strictly convex: $\nabla^2 f(x) \succ 0, \ \forall x \in C \Rightarrow f$ is strictly convex.
Note: f is strictly convex $\nRightarrow \nabla^2 f(x) \succ 0$.

Example 4. $f(x) = x^4$ *(strictly convex),* $\frac{d^2 f(x)}{dx^2} = 12x^2 (= 0 \ at \ x = 0)$

A function f is a **concave function** iff $-f$ is a convex function.
Convex set graph:



(a) Convex set (b) Non-convex set

Figure 1:

Claim 1. *Suppose f is a convex function over $\mathbb{R}^n$ and define the set*

$$C = \{x \in \mathbb{R}^n | f(x) \leq a\}, a \in \mathbb{R}$$

then C is a convex set.

Claim 2. *If $f_1, f_2, ..., f_k$ are convex functions over convex set &,*

1. $f_{sum}(x) = \sum_{i=1}^{k} f_i(x)$ *is convex over &*
2. $f_{max}(x) = \max_{i=1,...,k} f_i(x)$ *is convex over &*

Proof.
(2)

$$\begin{aligned} f_{max}(\alpha x + (1-\alpha)y) &= \max_{i=1,...,k} f_i(\alpha x + (1-\alpha)y) \\ &\leq \max_{i=1,...,k} [\alpha f_i(x) + (1-\alpha) f_i(y)] \\ &\leq \max_{i=1,...,k} \alpha f_i(x) + \max_{i=1,...,k} (1-\alpha) f_i(y) \\ &= \alpha f_{max}(x) + (1-\alpha) f_{max}(y) \end{aligned}$$

$\square$

2.2 Convex⇒Stationary point is global-min

Proposition 1. *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. If f is strictly convex,then min is unique.
(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.
Proof based on a property: If f is differentiable over C (open), then f is convex iff

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

□

Corollary 3. *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.

2.3 Unconstrained Quadratic Optimization

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

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

$$\nabla f(\mathbf{w}) = \mathbf{Q}\mathbf{w} - \mathbf{b},\ \nabla^2 f(\mathbf{w}) = \mathbf{Q}$$

(i) $\mathbf{Q} \succeq 0 \Leftrightarrow f$ is convex.

(ii) $\mathbf{Q} \succ 0 \Leftrightarrow f$ is strictly convex.

(iii) $\mathbf{Q} \preceq 0 \Leftrightarrow f$ is concave.

(iv) $\mathbf{Q} \prec 0 \Leftrightarrow f$ is strictly concave.

- Necessary condition for (local) optimality

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

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

Example 5. *Toy Problem 1:* $\min_{x,y\in\mathbb{R}} f(x,y) \triangleq x^2 + y^2 + \alpha xy$.

1. Step 1: First order condition: $2x^* + \alpha y^* = 0, 2y^* + \alpha x^* = 0$.
   - We get $4x^* = -2\alpha y^* = \alpha^2 x^*$. So $\left(4-\alpha^2\right) x^* = 0$.
   - 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.

2. 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 $|\alpha| \le 2$, then $\lambda_i \ge 0, \forall i$. Thus f is convex. Any stat-pt is global-min.

   - If $|\alpha| > 2$, at least one $\lambda_i < 0$, thus f is not convex.

3. Step 3 (can be skipped now): For non-convex case ($|\alpha| > 2$), prove no lower bound.
   $f(x, y) = (x + \alpha y/2) + \left(1 - \alpha^2/4\right) y^2$. Pick $y = M, x = -\alpha M/2$, then $f(x, y) = \left(1 - \alpha^2/4\right) M^2 \to -\infty$ 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.

Example 6. *Linear Regression*

minimize $f(\mathbf{w}) = \frac{1}{2} \left\|\mathbf{X}^T \mathbf{w} - \mathbf{y}\right\|^2$ subject to $\mathbf{w} \in \mathbb{R}^d$
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; Case 1 never happens!

Claim: Linear regression problem is always convex; it has global-min.
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 $n \ge d$), then there is a unique stationary point $x = \left(A^\top A\right)^{-1} A^\top b$. 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.

2.4 Strongly Convexity

2.4.1 μ-Strongly Convex: $\langle \nabla f(w) - \nabla f(v), w - v \rangle \ge \mu \|w - v\|^2$

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

$$\langle \nabla f(w) - \nabla f(v), w - v \rangle \ge \mu \|w - v\|^2, \quad \forall w, v \in C.$$

2.4.2 μ-strongly convex $\Leftrightarrow \nabla^2 f(x) \succeq \mu I$

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

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

Definition 3. *A twice continuously differentiable function is* <u>*strongly convex*</u> *if*

$$\exists m > 0 \ s.t. \ \nabla^2 f(x) \succeq mI \quad \forall x$$

which is also called $m-$strongly convex.

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

2.4.3 Lemma: Strongly convexity $\Rightarrow$ Strictly convexity

Lemma 1. *Strongly convexity $\Rightarrow$ Strictly convexity.*

Proof.

$$\begin{aligned}\nabla^2 f(x) \succeq mI \Rightarrow \nabla^2 f(x) - mI &\succeq 0 \\ \Rightarrow \forall z \neq 0 \quad z^T(\nabla^2 f(x) - mI)z &\geq 0 \\ \Rightarrow z^T \nabla^2 f(x) z \geq m z^T z &> 0\end{aligned}$$

□

Note: converse is not true: e.g. $f(x) = x^4$ is strictly convex but $\nabla^2 f(0) = 0$

2.4.4 Lemma: $\nabla^2 f(x) \succeq mI \Rightarrow f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2$

Lemma 2. $\nabla^2 f(x) \succeq mI \quad \forall x$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2$$

Proof. By Taylor's Theorem,

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f((1-\beta)x + \beta y)(y-x), \quad \text{for some } \beta \in [0,1] \\ &\geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T m(y-x) \\ &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \end{aligned}$$

□

3 Gradient Methods

Definition 4 (Iterative Descent)**.** *Start at some point x_0, and successively generate $x_1, x_2, ..$ s.t.*

$$f(x_{k+1}) < f(x_k) \quad k = 0, 1, ...$$

Definition 5 (**General Gradient Descent Algorithm**)**.** *Assume that $\nabla f(x_k) \neq 0$. Then*

$$x_{k+1} = x_k + \alpha_k d_k$$

where d_k is s.t. d_k has a positive projection along $-\nabla f(x_k)$,

$$\nabla f(x_k)^T d_k < 0 \equiv -\nabla f(x_k)^T d_k > 0$$

- If $d_k = -\nabla f(x_k)$ we get **steepest descent**.
- Often d_k is constructed using matrix $D_k \succ 0$

$$d_k = -D_k \nabla f(x_k)$$

3.1 Steepest Descent

We want the x_k that decreases the function most.

Proposition 2. $-\nabla f(x_k)$ *is the direction deceases the function most.*

Proof. Suppose the direction is $v \in \mathbb{R}^n, v \neq 0$.

$$f(x + \alpha v) = f(x) + \alpha v^T \nabla f(x) + O(\alpha)$$

The rate of change of f along direction v:

$$\lim_{\alpha \to 0} \frac{f(x + \alpha v) - f(x)}{\alpha} = v^T \nabla f(x)$$

By Cauchy-schwarz inequality,

$$|v^T \nabla f(x)| \leq \|v\| \|\nabla f(x)\|$$

Equation holds when $v = \beta \nabla f(x)$. Hence, $-\nabla f(x)$ is the direction decreases the function most. $\square$

Definition 6 (**Steepest Descent Algorithm**)**.**

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

α_k *is the step size, which need to choose carefully.*

3.2 Methods for Choosing α_k

Method (1): Fixed step size: $\alpha_k = \alpha$ (can have issue with *convergence*)

Method (2): **Optimal Line Search**: choose α_k to optimize the value of next iteration, i.e. solve

$$\min_{\alpha \geq 0} f(x_k + \alpha d_k)$$

(may be *difficult in practice*)

Method (3): **Armijo's Rule** (successive step size reduction):

$$f(x_k + \alpha_k d_k) = f(x_k) + \alpha_k \nabla f(x_k)^T d_k + O(\alpha_k)$$

Since $\nabla f(x_k)^T d_k < 0$, f decreases when α_k is sufficiently small. But we also don't want α_k to be too small (slow).

3.3 Optimal(Exact) Line Search

Example 7. *(False $\times$) The gradient descent algorithm with an exact line search always finds the minimum of a strictly convex quadratic function in exactly one iteration.*

Note: the moving direction is restricted to the gradient.
Counterexample: False. It is not necessary that the gradient at x_0 towards the exact solution. For example, let $f(x) = \frac{1}{2} x^\top Q x + x^\top b$ where $Q = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Clearly we have $x^* = \begin{pmatrix} -1/2 \\ 1 \end{pmatrix}$. If we start with $x_0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, by using exact line search, the step size $\alpha = \arg\min f(x_0 - \alpha \nabla f(x_0)) = 10/19$. Hence $x_1 = x_0 - \alpha \nabla f(x_0) = \begin{pmatrix} -11/19 \\ 28/19 \end{pmatrix} \neq x^*$.

3.4 Armijo's Rule

(i) Initialize $\alpha_k = \tilde{\alpha}$. Let $\sigma, \beta \in (0,1)$ be prespecified paramenters.

(ii) If $f(x_k) - f(x_k + \alpha_k d_k) \geq -\sigma\alpha_k \nabla f(x_k)^T d_k$, stop.

(Which shows $f(x_k + \alpha_k d_k)$ is at least smaller than $f(x_k)$ in a degree that correlated with $\nabla f(x_k)^T d_k$)

(iii) Else, set $\alpha_k = \beta\alpha_k$ and go back to step 2. (use a smaller α_k)

Termination at smallest integer m s.t.

$$f(x_k) - f(x_k + \beta^m \tilde{\alpha} d_k) \geq -\sigma\beta^m \tilde{\alpha} \nabla f(x)^T d_k$$

In Bersekas's book: $\sigma \in [10^{-5}, 10^{-1}], \beta \in [\frac{1}{10}, \frac{1}{2}]$.

As σ, β are smaller, the algorithm is quicker.

3.5 Armijo's Rule for Steepest Descent

$\alpha_k = \tilde{\alpha}\beta^{m_k}$, where m_k is smallest m s.t.

$$f(x_k) - f(x_k - \tilde{\alpha}\beta^m \nabla f(x_k)) \geq \sigma\tilde{\alpha}\beta^m \|\nabla f(x_k)\|^2$$

Proposition 3. *Assume* $\inf_x f(x) > -\infty$. *Then every limit point of* $\{x_k\}$ *for steepest descent with Armijo's rule is a stationary point of* f.

Proof. Assume that $\bar{x}$ is a limit point of $\{x_k\}$ s.t. $\nabla f(\bar{x}) \neq 0$.

- Since $\{f(x_k)\}$ is monotonically non-increasing and bounded below, $\{f(x_k)\}$ converges.
- f is continuous $\Rightarrow$ $f(\bar{x})$ is a limit point of $\{f(x_k)\}$ $\Rightarrow \lim_{k\to\infty} f(x_k) = f(\bar{x}) \Rightarrow f(x_k) - f(x_{k+1}) \to 0$
- By definition of Armijo's rule:

$$f(x_k) - f(x_{k+1}) \geq \sigma\alpha_k \|\nabla f(x_k)\|^2$$

Hence, $\sigma\alpha_k\|\nabla f(x_k)\|^2 \to 0$.

Since $\nabla f(\bar{x}) \neq 0$, $\lim_{k\to\infty} \alpha_k = 0$

$$ln\alpha_k = ln(\tilde{\alpha}\beta^{m_k}) = ln\tilde{\alpha} + m_k ln\beta \Rightarrow m_k = \frac{ln\alpha_k - ln\tilde{\alpha}}{ln\beta} \Rightarrow \lim_{k\to\infty} m_k = \infty$$

Exist $\bar{k}$ s.t. $m_k > 1, \forall k > \bar{k}$

$$f(x_k) - f(x_k - \frac{\alpha_k}{\beta}\nabla f(x_k)) < \sigma\frac{\alpha_k}{\beta}\|\nabla f(x_k)\|^2, \forall k > \bar{k}$$

By Taylor's Theorem,

$$f(x_k - \frac{\alpha_k}{\beta}\nabla f(x_k)) = f(x_k) - \nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \frac{\alpha_k}{\beta}\nabla f(x_k)$$

for some $\bar{\alpha}_k \in (0, \alpha_k)$

Hence,

$$\nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \frac{\alpha_k}{\beta}\nabla f(x_k) < \sigma\frac{\alpha_k}{\beta}\|\nabla f(x_k)\|^2$$
$$\nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \nabla f(x_k) < \sigma\|\nabla f(x_k)\|^2, \forall k > \bar{k}$$
$$\text{As } \alpha_k \to 0 \Rightarrow \bar{\alpha}_k \to 0$$
$$\|\nabla f(x_k)\|^2 < \sigma\|\nabla f(x_k)\|^2$$

Which contradicts to $\sigma < 1$.

□

4 Convergence of GD with Constant Stepsize

4.1 Lipschitz Gradient (L-Smooth)

Definition 7 (Lipschitz Continunity)**.** *A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is called Lipschitz (continuous) if $\exists L > 0$ s.t.*

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

L is Lipschitz constant. g is called L-smooth.

Definition 8 (Lipschitz Gradient)**.** *$\nabla f(x)$ is Lipschitz if $\exists L > 0$ s.t.*

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

Example 8.

1. *$f(x) = \|x\|^4$, $\nabla f(x) = 4\|x\|^2 x$*
   Test $\|\nabla f(x) - \nabla f(-x)\| \leq L\|2x\|$, $8\|x\|^2\|x\| \leq 2L\|x\|$ which doesn't hold when $\|x\|^2 > \frac{L}{4}$.
2. *If f is twice continuously differentiable with $\nabla^2 f(x) \succeq -MI$ and $\nabla^2 f(x) \preceq MI$ then $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|, \forall x, y \in \mathbb{R}^n$. ($A \succeq B$ means $A - B \succeq 0$, $A \preceq B$ means $A - B \preceq 0$)*

4.1.1 Theorem: $-MI \preceq \nabla^2 f(x) \preceq MI \Rightarrow \nabla f(x)$ is Lipschitz with constant M

Theorem 7. *$-MI \preceq \nabla^2 f(x) \preceq MI, \forall x \Rightarrow \|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|, \forall x, y$*

Proof. For symmetric A,

1. $x^T A x \leq \lambda_{\max}(A)\|x\|^2$
2. $\lambda_i(A^2) = \lambda_i^2(A)$
3. $-MI \preceq A \preceq MI \Rightarrow \lambda_{\min}(A) \geq -M, \lambda_{\max}(A) \leq M$

Define $g(t) = \frac{\partial f}{\partial x_i}(x + t(y-x))$. Then

$$\begin{aligned}
g(1) &= g(0) + \int_0^1 g'(s)ds \\
\Rightarrow \frac{\partial f(y)}{\partial x_i} &= \frac{\partial f(x)}{\partial x_i} + \int_0^1 \sum_{j=1}^n \frac{\partial^2 f(x + s(y-x))}{\partial x_i \partial x_j}(y_j - x_j)ds \\
\nabla f(y) &= \nabla f(x) + \int_0^1 \nabla^2 f(x + s(y-x))(y-x)ds \\
\|\nabla f(y) - \nabla f(x)\| &= \| \int_0^1 \nabla^2 f(x + s(y-x))(y-x)ds \| \\
&\leq \int_0^1 \|\nabla^2 f(x + s(y-x))(y-x)\| ds \\
&= \int_0^1 \sqrt{(y-x)^T [\nabla^2 f(x + s(y-x))]^2 (y-x)} ds \\
&\text{(Set } H = \nabla^2 f(x + s(y-x))\text{)} \\
&\leq \int_0^1 \sqrt{\lambda_{\max}(H^2)\|y-x\|^2} ds \\
&\leq M\|y-x\|
\end{aligned}$$

□

4.1.2 Descent Lemma: $\nabla f(x)$ is Lipschitz with constant $L \Rightarrow f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$

Lemma 3 (Descent Lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with a Lipschitz gradient with Lipschitz constant L. Then*

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}L\|y-x\|^2$$

Proof. Let $g(t) = f(x + t(y-x))$. Then $g(0) = f(x)$ and $g(1) = f(y)$, $g(1) = g(0) + \int_0^1 g'(t)dt$. Where $g'(t) = \nabla f(x + t(y-x))^T(y-x)$

$$\begin{aligned}
\Rightarrow f(y) &= f(x) + \int_0^1 \nabla f(x + t(y-x))^T (y-x)dt \\
&= f(x) + \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))^T (y-x)dt + \nabla f(x)^T(y-x) \\
&\leq f(x) + \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\|\|y-x\|dt + \nabla f(x)^T(y-x) \\
&\leq f(x) + L\int_0^1 \|t(y-x)\|\|y-x\|dt + \nabla f(x)^T(y-x) \\
&= f(x) + \frac{1}{2}L\|y-x\|^2 + \nabla f(x)^T(y-x)
\end{aligned}$$

□

4.1.3 Co-coercivity Condition: $(\nabla f(x)-\nabla f(y))^T(x-y)\geq \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2$

Theorem 8 (Co-coercivity Condition)**.** *Let f be convex and continuously differentiable. Let f be L-smooth. Then*

$$(\nabla f(x)-\nabla f(y))^T(x-y)\geq \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2$$

Proof. Let $y\in\mathbb{R}^n$, and define $g(x)=f(x)-\nabla f(y)^Tx$. Then $\nabla g(y)=\nabla f(y)-\nabla f(y)=0$ and $\nabla^2 g(y)=\nabla^2 f(y)\succeq 0$, i.e. y minimize g. Because $g(y)\leq g(\cdot)$, $g(y)\leq g(x-\frac{1}{L}\nabla g(x))$ According to the descent lemma,

$$\begin{aligned}
g(x-\frac{1}{L}\nabla g(x)) &= f(x-\frac{1}{L}\nabla g(x))-\nabla f(y)^T(x-\frac{1}{L}\nabla g(x))\\
&\leq f(x)+\frac{L}{2}\|-\frac{1}{L}\nabla g(x)\|^2+\nabla f(x)^T(-\frac{1}{L}\nabla g(x))-\nabla f(y)^T(x-\frac{1}{L}\nabla g(x))\\
&\leq f(x)+\frac{1}{2L}\|\nabla g(x)\|^2-(\nabla f(x)-\nabla f(y))^T\frac{1}{L}\nabla g(x)-\nabla f(y)^Tx\\
&= f(x)-\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2-\nabla f(y)^Tx\\
&= g(x)-\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
g(y)\leq g(x-\frac{1}{L}\nabla g(x)) &= g(x)-\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2\\
\Rightarrow g(y)-g(x)=f(y)-\nabla f(y)^Ty-f(x)-\nabla f(y)^Tx &\leq -\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2
\end{aligned}$$

We can interchange x, y,

$$\begin{cases} f(y)-\nabla f(y)^Ty-f(x)-\nabla f(y)^Tx\leq -\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2\\ f(x)-\nabla f(x)^Tx-f(y)-\nabla f(x)^Ty\leq -\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2\end{cases}$$

Add these two inequalitys together,

$$(\nabla f(x)-\nabla f(y))^T(x-y)\geq \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2$$

□

4.2 Convergence of Steepest Descent with Fixed Stepsize

4.2.1 Theorem: f has Lipschitz gradient $\Rightarrow \{x_k\}$ converges to stationary point

Theorem 9. *Consider the GD algorithm*

$$x_{k+1}=x_k-\alpha\nabla f(x_k),\quad k=0,1,...$$

Assume that f has Lipschitz gradient with a Lipschitz gradient with Lipschitz constant L. Then if α is sufficiently small ($\alpha\in(0,\frac{2}{L})$) and $f(x)\geq f_{\min}$ for all $x\in\mathbb{R}^n$,
(1). $f(x_{k+1})\leq f(x_k)-\alpha(1-\frac{L\alpha}{2})\|\nabla f(x_k)\|^2$
(2). $\sum_{k=0}^N\|\nabla f(x_k)\|^2\leq \frac{f(x_0)-f_{\min}}{\alpha(1-\frac{L\alpha}{2})}$
(3). every limit point of $\{x_k\}$ is a stationary point of f.

Proof. Applying the descent lemma,

$$\begin{aligned}
f(x_{k+1}) &\le f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k) + \frac{L}{2}\alpha^2 \|\nabla f(x_k)\|^2 \\
&= f(x_k) + \alpha(\frac{L\alpha}{2} - 1)\|\nabla f(x_k)\|^2
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 &\le f(x_k) - f(x_{k+1}) \\
\alpha \sum_{k=0}^{N} (1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 &\le f(x_0) - f(x_{N+1}) \\
&\le f(x_0) - f_{\min}
\end{aligned}$$

If $\alpha \in (0, \frac{2}{L})$, i.e. $\alpha(1 - \frac{L\alpha}{2})$,

$$\begin{aligned}
&\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \le \frac{f(x_0) - f_{\min}}{\alpha(1 - \frac{L\alpha}{2})} < \infty, \forall N \\
&\Rightarrow \lim_{k\to\infty} \nabla f(x_k) = 0
\end{aligned}$$

If $\bar{x}$ is a limit point of $\{x_k\}$, $\lim_{k\to\infty} x_k = \bar{x}$.
By continunity of ∇f, $\nabla f(\bar{x}) = 0$ □

Example 9. *$f(x) = \frac{1}{2}x^2, x \in \mathbb{R}$, $\nabla f(x) = x$, Lipschitz with $L = 1$.*

$$\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) \\
&= x_k(1 - \alpha)
\end{aligned}$$

$0 < \alpha < \frac{2}{L} = 2$ is needed for convergence.

Test (1) $\alpha = 1.5$ Then $x_{k+1} = x_k(-0.5)$,

$$\Rightarrow x_k = x_0(-0.5)^k \to 0 \text{ as } k \to \infty$$

Test (2) $\alpha = 2.5$ Then $x_{k+1} = x_k(-1.5)$.

$$\Rightarrow x_k = x_0(-1.5)^k \Rightarrow |x_k| \to \infty$$

Test (3) $\alpha = 2$ Then $x_{k+1} = -x_k$.

$$\Rightarrow x_k = (-1)^k x_0 \Rightarrow \text{ oscillation between } -x_0, x_0$$

Example 10. *What if gradient is not Lipschitz? e.g. $f(x) = x^4, x \in \mathbb{R}$, $\nabla f(x) = 4x^3$, $x = 0$ is the only stationary point (global-min)*

$$x_{k+1} = x_k - 4\alpha x_k^3 = x_k(1 - 4\alpha x_k^2)$$

- $|x_1| = |x_0|$, then $|x_k| = |x_0|$ for all k, and $\{x_k\}$ stays bounded away from 0, except if $x_0 = 0$

- $$\begin{aligned}|x_1| < |x_0| &\Leftrightarrow |x_0||1-4\alpha x_0^2| < |x_0| \\ &\Leftrightarrow -1 < 1-4\alpha x_0^2 < 1 \\ &\Leftrightarrow 0 < x_0^2 < \frac{1}{2\alpha} \Leftrightarrow 0 < |x_0| < \frac{1}{\sqrt{2\alpha}}\end{aligned}$$

- Therefore, if $|x_1| < |x_0|$, then $|x_1| < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_2| < |x_1|, ..., |x_{k+1}| < |x_k|, \forall k \Rightarrow \{|x_k|\}$convergences

- And if $|x_1| > |x_0|$, then $|x_{k+1}| > |x_k|$ for all k and $\{x_k\}$ stays bounded away from 0.

Claim 3. $0 < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_k| \to 0$

Proof. Suppose $|x_k| \to c > 0$. Then $\frac{|x_{k+1}|}{|x_k|} \to 1$
But $\frac{|x_{k+1}|}{|x_k|} = |1-4\alpha x_k^2| \to |1-4\alpha c^2|$. Thus $|1-4\alpha c^2| = 1 \Rightarrow c = \frac{1}{\sqrt{2\alpha}}$, which contradicts to $c < |x_0| < \frac{1}{\sqrt{2\alpha}}$, hence $c = 0$

□

4.3 Convergence of GD for convex functions

4.3.1 Theorem: f is convex and has Lipschitz gradient $\Rightarrow f(x_k)$ converges to global-min value with rate $\frac{1}{k}$

Theorem 10. *Consider the GD algorithm*

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \quad k = 0, 1, ...$$

Assume that f has Lipschitz gradient with Lipschitz constant L. Further assume that

(a) f is a convex function.

(b) $\exists x^*$ s.t. $f(x^*) = \min f(x)$

Then for sufficiently small α:

(i) $\lim_{k\to\infty} f(x_k) = \min f(x) = f(x^*)$

(ii) $f(x_k)$ converges to $f(x^*)$ at rate $\frac{1}{k}$.

Proof.

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - \alpha\nabla f(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 - 2\alpha\nabla f(x)^T(x_k - x^*)\end{aligned}$$

By convexity,

$$\begin{aligned} f(x^*) &\geq f(x_k) + \nabla f(x_k)^T(x^* - x_k) \\ \Rightarrow \nabla f(x_k)^T(x^* - x_k) &\leq f(x^*) - f(x_k)\end{aligned}$$

Thus,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 + 2\alpha(f(x^*) - f(x_k)) \\ \Rightarrow 2\alpha(f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 \\ 2\alpha\sum_{k=0}^{N}(f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_{N+1} - x^*\|^2 + \alpha^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \\ &\leq \|x_0 - x^*\|^2 + \alpha^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2\end{aligned}$$

According to previous theorm, if $\alpha \in (0, \frac{2}{L})$, $\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \le \frac{f(x_0)-f(x^*)}{\alpha(1-\frac{L\alpha}{2})}$ and

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &\le -\alpha(1-\frac{L\alpha}{2})\|\nabla f(x_k)\|^2 \le 0 \\
\Rightarrow f(x_N) &\le f(x_k), \quad \forall k = 0, 1..., N \\
\Rightarrow \sum_{k=0}^{N}(f(x_k) - f(x^*)) &\ge (N+1)(f(x_N) - f(x^*)) \\
f(x_N) - f(x^*) &\le \frac{1}{N+1}\sum_{k=0}^{N}(f(x_k) - f(x^*)) \\
&\le \frac{1}{2\alpha(N+1)}(\|x_0 - x^*\|^2 + \alpha^2 \frac{f(x_0)-f(x^*)}{\alpha(1-\frac{L\alpha}{2})}) \\
&\to 0 \text{ as } N \to \infty
\end{aligned}
$$

The rate of convergence is $\frac{1}{N}$.
To make $f(x_N) - f(x^*) < \varepsilon$, we need $N \sim O(\frac{1}{\varepsilon})$. □

Note: Armijo's rule also convergencesat rate $\frac{1}{N}$ if ∇f is Lipschitz, without priot knowledge of L. But need $r \in [\frac{1}{2}, 1)$

4.4 Convergence of GD for strongly convex functions

Strong convexity with parameter m, along with $M-$Lipschitz gradient assumption (with $M \ge m$) According to the lemmas we proved before

$$\frac{m}{2}\|y-x\|^2 \le f(y) - f(x) - \nabla^T f(x)(y-x) \le \frac{M}{2}\|y-x\|^2$$

4.4.1 Theorem: Strongly convex, Lipschitz gradient $\Rightarrow \{x_k\}$ converges to global-min geometrically

Theorem 11. *If f has Lipschitz gradient with Lipschitz constant M and strongly convex with parameter m, $\{x_k\}$ converges to x^* **geometrically**.*

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \alpha\nabla f(x_k) - x^*\|^2 \\
(\nabla f(x^*) = 0) \qquad &= \|(x_k - x^*) - \alpha(\nabla f(x_k) - \nabla f(x^*))\|^2 \\
&= \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k) - \nabla f(x^*)\|^2 - 2\alpha(x_k - x^*)^T(\nabla f(x_k) - 0) \\
(\nabla f \text{ is M-Lipschitz}) \qquad &\le \|x_k - x^*\|^2 + \alpha^2 M^2\|x_k - x^*\|^2 + 2\alpha(x^* - x_k)^T\nabla f(x_k) \\
(\text{Strong convexity with } m) \qquad &\le \|x_k - x^*\|^2 + \alpha^2 M^2\|x_k - x^*\|^2 + 2\alpha(f(x^*) - f(x_k) - \frac{m}{2}\|x^* - x_k\|^2) \\
&= (1 + \alpha^2 M^2 - \alpha m)\|x_k - x^*\|^2 + 2\alpha(f(x^*) - f(x_k))
\end{aligned}
$$

By strong convexity of f

$$
\begin{aligned}
f(x_k) &\ge f(x^*) + \nabla^T f(x^*)(x_k - x^*) + \frac{m}{2}\|x_k - x^*\|^2 \\
&= f(x^*) + \frac{m}{2}\|x_k - x^*\|^2 \\
\Rightarrow f(x^*) - f(x_k) &\le -\frac{m}{2}\|x_k - x^*\|^2
\end{aligned}
$$

Then,

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq (1 + \alpha^2 M^2 - \alpha m)\|x_k - x^*\|^2 + 2\alpha(-\frac{m}{2}\|x_k - x^*\|^2) \\
&\leq (1 + \alpha^2 M^2 - 2\alpha m)\|x_k - x^*\|^2 \\
&\leq (1 + \alpha^2 M^2 - 2\alpha m)^{k+1}\|x_0 - x^*\|^2 \\
\Rightarrow \|x_N - x^*\|^2 &\leq (1 + \alpha^2 M^2 - 2\alpha m)^N\|x_0 - x^*\|^2
\end{aligned}$$

If $\alpha \in (0, \frac{2m}{M^2})$, $1 + \alpha^2 M^2 - 2\alpha m < 1$. Then $x_N \to x^*$ **geometrically** as $N \to \infty$.
Note: Just having $0 < \alpha < \frac{2}{M}$ doesn't guarantee geometric convergence to x^*. e.g. $\alpha = \frac{1}{M} \Rightarrow 1 + \alpha^2 M^2 - 2m\alpha = 2(1 - \frac{m}{M}) \geq 1$ if $\frac{m}{M} \leq 0.5$
To get the highest convergence rate:

$$\begin{aligned}
1 + \alpha^2 M^2 - 2m\alpha &= (\alpha M)^2 - 2\alpha M\frac{m}{M} + 1 \\
&= (\alpha M - \frac{m}{M})^2 + 1 - \frac{m^2}{M^2}
\end{aligned}$$

Which is minimized by setting

$$\alpha = \alpha^* = \frac{m}{M^2}$$

$$\min_{\alpha > 0} 1 + \alpha^2 M^2 - 2m\alpha = 1 - \frac{m^2}{M^2} \in [0, 1)$$

Since $M > m$, $\alpha^* = \frac{m}{M^2} < \frac{1}{M} < \frac{2}{M}$.
With $\alpha = \alpha^*$,

$$\|x_N - x^*\|^2 \leq (1 - \frac{m^2}{M^2})^N\|x_0 - x^*\|^2$$

$\frac{M}{m}$ is called the **condition number**.

- If $\frac{M}{m} >> 1$, then $1 - \frac{m^2}{M^2}$ is close to 1 and convergence is slow.
- If $\frac{M}{m} = 1$, $\alpha^* = \frac{1}{M}$, and $x_N = x^*, \forall N \geq 1$. (Convergence in one step.)

Note that since $\nabla f(x^*) = 0$,

$$\begin{aligned}
f(x_N) - f(x^*) &\leq \frac{M}{2}\|x_N - x^*\|^2 \\
&\leq (1 - \frac{m^2}{M^2})^N\frac{M}{2}\|x_0 - x^*\|^2
\end{aligned}$$

To make $f(x_N) - f(x^*) < \varepsilon$, we only need $N \sim O(log\frac{1}{\varepsilon})$ - called "linear" convergence.

4.4.2 Example

Example 11. $f(x) = \frac{1}{2}x^T Qx + b^T x + c, \quad Q \succ 0,\ \nabla^2 f(x) = Q.$

Let $\lambda_{\min}$ and $\lambda_{\max}$ be the min and max eigenvalue of Q. Then we know

$$\lambda_{\min}\|z\|^2 \leq z^T Qz \leq \lambda_{\max}\|z\|^2$$

Thus for all $z \in \mathbb{R}^n$

$$z^T(Q - \lambda_{\min}I)z \geq 0 \Rightarrow Q \succeq \lambda_{\min}I$$

Similarly, $Q \preceq \lambda_{\max}I$. Thus

$$\lambda_{\min}I \preceq \nabla^2 f(x) \preceq \lambda_{\max}I$$

$\lambda_{\min} I \preceq \nabla^2 f(x) \Leftrightarrow f$ is $\lambda_{\min}$-strongly convex; $\nabla^2 f(x) \preceq \lambda_{\max} I$ is a sufficient condition for f is $\lambda_{\max}$-smooth.
The condition number $= \frac{\lambda_{\max}}{\lambda_{\min}}$
Special Case: $Q = \mu I, \quad \mu > 0$, $\lambda_{\min} = \lambda_{\max} = \mu = m = M$.
$f(x) = \frac{\mu}{2}\|x\|^2 + b^T x + c$, $\nabla f(x) = \mu x + b$, $x^* = -\frac{b}{\mu}$, $\alpha^* = \frac{m}{M^2} = \frac{1}{\mu}$,

$$x_1 = x_0 - \alpha^* \nabla f(x_0) = x_0 - \frac{1}{\mu}(\mu x_0 + b) = -\frac{b}{\mu} = x^*$$

Convergence in one step!

4.5 Convergence of Gradient Descent on Smooth Strongly-Convex Functions

Still consider the constant stepsize gradient method

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

Lemma 4. *Suppose the sequences $\{\xi_k \in \mathbb{R}^p : k = 0, 1, \ldots\}$ and $\{u_k \in \mathbb{R}^p : k = 0, 1, 2, \ldots\}$ satisfy $\xi_{k+1} = \xi_k - \alpha u_k$. In addition, assume there is a martix M, the following inequality holds for all k*

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0$$

If there exist $0 < \rho < 1$ and $\lambda \geq 0$ such that

$$\begin{bmatrix} \left(1-\rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M$$

is a negative semidefinite matrix, then the sequence $\{\xi_k : k = 0, 1, \ldots\}$ satisfies $\|\xi_k\| \leq \rho^k \|\xi_0\|$.

Proof. The key relation is

$$\|\xi_{k+1}\|^2 = \|\xi_k - \alpha u_k\|^2 = \|\xi_k\|^2 - 2\alpha(\xi_k)^T u_k + \alpha^2\|u_k\|^2 = \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}$$

Since $\begin{bmatrix} \left(1-\rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M$ is negative semidefinite, we have

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \left(\begin{bmatrix} \left(1-\rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M \right) \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Expand the inequality,

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} -\rho^2 I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Applying the key relation

$$\|\xi_{k+1}\|^2 - \rho^2\|\xi_k\|^2 + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

$$\|\xi_{k+1}\|^2 - \rho^2\|\xi_k\|^2 \leq -\lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Hence, $\|\xi_{k+1}\| \leq \rho\|\xi_k\|$ for all k. Therefore, we have $\|\xi_k\| \leq \rho^k \|\xi_0\|$. $\square$

Theorem 12. *Suppose f is L-smooth and m-strongly convex. Let x^* be the unique global min. Given a stepsize α,if there exists $0 < \rho < 1$ and $\lambda \geq 0$ such that*

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix}$$

is a negative semidefinite matrix, then the gradient method satisfies

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

Proof. We set f is L-smooth and m-strongly convex,
According to the definition of m-strongly convex

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq m\|x-y\|^2$$

And the co-coercivity condition, if f is $L-$smooth,

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

Set $g(x) = f(x) - \frac{m}{2}\|x\|^2$, $\nabla g(x) = \nabla f(x) - mx$.

$$\begin{aligned} f \text{ is } L\text{-smooth} &\Leftrightarrow \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \\ &\Leftrightarrow \|\nabla g(x) - \nabla g(y)\| \leq (L-m)\|x-y\| \\ &\Leftrightarrow g \text{ is } L-m\text{-smooth} \end{aligned}$$

Hence,

$$\begin{aligned} (\nabla g(x) - \nabla g(y))^T(x-y) &\geq \frac{1}{L-m}\|\nabla g(x) - \nabla g(y)\|^2 \\ (\nabla f(x) - \nabla f(y) - m(x-y))^T(x-y) &\geq \frac{1}{L-m}\|\nabla f(x) - \nabla f(y) - m(x-y)\|^2 \\ (L-m)[(\nabla f(x) - \nabla f(y))^T(x-y) - m\|x-y\|^2] & \\ \geq \|\nabla f(x) - \nabla f(y)\|^2 + m^2\|(x-y)\|^2 &- 2m(\nabla f(x) - \nabla f(y))^T(x-y) \\ (L+m)(\nabla f(x) - \nabla f(y))^T(x-y) &\geq mL\|x-y\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 \\ \Rightarrow (\nabla f(x) - \nabla f(y))^T(x-y) &\geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2 \end{aligned}$$

Which can be rewritten as

$$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0$$

Let $y = x^*$ and $\nabla f(y) = \nabla f(x^*) = 0$

$$\begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix} \geq 0$$

Set $\xi_k = x_k - x^*$ and $u_k = \nabla f(x_k)$. And $\xi_{k+1} = x_{k+1} - x^* = x_k - \alpha\nabla f(x_k) - x^* = \xi_k - \alpha u_k$

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0$$

Choose $M = \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix}$. Then prove by previous lemma. □

Now we apply the theorem to obtain the convergence rate ρ for the gradient method with various stepsize choices.

- Case 1: If we choose $\alpha = \frac{1}{L}, \rho = 1 - \frac{m}{L}$, and $\lambda = \frac{1}{L^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} -\frac{m^2}{L^2} & \frac{m}{L^2} \\ \frac{m}{L^2} & -\frac{1}{L^2} \end{bmatrix} = \frac{1}{L^2} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix}$$

  The right side is clearly negative semidefinite due to the fact that $\begin{bmatrix} a \\ b \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = -(ma-b)^2 \leq 0$. Therefore, the gradient method with $\alpha = \frac{1}{L}$ converges as

$$\|x_k - x^*\| \leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|$$

- Case 2: If we choose $\alpha = \frac{2}{m+L}, \rho = \frac{L-m}{L+m}$, and $\lambda = \frac{2}{(m+L)^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

  The zero matrix is clearly negative semidefinite. Therefore, the gradient method with $\alpha = \frac{2}{m+L}$ converges as

$$\|x_k - x^*\| \leq \left(\frac{L-m}{L+m}\right)^k \|x_0 - x^*\|$$

Notice $L \geq m > 0$ and hence $1 - \frac{m}{L} \geq \frac{L-m}{L+m}$. This means the gradient method with $\alpha = \frac{2}{m+L}$ converges slightly faster than the case with $\alpha = \frac{1}{L}$. However, m is typically unknown in practice. The step choice of $\alpha = \frac{1}{L}$ is also more robust. The most popular choice for α is still $\frac{1}{L}$.
We can further express ρ as a function of α. To do this, we need to choose λ carefully for a given α. If we choose λ reasonably, we can show the best value for ρ that we can find is $\max\{|1-m\alpha|, |L\alpha - 1|\}$.

4.6 From convergence rate to iteration complexity

The convergence rate ρ naturally leads to an iteration number T guaranteeing the algorithm to achieve the so-called ε-**optimality**, i.e. $\|x_T - x^*\| \leq \varepsilon$.
To guarantee $\|x_T - x^*\| \leq \varepsilon$, we can use the bound $\|x_T - x^*\| \leq \rho^T \|x_0 - x^*\|$. If we choose T such that $\rho^T \|x_0 - x^*\| \leq \varepsilon$, then we guarantee $\|x_T - x^*\| \leq \varepsilon$. Denote $c = \|x_0 - x^*\|$. Then $c\rho^k \leq \varepsilon$ is equivalent to

$$\log c + k \log \rho \leq \log(\varepsilon)$$

Notice $\rho < 1$ and $\log \rho < 0$. The above inequality is equivalent to

$$k \geq \log\left(\frac{\varepsilon}{c}\right) / \log \rho = \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$$

So if we choose $T = \log\left(\frac{c}{\varepsilon}\right)/(-\log \rho)$, we guarantee $\|x_T - x^*\| \leq \varepsilon$. Notice $\log \rho \leq \rho - 1 < 0$ (this can be proved using the concavity of log function and we will talk about concavity in later lectures), so $\frac{1}{1-\rho} \geq -\frac{1}{\log \rho}$ and we can also choose $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) \geq \log\left(\frac{c}{\varepsilon}\right)/(-\log \rho)$ to guarantee $\|x_T - x^*\| \leq \varepsilon$.
Another interpretation for $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho)$ is that a first-order Taylor expansion of $-\log \rho$ at $\rho = 1$ leads to $-\log \rho \approx 1 - \rho$. So $\log\left(\frac{c}{\varepsilon}\right)/(-\log \rho)$ is roughly equal to $\log\left(\frac{c}{\varepsilon}\right)/(1-\rho)$ when ρ is close to 1 .
Clearly the smaller T is, the more efficient the optimization method is. The iteration number T describes the "ε-optimal iteration complexity" of the gradient method for smooth strongly-convex objective functions.

- For the gradient method with $\alpha = \frac{1}{L}$, we have $\rho = 1 - \frac{m}{L} = 1 - \frac{1}{\kappa}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) = \kappa \log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ .2 Here we use the big O notation to highlight the dependence on κ and ε and hide the dependence on the constant c.

- For the gradient method with $\alpha = \frac{2}{L+m}$, we have $\rho = \frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) = \frac{\kappa+1}{2}\log\left(\frac{c}{\varepsilon}\right)$. Although $\frac{\kappa+1}{2} \leq \kappa$, we still have $\frac{\kappa+1}{2}\log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$. Therefore, the stepsize $\alpha = \frac{2}{m+L}$ can only improve the constant C hidden in the big O notation of the iteration complexity. People call this "improvement of a constant factor".

- In general, when ρ has the form $\rho = 1 - 1/(a\kappa + b)$, the resultant iteration complexity is always $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$.

There are algorithms which can significantly decrease the iteration complexity for unconstrained optimization problems with smooth strongly-convex objective functions. For example, Nesterov's method can decrease the iteration complexity from $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ to $O\left(\sqrt{\kappa} \log\left(\frac{1}{\varepsilon}\right)\right)$. Momentum is used to accelerate optimization as:

$$x_{k+1} = x_k - \alpha \nabla f\left((1+\beta)x_k - \beta x_{k-1}\right) + \beta\left(x_k - x_{k-1}\right).$$

5 Newton's Method

One dimential:
Finding solution to non-linear equation:

$$g(x^*) = 0$$

with $g : \mathbb{R} \to \mathbb{R}$. Given x_k, find x_{k+1} to solve x^*.

$$0 = g(x_{k+1}) \approx g(x_k) + g'(x_k)(x_{k+1} - x_k)$$

Assuming $g'(x_k) \neq 0$, set

$$x_{k+1} = x_k - (g'(x_k))^{-1} g(x_k)$$

5.1 Generalization to Optimization

In optimization, the goal is to get to x s.t. $\nabla f(x) = 0$.
Given x_k, we want to find x_{k+1} s.t. $\nabla f(x_{k+1}) = 0$.
Taylor's Approx:

$$\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)$$

Set

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

, which can be viewed as GD with $\alpha_k = 1$ and $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$
If $\nabla^2 f(x_k) \succeq 0$, then $\nabla f(x_k)^T d_k \geq 0$.

5.2 A New Interpretation of Newton's Method

Since $f(x) \approx f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$, at each step k, we can solve a quadratic minimization problem,

$$x_{k+1} = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \{ f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k) \}$$

5.3 Convergence of Newton's Method

Let x^* be s.t. $\nabla f(x^*) = 0$, then

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla f(x_k)\| \\ &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}(\nabla f(x_k) - \nabla f(x^*))\|\end{aligned}$$

By Taylor's theorem,

$$\nabla f(x_k) = \nabla f(x^*) + \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*) \text{ for some } \beta \in [0, 1]$$

Thus,

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*)\| \\ &= \|(\nabla^2 f(x_k))^{-1}(\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k))(x_k - x^*)\| \\ &\leq \|(\nabla^2 f(x_k))^{-1}\|\|\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k)\|\|x_k - x^*\|\end{aligned}$$

We use 1-norm $\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ here, $\|A\| \geq \frac{\|Ax\|}{\|x\|} \Rightarrow \|Ax\| \leq \|A\|\|x\|$.
Easy to prove, for symmetric $A \succeq 0$, $\|A\| = \lambda_{\max}(A)$, $\|A^{-1}\| = \lambda_{\max}(A^{-1}) = \lambda_{\min}^{-1}(A)$

- Now suppose <u>f **is loacl m-strongly convex near** x^*</u>, then

$$\begin{aligned}\nabla^2 f(x^*) &\succeq mI \text{ with } m > 0 \\ \Rightarrow \lambda_{\min}(\nabla^2 f(x^*)) &\geq m > 0 \\ \Rightarrow \lambda_{\min}^{-1}(\nabla^2 f(x^*)) &\leq \frac{1}{m}\end{aligned}$$

- When f is not local strongly convex near x^*. Assuming $\nabla^2 f(x)$ is continuous, if $\|x_k - x^*\|$ is small, then $\lambda_{\min}(\nabla^2 f(x_k))$ is close to $\lambda_{\min}(\nabla^2 f(x^*))$ i.e $\lambda_{\min}(\nabla^2 f(x^*))$ should be greater than a constant $\lambda_{\min}(\nabla^2 f(x^*)) \geq \bar{\gamma} > 0$. Then,

$$\|\nabla^2 f(x_k)^{-1}\| = \lambda_{\min}^{-1}(\nabla^2 f(x_k)) \leq \frac{1}{\bar{\gamma}} = \gamma$$

Furthurmore, <u>assume that $\nabla^2 f$ **is L-Lipschitz in a neighborhood & of** x^*</u>, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \forall x, y \in \&$$

Thus,

$$\begin{aligned}\|x_{k+1} - x^*\| &\leq \|(\nabla^2 f(x_k))^{-1}\|\|\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k)\|\|x_k - x^*\| \\ &\leq \gamma L\|x^* + \beta(x_k - x^*) - x_k\|\|x_k - x^*\| \\ &\leq \gamma L\|(\beta - 1)(x_k - x^*)\|\|x_k - x^*\| \\ \text{(Since } \beta \in [0,1]\text{)} \quad &\leq \gamma L\|x_k - x^*\|^2\end{aligned}$$

Hence,

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

Now suppose <u>x_0 **is close enough to** x^*</u> s.t.

$$\gamma L\|x_0 - x^*\| = \sigma < 1$$

Then,

$$\begin{aligned}
\|x_1 - x^*\| &\le \sigma \|x_0 - x^*\| \\
\|x_2 - x^*\| &\le \gamma L \|x_1 - x^*\|^2 \\
&\le \gamma L \sigma^2 \|x_0 - x^*\|^2 = \sigma^3 \|x_0 - x^*\| \\
\|x_3 - x^*\| &\le \gamma L \|x_2 - x^*\|^2 \\
&\le \gamma L \sigma^6 \|x_0 - x^*\|^2 = \sigma^7 \|x_0 - x^*\| \\
&\dots \\
\|x_N - x^*\| &\le \sigma^{2^N - 1} \|x_0 - x^*\|
\end{aligned}$$

Assuming **∇f is M-Lipschitz in neighborhood of x^***,

$$\begin{aligned}
f(x_N) - f(x^*) &\le \nabla f(x^*)(x_N - x^*) + \frac{M}{2}\|x_N - x^*\|^2 \\
&\le \frac{M}{2}\sigma^{(2^{N+1}-2)}\|x_N - x^*\|^2
\end{aligned}$$

Thus to make $f(x_N) - f(x^*) < \varepsilon$, need $N \sim O(log(log(\frac{1}{\varepsilon})))$
We call it **order-2 or super-linear convergence**.

5.4 Note: Cons and Pros

- Newton's Method is super-fast close to local min if function strongly convex around min.
- If the function is quadratic, Newton's method converges in one step.

$$\begin{aligned}
f(x) = \frac{1}{2}x^T Q x + bx + c, \quad Q \succ 0. \\
\nabla f(x) = Qx + b, \nabla^2 f(x) = Q. \\
\text{Global min } x^* \text{ satisfies } Qx^* + b = 0 \Rightarrow x^* = -Q^{-1}b
\end{aligned}$$

  Newton's method: for any $x_0 \in \mathbb{R}^n$,

$$\begin{aligned}
x_1 &= x_0 - (\nabla^2 f(x_0))^{-1}\nabla f(x_0) \\
&= x_0 - Q^{-1}(Qx_0 + b) = -Q^{-1}b = x^*
\end{aligned}$$

  Intuition: when f is a quadratic function, $\nabla^3 f(x) = 0, \forall x$. Hence, $f(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$, the minimization problem will get the min in one step.

- But Newton's method has several **drawbacks**:
  (1) Newton's method requires the matrix inversion step, and this is quite expensive. So the per step cost for Newton's method is higher.
  (2) Newton's method has faster local convergence but may diverge if initialized from some place far from the optimal point.
  (3) $\nabla^2 f(x)^{-1}$ may fail to exist, i.e. $\nabla^2 f(x)$ is singular, e.g. linear f.
  (4) It is not necessarily a general GD method since $\nabla^2 f(x_k)$ may not be $\succ 0$.
  (5) It is not a descent method, $f(x_{k+1})$ may be $> f(x_k)$.
  (6) It may stop at local max or saddlepoints.

5.5 Modifications to ensure global convergence

(a) Try Newton's method. If either $\nabla^2 f(x_k)$ is singular or $f(x_{k+1}) > f(x_k)$ then use (b).

(b) Find δ_k s.t.

$$(\delta_k I + \nabla^2 f(x_k)) \succ 0$$

and

$$\lambda_{\min}(\delta_k I + \nabla^2 f(x_k)) \succeq \Delta > 0$$

so that $\delta_k I + \nabla^2 f(x_k)$ is easily invertible.

Then set $d_k = -(\delta_k I + \nabla^2 f(x_k))^{-1} \nabla f(x_k)$. This ensures that $\nabla^T f(x_k) d_k < 0$.

Then we use $x_{k+1} = x_k + \alpha_k d_k$ with α_k chosen using Armijo's Rule.

If at any point $\nabla^2 f(x_k) \succ 0$, go back to Newton's method and check if $f(x_{k+1}) < f(x_k)$. Continue Newton's method as long as $\nabla^2 f(x_k) \succ 0$ and $f(x_{k+1}) < f(x_k)$.

5.6 Quasi-Newton Methods

Estimating Hessian $\nabla^2 f(x_k)$ is expensive, so we use some simplier matrix H_k instead.
Quasi-Newton method have the iteration form:

$$x_{k+1} = x_k - \alpha_k H_k^{-1} \nabla f(x_k)$$

where H_k is some estimated version of $\nabla^2 f(x_k)$, and the stepsize α_k is typically determined by Armijo rule.
Previously, we approximate $f(x)$ by

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

Now, we define the form by H_k

$$g(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T H_k (x - x_k)$$

We hope $g(x) \approx f(x)$ and optimize g for this step. We enforce

(1) $\nabla f(x_k) = \nabla g(x_k)$ (Automatically satisfied)

(2) $\nabla f(x_{k-1}) = \nabla g(x_{k-1}) \quad \Leftrightarrow$

$$H_k(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1})$$

The condition (2) is called the <u>secant equation</u>.
There are infinitely many H_k satisfying this condition. Various choices of H_k lead to different Quasi-Newton methods. We discuss the BFGS method.

5.6.1 BFGS Method

We need H_k to be constructed in a way that it can be efficiently computed.
We want H_k to have two properties:

(1) H_k can be computed by some iterative formula

$$H_k = H_{k-1} + M_{k-1}$$

(2) H_k is positive definite (at least guarantee that the BFGS method is a descent method, i.e. $f(x_{k+1}) \leq f(x_k)$).

We can choose $H_0 > 0$ and then guarantee $M_k \geq 0$.
Rank-2 BFGS Method:

$$H_{k+1} = H_k + a_k v_k v_k^T + b_k u_k u_k^T$$

where $v_k \in \mathbb{R}^p$ and $u_k \in \mathbb{R}^p$ are some vectors. If $H_0 > 0$, the above iterative formula can guarantee H_k to be positive definite.
How can we choose v_k and u_k to guarantee the secant equation $H_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$? Let's denote $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. The secant equation: $H_{k+1} s_k = y_k$, then substitute it into the above formula,

$$\begin{aligned}
& y_k = H_{k+1} s_k = H_k s_k + a_k v_k v_k^T s_k + b_k u_k u_k^T s_k \\
& \Leftrightarrow y_k - H_k s_k = a_k (v_k^T s_k) v_k + b_k (u_k^T s_k) u_k
\end{aligned}$$

To let the above equation be satisfied. We let $v_k = y_k$, $u_k = H_k s_k$, $a_k = \frac{1}{y_k^T s_k}$, and $b_k = -\frac{1}{s_k^T H_k s_k}$. Then, the iteration formula becomes

$$H_{k+1} = H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k}$$

where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

This is exactly the BFGS method.
Since we implement the BFGS method as

$$x_{k+1} = x_k - \alpha_k H_k^{-1} \nabla f(x_k)$$

It will be better to compute H_k^{-1} directly instead of H_k.

$$\begin{aligned}
H_{k+1}^{-1} &= \left(H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} \right)^{-1} \\
&= \left(H_k + [H_k s_k \ y_k] \begin{bmatrix} -\frac{1}{s_k^T H_k s_k} & 0 \\ 0 & \frac{1}{y_k^T s_k} \end{bmatrix} \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} \right)^{-1} \\
& \text{(by woodbury formula)} \\
&= H_k^{-1} - H_k^{-1} [H_k s_k \ y_k] \left(\begin{bmatrix} -\frac{1}{s_k^T H_k s_k} & 0 \\ 0 & \frac{1}{y_k^T s_k} \end{bmatrix}^{-1} + \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} H_k^{-1} [H_k s_k \ y_k] \right)^{-1} \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} H_k^{-1} \\
&= H_k^{-1} - [s_k \ H_k^{-1} y_k] \begin{bmatrix} 0 & s_k^T y_k \\ y_k^T s_k & y_k^T (s_k + H_k^{-1} y_k) \end{bmatrix}^{-1} \begin{bmatrix} s_k^T \\ y_k^T H_k^{-1} \end{bmatrix} \\
&= H_k^{-1} - [s_k \ H_k^{-1} y_k] \begin{bmatrix} -\frac{y_k^T s_k + y_k^T H_k^{-1} y_k}{y_k^T s_k s_k^T y_k} & \frac{1}{y_k^T s_k} \\ \frac{1}{y_k^T s_k} & 0 \end{bmatrix} \begin{bmatrix} s_k^T \\ y_k^T H_k^{-1} \end{bmatrix} \\
&= H_k^{-1} - \frac{H_k^{-1} y_k s_k^T}{y_k^T s_K} - \frac{s_k y_k^T H_k^{-1}}{y_k^T s_K} + \frac{s_k s_k^T}{y_k^T s_K} + \frac{s_k y_k^T H_k^{-1} y_k s_k^T}{(y_k^T s_k)^2} \\
&= \left(I - \frac{s_k y_k^T}{y_k^T s_k} \right) H_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k} \right) + \frac{s_k s_k^T}{y_k^T s_k}
\end{aligned}$$

$$H_{k+1}^{-1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}$$

is the iteration computation H_k^{-1} of BFGS method. where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

5.7 Trust-Region Method

$$x_{k+1} = \operatorname*{argmin}_{\|x-x_k\| \le \Delta_k} \{f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)\}$$

This method can escape addle points under some assumptions.

5.8 Cubic Regularization

Contain higher order term $\|x - x_k\|^3$ to the quadratic estimation.

6 Neural Networks

6.1 Neuron

Neuron Neuron is a non-linear function, which takes $\mathfrak{Z} \in \mathbb{R}$ as input and produce $\sigma(\mathfrak{Z}) \in \mathbb{R}$ as output.

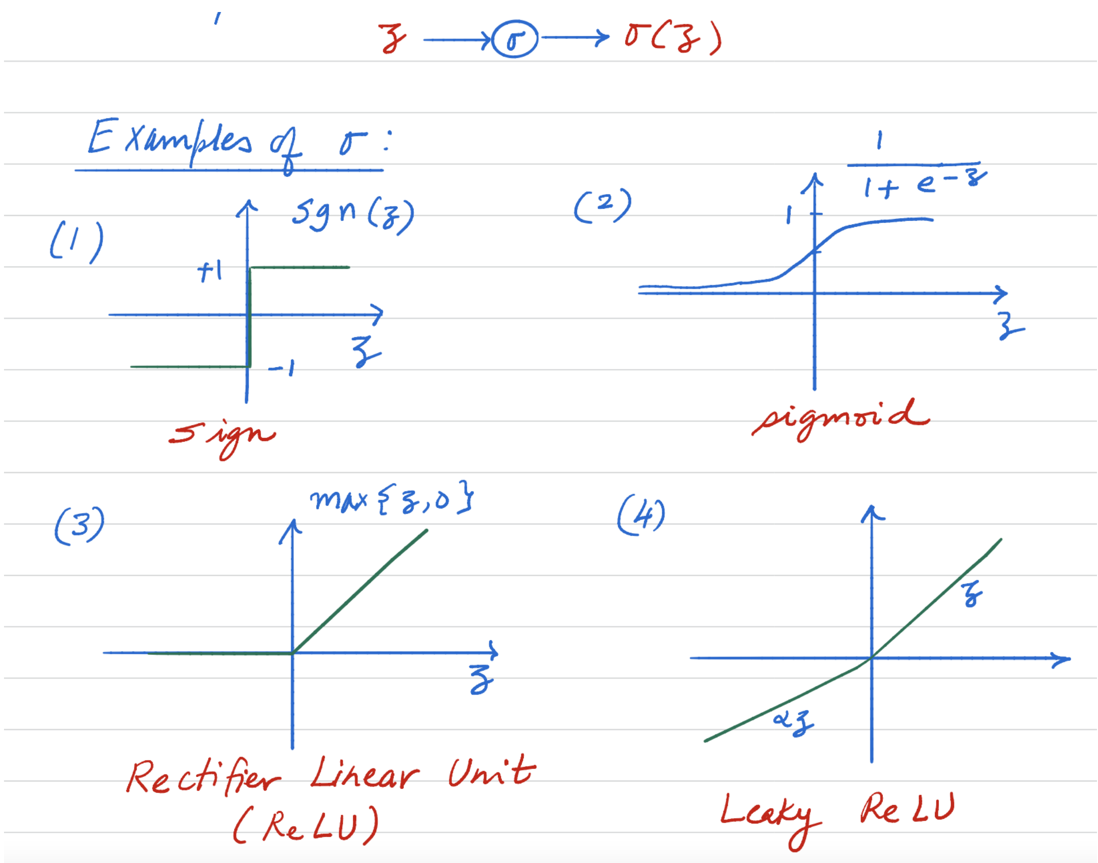


Figure 2: Neuron Examples

Vector Input

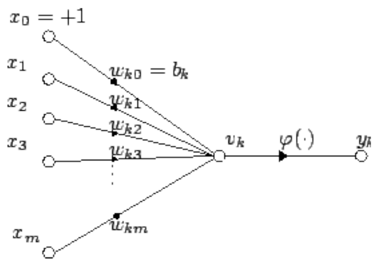


Figure 3: Vector Input

The output is

$$\sigma(\omega^T x + b)$$

6.2 Multilayer Neural Network

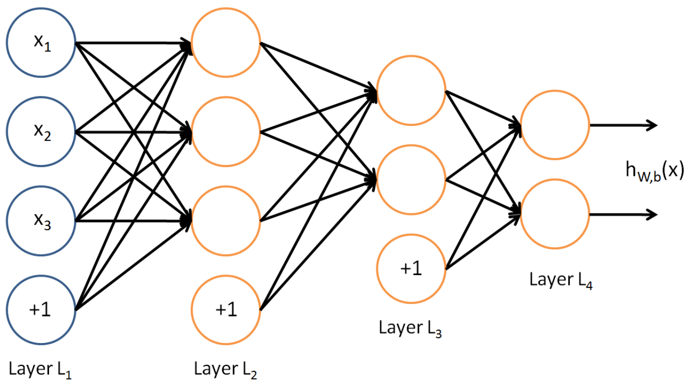


Figure 4: Multilayer Neural Network

- Number of neurons in each layer can be different.
- All weights on edge connecting layers $m-1$ and m is matrix $W^{(m)}$, with $w_{ij}^{(m)}$ being the weight connecting output j of layer $m-1$ with neuron i of layer m.
- Input to network is vector x; output of layer m is vector $y^{(m)}$

$$
\begin{aligned}
y_i^{(1)} &= \sigma(x_i^{(1)}), \text{ with } x_i^{(1)} = \sum_j w_{ij}^{(1)} x_j + b_i^{(1)} \\
y^{(1)} &= \sigma(x^{(1)}), \text{ with } x^{(1)} = W^{(1)} x + b^{(1)} \\
y^{(2)} &= \sigma(x^{(2)}), \text{ with } x^{(2)} = W^{(2)} y^{(1)} + b^{(2)} \\
&\vdots \\
y^{(M)} &= \sigma(x^{(M)}), \text{ with } x^{(M)} = W^{(M)} y^{(M-1)} + b^{(M)}
\end{aligned}
$$

We want to find the weights $W^{(1)}, \cdots, W^{(M)}, b^{(1)}, \cdots, b^{(M)}$ so that the output of last layer

$$\hat{y} = y^{(M)} \approx f^*(x) = y$$

$f^*(x)$ is the unknown thing we need to predict.
We use labelled training data, i.e.

$$(x[1], y[1]), (x[2], y[2]), \cdots (x[N], y[N])$$

Minimize the "empirical" loss on training data.

$$J = \sum_{i=1}^{N} L(y[i], \hat{y}[i])$$

where $\bar{y}[i]$ is the output of NN whose input is $x[i]$.

- L is the function of $W^{(1)}, \cdots, W^{(M)}, b^{(1)}, \cdots, b^{(M)}$ to measure the loss. e.g. the square loss

$$L(y, \hat{y}) = (y - \hat{y})^2$$

- We wish to minimize J using a gradient descent procedure.
- To compute gradient we need:

$$\frac{\partial L}{\partial w_{ij}^{(l)}} \text{ for each } l, i, j; \quad \frac{\partial L}{\partial b_i^{(l)}} \text{ for each } l, i.$$

6.3 Back Propagation Algorithm

Recall $y_i^{(m)} = \sigma(x_i^{(m)})$, $x_i^{(m)} = \sum_j w_{ij}^{(m)} y_j^{(m-1)} + b_i^{(m)}$

$$\frac{\partial L}{\partial w_{ij}^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial w_{ij}^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial w_{ij}^{(m)}}$$

$$\frac{\partial L}{\partial b_i^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial b_i^{(m)}}$$

For large M,

- $\frac{\partial L}{\partial y_i^{(M)}}$ is easy to compute.
- $\frac{\partial y_i^{(M)}}{\partial x_i^{(M)}} = \frac{\partial \sigma(x_i^{(M)})}{\partial x_i^{(M)}} = \sigma'(x_i^{(M)})$, (assuming σ differentiable).
- $\frac{\partial x_i^{(M)}}{\partial w_{ij}^{(M)}} = y_j^{(M-1)}$

Thus,

$$\frac{\partial L}{\partial w_{ij}^{(M)}} = \frac{\partial L}{\partial y_i^{(M)}} \cdot \sigma'(x_i^{(M)}) \cdot y_j^{(M-1)}$$

Similarly,

$$\begin{aligned}\frac{\partial L}{\partial b_i^{(M)}} &= \frac{\partial L}{\partial y_i^{(M)}} \cdot \frac{\partial y_i^{(M)}}{\partial x_i^{(M)}} \cdot \frac{\partial x_i^{(M)}}{\partial b_i^{(M)}} \\ &= \frac{\partial L}{\partial y_i^{(M)}} \cdot \sigma'(x_i^{(M)})\end{aligned}$$

<u>**For** $1 \le m < M$</u>, in this situation $\frac{\partial L}{\partial y_i^{(m)}}$ is not easy to compute. Note that $x^{(m+1)} = W^{(m+1)} y^{(m)} + b^{(m+1)}$.

$$\begin{aligned}\frac{\partial L}{\partial y_i^{(m)}} &= \sum_k \frac{\partial L}{\partial x_k^{(m+1)}} \cdot \frac{\partial x_k^{(m+1)}}{\partial y_i^{(m)}} \\ &= \sum_k \frac{\partial L}{\partial y_k^{(m+1)}} \cdot \frac{\partial y_k^{(m+1)}}{\partial x_k^{(m+1)}} \cdot \frac{\partial x_k^{(m+1)}}{\partial y_i^{(m)}} \\ &= \sum_k \frac{\partial L}{\partial y_k^{(m+1)}} \cdot \sigma'(x_k^{(m+1)}) \cdot w_{ki}^{(m+1)}\end{aligned}$$

Then use this form to compute,

$$\begin{aligned}\frac{\partial L}{\partial w_{ij}^{(m)}} &= \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial w_{ij}^{(m)}} \\ &= \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)}) \cdot y_j^{(m-1)}\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial L}{\partial b_i^{(m)}} &= \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial b_i^{(m)}} \\ &= \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)})\end{aligned}$$

Summary

1. Compute $\frac{\partial L}{\partial y_i^{(M)}}$.
2. Use
$$\frac{\partial L}{\partial y_i^{(m)}} = \sum_k \frac{\partial L}{\partial y_k^{(m+1)}} \cdot \sigma'(x_k^{(m+1)}) \cdot w_{ki}^{(m+1)}$$
compute $\frac{\partial L}{\partial y_i^{(m)}}$ for $m = 1, 2..., M-1$.
3. Compute
$$\frac{\partial L}{\partial w_{ij}^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)}) \cdot y_j^{(m-1)}$$
for $m = 1, 2..., M$.
4. Compute
$$\frac{\partial L}{\partial b_i^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)})$$
for $m = 1, 2..., M$.

6.4 Other Methods

Stochastic Gradient Descent (SGD)
Subgradient Method

7 Constrained Optimization and Gradient Projection

7.1 Constrained Optimization: Basic

7.1.1 Def: Optimality

$$\min_{x\in\&} f(x)$$

where & is a non-empty closed and convex subset of $\mathbb{R}^n$.
Assume f is continuously differentaible on &.

Definition 9. *x^* is a local min of f over & if $\exists\varepsilon > 0$ s.t. $f(x^*) \leq f(x) \quad \forall x \in \&$ with $\|x - x^*\| < \varepsilon$.*
x^* is global min of f over & if $f(x^*) \leq f(x) \quad \forall x \in \&$.

7.1.2 Prop: local-min$\Rightarrow \nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in \& \Leftrightarrow$ global-min in convex

Proposition 4 (optimality conditions)**.**

(a) (Necessary Conditions for local-min) If x^* is a local min of f over &, then

$$\nabla f(x^*)^T(x - x^*) \geq 0 \quad \forall x \in \&$$

(b) (Sufficient and Necessary Condition for global-min of convex f) If f is convex over &, then above condition is also sufficient for x^* to be a global-min.

Proof.

(a) Suppose x^* is a local-min, and $\nabla f(x^*)^T(x - x^*) < 0$ for some $x \in \&$.
Let $g(\varepsilon) = f(x^* + \varepsilon(x - x^*))$, then $g'(\varepsilon) = \nabla f(x^* + \varepsilon(x - x^*))^T(x - x^*)$.
By MVT(middle value theorem), $g(\varepsilon) = g(0) + \varepsilon g'(\beta\varepsilon)$ for some $\beta \in [0, 1]$

$$\Rightarrow f(x^* + \varepsilon(x - x^*)) = f(x^*) + \varepsilon\nabla f(x^* + \beta\varepsilon(x - x^*))^T(x - x^*) \quad \text{for some } \beta \in [0, 1]$$

Since ∇f is continuous, we have that for all sufficient small $\varepsilon > 0$, $\nabla f(x^* + \beta\varepsilon(x - x^*))^T(x - x^*) < 0 \Rightarrow f(x^* + \varepsilon(x - x^*)) = f(x^*)$
Since $x^* + \varepsilon(x - x^*) = \varepsilon x + (1 - \varepsilon)x^* \in \&$, then x^* can't be a local-min over & $\rightarrow$ contradiction.

(b) Convexity of f over & $\Rightarrow f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*), \quad \forall x \in \&$.
Thus,

$$\begin{aligned}
&\nabla f(x^*)^T(x - x^*), \quad \forall x \in \& \\
\Rightarrow &f(x) \geq f(x^*) \quad \forall x \in \& \\
\Rightarrow &x^* \text{ is a global min of } f \text{ over } \&
\end{aligned}$$

$\square$

7.1.3 Def: Interior Point

Definition 10. *y is an interior point of $\&$ if $\exists \varepsilon > 0$ s.t.*

$$B_\varepsilon = \{x : \|y - x\| < \varepsilon\} \subset \&$$

Remark: If x^* is an interior point of $\&$, then

$$\text{"}x^* \text{ is local min"} \Rightarrow \text{"}\nabla f(x^*) = 0\text{"}$$
$$\text{If } f \text{ is convex, "}x^* \text{ is global min"} \Leftrightarrow \text{"}\nabla f(x^*) = 0\text{"}$$

7.2 Constrained Optimization Example

$$\begin{array}{c} \max\limits_{x\in\&} \quad x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n} \\ \& = \{x : \sum_{i=1}^n x_i = 1, x_i \geq 0, i = 1, 2, ..., n\} \\ a_i, i = 1, 2, ..., n \text{ are given positive scalars} \end{array}.$$

equivalent to

$$\min_{x\in\&} \quad f(x)$$

with $f(x) = -\sum a_i \ln x_i$

$$\nabla f(x) = \left(-\frac{a_1}{x_1}, -\frac{a_2}{x_2}, ..., -\frac{a_n}{x_n}\right)$$
$$\nabla^2 f(x) = diag\left(\frac{a_1}{x_1^2}, \frac{a_2}{x_2^2}, ..., \frac{a_n}{x_n^2}\right) \succ 0$$
$$\Rightarrow f \text{ is strictly convex.}$$
$$x^* \in \& \text{ is (unique) min} \Leftrightarrow \nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in \&.$$
$$\Leftrightarrow -\sum_{i=1}^n \frac{a_i}{x_i^*}(x - x^*) \geq 0 \quad \forall x \in \&.$$
$$\Leftrightarrow -\sum_{i=1}^n a_i \frac{x_i}{x_i^*} + \sum_{i=1}^n a_i \geq 0 \quad \forall x \in \&.$$

Guess: $x_i^* = \frac{a_i}{\sum_{i=1}^n a_i}$. Then,

$$-\sum_{i=1}^n a_i \frac{x_i}{x_i^*} + \sum_{i=1}^n a_i = 0, \quad \forall x \in \&$$

Thus $x^* = \frac{a_i}{\sum_{i=1}^n a_i}$ is unique min.

7.3 Projection onto Closed Convex Set

7.3.1 Def: Projection $[z]^\&$

Definition 11. *Let $\&$ be a closed convex subset of $\mathbb{R}^n$. Then, for $z \in \mathbb{R}^n$, the projection of z on $\&$ is denoted by $[z]^\&$ and is given by*

$$[z]^\& = \arg\min_{y\in\&} \|z - y\|^2$$

i.e. Find the min distance from $\&$ to z

Note: $[z]^\&$ exists and is unique in convex $\&$, however, when $\&$ is not convex, $[z]^\&$ may not be unique.

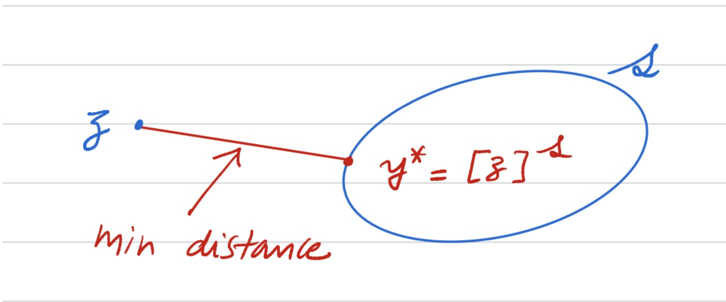


Figure 5: Projection onto Closed Convex Set

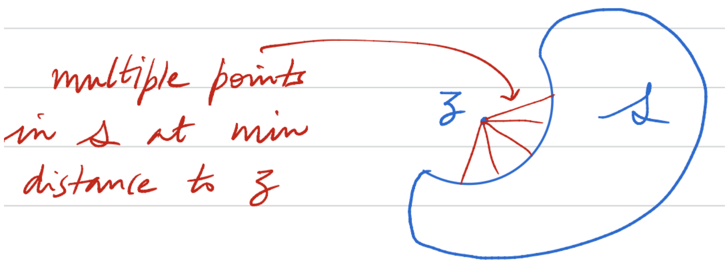


Figure 6: Projection onto Closed non-Convex Set

7.3.2 Prop: <u>unique</u> projection $[z]^{\&}$ on <u>closed convex</u> subset of $\mathbb{R}^n$

Proposition 5 (Existence and Uniqueness of Projection). *Let & be a <u>closed convex</u> subset of $\mathbb{R}^n$. Then, for every $z \in \mathbb{R}^n$, there exists a unique $[z]^{\&}$.*

Proof. Nee to show that $\min_{y\in\&} \|z-y\|^2$ exists and is unique.
Let x be some element of &. Then

$$\begin{aligned}&\text{minimizing } \|z-y\|^2 \text{ over all } y \in \& \\ \equiv&\text{minimizing } \|z-y\|^2 \text{ over the set } A = \{y \in \& : \|z-y\|^2\}\end{aligned}$$

$g(y) = \|z-y\|^2$ is strictly convex on set & $\Rightarrow$ A is a convex set and g is convex on A.
Also g is continuous $\Rightarrow$ A is closed.
Finally, $y \in A \Rightarrow \|y\|^2 = \|y-z+z\|^2 \leq \|y-z\|^2 + \|z\|^2 \leq \|z-x\|^2 + \|z\|^2 \Rightarrow A$ is bounded.
Thus, $g(y) = \|z-y\|^2$ is strictly convex over set A, which is compact.
Therefore, $\min_{y\in\&} \|\& - y\|^2 = \min_{y\in A} \|\& - y\|^2$ exists (Weierstrass' Theorem) and is unique (strict convexity). □

7.3.3 Projection Theorem: $x = [z]^{\&}$ is projection on <u>closed conex</u> subset of $\mathbb{R}^n \Leftrightarrow (z-x)^T(y-x) \leq 0, \forall y \in \&$

Proposition 6 (Necessary and Sufficient Condition for Projection). *Let & be a <u>closed conex</u> subset of $\mathbb{R}^n$. Then,*

$$\begin{aligned}[z]^{\&} = y^* &\Leftrightarrow (y^*-z)^T(y-y^*) \geq 0, \quad \forall y \in \&. \\ &\Leftrightarrow (z-y^*)^T(y-y^*) \leq 0, \quad \forall y \in \&.\end{aligned}$$

Proof. $[z]^{\&} = \operatorname{argmin}_{y\in\&} g(y)$, with $g(y) = \|z-y\|^2$ (which is strictly convex), $\nabla g(y) = 2(y-z)$.

By the optimality conditions,

$$
\begin{aligned}
&y^* \text{ is the unique minimizer of } g(y) \text{ over } \& \\
\Leftrightarrow &\nabla g(y^*)^T (y - y^*) \geq 0 \quad \forall y \in \& \\
\Leftrightarrow &(y^* - z)^T (y - y^*) \geq 0, \quad \forall y \in \&. \\
\Leftrightarrow &(z - y^*)^T (y - y^*) \leq 0, \quad \forall y \in \&.
\end{aligned}
$$

□

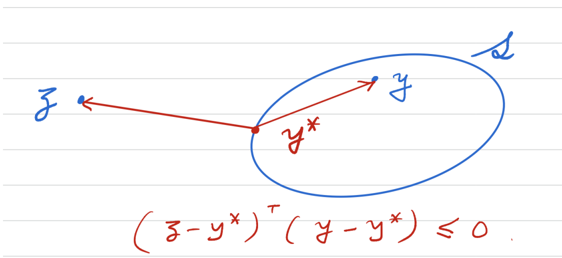


Figure 7: Necessary and Sufficient Condition for Projection

7.3.4 Prop: Projection is non-expansive $\|[x]^\& - [z]^\&\| \leq \|x - z\|, \forall x, z \in \mathbb{R}^n$

Proposition 7 (Projection is non-expansive). *Let & be a* <u>*closed convex*</u> *subset of* $\mathbb{R}^n$*. Then for* $x, z \in \mathbb{R}^n$

$$\|[x]^\& - [z]^\&\| \leq \|x - z\| \quad \forall x, z \in \mathbb{R}^n$$

Proof. From previous theorem, we know

$$
\begin{aligned}
(1). \quad ([x]^\& - x)^T (y - [x]^\&) \geq 0, \quad \forall y \in \&. \\
(2). \quad ([z]^\& - z)^T (y - [z]^\&) \geq 0, \quad \forall y \in \&.
\end{aligned}
$$

set $y = [z]^\&$ in (1) and $y = [x]^\&$ in (2), and adding,

$$
\begin{aligned}
&([z]^\& - [x]^\&)^T ([x]^\& - x + z - [z]^\&) \geq 0 \\
\Rightarrow &([z]^\& - [x]^\&)^T (z - x) \geq \|[z]^\& - [x]^\&\|^2
\end{aligned}
$$

Applying Cauchy-schwary inequality,

$$
\begin{aligned}
&\|[z]^\& - [x]^\&\|^2 \leq \|[z]^\& - [x]^\&\| \|z - x\| \\
&\|[z]^\& - [x]^\&\| \leq \|z - x\|
\end{aligned}
$$

□

7.4 Projection on (Linear) Subspaces of $\mathbb{R}^n$

7.4.1 Orthogonality Principle in subspaces of $\mathbb{R}^n$: $(z - y^*)^T x = 0, \forall x \in \&$

Suppose & is a linear subspace of $\mathbb{R}^n$, any linear combination of points in & is also in &. Note that & is <u>closed and convex</u>.

Then, for $z \in \mathbb{R}^n$, $[z]^{\&} = y^*$ satisfies:

$$(z - y^*)^T (y - y^*) \leq 0, \quad \forall y \in \&.$$

According to the property of subsapce, we can infer that

$$(z - y^*)^T x \leq 0, \quad \forall x \in \&.$$

$-x$ also in $\&$, $-x \in \& \Rightarrow$

$$(z - y^*)^T x \geq 0, \quad \forall x \in \&.$$

Then we can infer that

$$(z - y^*)^T x = 0, \quad \forall x \in \&.$$

which is called orthogonality principle.

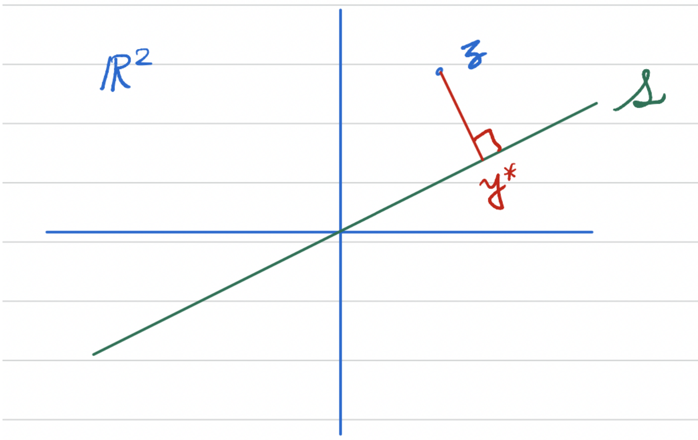


Figure 8: Point from $\mathbb{R}^2$ to $\mathbb{R}$

7.5 Gradient Projection Method

$\min_{x \in \&} f(x)$, $\&$ is convex and closed.

$$x_{k+1} = [x_k + \alpha_k d_k]^{\&}$$

Special Case: Fixed step-size, steepest descent

$$x_{k+1} = [x_k - \alpha \nabla f(x_k)]^{\&} \tag{1}$$

7.5.1 Def: fixed point in fixed step-size steepest descent method, $\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$

Definition 12. *$\tilde{x}$ is a fixed (stationary) point of iteration in (1) if*

$$\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$$

7.5.2 Prop: $L-$smooth, $0 < \alpha < \frac{2}{L} \Rightarrow$ limit point is a fixed point (in fixed step-size steepest descent method)

Proposition 8. *If f has $L-$Lipschitz gradient and $0 < \alpha < \frac{2}{L}$, every limit point of (1) is a fixed point of (1).*

Proof. By the Descent Lemma,

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \tag{2}$$

By the necessary and sufficient condition for projection,

$$(x_k - \alpha\nabla f(x_k) - x_{k+1})^T(x - x_{k+1}) \leq 0, \quad \forall x \in \&$$

Set $x = x_k$ above

$$\Rightarrow \alpha\nabla f(x_k)^T(x_{k+1} - x_k) \leq -\|x_k - x_{k+1}\|^2 \tag{3}$$

According to (2) and (3),

$$f(x_{k+1}) - f(x_k) \leq (\frac{L}{2} - \frac{1}{\alpha})\|x_k - x_{k+1}\|^2$$

where $\frac{L}{2} - \frac{1}{\alpha} < 0$
If $\{x_k\}$ has limit point $\bar{x}$, $LHS \overset{k\to\infty}{\longrightarrow} 0$

$$\|x_{k+1} - x_k\| \overset{k\to\infty}{\longrightarrow} 0 \Rightarrow [\bar{x} - \alpha\nabla f(\bar{x})]^{\&} = \bar{x}$$

□

7.5.3 Prop: x is minimizer in convex func $\Leftrightarrow$ fixed point (in fixed step-size steepest descent method)

Proposition 9. *If f is convex, then x^* is a minimizer of f over $\&$ $\Leftrightarrow x^* = [x^* - \alpha\nabla f(x^*)]^{\&}$ (i.e., x^* is a fixed point of (1))*

Proof.

$$\begin{aligned} x^* \text{ is minimizer of convex } f \text{ over } \& &\Leftrightarrow \nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in \& \\ &\Leftrightarrow -\alpha\nabla f(x^*)^T(x - x^*) \leq 0, \forall x \in \& \\ &\Leftrightarrow (x^* - \alpha\nabla f(x^*) - x^*)^T(x - x^*) \leq 0, \forall x \in \& \\ \text{(By Projection Theorem)} &\Leftrightarrow [x^* - \alpha\nabla f(x^*)]^{\&} = x^* \end{aligned}$$

□

7.5.4 Thm: Convergence of Gradient Projection: Convex, $L-$smooth, $0 < \alpha < \frac{2}{L} \Rightarrow f(x_k) \to f(x^*)$ at rate $\frac{1}{k}$

Theorem 13. *If f is convex and $L-$Lipschitz gradient, it can be shown that for $0 < \alpha < \frac{2}{L}$*

$$f(x_k) \to f(x^*) \text{ at rate } \frac{1}{k} \text{(same as unconstrainted)}$$

7.5.5 Thm: Strongly convex, Lipschitz gradient $\Rightarrow \{x_k\}$ converges to x^* geometrically

Theorem 14. *If f has Lipschitz gradient with Lipschitz constant M and strongly convex with parameter m, $\{x_k\}$ converges to x^** ***geometrically****.*

Proof. $M-$smooth $\Rightarrow$

$$\|\nabla f(x) - \nabla f(y)\| \le M\|x-y\|, \quad \forall x, y \in \&$$

$m-$strongly convex $\Rightarrow$

$$\nabla^2 f(x) \succeq mI, \quad \forall x \in \&$$

$$(x-y)^T(\nabla f(x) - \nabla f(y)) \ge m\|x-y\|^2 \quad \forall x, y \in \&$$

Let x^* be the (unique) min of f over &

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|[x_k - \alpha\nabla f(x_k)]^{\&} - x^*\|^2 \\
(x^* \text{ is fixed point}) \qquad &= \|[x_k - \alpha\nabla f(x_k)]^{\&} - [x^* - \alpha\nabla f(x^*)]^{\&}\|^2 \\
(\text{non-expansive}) \qquad &\le \|(x_k - \alpha\nabla f(x_k)) - (x^* - \alpha\nabla f(x^*))\|^2 \\
&= \|(x_k - x^*) - \alpha(\nabla f(x_k) - \nabla f(x^*))\|^2 \\
&= \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k) - \nabla f(x^*)\|^2 - 2\alpha(x_k - x^*)^T(\nabla f(x_k) - \nabla f(x^*)) \\
(\nabla f \text{ is M-Lipschitz}) \qquad &\le \|x_k - x^*\|^2 + \alpha^2 M^2\|x_k - x^*\|^2 - 2\alpha(x_k - x^*)^T(\nabla f(x_k) - \nabla f(x^*)) \\
(m - \text{strong convexity}) \qquad &\le \|x_k - x^*\|^2 + \alpha^2 M^2\|x_k - x^*\|^2 - 2\alpha m\|x_k - x^*\|^2 \\
&= (1 + \alpha^2 M^2 - 2\alpha m)\|x_k - x^*\|^2
\end{aligned}$$

$$\|x_{k+1} - x^*\|^2 \le (1 + \alpha^2 M^2 - 2\alpha m)\|x_k - x^*\|^2$$

If $|1 + \alpha^2 M^2 - 2\alpha m| < 1$. Then $x_N \to x^*$ **geometrically** as $N \to \infty$. (Same as unconstrained case) □

8 Optimization with Equality Constraints

8.1 Basic

$$\begin{aligned}
\min \quad & f(x) \\
s.t. \quad & h(x) = 0
\end{aligned}$$

where $h(x) = 0$ is a combination of $\begin{cases} h_1(x) = 0 \\ h_2(x) = 0 \\ \quad\vdots \\ h_m(x) = 0 \end{cases}$. $f : \mathbb{R}^n \to \mathbb{R}, h_i : \mathbb{R}^n \to \mathbb{R}, h : \mathbb{R}^n \to \mathbb{R}^m$

Note: we usually assume 1. h is coninuous, then $H = \{x : h(x) = 0\}$ is closed but may not convex; 2. h_i are consistent, i.e., H is non-empty.

8.2 Lagrange Mutiplier Theorem

8.2.1 First-order necessary condition: $\exists\lambda, \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0$

We say the optimal solution is regular if $\nabla h_i(x^*), i = 1, ..., m$ are linearily independent.

Theorem 15 (Lagrange Mutiplier Theorem: First-order necessary condition)**.** *Let x^* be a local-min of $f(x)$ subject to $h(x) = 0$. Assume that $\nabla h_i(x^*), i = 1, ..., m$ are linearily independent. Then $\exists$ a unqiue $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$ s.t.*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0 \tag{4}$$

Remark: Since $x \in \mathbb{R}^n$, $m = n$ makes the equation trivial. We only consider $m < n$.

Proof. Consider sequence of functions:

$$g^{(k)}(x) = f(x) + \frac{k}{2}\|h(x)\|^2 + \frac{\alpha}{2}\|x - x^*\|^2, \quad k = 1, 2, ...$$

x^* local min of $f(x)$ over $H = \{x : h(x) = 0\} \Rightarrow \exists \varepsilon > 0$ s.t. $f(x^*) \leq f(x)$ for all $x \in H \cap \&$, where $\& = \{x : \|x - x^*\| \leq \varepsilon\}$.
According to the Weierstrass's Theorem, $\& \Rightarrow$ Optimal solution to $\min_{x\in\&} g^{(k)}(x)$ exists, we denote the optimal solution to be $x^{(k)}$.

Lemma 5. $x^{(k)} \to x^*$ *as* $k \to \infty$

Proof.

$$\begin{aligned} g^{(k)}(x^{(k)}) &= f(x^{(k)}) + \frac{k}{2}\|h(x^{(k)})\|^2 + \frac{\alpha}{2}\|x^{(k)} - x^*\|^2, \quad k = 1, 2, ... \\ &\leq g^{(k)}(x^*) = f(x^*) \end{aligned}$$

Now since $f(x^{(k)})$ is bounded over $\&$, $\forall k$, we must have $\lim_{k\to\infty} \|h(x^{(k)})\| = 0$. Thus, every limit point of $x^{(k)}$, $\bar{x}$ must satisfy $h(\bar{x}) = 0$, i.e., $\bar{x} \in H$.

$$\begin{aligned} & \Rightarrow f(\bar{x}) + \frac{\alpha}{2}\|\bar{x} - x^*\|^2 \leq f(x^*) \\ (x^* \text{ is local-min, i.e., } f(x^*) \leq f(\bar{x})) \quad & \Rightarrow \bar{x} = x^* \end{aligned}$$

Thus $\lim_{k\to\infty} x^{(k)} = x^*$. □

According to the lemma, $x^{(k)}$ is an interior point of $\&$ for k sufficiently large.
$\Rightarrow \nabla g^{(k)}(x^{(k)}) = 0$ for k sufficiently large.

$$\begin{aligned} g^{(k)}(x) &= f(x) + \frac{k}{2}\|h(x)\|^2 + \frac{\alpha}{2}\|x - x^*\|^2 \\ \nabla g^{(k)}(x) &= \nabla f(x) + k\sum_{i=1}^{m} h_i(x)\nabla h_i(x) + \alpha(x - x^*) \end{aligned}$$

Let $\nabla h(x)$ denote the combination of $\nabla h_i(x), i = 1, ..., m$

$$\begin{aligned} 0 =& \nabla g^{(k)}(x^{(k)}) = \nabla f(x^{(k)}) + k\nabla h(x^{(k)})h(x^{(k)}) + \alpha(x^{(k)} - x^*) \\ \Rightarrow& k\nabla h(x^{(k)})h(x^{(k)}) = -(\nabla f(x^{(k)}) + \alpha(x^{(k)} - x^*)) \\ \Rightarrow& kh(x^{(k)}) = -\left(\nabla h(x^{(k)})\right)^+ (\nabla f(x^{(k)}) + \alpha(x^{(k)} - x^*)) \\ \Rightarrow& \lim_{k\to\infty} kh(x^{(k)}) = -\left(\nabla h(x^*)\right)^+ \nabla f(x^*) \triangleq \lambda \\ & \text{(Uniqueness from uniqueness of limit)} \end{aligned}$$

Where $(\nabla h(x^*))^+$ is the pseudo-inverse of $\nabla h(x^*) = (\nabla h(x^*)^T \nabla h(x^*))^{-1} \nabla h(x^*)^T$ Then

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0$$

□

8.2.2 Second-order necessary condition: $z^T\left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*)\right) z \geq 0, \forall z \in V(x^*)$

Theorem 16. *With the unique $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$ satisfies $\nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0$, the second-order necessary condition is*

$$z^T \left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*) \right) z \geq 0, \quad \forall z \in V(x^*)$$

where $V(x^*) = \{z : \nabla h_i(x^*)^T z = 0, i = 1, ..., m\}$.

Proof.

$$\nabla g^{(k)}(x) = \nabla f(x) + k \sum_{i=1}^m h_i(x) \nabla h_i(x) + \alpha(x - x^*)$$

$$\nabla^2 g^{(k)}(x) = \nabla^2 f(x) + k \sum_{i=1}^m \nabla h_i(x) \nabla h_i(x)^T + k \sum_{i=1}^m h_i(x) \nabla^2 h_i(x) + \alpha I$$

Since $x^{(k)}$ is the optimal value of unconstrained minimization of $g^{(k)}(x)$, we have $\nabla^2 g^{(k)}(x^{(k)}) \succeq 0$. Then,

$$\nabla^2 f(x) + k \sum_{i=1}^m \nabla h_i(x) \nabla h_i(x)^T + k \sum_{i=1}^m h_i(x) \nabla^2 h_i(x) + \alpha I \succeq 0 \tag{5}$$

Consider $z \in V(x^*) = \{z : \nabla h_i(x^*)^T z = 0, i = 1, ..., m\}$. Let

$$\begin{aligned} z^{(k)} &= z - \nabla h(x^{(k)}) \left(\nabla h(x^{(k)})^T \nabla h(x^{(k)}) \right)^{-1} \nabla h(x^{(k)})^T z \\ &= z - \nabla h(x^{(k)}) \left(\nabla h(x^{(k)}) \right)^+ z \end{aligned}$$

Multiply $\nabla h(x^{(k)})$

$$\nabla h(x^{(k)})^T z^{(k)} = 0$$

(5) implies that

$$\begin{aligned} &(z^{(k)})^T \left(\nabla^2 f(x^{(k)}) + k \sum_{i=1}^m \nabla h_i(x^{(k)}) \nabla h_i(x^{(k)})^T + k \sum_{i=1}^m h_i(x^{(k)}) \nabla^2 h_i(x^{(k)}) + \alpha I \right) z^{(k)} \\ =&(z^{(k)})^T \left(\nabla^2 f(x^{(k)}) + k \sum_{i=1}^m h_i(x^{(k)}) \nabla^2 h_i(x^{(k)}) + \alpha I \right) z^{(k)} \geq 0 \end{aligned}$$

As $k \to \infty$, $x^{(k)} \to x^*$, $kh(x^{(k)}) \to -\left(\nabla h(x^*)\right)^+ \nabla f(x^*) \triangleq \lambda$ (is proved in first-order necessary condition part), and $z^{(k)} \to z$, then

$$z^T \left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*) + \alpha I \right) z \geq 0, \quad \forall z \in V(x^*)$$

Taking $\alpha \to 0$,

$$z^T \left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*) \right) z \geq 0, \quad \forall z \in V(x^*)$$

□

8.2.3 Sufficient Condition: $\exists\lambda$ **1.** $\nabla f(x^*)+\sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0$ **2.** $z^T\left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*)\right) z > 0, \forall z \in V(x^*), z \neq 0$

Theorem 17. *Sufficient condition: For x^* that is feasible and regular, if $\exists\lambda$ s.t.*

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0$$

and

$$z^T\left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*)\right) z > 0, \quad \forall z \in V(x^*), z \neq 0$$

Then x^* is a (strict) local min for

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & h(x) = 0 \end{aligned}$$

Note:

(1) If f is convex and $H = \{x : h(x) = 0\}$ is convex and closed. Therefore x^* is also global-min.

(2) If f is coercive and $H = \{x : h(x) = 0\}$ is closed. Therefore f can attain its global-min on &, we just find the minimal local-min.

8.2.4 Lagrangian Function

Lagrangian Function:

$$L(x, \lambda) \triangleq f(x) + \sum_{i=1}^m \lambda_i h_i(x)$$

Then the necessary condition can be rewrote to

$$\begin{array}{ll} \text{First-order:} & \left.\begin{array}{rl} \nabla_x L(x^*, \lambda) & = 0 \\ h(x^*) = \nabla_\lambda L(x^*, \lambda) & = 0 \end{array}\right\} m+n \text{ equations in total} \\ \text{Second-order:} & z^T \nabla^2_{xx} L(x^*, \lambda) z \geq 0, \quad \forall z \in V(x^*) \end{array}$$

The sufficient condition can be rewrote to

$$\begin{array}{ll} \text{First-order:} & \nabla_x L(x^*, \lambda) = 0 \\ \text{Second-order:} & z^T \nabla^2_{xx} L(x^*, \lambda) z > 0, \quad \forall z \in V(x^*), z \neq 0 \end{array}$$

8.2.5 Example

Example 12.

$$\begin{aligned} \min \quad & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ s.t. \quad & x_1 + x_2 + x_3 = 3 \end{aligned} \quad .$$

$$\begin{aligned} h(x) &= x_1 + x_2 + x_3 - 3 \\ L(x, \lambda) &= \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 - 3) \\ \nabla_x L(x, \lambda) &= [x_1 + \lambda, x_2 + \lambda, x_3 + \lambda]^T \\ \nabla^2_{xx} L(x, \lambda) &= I_{3\times 3} \end{aligned}$$

First order condition:

$$\begin{aligned}\nabla_x L(x^*, \lambda) &= [x_1^* + \lambda, x_2^* + \lambda, x_3^* + \lambda]^T = 0 \\ &\Rightarrow x^* = [-\lambda, -\lambda, -\lambda] \\ &h(x^*) = x_1^* + x_2^* + x_3^* - 3 = 0 \\ &\Rightarrow x^* = [1, 1, 1]\end{aligned}$$

And $\nabla_{xx}^2 L(x, \lambda) \succ 0$, then $x^* = [1, 1, 1]$ is local-min.
Note: Since $f(x) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$ is coercive on $H = \{x : x_1 + x_2 + x_3 = 3\}$ and $H = \{x : x_1 + x_2 + x_3 = 3\}$ is closed $\Rightarrow$ f achieves its global min on H. $x^* = [1, 1, 1]$ is the unique local min, so it is also global-min.

8.2.6 Sensitivity Analysis $f(x^*(u)) = f(x^*) - \lambda^T u + O(\|u\|)$

As constants of constraints change, how will the optimal value change?

$$\begin{aligned}\min \quad & f(x) \\ s.t. \quad & h(x) = u\end{aligned}$$

Claim 4.

$$f(x^*(u)) = f(x^*) - \lambda^T u + O(\|u\|)$$

Proof. Let $p(u) = f(x^*(u))$, $p(0) = f(x^*(0)) = f(x^*)$
First-order necessary condition:

$$\begin{aligned}\nabla f(x^*(u)) + \sum_{i=1}^{m} \lambda_i(u) \nabla h_i(x^*(u)) = 0 \\ \Rightarrow \frac{\partial f(x^*(u))}{\partial x_k} = -\sum_{i=1}^{m} \lambda_i(u) \frac{\partial h_i(x^*(u))}{\partial x_k}\end{aligned}$$

$$\begin{aligned}\frac{\partial p(u)}{\partial u_j} = \frac{\partial f(x^*(u))}{\partial u_j} &= \sum_{k=1}^{n} \frac{\partial f(x^*(u))}{\partial x_k} \frac{\partial x_k^*(u)}{\partial u_j} \\ &= -\sum_{k=1}^{n} \sum_{i=1}^{m} \lambda_i(u) \frac{\partial h_i(x^*(u))}{\partial x_k} \frac{\partial x_k^*(u)}{\partial u_j} \\ &= -\sum_{i=1}^{m} \lambda_i(u) \frac{\partial h_i(x^*(u))}{\partial u_j} \\ (h_i = u_i) \qquad &= \lambda_j(u)\end{aligned}$$

Then we can conclude that

$$\begin{aligned}\nabla p(u) &= -\lambda(u) \\ \Rightarrow \nabla p(0) &= -\lambda \\ \Rightarrow f(x^*(u)) &= f(x^*) + \nabla p(0)(u - 0) + O(\|u\|) \\ &= f(x^*) - \lambda^T u + O(\|u\|)\end{aligned}$$

□

8.2.7 Linear Constraints

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & Ax = b \end{aligned}$$

where $A = [a_1, a_2, ..., a_m]^T$, $b = [b_1, b_2, ..., b_m]^T$. $h_i(x) = a_i^T x - b_i, i = 1, ..., m$, $\nabla h_i(x) = a_i$.

$$\begin{aligned} L(x, \lambda) &= f(x) + \sum_{i=1}^{m} \lambda_i (a_i^T x - b_i) \\ \nabla_x L(x, \lambda) &= \nabla f(x) + \sum_{i=1}^{m} \lambda_i a_i \\ \nabla_{xx}^2 L(x, \lambda) &= \nabla^2 f(x) \end{aligned}$$

Then, the first order condition can be reworte to

$$\begin{aligned} &\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0 \\ \Leftrightarrow &\sum_{i=1}^{m} \lambda_i a_i = -\nabla f(x^*) \\ \Leftrightarrow &A^T \lambda = -\nabla f(x^*) \end{aligned}$$

Also, $Ax^* = b$. If $m = n$, there will unique $x^* = A^{-1}b$. If $m < n$,

Step (1) We compute first-order condition:

$$\left.\begin{array}{rl} A^T \lambda + \nabla f(x^*) & = 0 \\ Ax^* & = b \end{array}\right\} m + n \text{ equations}$$

Step (2) Check sufficient conditions (second-order): Check:

$$z^T \nabla^2 f(x) z > 0$$

$\forall z \neq 0$ s.t. $a_i^T z = 0, i = 1, ..., m$

Example 13.

$$\begin{aligned} \min \quad & -(x_1x_2 + x_2x_3 + x_1x_3) \\ s.t. \quad & x_1 + x_2 = 2 \\ & x_2 + x_3 = 1 \end{aligned}$$

$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\nabla f(x) = -\begin{bmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_1 + x_2 \end{bmatrix}$

Solve

$$\left\{\begin{array}{rl} A^T \lambda + \nabla f(x^*) & = 0 \\ Ax^* & = b \end{array}\right. \Rightarrow x_1^* = 2, x_2^* = 0, x_3^* = 1$$

Check second order condition for $x^* = (2, 0, 1)$:

$$\begin{aligned} \nabla^2_{xx} L(x, \lambda) &= \nabla^2 f(x) \\ &= \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} \end{aligned}$$

$\nabla^2_{xx} L(x, \lambda)$ is not PD or PSD. But we only need $z^T \nabla^2 f(x) z > 0 \ \forall z \neq 0$ s.t. $a_i^T z = 0, i = 1, 2$

$$Az = \begin{bmatrix} z_1 + z_2 \\ z_2 + z_3 \end{bmatrix} = 0 \Rightarrow \begin{cases} z_1 = -z_2 \\ z_3 = -z_2 \end{cases}$$

$$\begin{aligned} z^T \nabla^2 f(x) z &= -2(z_1 z_2 + z_2 z_3 + z_1 z_3) \\ &= 2z_2^2 > 0 \quad z \neq 0 \end{aligned}$$

Thus $x^* = (2, 0, 1)$ is a local-min.

$$\begin{aligned} f(x) &= -(x_1 x_2 + x_2 x_3 + x_1 x_3) \\ \text{with } h(x) = 0 \quad &= x_2^2 - 2 \end{aligned}$$

f is coercive on $H = \{x : h(x) = 0\}$ which is closed. Then f achieves its global min on H. Hence, $x^* = (2, 0, 1)$ is the global min of $f(x)$ on H.

9 Optimization with Inequality Constraints

9.1 Basic

Inequality Constraints Problem (ICP)

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & h(x) = 0 \\ & g(x) \leq 0 \end{aligned}$$

where $h(x) = 0$ is a combination of $h_i(x) = 0, i = 1, ..., m$ and $g(x) \leq 0$ is a combination of $g_j(x) \leq 0, j = 1, ..., r$. $f : \mathbb{R}^n \to \mathbb{R}, h : \mathbb{R}^n \to \mathbb{R}^m, g : \mathbb{R}^n \to \mathbb{R}^r$

9.1.1 Active vs. Inactive Inequality Constraints

The constraint $g_j(x) \leq 0$ is said to be active at x if $g_j(x) = 0$, and inactive if $g_j(x) < 0$. We set the set of active inequality constraints $A(x) = \{j \in \{1, ..., r\} : g_j(x) = 0\}$

9.1.2 ICP → ECP

Claim: If x^* is a local min for ICP, then x^* is also a local min for ECP:

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & h_i(x) = 0 \quad i = 1, ..., m \\ & g_j(x) = 0 \quad j \in A(x^*) \end{aligned}$$

If x^* is regular for the ECP, i.e., if $\nabla h_i(x^*), i = 1, ..., m$ and $\nabla g_j(x^*), j = 1, ..., r$ are linearly independent, then $\exists\ \lambda_i, i = 1, ..., m$, $\mu_j, j \in A(x^*)$ s.t.

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j \nabla g_j(x^*) = 0$$

9.1.3 Intuition $\mu_j \geq 0, \forall j \in A(x^*)$

Consider $g_j(x) \leq 0$ is changed to $g_j(x) \leq u_j$. Then $f(x^*(u_j)) \leq f(x^*)$ because of the larger set. Then, if $j \in A(x^*)$, then $g(x^*) = 0$,

$$\begin{aligned} f(x^*(u_j)) &= f(x^*) - \mu_j u_j + O(u_j) \\ \Rightarrow -\mu_j u_j + O(u_j) &= f(x^*(u_j)) - f(x^*) \leq 0 \end{aligned}$$

Dividing by u_j and letting $u_j \to 0 \Rightarrow \mu_j \geq 0$

9.1.4 Complementary Slackness

$$\begin{aligned} &\mu_j = 0 \quad \forall j \notin A(x^*) \\ \Leftrightarrow &\mu_j = 0 \quad \text{whenever } g_j(x^*) < 0, j = 1, ..., r \\ \Leftrightarrow &\mu_j g_j(x^*) = 0, \quad j = 1, ..., r \end{aligned}$$

9.2 Karush–Kuhn–Tucker (KKT) Necessary Conditions

Largrangian Function for ICP:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x)$$

Proposition 10 (KKT)**.** *Let x^* be a local min of (ICP) and assume that x^* is regular for (ECP). Then $\exists$ unique largrange multipliers $\lambda = (\lambda_1, ..., \lambda_m)$ and $\mu = (\mu_1, ..., \mu_r)$ s.t.*

$$\begin{aligned} \nabla_x L(x^*, \lambda, \mu) = 0 \\ \mu_j \geq 0, \ j = 1, ..., r \\ \mu_j = 0, \ \forall j \notin A(x^*) \end{aligned}$$

If f, h_i, g_j are twice continuously differentiable, then

$$y^T \nabla^2_{xx} L(x^*, \lambda, \mu) y \geq 0$$

$\forall y \in \mathbb{R}^n$, $\nabla h_i(x^*)^T y = 0, i = 1, ..., m$, $\nabla g_j(x^*)^T y = 0, \forall j \in A(x^*)$

Proof. Convert (ICP) to:

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & h_i(x) = 0 \quad i = 1, ..., m \\ & g_j(x) + z_j^2 = 0 \quad j = 1, ..., r \end{aligned}$$

Auxilliary valiables $z = (z_1, ..., z_r), z_j \geq 0$.
Let x^* be a local min. For (ICP), then (x^*, z^*) is a local min.

$$z_j^* = (-g_j(x^*))^{\frac{1}{2}}, \quad j = 1, ..., r$$

We can consider the optimization problem over x and $z = (z_1, ..., z_r)$. Define Lagrangian:

$$L(x, z, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) + \sum_{j=1}^{r} \mu_j z_j^2$$

(1) <u>First-order necessary condition</u>

From first-order necessary condition of optimization over x and z: assuming (x^*, z^*) is regular,

$$\nabla L(x^*, z^*, \lambda, \mu) = 0 \quad \text{Note: } \nabla \text{ r.t. both } x \text{ and } z$$
$$\Rightarrow \begin{cases} \nabla_x f(x^*) + \sum_{i=1}^m \lambda_i \nabla_x h_i(x^*) + \sum_{j=1}^r \mu_j \nabla_x g_j(x^*) = 0 \\ \sum_{j=1}^r \mu_j \nabla_z (z_j^2)|_{z=z^*} = 0 \end{cases}$$
$$\Rightarrow \begin{cases} \nabla_x L(x^*, \lambda, \mu) = 0 \\ \mu_j z_j^* = 0, \quad j = 1, .., r \end{cases}$$

Since $z_j^* = (-g_j(x^*))^{\frac{1}{2}} > 0, \forall j \notin A(x^*) \Rightarrow \mu_j = 0$ for $j \notin A(x^*)$

(2) <u>Second-order necessary condition</u>

$$\begin{bmatrix} y \\ \omega \end{bmatrix}^T \nabla^2 L(x^*, z^*, \lambda, \mu) \begin{bmatrix} y \\ \omega \end{bmatrix} \geq 0$$

$\forall y \in \mathbb{R}^n$, $\nabla_x h_i(x^*)^T y = 0, i = 1, ..., m$, $\nabla_x g_j(x^*)^T y + 2z_j^* \omega_j = 0, j = 1, ..., r$

$$\nabla^2 L(x, z, \lambda, \mu) = \begin{bmatrix} \nabla_{xx}^2 L(x, \lambda, \mu) & 0 & 0 & \cdots & 0 \\ 0 & 2\mu_1 & 0 & \cdots & 0 \\ 0 & 0 & 2\mu_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 2\mu_r \end{bmatrix}_{(n+r)\times(n+r)}$$

For every $j \in A(x^*)$ select (y, ω) with $y = 0, \omega_j \neq 0$ and $\omega_k = 0, \forall k \neq j$.

Then (y, ω) satisfies $\nabla h_i(x^*)^T y = 0, i = 1, ..., m$ and $\nabla g_k(x^*)^T y + 2z_k^* \omega_k = 0, \forall k \neq j$ and $\nabla g_j(x^*)^T y + 2z_j^* \omega_j = 0$ (since $z_j^* = 0, \forall j \in A(x^*)$) Thus, for the choice of (y, ω)

$$\begin{bmatrix} y \\ \omega \end{bmatrix}^T \nabla^2 L(x^*, z^*, \lambda, \mu) \begin{bmatrix} y \\ \omega \end{bmatrix} = 2\mu_j \omega_j^2 \geq 0 \Rightarrow \mu_j \geq 0$$

Similarly we can show $\mu_j \geq 0, \forall j \in A(x^*)$

Now choose $y \in \mathbb{R}^n$ s.t.

$$\nabla_x h_i(x^*)^T y = 0, \quad i = 1, ..., m, \quad \nabla_x g_j(x^*)^T y = 0, \forall j \in A(x^*)$$

and ω s.t.

$$\omega_j = \begin{cases} 0 & \text{if } j \in A(x^*) \\ -\frac{\nabla_x g_j(x^*)^T y}{2z_j^*} & \text{if } j \notin A(x^*) \end{cases}$$

Thus $\nabla_x g_j(x^*)^T y + 2z_j^* \omega_j = 0 \quad \forall j = 1, ..., r$.

Since $\omega_j = 0, \forall j \in A(x^*)$ and $\mu_j = 0, \forall j \notin A(x^*) \Rightarrow \mu_j \omega_j = 0, \forall j = 1, ..., r$.

Then, for the above choive of (y, ω):

$$\begin{bmatrix} y \\ \omega \end{bmatrix}^T \nabla^2 L(x^*, z^*, \lambda, \mu) \begin{bmatrix} y \\ \omega \end{bmatrix} \geq 0$$
$$\Rightarrow y^T \nabla_{xx}^2 L(x^*, \lambda, \mu) y \geq 0$$

$\square$

9.3 Karush–Kuhn–Tucker (KKT) Sufficient Conditions

Proposition 11 (KKT)**.** *Suppose x^*, λ, μ satify the first order necessary condition i.e., $\exists$ unique largrange multipliers $\lambda = (\lambda_1, ..., \lambda_m)$ and $\mu = (\mu_1, ..., \mu_r)$ s.t.*

$$\nabla_x L(x^*, \lambda, \mu) = 0$$
$$\mu_j \geq 0, \ \ j = 1, ..., r$$
$$\mu_j = 0, \ \ \forall j \notin A(x^*)$$

and in addition

$$\mu_j > 0, \quad \forall j \in A(x^*)$$

and

$$y^T \nabla^2_{xx} L(x^*, \lambda, \mu) y > 0$$

$\forall y \neq 0$, $\nabla h_i(x^*)^T y = 0, i = 1, ..., m$, $\nabla g_j(x^*)^T y = 0, \forall j \in A(x^*)$
Then x^* is a (strict) local min of (ICP).

Example 14.

$$\begin{aligned} \min \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ s.t. \quad & x_1^2 + x_2^2 \leq 5 \\ & 3x_1 + x_2 \leq 6 \end{aligned}$$

$$f(x) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2$$
$$g_1(x) = x_1^2 + x_2^2 - 5, \quad g_2(x) = 3x_1 + x_2 - 6$$
$$\nabla f(x) = \begin{bmatrix} 4x_1 + 2x_2 - 10 \\ 2x_1 + 2x_2 - 10 \end{bmatrix}, \nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}, \nabla g_2(x) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

- <u>KKT First-order Necessary Condition</u>: $g_1(x^*) \leq 0$, $g_2(x^*) \leq 0$, assuming x^* regular,

$$\nabla L(x^*, \mu) = \nabla f(x^*) + \mu_1 \nabla g_1(x^*) + \mu_2 \nabla g_2(x^*) = 0$$
$$\begin{bmatrix} (4 + 2\mu_1)x_1^* + 2x_2^* + 3\mu_2 - 10 \\ 2x_1^* + (2 + 2\mu_1)x_2^* + \mu_2 - 10 \end{bmatrix} = 0$$
$$\mu_1 \geq 0, \quad \mu_2 \geq 0$$
$$\mu_1 g_1(x^*) = 0, \quad \mu_2 g_2(x^*) = 0$$
$$\mu_1((x_1^*)^2 + (x_2^*)^2 - 5) = \mu_2(3x_1^* + x_2^* - 6) = 0$$

We don't know which constraints are active. We *check all possibilities.*

(1) (1 is inactive, 2 is inactive): $\mu_1 = \mu_2 = 0$ (no need to check for regularity)

$$\nabla f(x) = 0 \Rightarrow x_1 = 0, x_2 = 5$$

which contradicts to $x_1^2 + x_2^2 \leq 5$.

(2) (1 is inactive, 2 is active), i.e., $\mu_1 = 0$

$$\nabla g_2(x) = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \Rightarrow \text{ all feasible } x \text{ are regular}$$

$$\nabla L(x, \mu) = \nabla f(x) + \mu_2 \nabla g_2(x) = 0$$
$$\Rightarrow \left.\begin{aligned} 4x_1 + 2x_2 + 3\mu_2 - 10 &= 0 \\ 2x_1 + 2x_2 + \mu_2 - 10 &= 0 \\ g_2(x) = 0 \Rightarrow 3x_1 + x_2 - 6 &= 0 \end{aligned}\right\} \Rightarrow \begin{cases} x_1 &= \frac{2}{5} \\ x_2 &= \frac{24}{5} \\ \mu_2 &= -\frac{2}{5} \end{cases}$$

But $\mu_2 < 0$ not allowed $\Rightarrow$ solution invalid.

(3) (1 is active, 2 is inactive): $\mu_2 = 0$

$$\nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \Rightarrow \text{ all } x \neq 0 \text{ are regular}$$

$$\begin{aligned} \nabla L(x,\mu) &= \nabla f(x) + \mu_1 \nabla g_1(x) = 0 \\ &\Rightarrow \left.\begin{array}{rl} (4+2\mu_1)x_1 + 2x_2 - 10 &= 0 \\ 2x_1 + (2+2\mu_1)x_2 - 10 &= 0 \\ g_1(x) = 0 \Rightarrow x_1^2 + x_2^2 - 5 &= 0 \end{array}\right\} \Rightarrow \left\{\begin{array}{ll} x_1 &= 1 \\ x_2 &= 2 \\ \mu_1 &= 1 \end{array}\right. \end{aligned}$$

check: $x^* = (1,2)$ is regular (since $\neq 0$)

$$g_1(x^*) = 1 + 4 - 5 = 0$$
$$g_2(x^*) = 3 + 2 - 6 = -1 < 0$$

Hence, $x^* = (1,2)$, $\mu = (1,0)$ satisfy 1^{st} order KKT.

(4) (1 is active, 2 is active)

$\nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$ and $\nabla g_2(x) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ are linearly independent if as $x \neq 0$, $x_1 \neq 3x_2$. But $x = 0$ doesn't satisfy g But $x = 0$ doesn't satisfy either $g_1(x) = 0$ or $g_2(x) = 0$, and $x_1 = 3x_2$ can't satisfy $g_1(x) = 0, g_2(x) = 0$ at the same time. Hence, all feasible x in this case are regular.

$$\nabla L(x,\mu) = \nabla f(x) + \mu_1 \nabla g_1(x) + \mu_2 \nabla g_2(x) = 0$$
$$\Rightarrow \left.\begin{array}{rl} (4+2\mu_1)x_1 + 2x_2 + 3\mu_2 - 10 &= 0 \\ 2x_1 + (2+2\mu_1)x_2 + \mu_2 - 10 &= 0 \\ g_1(x) = 0 \Rightarrow x_1^2 + x_2^2 - 5 &= 0 \\ g_2(x) = 0 \Rightarrow 3x_1 + x_2 - 6 &= 0 \end{array}\right\} \Rightarrow \left\{\begin{array}{ll} x_1 &= 2.2 \\ x_2 &= -0.5 \\ \mu_1 &= -2.4 \\ \mu_2 &= 4.2 \end{array}\right. \text{ or } \left\{\begin{array}{ll} x_1 &= 1.4 \\ x_2 &= 1.7 \\ \mu_1 &= 1.4 \\ \mu_2 &= -1.0 \end{array}\right.$$

both not valid since μ_1, μ_2 are required to be greater than 0.

Hence, the only candidate that satisfies the first order condition is $x^* = (1,2)$, $\mu = (1,0)$

- KKT Second-order Sufficient Condition: $x^* = (1,2)$, $\mu = (1,0)$

$$\nabla^2 L(x^*,\mu) = \begin{bmatrix} 6 & 2 \\ 2 & 4 \end{bmatrix} \succ 0$$

$\mu_1 > 0$.

$\Rightarrow x^*$ is a local min by sufficient condition.

- Constraint set: $\& = \{x : g_1(x) \leq 0, g_2 \leq 0\} = \{x : x_1^2 + x_2^2 \leq 5\} \cap \{x : 3x_1 + x_2 \leq 6\}$ which is compact. So, by WT global minimum exists which is $x^* = (1,2)$.

9.4 General Sufficiency Condition

With possible additional constraints

$$\begin{array}{ll} \min & f(x) \\ s.t. & x \in \& \leftarrow \text{possible additional constraints} \\ & h(x) = 0 \\ & g(x) \leq 0 \end{array}$$

Theorem 18. *Let* $L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x)$.
Suppose (x^*, λ, μ) *satisfy:*

$$\begin{aligned} h_i(x^*) &= 0, i = 1, ..., m \\ g_j(x^*) &\leq 0, j = 1, ..., r \\ \mu_j &\geq 0, j = 1, ..., r \\ \mu_j g_j(x^*) &= 0, j = 1, ..., r \end{aligned}$$

and

$$L(x^*, \lambda, \mu) = \min_{x \in \&} L(x, \lambda, \mu)$$

Then, x^* *is a* <u>*global min*</u> *of this ICP.*
Note: *If* & *is a* <u>*convex*</u> *set, and* f *is* <u>*convex*</u>*,* h_i *are affine (linear + constant),* g_j *are convex over* &*, then* $L(x, \lambda, \mu)$ *is convex* $\Rightarrow \nabla L(x^*, \lambda, \mu) = 0$ *is sufficient for* x^* *to be global min for this ICP.*

Example 15. *Application of General Sufficiency Condition of former example:*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ s.t. \quad & x_1^2 + x_2^2 \leq 5 \\ & 3x_1 + x_2 \leq 6 \end{aligned}$$

$$f(x) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2$$
$$g_1(x) = x_1^2 + x_2^2 - 5, \quad g_2(x) = 3x_1 + x_2 - 6$$

For $x^* = (1, 2)$ and $\mu = (1, 0)$, $L(x, \mu) = f(x) + g_1(x)$ is convex and $\nabla L(x^*, \mu) = 0$

$$\Rightarrow L(x^*, \mu) = \min_{x \in \mathbb{R}^2} L(x, \mu)$$

Then, by genral sufficiency condition, x^* is global min.

9.5 Barrier Method

Computationed method to solve inequality constrained problems.

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in \& \\ & g(x) \leq 0 \end{aligned}$$

where & is closed set.

Barrier Function

$B(x)$ is a function that is continuous and $\to \infty$ as any $g_j(x) \to 0$

Example 16.

$$B(x) = -\sum_{j=1}^{r} \ln(-g_j(x))$$
$$B(x) = -\sum_{j=1}^{r} \frac{1}{g_j(x)}$$

Note: that if $g_j(x)$ is <u>convex</u> for all j, then both of these barrier functions are <u>convex</u>.
In Barrier Method, choose sequence $\{\varepsilon_k\}$ s.t.

$$0 < \varepsilon_{k+1} < \varepsilon_k, \quad k = 0, 1, ...$$

and $\varepsilon_k \to 0$ as $k \to \infty$.
Define feasible set $F = \& \cap \{g_j(x) \leq 0, \forall j\}$. Note F is a closed set since & and $\{g_j(x) \leq 0, \forall j\}$ are closed.
Let $x^{(k)}$ be a solution to

$$\min_{x \in F \cap \text{dom}(B)} f(x) + \varepsilon_k B(x)$$

Since $B(x) \to \infty$ as one $g_j(x) \to 0$ which is on the boundary of F.
$x^{(k)}$ must be an interior point of F

$$\Rightarrow \nabla f(x^{(k)}) + \varepsilon_k \nabla B(x^{(k)}) = 0$$

Therefore, if we have a initial point in the interior of F, we can choose step size of any unconstrained GD method to stay in interior of F for all iterations and solve the ICP. (Because barrier function $B(x)$ will prevent us from reaching boundary)
As $k \to \infty$, $\varepsilon_k \to 0$, and barrier $\varepsilon_k B(x)$ becomes inconsequential, and we expect $x^{(k)}$ to approcah minimum of original problem.

Proposition 12. *Every limit point $\bar{x}$ of $\{x^{(k)}\}$ is a global min of the ICP.*

Proof. Let $\bar{x} = \lim_{k\to\infty, k\in\mathcal{K}} x^{(k)}$, since $x^{(k)} \in F$ forall k, and F is closed, $\bar{x} \in F$.
Suppose x^* is a global min of ICP and x^* is in interior of F, and $f(x^*) < f(\bar{x})$, i.e., $\bar{x}$ is not global min for ICP.
Then, by definition of $x^{(k)}$, $f(x^{(k)}) + \varepsilon_k B(x^{(k)}) \leq f(x^*) + \varepsilon_k B(x^*)$
Taking limit as $k \to \infty$, $k \in \mathcal{K}$,

$$f(\bar{x}) + \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(\bar{x}) \leq f(x^*) + \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(x^*) = f(x^*)$$
$$(\text{Since } |B(x^*)| < \infty, \varepsilon_k \to 0 \text{ as } k \to \infty)$$

If $\bar{x}$ is in interior of F, then $|B(\bar{x})| < \infty \Rightarrow \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(x^{(k)}) = 0$
If $\bar{x}$ is on boundary of F, then $|B(\bar{x})| \to \infty \Rightarrow \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(x^{(k)}) \geq 0$
Therefore, $f(\bar{x}) < f(x^*)$ is contradiced.
If x^* is not in interior of F, we can assume that $\exists$ an interior point $\bar{x}$ which can be made arbitrarily close to x^*. ☐

9.6 An Exmaple Using KKT or Barrier

Example 17.

$$\min \quad f(x) = \frac{1}{2}(x_1^2 + x_2^2)$$
$$s.t. \ x_1 \geq 2$$

9.6.1 Solution using KKT conditions

$$\begin{aligned} g(x) &= -x_1 + 2 \\ \nabla g(x) &= (-1, 0) \quad \text{All feasible } x \text{ are regular} \\ \nabla f(x) &= (x_1, x_2) \\ L(x, \mu) &= f(x) + \mu g(x) \\ \nabla L(x, \mu) &= \nabla f(x) + \mu \nabla g(x) = (x_1 - \mu, x_2) \end{aligned}$$

Case 1: constraint inactive, i.e., $\mu = 0$

$$\nabla L(x,\mu) = 0 \Rightarrow x = (0,0)$$

Doesn't satisfy $x_1 \geq 2$. This case is infeasible.

Case 2: constraint active,

$$\begin{aligned} \nabla L(x,\mu) = 0 \Rightarrow x_1 - \mu = 0, x_2 = 0 \\ g(x) = 0 \Rightarrow x_1 = 2 \\ \Rightarrow x^* = (2,0), \mu = 2 \end{aligned}$$

It satisfies the first-order KKT condition.

Since $L(x,\mu)$ is strictly convex on $\mathbb{R}^2$, $x^* = (2,0)$ is the global-min.

9.6.2 Solution using logarithmic barrier

$B(x) = -\ln(-g(x)) = -\ln(x_1 - 2)$

$$\begin{aligned} \text{Set } G^{(k)}(x) &= f(x) + \varepsilon_k B(x) \\ &= \frac{1}{2}(x_1^2 + x_2^2) - \varepsilon_k \ln(x_1 - 2) \\ (G^{(k)}(x) &\text{ is convex in } x \text{ over } \{x : x > 2\}) \\ \nabla G^{(k)}(x) = 0 &\Rightarrow \ x_1 - \frac{\varepsilon_k}{x_1 - 2} = 0, \ x_2 = 0 \\ &\Rightarrow \ x^{(k)} = (1 + \sqrt{1 + \varepsilon_k}, 0) \\ \text{as } k \to \infty, \varepsilon_k \to 0 &\text{ and } x^{(k)} \to (2,0) = x^* \end{aligned}$$

9.7 Penalty Method (For ECP)

Computational method for equality constraints.

$$\begin{aligned} \min \quad & f(x) \\ s.t. \ & x \in \& \\ & h_i(x) = 0, \quad i = 1, ..., m \end{aligned}$$

Algorithm

(1) Choose an increasing positive sequence $\{c_k\}$ s.t. $c_k \to \infty$ as $k \to \infty$.

(2) Solve for $x^{(k)}$ to:

$$\min_{x \in \&} \quad f(x) + c_k \|h(x)\|^2$$

$$\text{Note: } \|h(x)\|^2 = \sum_{i=1}^{m} (h_i(x))^2$$

Proposition 13. *Every limit point $\bar{x}$ of $\{x^{(k)}\}$ is a global min of the ECP if $\&$ is closed.*

Proof. Let $\bar{x} = \lim_{k\to\infty, k\in\mathcal{K}} x^{(k)}$

$$\begin{aligned} f^* = \min_{x\in\&, h(x)=0} f(x) &= \min_{x\in\&, h(x)=0} f(x) + c_k\|h(x)\|^2 \\ &\geq \min_{x\in\&} f(x) + c_k\|h(x)\|^2 \\ &= f(x^{(k)}) + c_k\|h(x^{(k)})\|^2 \\ \Rightarrow c_k\|h(x^{(k)})\|^2 \leq f^* - f(x^{(k)}) \end{aligned}$$

By continunity of f, $\lim_{k\to\infty, k\in\mathcal{K}} f(x^{(k)}) = f(\bar{x})$.
Thus, as $k \to \infty$, $k \to \mathcal{K}$, $f^* - f(x^{(k)}) = f^* - f(\bar{x})$ which is finite.
Since $c_k \to \infty$ as $k \to \infty$, $k \to \mathcal{K}$,

$$\lim_{k\to\infty, k\in\mathcal{K}} \|h(x^{(k)})\|^2 = 0$$

By continunity of h,

$$\lim_{k\to\infty, k\in\mathcal{K}} \|h(x^{(k)})\|^2 = \|h(\bar{x})\|^2 = 0 \Rightarrow h(\bar{x}) = 0$$

Now, since & is closed, and $x^{(k)} \in \&$ for all k, $\bar{x} \in \&$ as well.

$$\begin{aligned} &f^* - f(x^{(k)}) \geq c_k\|h(x^{(k)})\|^2 \geq 0 \\ \Rightarrow &f(\bar{x}) = \lim_{k\to\infty, k\in\mathcal{K}} f(x^{(k)}) \leq f^* \end{aligned}$$

Since $\bar{x}$ is feasible ($\bar{x} \in \&$ and $h(\bar{x}) = 0$) and $f(\bar{x}) \leq f^*$, $\Rightarrow$ $\bar{x}$ is a global min of the ECP. □

10 Duality

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in \& \\ & h(x) = 0 \\ & g(x) \leq 0 \end{aligned}$$

$$\begin{aligned} L(x, \lambda, \mu) &= f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) \\ &= f(x) + \lambda^T h(x) + \mu^T g(x) \end{aligned}$$

Dual Function

$D(\lambda, \mu) = \min_{x\in\&} L(x, \lambda, \mu)$ on convex set $G = \{(\lambda, \mu) : \lambda \in \mathbb{R}^m, \mu_j \geq 0, j = 1, ..., r\}$

Proposition 14. *$D(\lambda, \mu)$ is concave on G*

Proof. Let (λ, μ) and $(\tilde{\lambda}, \tilde{\mu}) \in G$.
For $\alpha \in [0, 1]$,

$$\begin{aligned} &D(\alpha\lambda + (1-\alpha)\tilde{\lambda}, \alpha\mu + (1-\alpha)\tilde{\mu}) \\ =&\min_{x\in\&} f(x) + (\alpha\lambda + (1-\alpha)\tilde{\lambda})^T h(x) + (\alpha\mu + (1-\alpha)\tilde{\mu})^T g(x) \\ =&\min_{x\in\&} \alpha[f(x) + \lambda^T h(x) + \mu^T g(x)] + (1-\alpha)[f(x) + \tilde{\lambda}^T h(x) + \tilde{\mu}^T g(x)] \\ \geq&\min_{x\in\&} \alpha[f(x) + \lambda^T h(x) + \mu^T g(x)] + \min_{x\in\&}(1-\alpha)[f(x) + \tilde{\lambda}^T h(x) + \tilde{\mu}^T g(x)] \\ =&\alpha D(\lambda, \mu) + (1-\alpha)D(\tilde{\lambda}, \tilde{\mu}) \end{aligned}$$

□

10.1 Weak Duality Theorem: $\max_{(\lambda,\mu)\in G} D(\lambda,\mu) \le \min_{x\in F} f(x)$

Define the feasibility set $F = \{x : x \in \&, h(x) = 0, g(x) \le 0\}$

Proposition 15.

$$\max_{(\lambda,\mu)\in G} D(\lambda,\mu) \le \min_{x\in F} f(x)$$

Proof. For $(\lambda,\mu) \in G$, $x \in F$

$$L(x,\lambda,\mu) = f(x) + \lambda^T h(x) + \mu^T g(x) \le f(x)$$

$$\begin{aligned} &\Rightarrow \min_{x\in F} L(x,\lambda,\mu) \le f(x), \quad \forall x \in F \\ &\min_{x\in F} L(x,\lambda,\mu) \le \min_{x\in F} f(x) = f^* \end{aligned}$$

Since $F \subseteq \&$,

$$\begin{aligned} &\min_{x\in\&} L(x,\lambda,\mu) \le f^* \\ &\text{i.e.} \quad D(\lambda,\mu) \le f^*, \quad \forall(\lambda,\mu) \in G \\ \Rightarrow &\max_{(\lambda,\mu)\in G} D(\lambda,\mu) \le f^* \end{aligned}$$

□

10.2 Strong Duality Theorem: under some conditions, $\max_{(\lambda,\mu)\in G} D(\lambda,\mu) = \min_{x\in F} f(x)$

Under some conditions, equality holds, i.e.

$$\underbrace{\max_{(\lambda,\mu)\in G} D(\lambda,\mu)}_{\text{dual problem}} = \underbrace{\min_{x\in F} f(x)}_{\text{primal problem}}$$

Proposition 16. *Suppose f is <u>convex</u>, h_i are <u>affine</u>, g_j are <u>convex</u>, and <u>$\& = \mathbb{R}^n$</u>. If x^* is an optimal solution for primal problem, x^* is regular, and (λ^*,μ^*) are corresponding Largrange multipliers, then strong duality holds and (λ^*,μ^*) maximize $D(\lambda,\mu)$.*

$$\max_{(\lambda,\mu)\in G} D(\lambda,\mu) = D(\lambda^*,\mu^*) = f(x^*) = \min_{x\in F} f(x)$$

Proof. Under regularity assumption, using first-order KKT necessary conditions,

$$\begin{aligned} \nabla_x L(x^*,\lambda^*,\mu^*) &= 0 \\ \mu^{*T} g(x^*) &= 0 \end{aligned}$$

Since f is <u>convex</u>, h_i are <u>affine</u>, g_j are <u>convex</u>,
$L(x^*,\lambda^*,\mu^*)$ is convex in x. Thus,

$$\begin{aligned} L(x^*,\lambda^*,\mu^*) &= \min_{x\in\mathbb{R}^n} L(x,\lambda^*,\mu^*) \\ &\le \max_{(\lambda,\mu)\in G} \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \end{aligned}$$

Furthermore,

$$\begin{aligned}
L(x^*, \lambda^*, \mu^*) &= f(x^*) + \lambda^{*T} \underbrace{h(x^*)}_{=0} + \underbrace{\mu^{*T} g(x^*)}_{=0} \\
&= f(x^*) \\
&\geq f(x^*) + \lambda^T h(x^*) + \mu^T g(x^*) \quad \forall (\lambda, \mu) \in G \\
&= L(x^*, \lambda, \mu) \quad \forall (\lambda, \mu) \in G \\
\Rightarrow \quad L(x^*, \lambda^*, \mu^*) &\geq \max_{(\lambda,\mu)\in G} L(x^*, \lambda, \mu) \\
&\geq \min_{x\in\mathbb{R}^n} \max_{(\lambda,\mu)\in G} L(x, \lambda, \mu)
\end{aligned}$$

Hence,

$$\min_{x\in\mathbb{R}^n} \max_{(\lambda,\mu)\in G} L(x, \lambda, \mu) \leq L(x^*, \lambda^*, \mu^*) \leq \max_{(\lambda,\mu)\in G} \min_{x\in\mathbb{R}^n} L(x, \lambda, \mu)$$

Lemma 6. *Consider function $g(y, z)$, $y \in \mathbb{Y}$, $z \in \mathbb{Z}$.*

$$\max_{z\in\mathbb{Z}} \min_{y\in\mathbb{Y}} g(y, z) \leq \min_{y\in\mathbb{Y}} \max_{z\in\mathbb{Z}} g(y, z)$$

Proof. Set $f(z) = \min_{y\in\mathbb{Y}} g(y, z)$

$$\begin{aligned}
f(z) &\leq g(y, z) \quad \forall y \in \mathbb{Y} \\
\Rightarrow \max_{z\in\mathbb{Z}} f(z) &\leq \max_{z\in\mathbb{Z}} g(y, z) \quad \forall y \in \mathbb{Y} \\
\Rightarrow \max_{z\in\mathbb{Z}} f(z) &\leq \min_{y\in\mathbb{Y}} \max_{z\in\mathbb{Z}} g(y, z) \\
\Rightarrow \max_{z\in\mathbb{Z}} \min_{y\in\mathbb{Y}} g(y, z) &\leq \min_{y\in\mathbb{Y}} \max_{z\in\mathbb{Z}} g(y, z)
\end{aligned}$$

□

By the lemma,

$$\min_{x\in\mathbb{R}^n} \max_{(\lambda,\mu)\in G} L(x, \lambda, \mu) = \max_{(\lambda,\mu)\in G} \min_{x\in\mathbb{R}^n} L(x, \lambda, \mu) = L(x^*, \lambda^*, \mu^*)$$

$$\begin{aligned}
\max_{(\lambda,\mu)\in G} L(x, \lambda, \mu) &= \max_{(\lambda,\mu)\in G} f(x) + \lambda^T h(x) + \mu^T g(x) \\
&= \begin{cases} \infty & \text{if } x \notin F \\ f(x) & \text{if } x \in F \end{cases} \\
\Rightarrow \min_{x\in\mathbb{R}^n} \max_{(\lambda,\mu)\in G} L(x, \lambda, \mu) &= \min_{x\in F} f(x)
\end{aligned}$$

Also,

$$\max_{(\lambda,\mu)\in G} \min_{x\in\mathbb{R}^n} L(x, \lambda, \mu) = \max_{(\lambda,\mu)\in G} D(\lambda, \mu)$$

Hence,

$$\max_{(\lambda,\mu)\in G} D(\lambda, \mu) = \min_{x\in F} f(x)$$

Furthermore,

$$\max_{(\lambda,\mu)\in G} D(\lambda, \mu) = L(x^*, \lambda^*, \mu^*) = D(\lambda^*, \mu^*)$$

i.e., (λ^*, μ^*) maximize $D(\lambda, \mu)$ □

> **Note:** If the optimization problem is a linear program and its is feasible, then strong duality holds (always)
> Two ways to prove: 1) Simplex method; 2) Farkas Lemma.

10.2.1 Slater's sufficient condition for strong duality

Proposition 17 (Slater's condition). *If (1) the primal problem is convex (2) it is strictly feasible, that is, there exists x in the relative interior of & such that $g_j(x) < 0, \forall j$ and $h_i(x) = 0, \forall i$. Then, strong duality holds.*

10.2.2 Example

We showed that $x^* = (\frac{3}{2}, \frac{1}{2})$ is the global min with $\mu^* = (1, 0)$ of

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 - 4x_1 - 2x_2 + 2 \\ s.t. \quad & x_1 + x_2 \leq 2 \\ & x_1 + 2x_3 \leq 3 \end{aligned}$$

The optimal value $f(x^*) = -\frac{5}{2}$. What is the dual of the convex program?

$$\begin{aligned} L(x, \mu) &= x_1^2 + x_2^2 - 4x_1 - 2x_2 + 2 + \mu_1(x_1 + x_2 - 2) + \mu_2(x_1 + 2x_2 - 3) \\ &= x_1^2 + x_2^2 + (\mu_1 + \mu_2 - 4)x_1 + (\mu_1 + 2\mu_2 - 2)x_2 + 2 - 2\mu_1 - 3\mu_2 \end{aligned}$$

$D(\mu) = \min_{x \in \mathbb{R}^2} L(x, \mu)$

$$\nabla_x L(x, \mu) = 0 \Rightarrow \begin{cases} 2x_1 + (\mu_1 + \mu_2 - 4) = 0 \\ 2x_2 + (\mu_1 + 2\mu_2 - 2) = 0 \end{cases} \Rightarrow \begin{cases} x_1 &= \frac{-(\mu_1 + \mu_2 - 4)}{2} \\ x_2 &= \frac{-(\mu_1 + 2\mu_2 - 2)}{2} \end{cases}$$

$$D(\mu) = -\left(\frac{\mu_1 + \mu_2 - 4}{2}\right)^2 - \left(\frac{\mu_1 + 2\mu_2 - 2}{2}\right)^2 + 2 - 2\mu_1 - 3\mu_2$$

<u>Dual Problem</u>

$$\max_{\mu_1 \geq 0, \mu_2 \geq 0} \quad D(\mu)$$

$$\nabla D(\mu) = \begin{bmatrix} -\mu_1 - \frac{3}{2}\mu_2 + 1 \\ -\frac{3}{2}\mu_1 - \frac{5}{2}\mu_2 + 1 \end{bmatrix}, \nabla^2 D(\mu) = \begin{bmatrix} -1 & -\frac{3}{2} \\ -\frac{3}{2} & -\frac{5}{2} \end{bmatrix} \prec 0 \Rightarrow D(\mu) \text{ is strictly concave.}$$

We can compute $\mu_1^* = 1$, $\mu_2^* = 0$ (optimum check is omitted), $D(\mu^*) = -\frac{5}{2} = f(x^*)$.

10.3 Dual of Linear Program

LP in "standard" form:

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & Ax \leq b, x \geq 0 \end{aligned}$$

$$\begin{aligned} D(\mu_1, \mu_2) &= \min_{x \in \mathbb{R}^n} c^T x + \mu_1^T (Ax - b) - \mu_2^T x \\ &= \min_{x \in \mathbb{R}^n} (c^T + \mu_1^T A - \mu_2^T) x - \mu_1^T b \\ &= \begin{cases} -\infty & \text{if } c^T + \mu_1^T A - \mu_2^T \neq 0 \\ -\mu_1^T b & \text{if } c^T + \mu_1^T A - \mu_2^T = 0 \end{cases} \end{aligned}$$

Note: $c^T + \mu_1^T A - \mu_2^T = 0 \Leftrightarrow A^T \mu_1 + c = \mu_2$
Hence, the dual probelm is:

$$\begin{aligned} & \max_{s.t.\ \mu_1 \geq 0, \mu_2 \geq 0, A^T \mu_1 + c = \mu_2} \quad -\mu_1^T b \\ = & \min_{s.t.\ \mu_1 \geq 0, A^T \mu_1 + c \geq 0} \mu_1^T b \end{aligned}$$

Hence, the dual probelm is:

$$\min \ \bar{x}^T b$$
$$s.t. \quad -A^T\bar{x} \leq c, \bar{x} \geq 0$$

It is also easy to show that the dual of dual is exactly the primal.

11 Augmented Lagrangian Method (adjusted penalty method)

11.1 Motivation

In penalty method, problem becomes **ill-conditioned** and the optimization becomes super slow if c_k is huge. (c_k is an increasing positive sequence s.t. $c_k \to \infty$ as $k \to \infty$)

Example 18. *When we apply penalty method to*

$$\begin{aligned} &\min \ x^T Q x \\ &s.t. \quad Ax = b \end{aligned}$$

where $Q \succ 0$, $A_{m\times n}, m < n$

$$\min \ x^T Q x + c_k \|Ax - b\|^2$$

where $\|Ax-b\|^2 = (Ax-b)^T(Ax-b) = x^TA^TAx - 2x^TA^Tb + b^Tb$ Note: $A^TA \succeq 0$. Since $m < N$, $rank(A^TA) \leq m < n \Rightarrow \lambda_{\min}(A^TA) = 0$
Consider the problem (P_k) corresponding to c_k

$$\min \ x^TQx + c_k(x^TA^TAx - 2x^TA^Tb + b^Tb)$$

i.e.

$$\min \ x^T(Q + c_kA^TA)x - 2c_kx^TA^Tb + c_kb^Tb$$

Since $Q \succ 0, c_k > 0, A^TA \succeq 0$, $(Q + c_kA^TA) \succ 0$. Hence, (P_k) is strictly convex optimization problem $\Rightarrow$

$$x^{(k)} = \left(\frac{Q}{c_k} + A^TA\right)^{-1} A^Tb$$

If we use gradient descent to solve (P_k), the rate of convergence depends on

$$\textbf{condition number} = \frac{\lambda_{\max}\left(\frac{Q}{c_k} + A^TA\right)}{\lambda_{\min}\left(\frac{Q}{c_k} + A^TA\right)}$$

As $c_k \to \infty$,

$$\lambda_{\min}\left(\frac{Q}{c_k} + A^TA\right) \approx \lambda_{\min}\left(A^TA\right) = 0$$

i.e. optimization probelm (P_k) becomes ill-conditioned as $k \to \infty$

11.2 Augemented Lagrangian Method

$$L_c(x, \lambda) = f(x) + \lambda^T h(x) + c\|h(x)\|^2$$

with $\lambda \in \mathbb{R}^n$, $c_k \to \infty$ as $k \to \infty$.
If $x^{(k)} \in \operatorname{argmin}_x L_{c_k}(x, \lambda)$, then every limit point $\bar{x}$ of $\{x^{(k)}\}$ is a global min for the (P).

What is the advantage of adding $\lambda^T h(x)$?

If x^*, λ^* satisfy the second-order sufficiency condition (for x^* being a strict local min for (P)), then $\exists \bar{c} > 0$ s.t. $\underline{x^* \text{ is a local min of } L_c(x, \lambda^*)}$ if $c \geq \bar{c}$.
(i.e. for some $\gamma > 0$ and $\varepsilon > 0$, $L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}\|x - x^*\|^2$ for all x s.t. $\|x - x^*\| < \varepsilon$)
Therefore, (P) can be restricted to minimum of $L_c(x, \lambda^*)$ if $c > \bar{c}$. We don't need to discuss the situation that $c \to \infty$.
Therefore if λ can be choosen close to λ^*, augmented Lagrangian method can work without $c_k \to \infty$ as $k \to \infty$. $c_k > \bar{c}$ is enough.

11.2.1 Method of Multipliers

Then, the question is <u>how to make λ close to λ^* without knowing λ^*?</u> **Duality!**

$$(P): \ \min f(x) \text{ s.t. } h(x) = 0$$

$$\text{Lagrangian: } L(x, \lambda) = f(x) + \lambda^T h(x)$$
$$\text{Dual: } D(\lambda) = \min_x L(x, \lambda)$$

Let $x(\lambda)$ be a minimizer of $L(x, \lambda)$. Then

$$\nabla_x f(x(\lambda)) + \sum_{i=1}^m \lambda_i \nabla_x h_i(x(\lambda)) = 0$$

and

$$D(\lambda) = f(x(\lambda)) + \sum_{i=1}^m \lambda_i h(x(\lambda))$$

$$\begin{aligned}
\frac{\partial f(x(\lambda))}{\partial \lambda_i} &= \sum_{j=1}^n \frac{\partial f(x(\lambda))}{\partial x_j} \cdot \frac{\partial x_j(\lambda)}{\partial \lambda_i} \\
&= \left[\frac{\partial x_1(\lambda)}{\partial \lambda_i} \cdots \frac{\partial x_n(\lambda)}{\partial \lambda_i} \right] \nabla_x f(x(\lambda))
\end{aligned}$$

Define

$$\nabla_\lambda x(\lambda) = \begin{bmatrix} \frac{\partial x_1(\lambda)}{\partial \lambda_1} & \cdots & \frac{\partial x_n(\lambda)}{\partial \lambda_1} \\ \vdots & \cdots & \vdots \\ \frac{\partial x_1(\lambda)}{\partial \lambda_m} & \cdots & \frac{\partial x_n(\lambda)}{\partial \lambda_m} \end{bmatrix}$$

Then $\nabla_\lambda f(x(\lambda)) = \nabla_\lambda x(\lambda) \nabla_x f(x(\lambda))$, $\nabla_\lambda h_i(x(\lambda)) = \nabla_\lambda x(\lambda) \nabla_x h_i(x(\lambda))$.

$$\begin{aligned}
\nabla_\lambda D(\lambda) &= \nabla_\lambda f(x(\lambda)) + \nabla_\lambda \left(\sum_{i=1}^m \lambda_i h(x(\lambda)) \right) \\
&= \nabla_\lambda x(\lambda) \cdot \nabla_x f(x(\lambda)) + \sum_{i=1}^m \lambda_i \nabla_\lambda x(\lambda) \nabla_x h_i(x(\lambda)) + h(x(\lambda)) \\
&= \nabla_\lambda x(\lambda) \underbrace{\left(\nabla_x f(x(\lambda)) + \sum_{i=1}^m \lambda_i \nabla_x h_i(x(\lambda)) \right)}_{=0 \text{ by optimality of } x(\lambda)} + h(x(\lambda))
\end{aligned}$$

$$\Rightarrow \nabla_\lambda D(\lambda) = h(x(\lambda))$$

$D(\lambda)$ is concave in $\lambda \Rightarrow$ we can use gradient ascent to find λ^*

$$\begin{aligned}\lambda^{(k+1)} &= \lambda^{(k)} + \alpha_k \nabla_\lambda D(\lambda^{(k)}) \\ &= \lambda^{(k)} + \alpha_k h(x(\lambda^{(k)}))\end{aligned}$$

This leads to method of multipliers:

$$\begin{aligned}x^{(k)} &\in \operatorname*{argmin}_x L_{c_k}(x, \lambda^{(k)}) \\ \lambda^{(k+1)} &= \lambda^{(k)} + c_k h(x^{(k)})\end{aligned}$$

can show that method of multipliers converges to min of (P) under certain conditions (don't need to take $c_k \to \infty$)

Example:

$$\begin{aligned}\min \quad & f(x) = \frac{1}{2}(x_1^2 + x_2^2) \\ s.t. \quad & x_1 = 1\end{aligned}$$

Obviously, $x^* = (1, 0)$ and $\lambda^* = -1$.

$$L_c(x, \lambda) = \frac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \frac{c}{2}(x_1 - 1)^2$$

$x^{(k)} \in \operatorname{argmin}_x L_{c_k}(x, \lambda^{(k)}) \Rightarrow x^{(k)} \in \operatorname{argmin}_x \frac{1}{2}({x_1}^2 + {x_2}^2) + \lambda(x_1 - 1) + \frac{c_k}{2}(x_1 - 1)^2$

$$\Rightarrow x^{(k)} = \left(\frac{c_k - \lambda^{(k)}}{c_k + 1}, 0 \right)$$

.

$$\lambda^{(k+1)} = \lambda^{(k)} + c_k h(x^{(k)}) = \lambda^{(k)} + c_k \left(\frac{c_k - \lambda^{(k)}}{c_k + 1} - 1 \right)$$

$$\Rightarrow \lambda^{(k+1)} = \frac{\lambda^{(k)}}{c_k + 1} - \frac{c_k}{c_k + 1}$$

$$\begin{aligned}(\lambda^{(k+1)} - \lambda^*) &= \lambda^{(k+1)} + 1 \\ &= \frac{\lambda^{(k)} - \lambda^*}{c_k + 1}\end{aligned}$$

As long as $c_k \geq \bar{c} > 0, \forall \bar{c} > 0$, $\lambda^{(k)} \to \lambda^*$ linearly since $\frac{1}{\bar{c}+1} < 1$
Thus $\lambda^{(k)} \to \lambda^* \Rightarrow x^{(k)} \to (1, 0) = x^*$

12 Sub-gradient Methods

Gradient descent methods require ∇f exists. What if ∇f doesn't exist at some point?
Recall that when ∇f exists
f is convex on $\&$ $\Leftrightarrow$ $f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y \in \&$ (the inequality is strict for strict convexity)

12.1 Sub-gradient

Definition 13. *For* <u>*convex*</u> f *on* $\mathbb{R}^n$, g *is called a* ***sub-gradient*** *of* f *at* $x \in \mathbb{R}^n$ *if*

$$f(y) \geq f(x) + g^T(y-x), \quad \forall y \in \mathbb{R}^n$$

<u>**Properties of Sub-gradient**</u>

1) Sub-gradient always exist at any point for convex functions.

2) If ∇f exists at a point x for convex f, sub-gradient is <u>unique and $= \nabla f(x)$</u>

3) Some definition for sub-gradient can be applied for non-convex f, but sub-gradient may not exist.

Example 19. $f(x) = |x|, x \in \mathbb{R}$

For $x \neq 0$, ∇f exists and $=$ sub-gradient.
For $x = 0$, any $g \in [-1, 1]$ is a sub-gradient.

Proof.

(1) For $y > 0$, $f(y) = y \geq f(0) + gy = gy, \forall g \in [-1, 1]$

(2) For $y < 0$, $f(y) = -y \geq f(0) + gy = gy, \forall g \in [-1, 1]$

$\square$

12.2 Sub-differential

Definition 14. *Set of all sub-gradient at* x *is called* ***sub-differential*** *at* x*, denoted* $\partial f(x)$.

Example 20. *For* $f(x) = |x|$,

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Example 21. *For* $f(x) = \max\{1, |x| - 1\}$. *(Note:* $f(x)$ *is convex since* $1, |x| - 1$ *are both convex.)*

$$\partial f(x) = \begin{cases} -1 & \text{if } x < -2 \\ [-1, 0] & \text{if } x = -2 \\ 0 & \text{if } -1 < x < 2 \\ [0, 1] & \text{if } x = 2 \\ 1 & \text{if } x > 2 \end{cases}$$

When $x = 2$,

$$f(y) = \max\{1, |y| - 1\} \geq f(2) + g(y-2) = 1 + g(y-2)$$

(1) <u>$y \geq 0$</u>: $0 \geq g(y-2)$ or $0 \geq (g-1)(y-2)$. If $y > 2$, $g \leq 1$; If $y = 2$, $\forall g$; If $0 \leq y < 2$, $g \geq 0$. $\Rightarrow g \in [0, 1]$

(2) <u>$y < 0$</u>: $0 \geq g(y-2)$ or $-y-2 \geq g(y-2)$, i.e. $g \geq 0$ or $g \leq \frac{2+y}{2-y}$ (satisifed by $g \in [0, 1]$)

12.3 More examples

Example 22. $f(x) = \|x\| = \sqrt{x^T x}$

- f is convex (by Triangle Inequality: $\|x\| + \|y\| \geq \|x + y\|$)
- For $x \neq 0$, $\nabla f(x)$ exists and

$$\partial f(x) = \nabla f(x) = \frac{1}{2\sqrt{x^T x}} \cdot 2x = \frac{x}{\|x\|}$$

- If $x = 0$, $\nabla f(x)$ doesn't exist.

Claim 5. $\partial f(0) = \{g \in \mathbb{R}^n : \|g\| \leq 1\}$

Proof. Need to show that for $\|g\| \leq 1$ and $\forall y \in \mathbb{R}^n$,

$$f(y) = \|y\| \geq f(0) + g^T(y - 0) = g^T y$$

But by Cauchy-Schwarz inequality, for $\|g\| \leq 1$,

$$g^T y \leq \|g\|\|y\| \leq \|y\|, \forall y \in \mathbb{R}^n$$

To estabilish the converse, suppose $\|g\| > 1$.
Then, setting $y = \frac{g}{\|g\|} \Rightarrow \|y\| = 1$ but $g^T y = \|g\| > 1 = \|y\|$ $\square$

Example 23. $f(x) = |x_1 - x_2| \leftarrow$ *convex*

If $x_1 > x_2$, $|x_1 - x_2| = x_1 - x_2$, and ∇f exists and $(1, -1)$
If $x_1 < x_2$, $|x_1 - x_2| = x_2 - x_1$, and ∇f exists and $(-1, 1)$

Claim 6. *If* $x_1 = x_2$, $\partial f(x) = \{(a, b) : a = -b, |a| \leq 1\}$

Proof. Suppose $x_1 = x_2 = c$. Then we need to show $\forall y \in \mathbb{R}^2$, (a, b) s.t. $a = -b$, $|a| \leq 1$

$$|y_1 - y_2| \geq f(c, c) + [a\ b] \begin{bmatrix} y_1 - c \\ y_2 - c \end{bmatrix} = ay_1 + by_2 - c(a + b) = a(y_1 - y_2)$$

Since $|a| < 1$, this inequality holds $\forall y \in \mathbb{R}^2$
To show the converse,

1. Suppose $a \neq -b$.

   If $c(a + b) < 0$, setting $y_1 = y_2 = 0$. $\Rightarrow |y_1 - y_2| = 0$, and $ay_1 + by_2 - c(a + b) = -c(a + b) > 0 = |y_1 - y_2|$, above inequality fails to hold.

   If $c(a+b) > 0$, setting $y_1 = y_2 = 2c$. $\Rightarrow |y_1 - y_2| = 0$, and $ay_1 + by_2 - c(a+b) = c(a+b) > 0 = |y_1 - y_2|$, above inequality fails to hold.

   If $c = 0$, setting $y_1 = y_2 = (a + b)$. $\Rightarrow |y_1 - y_2| = 0$, and $ay_1 + by_2 - c(a + b) = (a + b)^2 > 0 = |y_1 - y_2|$, above inequality fails to hold.

2. Suppose $a = -b$ with $|a| > 1$.

   If $a > 1$, setting $y_1 = y_2 + 1$; If $a < -1$, setting $y_1 = y_2 - 1$.

$\square$

12.4 First-order necessary conditions for optimality in terms of subgradient

Proposition 18. *For convex f, $f(x^*) = \min_x f(x) \Leftrightarrow 0 \in \partial f(x^*)$*

Proof. x^* is a minimizer $\Leftrightarrow$ $f(x^*) \leq f(y), \forall y \in \mathbb{R}^n \Leftrightarrow f(x^*) + 0^T(y - x^*) \leq f(y), \forall y \in \mathbb{R}^n \Leftrightarrow 0 \in \partial f(x^*)$ $\square$

12.5 Properties of Subgradients

Let f, f_1, f_2 be convex functions.

(a) **Scaling:** For scalar $a > 0$, $\partial(af) = a\partial f$, i.e., g is a subgradient of f at x if and only if ag is a subgradient of af at x.

(b) **Addition:** If g_1 is a subgradient of f_1 at x, and g_2 is a subgradient of f_2 at x, then $g_1 + g_2$ is subgradient of $f_1 + f_2$ at x.

(c) **Affine Combination:** Let $h(x) = f(Ax + b)$, with A being a square, invertible matrix. Then $\partial h(x) = A^T \partial f(Ax + b)$, i.e., g is a subgradient of f at $Ax + b$ if and only if $A^T g$ is a subgradient of h at x.

12.6 Sub-gradient Descent for Unconstrained Optimization

Assumptions:

(i) f is convex on $\mathbb{R}^n$.

(ii) $f^* = \inf_{x\in\mathbb{R}^n} f(x)$ exists and there exists an x^* s.t. $f(x^*) = f^*$.

(iii) For all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, $\|g\| \leq a$.

Subgradient Descent with constant step-size:

$$x_{k+1} = x_k - \alpha g_k, \quad g_k \in \partial f(x_k)$$

Analysis:

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \alpha g_k - x^*\|^2 \\
&= \|x_k - x^*\|^2 + \alpha^2 \|g_k\|^2 - 2\alpha g_k^T (x_k - x^*) \\
&\leq \|x_k - x^*\|^2 + \alpha^2 a^2 - 2\alpha g_k^T (x_k - x^*)
\end{aligned}$$

By the definition of g_k,

$$f(x_k) + g_k^T (x^* - x_k) \leq f(x^*) = f^*$$

$$\Rightarrow \|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \alpha^2 a^2 + 2\alpha(f^* - f(x_k))$$

$$f(x_k) - f^* \leq \frac{\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha^2 a^2}{2\alpha}$$

Define $f_N^* = \min\{f(x_0), f(x_1), ..., f(x_{N-1})\}$

$$\sum_{k=0}^{N-1} (f(x_k) - f^*) \geq \sum_{k=0}^{N-1} (f_N^* - f^*) = N(f_N^* - f^*)$$

Then,

$$\begin{aligned}
N(f_N^* - f^*) &\leq \sum_{k=0}^{N-1} \frac{\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha^2 a^2}{2\alpha} \\
&= \frac{\|x_0 - x^*\|^2 - \|x_N - x^*\|^2 + N\alpha^2 a^2}{2\alpha} \\
\Rightarrow \quad f_N^* &\leq f^* + \frac{1}{2\alpha N}\|x_0 - x^*\|^2 + \frac{\alpha a^2}{2} \\
\lim_{N\to\infty} f_N^* &\leq f^* + \frac{\alpha a^2}{2}
\end{aligned}$$

For α samll enough and N large enough f_N^* can be made as close to f^* as desired.