535520: Optimization Algorithms

Lecture 3 — Duality and Gradient Descent

Ping-Chun Hsieh (謝秉均)

September 16, 2024

Optimization: 3 Questions to Answer

1. Characterization: Sufficient / necessary conditions of an optimal solution?

(Our focus of part 1 today)

2. Algorithms: Iterative algorithms that find an optimal solution?

(Our focus of part 2 today)

3. Convergence: Do the iterates converge to an optimum? How fast?

This Lecture

1. Weak and Strong Duality

2. Karush-Kuhn-Tucker (KKT) Conditions

3. Gradient Descent

- Reading Material:
  - Chapters 2 & 5 of Dimitri Bertsekas's textbook "Nonlinear Programming"
  - Chapter 5.5-5.6 of Stephen Boyd's textbook "Convex Optimization"
  - Chapter 3 of Jorge Nocedel and Stephen Wright's textbook "Numerical Optimization"
  - Yuxin Chen's lecture note: https://yuxinchen2020.github.io/ele522_optimization/lectures/grad_descent_unconstrained.pdf

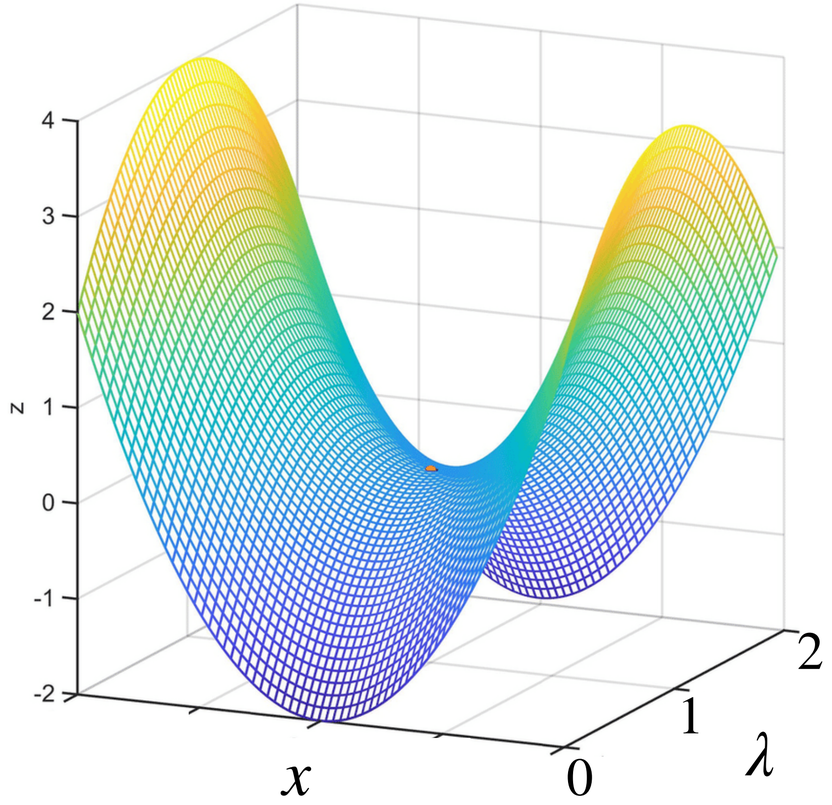
Review: A Natural Question to Ask

- Now we know that $p^* = \inf_{x \in X} \sup_{\lambda \geq 0} \mathscr{L}(x, \lambda)$

A natural question is: Do we have

$$\inf_{x \in X} \sup_{\lambda \geq 0} \mathscr{L}(x, \lambda) = \sup_{\lambda \geq 0} \inf_{x \in X} \mathscr{L}(x, \lambda)?$$

$$=: g(\lambda)$$



Dual function: $g := \mathbb{R}^m \to \mathbb{R}$

$$g(\lambda) := \inf_{x \in X} \mathscr{L}(x, \lambda)$$

Dual problem and dual value:

$$d^* := \sup_{\lambda \geq 0} g(\lambda) = \sup_{\lambda \geq 0} \inf_{x \in X} \mathscr{L}(x, \lambda)$$

1. Duality

Weak Duality

Recall that $p^* = \inf_{x \in X} \sup_{\lambda \geq 0} \mathscr{L}(x, \lambda)$

$$d^* := \sup_{\lambda \geq 0} g(\lambda) = \sup_{\lambda \geq 0} \inf_{x \in X} \mathscr{L}(x, \lambda)$$

Which one is larger?

Weak Duality (Formally)

Theorem (Weak Duality):

Given any primal and dual problem with optimal values p^*, d^*, we always have $p^* \geq d^*$

- **Remark**: One practical implication of weak duality is that if the primal problem is difficult, then we could solve the dual problem to get an "approximate solution" d^*

Geometric Interpretation

For ease of exposition, let's consider

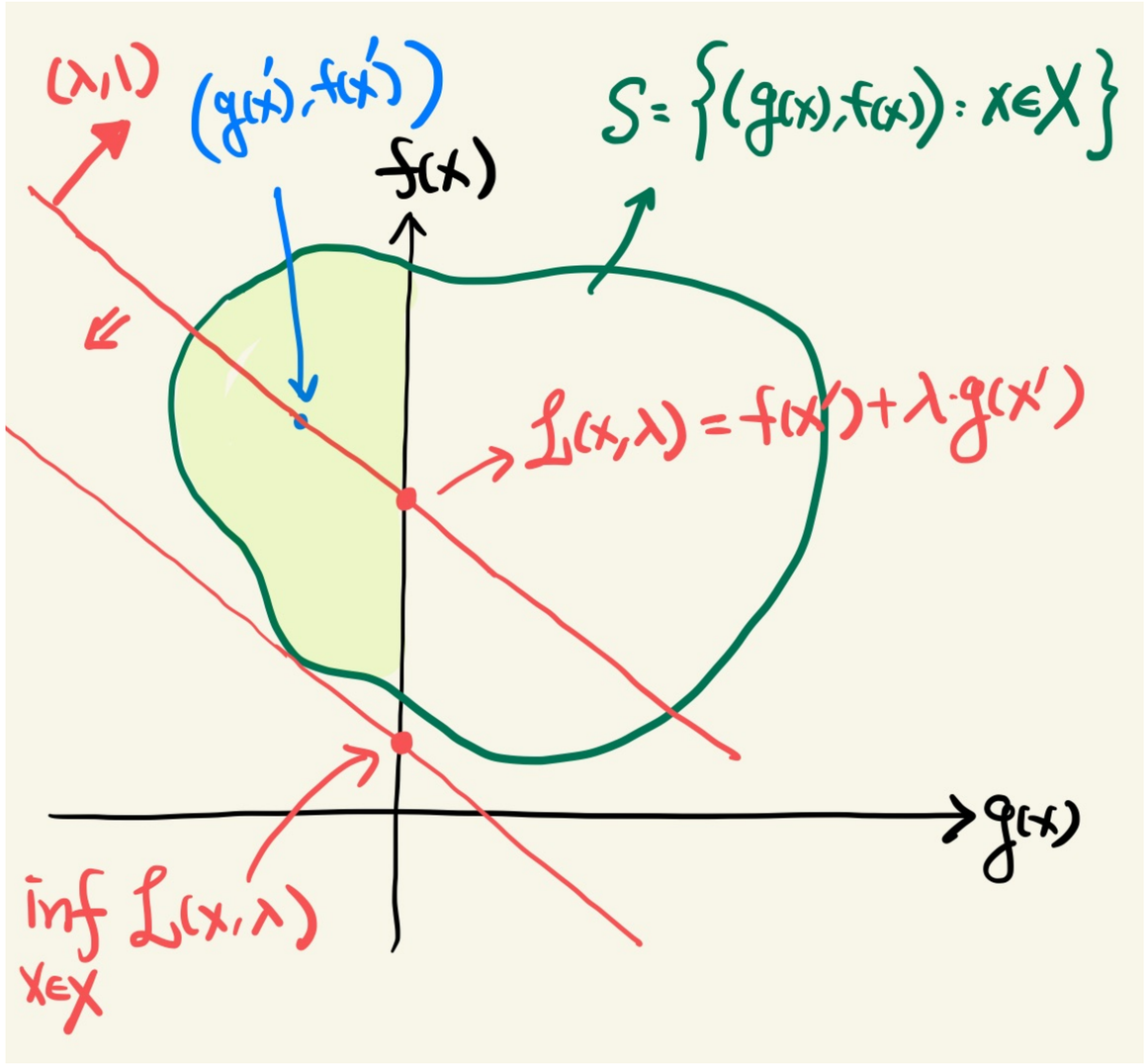
$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g_1(x) \leq 0 \end{aligned}$$

Visualize constraint-cost pair $(g_1(x), f(x))$

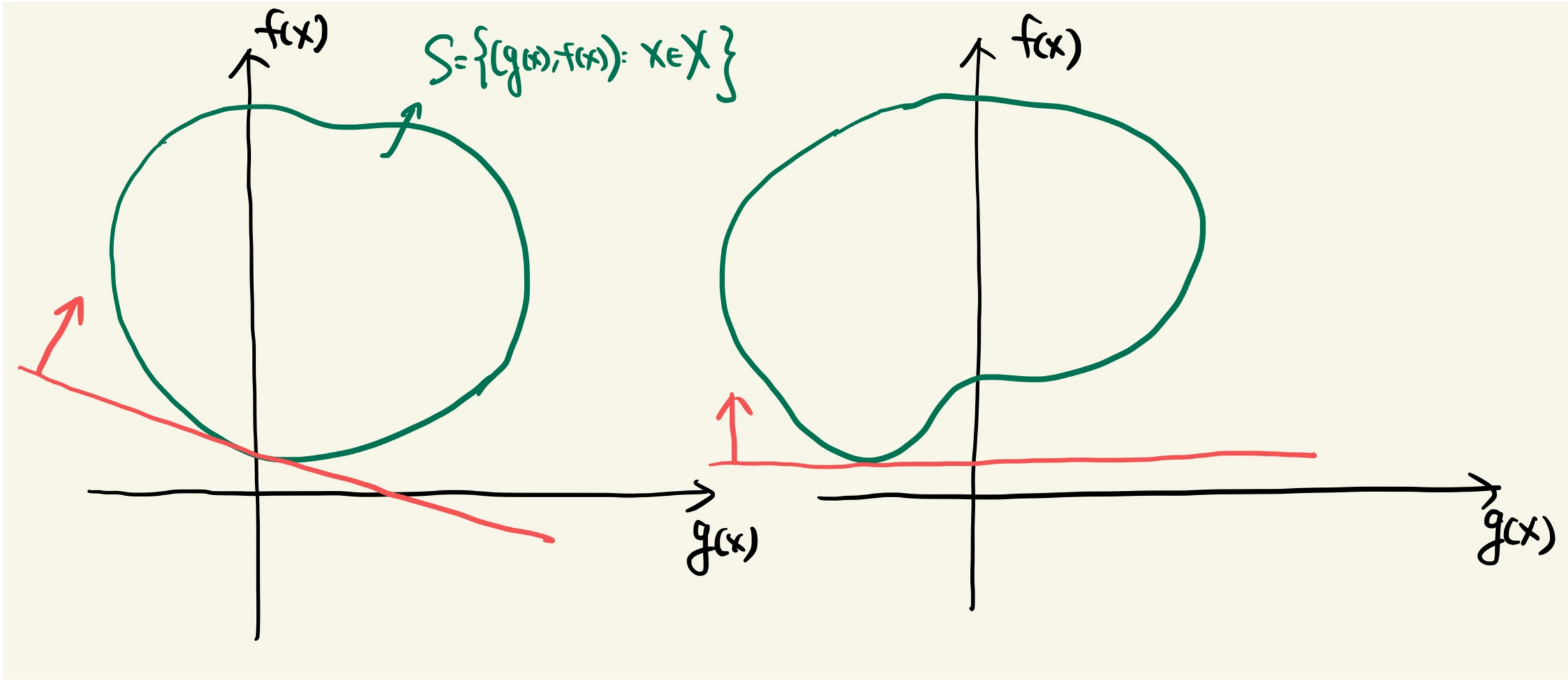
- Under each fixed λ, we have

$$\inf_{x \in X} \mathscr{L}(x, \lambda) = g(\lambda)$$

- To get d^*, we shall find a λ that gives us the maximum $g(\lambda)$



Geometric Interpretation: More Possible Scenarios



Proof of Weak Duality

Theorem (Weak Duality):

Given any primal and dual problem with optimal values p^*, d^*, we always have $p^* \geq d^*$

- Proof:

Step 1: We know $f(x') \geq \mathscr{L}(x', \lambda),$ for all $x' \in X$

Step 2: Therefore, for any $x \in X,$ we have $f(x) \geq \inf_{x' \in X} \mathscr{L}(x', \lambda) = g(\lambda)$

Step 3: By taking "inf" over x, we have $\inf_{x \in X} f(x) \geq g(\lambda)$

Step 4: By taking "sup" over λ, we have $\sup_{\lambda \geq 0} \inf_{x \in X} f(x) \geq \sup_{\lambda \geq 0} g(\lambda)$

Could we have $p^* = q^*$?

(If so, under what conditions?)

Strong Duality and Duality Gap

- Definition (Duality gap): Consider a primal problem (P) with its dual (D). Let p^* and d^* denote the primal optimum and the dual optimum, respectively. Then. The duality gap Δ is defined as

$$\Delta := p^* - d^*$$

- Definition (Strong duality): We say that strong duality holds if $\Delta = 0$.

- There are various sufficient conditions for *strong duality*!
- We will discuss two popular conditions: Slater's and KKT

Slater's Constraint Qualification (CQ)

Consider the following "convex problem"

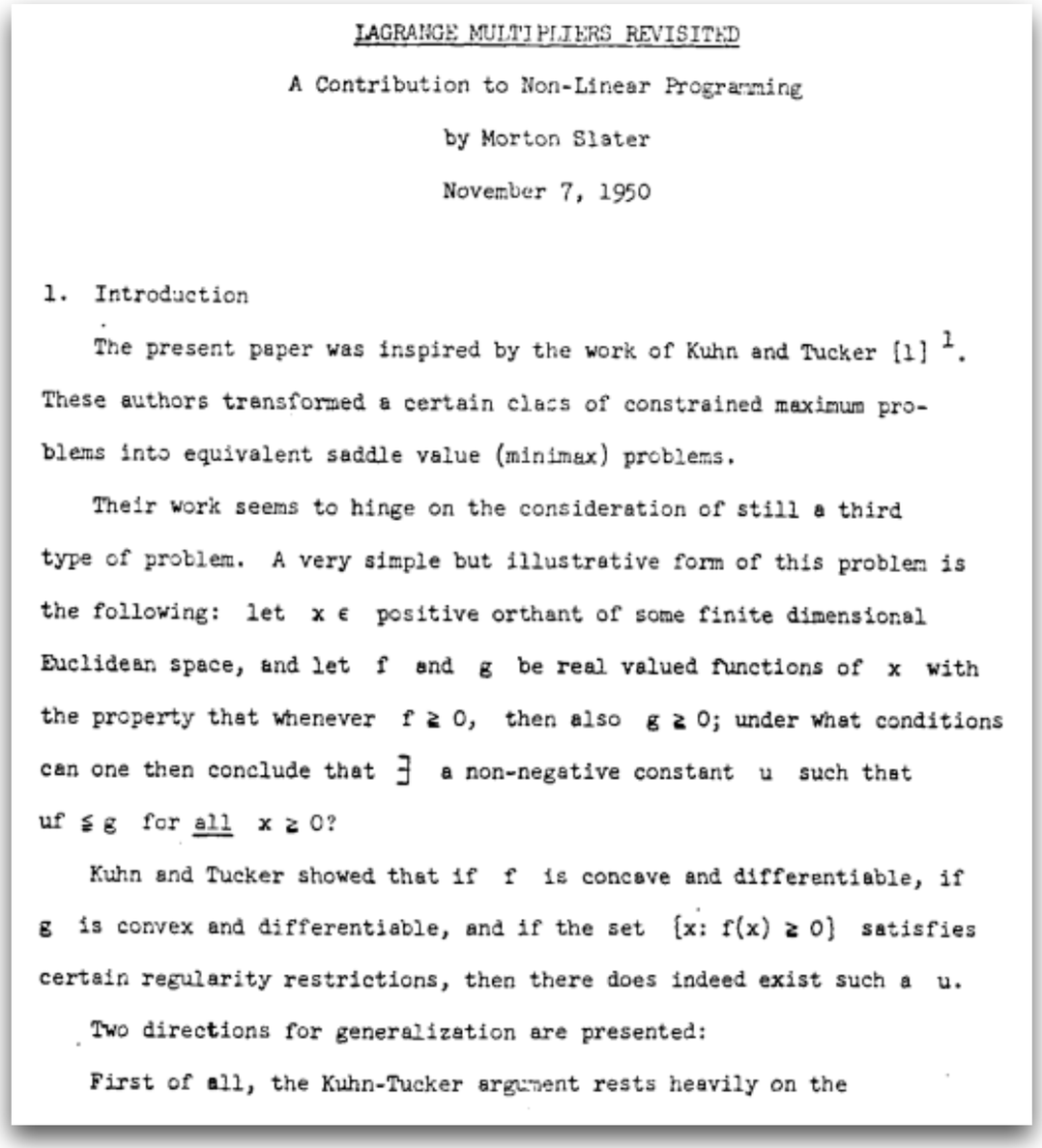
$$\min_{x} \quad f(x) \qquad (f(x) \text{ is a convex function})$$

$$\text{subject to} \quad g_i(x) \leq 0, \forall i \qquad (g_i(x)\text{'s are convex functions})$$

Theorem (Slater's Condition):

Given a convex problem, if *the problem is strictly feasible* (i.e., $g_i(x) < 0$ for all i), then *strong duality* holds.

(The proof can be found in Chapter 5.3.2 of Stephen Boyd's textbook)



LAGRANGE MULTIPLIERS REVISITED

A Contribution to Non-Linear Programming

by Morton Slater

November 7, 1950

1. Introduction

The present paper was inspired by the work of Kuhn and Tucker [1] [1]. These authors transformed a certain class of constrained maximum problems into equivalent saddle value (minimax) problems.

Their work seems to hinge on the consideration of still a third type of problem. A very simple but illustrative form of this problem is the following: let $x \in$ positive orthant of some finite dimensional Euclidean space, and let f and g be real valued functions of x with the property that whenever $f \geq 0$, then also $g \geq 0$; under what conditions can one then conclude that $\exists$ a non-negative constant u such that $uf \leq g$ for all $x \geq 0$?

Kuhn and Tucker showed that if f is concave and differentiable, if g is convex and differentiable, and if the set $\{x: f(x) \geq 0\}$ satisfies certain regularity restrictions, then there does indeed exist such a u.

Two directions for generalization are presented:

First of all, the Kuhn-Tucker argument rests heavily on the

"… Cowles Commission Discussion Papers are preliminary materials circulated privately to stimulate private discussion and are not ready for critical comment for appraisal in publications. …"

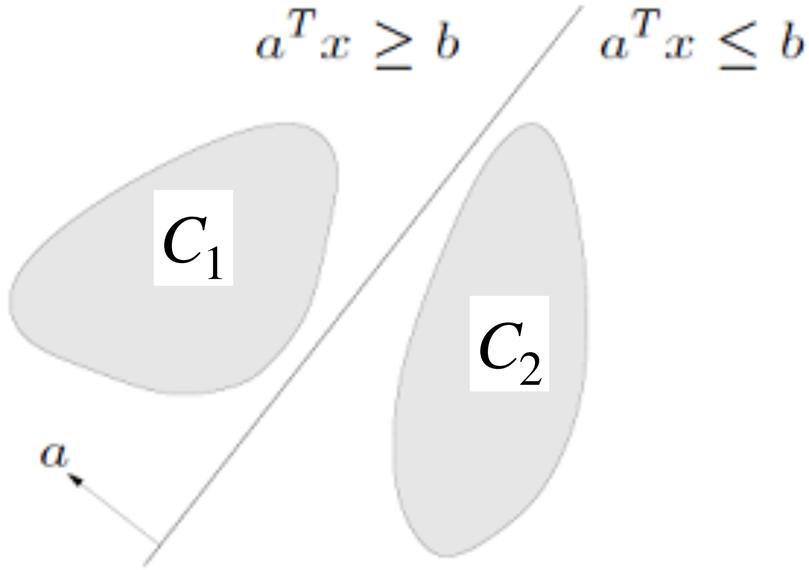
Morton Slater, "Lagrange Multipliers Revisited," *Cowles Commission Discussion Paper No. 403*, 1950

Separating Hyperplane Theorem (SHT)

Separating Hyperplane Theorem:

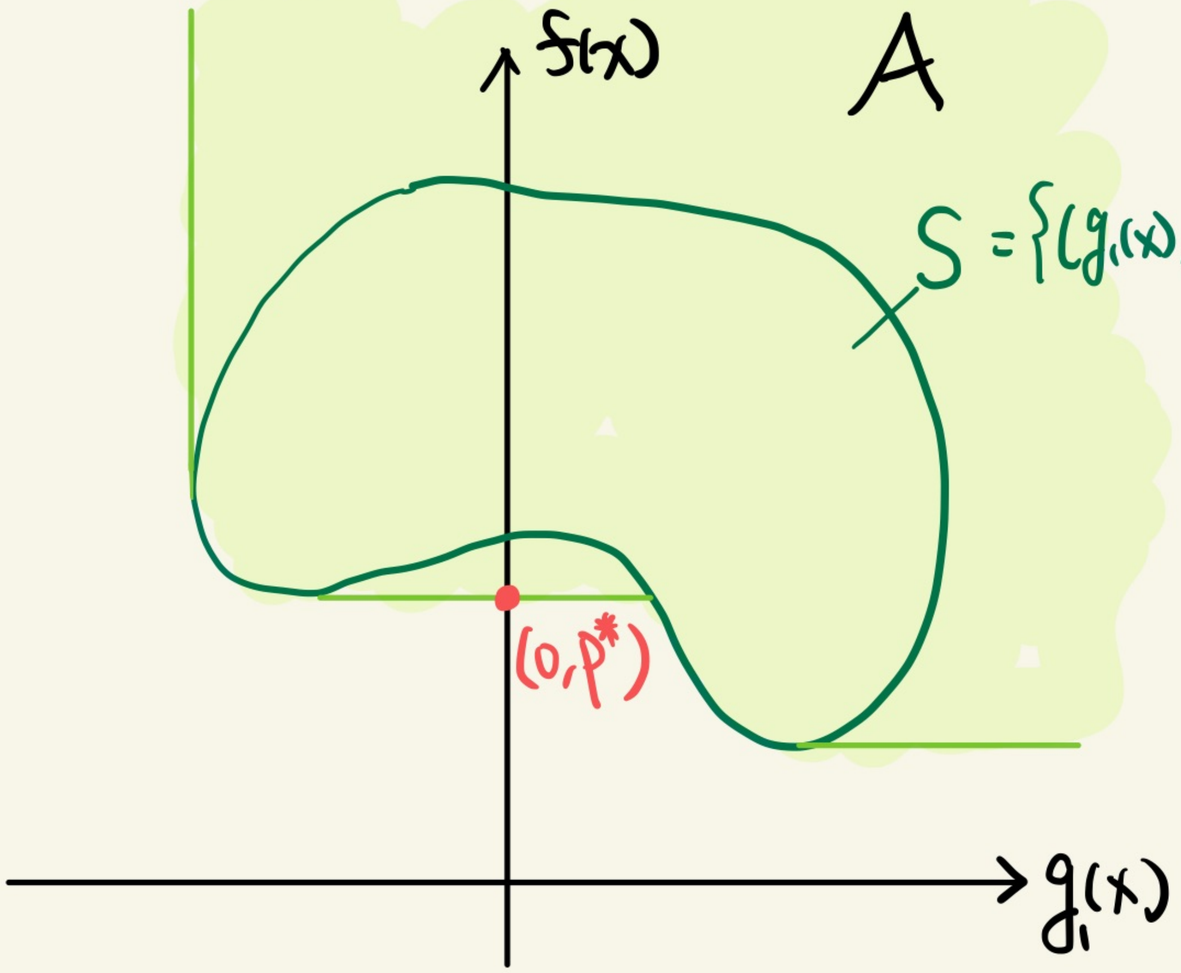
If C_1, C_2 are two "non-empty" and disjoint convex subsets of $\mathbb{R}^n$. There exists a hyperplance that seprates them, i.e., a vector $a \neq 0$ such that

$$a^\top x_1 \leq a^\top x_2, \ \ \forall x_1 \in C_1, x_2 \in C_2$$



(The proof can be found in Chapter 2.5 of Stephen Boyd's textbook)

A Useful Set $\mathcal{A}$



Recall the primal problem: $\min_x f(x)$

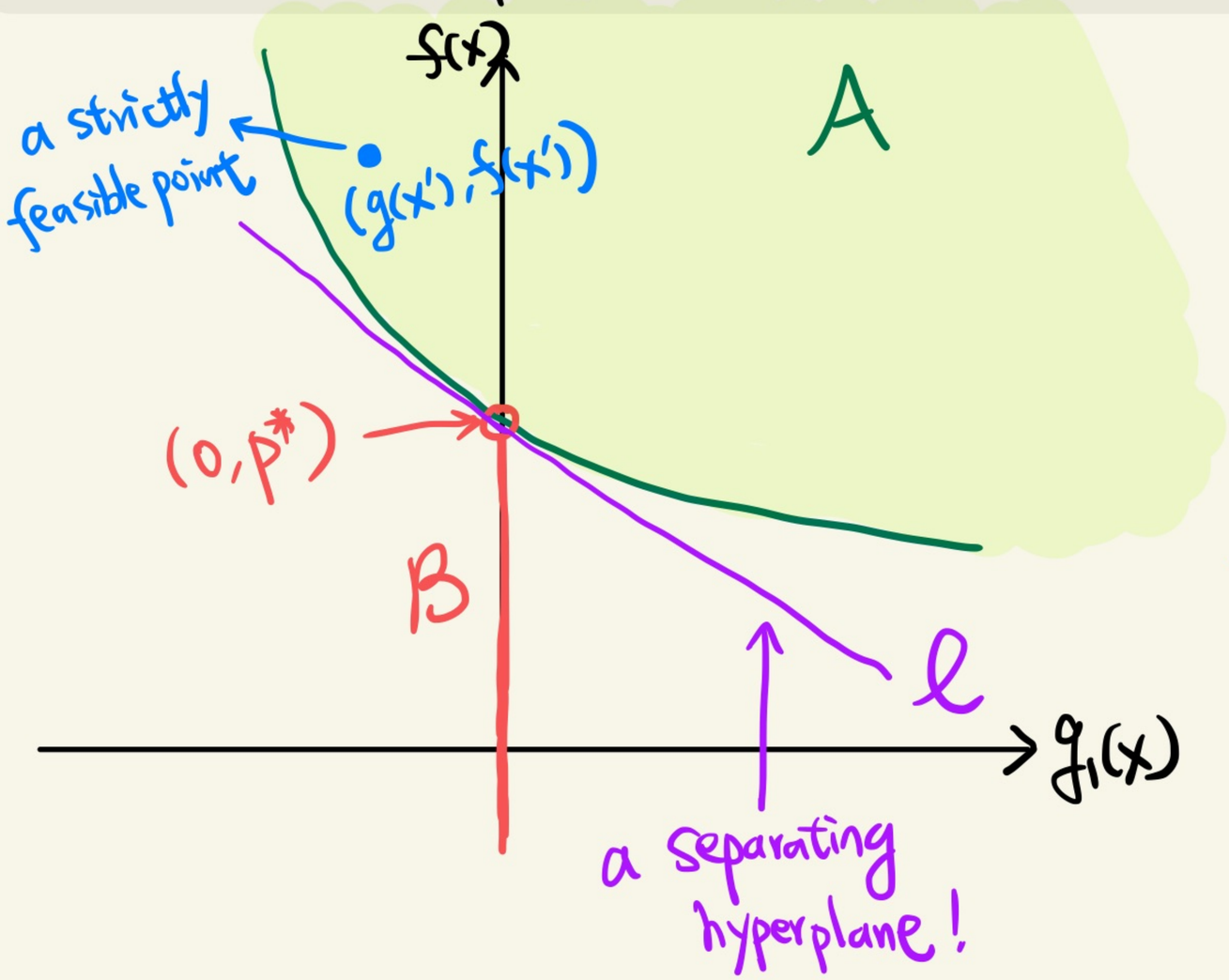
subject to $g_1(x) \leq 0$.

$$\mathcal{A} := \{(\alpha, \beta) : \exists\, x \in X \text{ such that } f(x) \leq \beta,\ g_1(x) \leq \alpha\}$$

Property: $\mathcal{A}$ is a convex set

if f and g_1 are convex function

Geometric Interpretation of Slater's Condition



$$A := \{(\alpha, \beta) : \exists x \in X \text{ such that } g_1(x) \le \alpha, f(x) \le \beta\}$$

$$B := \{(0, \beta) : \beta < p^*\}$$

Question: Are A and B convex sets?

Are A and B disjoint?

- In the 1-constraint case: Slater's condition implies that ℓ is not a vertical line.

2. KKT Conditions

Karush-Kuhn-Tucker (KKT) Optimality Conditions

(1) $g_i(x^*) \leqslant 0$, for all $i = 1, \cdots, m$ (Primal feasibility)

(2) $\lambda_i^* \geqslant 0$, for all $i = 1, \cdots, m$ (Dual feasibility)

(3) $\lambda_i^* g_i(x^*) = 0$, for all $i = 1, \cdots, m$ (Complementary slackness)

(4) $\nabla_x \mathscr{L}'(x, \lambda^*)\Big|_{x=x^*} = 0$ (Lagrangian stationarity)

Theorem (KKT as necessary conditions):

If strong duality holds and (x^*, λ^*) exists, KKT conditions must hold.

- **Remark**: This is true for any differentiable optimization problem

To Derive (3): Lagrangian Optimality Conditions (LOC)

Theorem (LOC):

Suppose strong duality holds and (x^*, λ^*) exists. Then, we have

$$x^* \in \arg\min_{x \in X} \mathscr{L}(x, \lambda^*)$$

- Comparison: Optimality condition FONC-C from Lecture 1

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

- Why is LOC useful?

Proof of Lagrangian Optimality Conditions (LOC)

Want to show: $x^* \in \arg\min_{x \in X} \mathscr{L}(x, \lambda^*)$

$$
\begin{aligned}
p^* &= f(x^*) && \ldots\ldots (\quad) \\
&= d^* && \ldots\ldots (\quad) \\
&= g(\lambda^*) && \ldots\ldots (\quad) \\
&= \min_{x \in X} \mathscr{L}(x, \lambda^*) && \ldots\ldots (\quad) \\
&\leq \mathscr{L}(x^*, \lambda^*) && \ldots\ldots (\quad) \\
&\leq f(x^*) && \ldots\ldots (\quad)
\end{aligned}
$$

From LOC to Complementary Slackness

Theorem (Complementary Slackness):

LOC implies that $\lambda_i^* g_i(x^*) = 0$, for all $i = 1, \cdots, m$

- Why the name “complementary” and “slackness”?

Proof:

Next Question: When are KKT conditions *sufficient*?

Convex problems!

Sufficiency of KKT Optimality Conditions

Theorem (Sufficiency of KKT):

Let f and g_i's be convex functions with a convex domain. If the KKT conditions (1)-(4) hold under $(\bar{x}, \bar{\lambda})$, then:

(i) $\bar{x}$ and $\bar{\lambda}$ are primal and dual optimal solutions

(ii) Strong duality holds

- Is this surprising to you?

Proof: Sufficiency of KKT Optimality Conditions

(1) $g_i(x^*) \leqslant 0$, for all $i = 1, \cdots, m$

(2) $\lambda_i^* \geqslant 0$, for all $i = 1, \cdots, m$

(3) $\lambda_i^* g_i(x^*) = 0$, for all $i = 1, \cdots, m$

(4) $\nabla_x \mathscr{L}'(x, \lambda^*)\Big|_{x=x^*} = 0$

Step 1: By (2), we know $\mathscr{L}(x, \bar{\lambda})$ is convex in x

Step 2: By (4), we know $\bar{x}$ is a minimizer of $\mathscr{L}(x, \bar{\lambda})$

Step 3: Moreover,

$$g(\bar{\lambda}) = \mathscr{L}(\bar{x}, \bar{\lambda}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}) = f(\bar{x})$$

What's the implication here?

Some Remarks

- KKT still are valid necessary conditions for strong duality even for non-convex problems

- If Slater's condition is satisfied, then KKT provide necessary and sufficient conditions for optimality.

- Note — The following statement is NOT true:
  - "For any convex optimization problem, strong duality always holds"

You will find a counterexample in HW0 :)

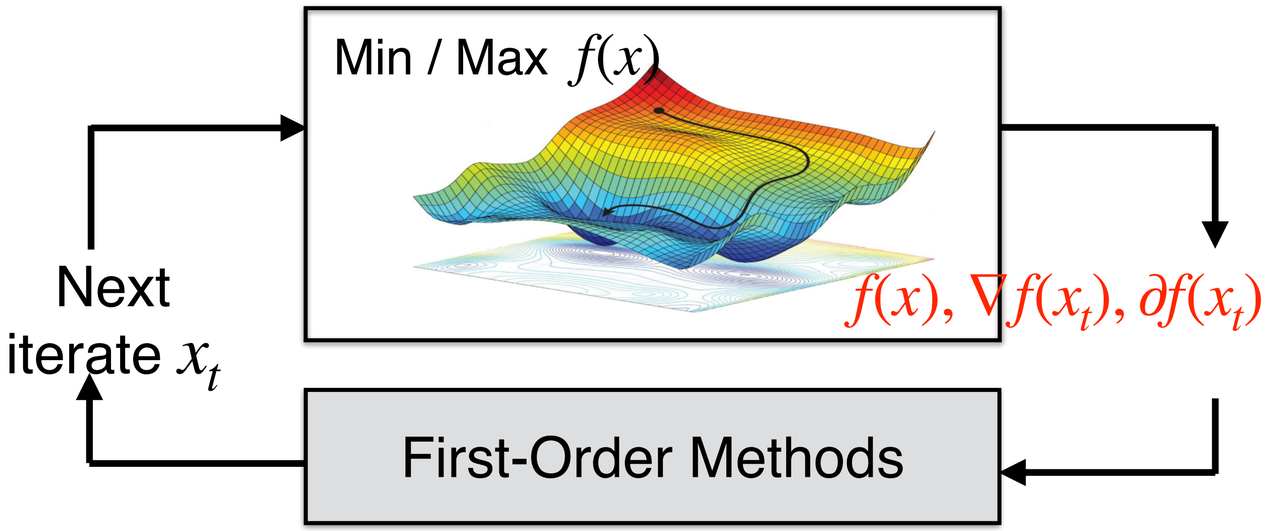
3. Gradient Descent
(For Unconstrained Problems)

First-Order Methods

$$\min_{x \in X} \; f(x)$$

- Question: Numerically, is it easy to find minimizers by "optimality conditions" (e.g., KKT)?

We prefer to solve the problem by "iterative methods"



Examples:

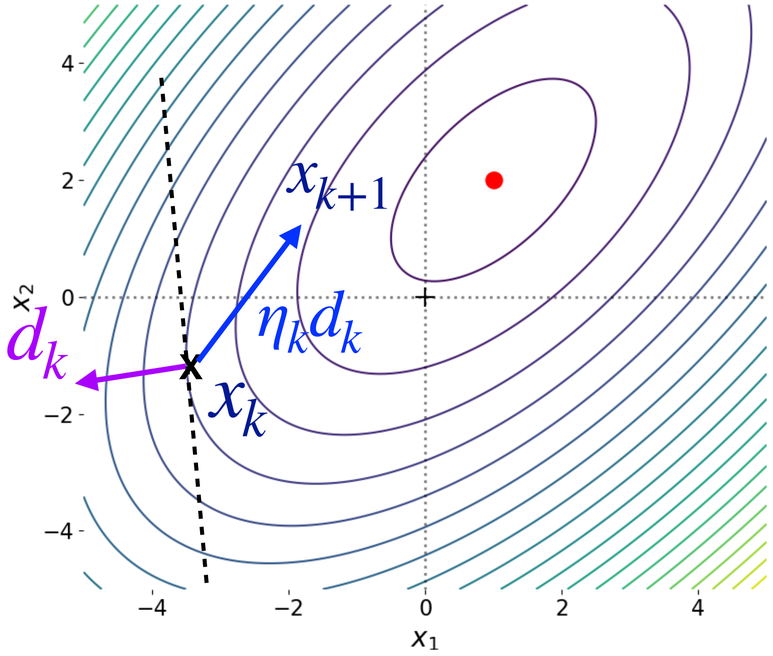
- Gradient descent: $x_{t+1} = x_t - \eta \nabla f(x_t)$
- Subgradient: $x_{t+1} = x_t - \eta g, \; g \in \partial f(x_t)$
- Heavy-ball momentum:

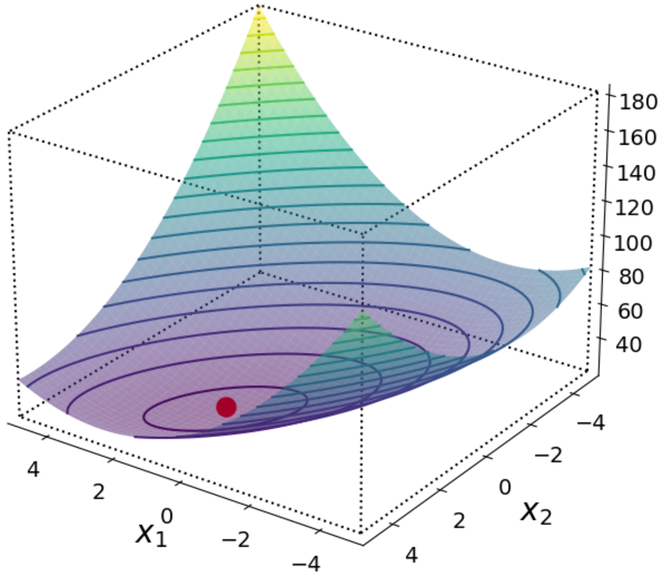
$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

- Quasi-Newton:

$$x_{t+1} = x_t - \eta D_t^{-1} \nabla f(x_t), \quad D_t \approx \nabla^2 f(x_t)$$

Descent Methods





- In the k-th iteration, update x by

$$x_{k+1} = x_k + \eta_k \cdot d_k$$

where $d_k \in \mathbb{R}^n$ is a direction "obtuse to $\nabla f(x_k)$", i.e.,

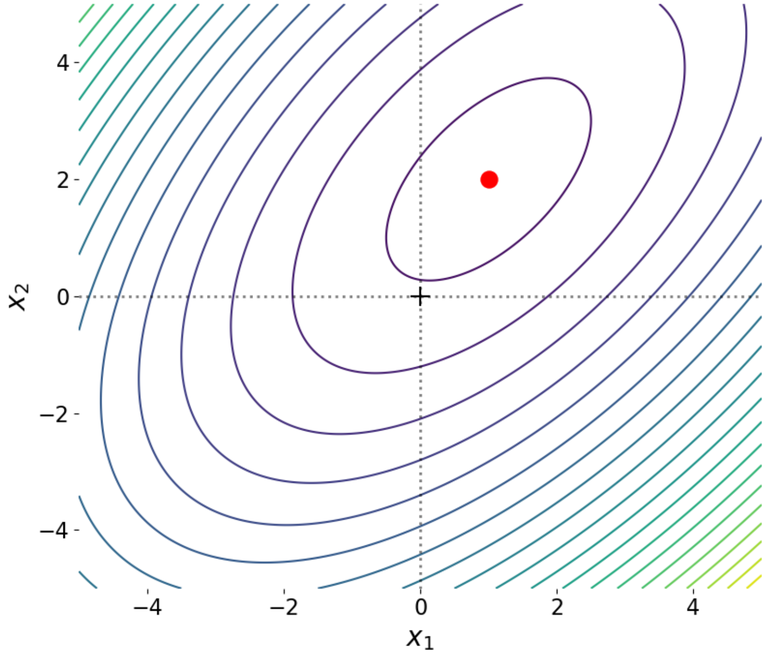
$$\nabla f(x_k)^\top d_k < 0$$

- The vector d_k is a "descent direction" since

$$f(x_{k+1}) = f(x_k) + \eta_k \nabla f(x_k)^\top d_k + o(\|\eta_k\|^2) < f(x_k)$$

(for sufficiently small step sizes)

Connecting "*Descent Methods*" and "*Directional Derivatives*"



Define

$$f'(x; d) := \lim_{\alpha \downarrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha} = \nabla f(x)^\top d$$

This is known as the "*directional derivative*"

Therefore, we know that d is a descent direction if and only if $f'(x; d) < 0$

Let’s start from Gradient Descent (GD) for a simple quadratic problem

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

GD can be traced back to (Augustin-Louis Cauchy, 1847)

Méthode générale pour la résolution des systèmes d'équations simultanées*

M. Augustine Cauchy

Comptes Rendus Hebd. Séances Acad. Sci. 25, 536–538.
OC I, 10 (383), 399–402.

Being given a system of simultaneous equations that the concern is to resolve, one begins ordinarily by reducing them to a single one, by aid of successive eliminations, save to resolve definitely, if it is able, the resulting equation. But it is important to observe, 1° that, in a great number of cases, the elimination is not able to be effected in any manner; 2° that the resulting equation is generally very complicated, even though the given equations are rather simple. For these two motives, one imagines that it would be very useful to understand a general method which may be able to serve to resolve directly a system of simultaneous equations. Such is that which I have obtained, and of which I am going to say some words here. I will limit myself for the moment to indicate the principles on which it is founded, proposing to myself to return with more details on the same subject, in a forthcoming Memoir.

Let first

$$u = f(x,y,z)$$

be a function of many variables $x, y, z, \ldots$ which never become negative and which remain continuous, at least between certain limits. In order to find the values of $x, y, z, \ldots$, which will verify the equation

(1) $$u = 0,$$

it will suffice to make decrease indefinitely the function u, until it vanishes. Now let

$$\mathrm{x}, \mathrm{y}, \mathrm{z}, \ldots$$

be particular values attributed to the variables $x, y, z, \ldots$; u the value corresponding to u; X, Y, Z, ... the values corresponding to $D_x u, D_y u, D_z u, \ldots$, and $\alpha, \beta, \gamma, \ldots$ some very small increments attributed to the particular values x, y, z, ... When one will put

$$x = \mathrm{x} + \alpha, \quad y = \mathrm{y} + \beta, \quad z = \mathrm{z} + \gamma, \ldots,$$

one will have sensibly

(2) $$u = f(\mathrm{x}+\alpha, \mathrm{y}+\beta, \ldots) = \mathrm{u} + \alpha\mathrm{X} + \beta\mathrm{Y} + \gamma\mathrm{Z} + \cdots$$

Augustin-Louis Cauchy, 1789-1857

https://cs.uwaterloo.ca/~y328yu/classics/cauchy-en.pdf

Convegence Rates of GD

	Constant step sizes	Time-varying step sizes
Quadratic problem		
Strongly convex and smooth		
PL condition		
Convex and smooth		

GD for Quadratic Problems

Let's motivate the convergence of GD using a quadratic objective function

$$\min_{x\in\mathbb{R}^n} \quad f(x) := \frac{1}{2}(x-x^*)^\top Q(x-x^*), \; Q \text{ is pd} \quad (\lambda_1(Q) \geq \cdots \geq \lambda_n(Q) > 0)$$

GD update: $x_{k+1} = x_k - \eta_k \nabla f(x_k) =$

Theorem (Convergence Rate of GD for Quadratic Problems):

Under the step size $\eta_t \equiv \eta = 2/(\lambda_1(Q) + \lambda_n(Q))$, then we have

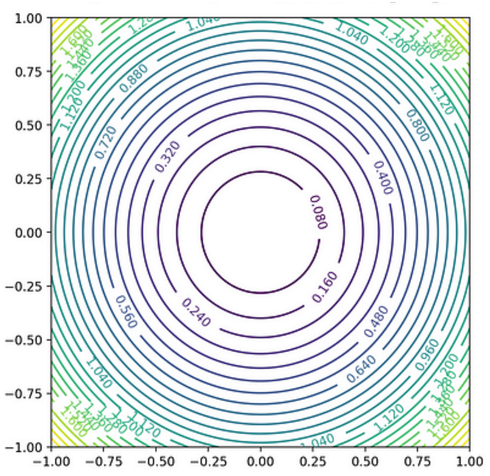
$$\|x_t - x^*\|_2 = \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)}\right)^t \cdot \|x_0 - x^*\|_2, \quad \forall t \in \mathbb{N}$$

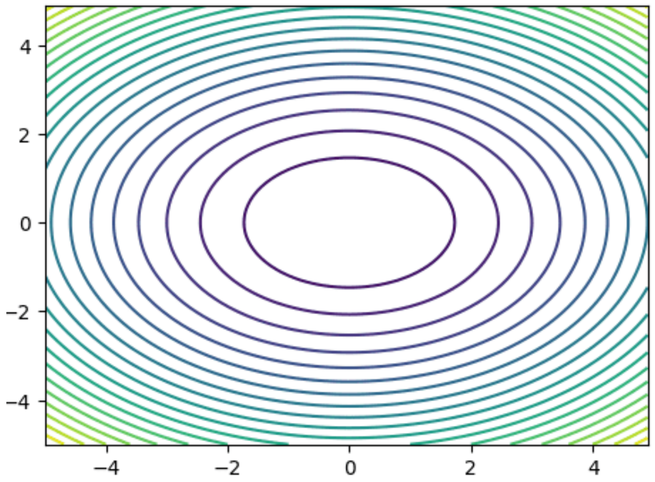
GD for Quadratic Problems (With Constant Step Sizes)

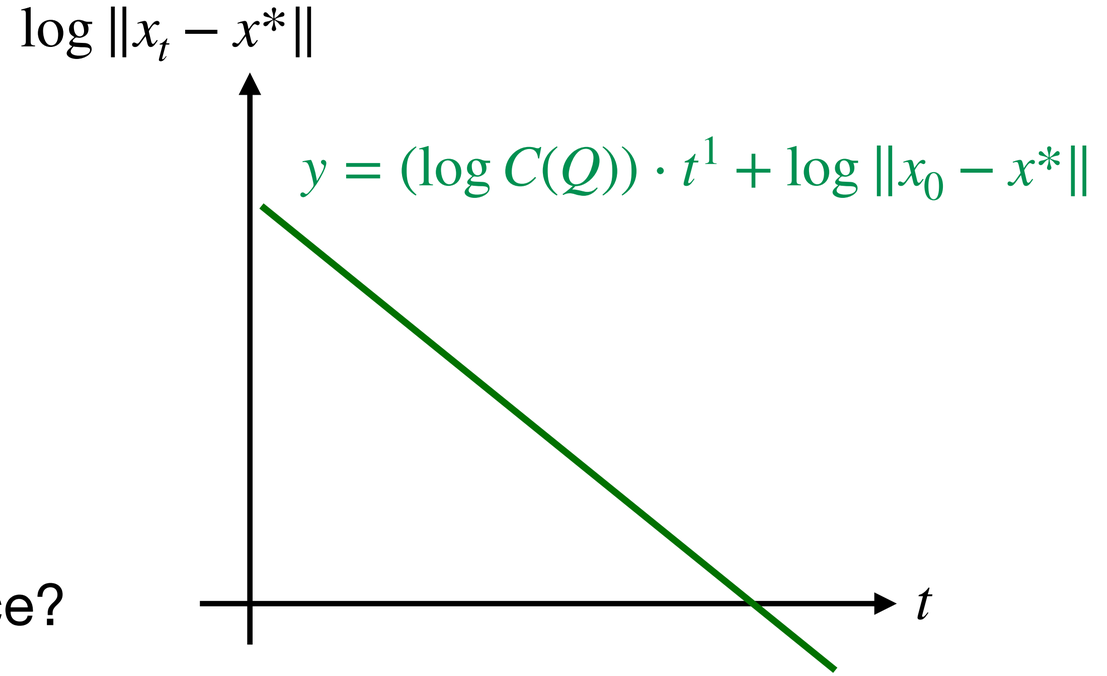
$$\|x_t - x^*\|_2 = \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)}\right)^t \cdot \|x_0 - x^*\|_2, \quad \forall t \in \mathbb{N}$$

$$= \left(\frac{1 - C(Q)}{1 + C(Q)}\right)^t \cdot \|x_0 - x^*\|_2 \qquad \text{where } C(Q) := \frac{\lambda_n(Q)}{\lambda_1(Q)} \text{ is "condition number"}$$

- Question: This is often called "geometric convergence" or "linear convergence"







How about "sub-linear" or "super-linear" convergence?

Terminology of Convergence Rates

- Let $e(x)$ denote the distance from optimality
  - Example: $e(x) = \|x - x^*\|$
  - Example: $e(x) = |f(x) - f(x^*)|$
- Rate of convergence: The limit of the ratio of successive errors

$$\lim_{k \to \infty} \frac{e(x_{k+1})}{e(x_k)} = \beta$$

  - If $\beta = 1$: We call it a sub-linear rate of convergence
  - If $\beta \in (0,1)$: We call it a linear rate of convergence
  - If $\beta = 0$: We call it a super-linear rate of convergence

Proof: Convergence of GD for Quadratic Problems

Step 1: Consider the GD update

$$x_{t+1} - x^* = \left(x_t - \eta \cdot \nabla f(x^t)\right) - x^* = (I_n - \eta Q) \cdot (x_t - x^*)$$

This implies that

$$\|x_{t+1} - x^*\|_2 \leq \|I_n - \eta \cdot Q\| \cdot \|x_t - x^*\|_2 \quad \text{..... (} \qquad \text{)}$$

What norm?

Step 2: Moreover, we have

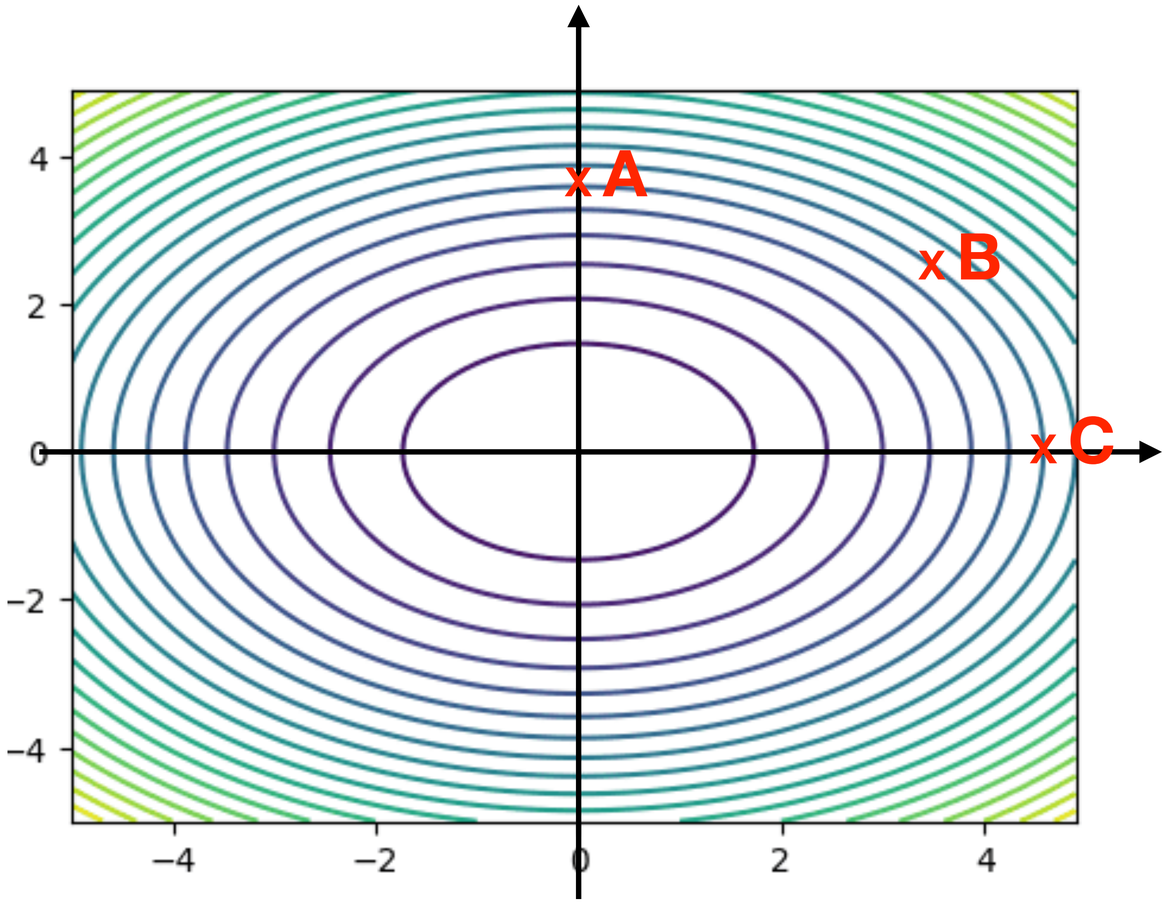
$$\|I_n - \eta Q\| = \max\left\{ |\,1 - \eta\lambda_1(Q)\,|, |\,1 - \eta\lambda_n(Q)\,| \right\}$$

$$= 1 - \frac{2\lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)}$$

Step 3: By repeating Step 1 recursively, we complete the proof.

Observations: GD with Constant Step Sizes

Consider $f(x) := \frac{1}{2} x^\top Q x$ with $Q = [[1,0],[0,10]]$



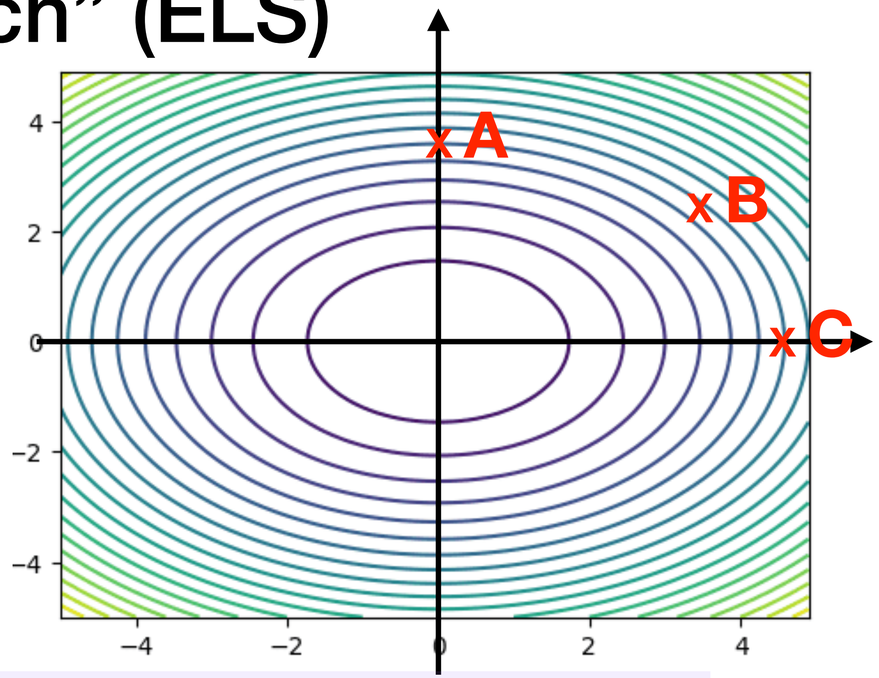
$$\eta = \frac{2}{\lambda_1(Q) + \lambda_n(Q)} = \frac{2}{11}$$

Can you find any special behavior of GD with constant step sizes?

Another Variant: GD with "Exact Line Search" (ELS)

To accelerate GD, we can choose the step sizes by

$$\eta_t = \arg\min_{\eta \geqslant 0} f\left(x_t - \eta \nabla f\left(x_t\right)\right)$$



Theorem (Convergence Rate of GD with ELS):

By applying GD with ELS to the quadractic problems, we have

$$f(x_t) - f(x^*) \leq \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(\theta) + \lambda_n(\theta)}\right)^{2t} \cdot f(x_0) - f(x^*), \quad \forall t \in \mathbb{N}$$

Remark: The convergence rate is actually the same of GD with constant η

Proof: Convergence of GD for Quadratic Problems (1/2)

Notation: $g_t \equiv \nabla f(x_t) = Q(x_t - x^*)$

Step 1: Under ELS, we have

$$\eta_t = \arg\min_{\eta \geqslant 0} f\left(x_t - \eta \nabla f\left(x_t\right)\right) = \frac{g_t^\top g_t}{g_t^\top Q g_t} \quad \text{(Why?)}$$

Step 2: Then, we can find $f(x_{t+1})$ as

$$\begin{aligned} f\left(x_{t+1}\right) &= \frac{1}{2}\left(x_t - \eta_i Q\left(x_t - x^*\right) - x^*\right)^\top Q\left(x_t - \eta_t Q\left(x_t - x^*\right) - x^*\right) \\ &= \frac{1}{2}\left(x_t - x^*\right)^\top Q\left(x_t - x^*\right) - \frac{\|g_t\|^4}{2 g_t^\top Q g_t} \\ &= \underbrace{\frac{1}{2}\left(x_t - x^*\right)^\top Q\left(x_t - x^*\right)}_{f(x_t)} \cdot \left(1 - \frac{\|g_t\|^4}{\left(g_t^\top Q g_t\right) \cdot \left(g_t^\top Q^{-1} \cdot Q \cdot Q^{-1} g\right)}\right) \end{aligned}$$

Proof: Convergence of GD for Quadratic Problems (2/2)

Step 3: To bound (A), we can use the "Kantorovich's inequality"

Lemma (Kantorovich's inequality):

Let Q be a symmetric and pd matrix. Then, for any $y \in \mathbb{R}\backslash\{0\}$,

$$\frac{\|y\|^4}{\left(y^\top Q y\right) \cdot \left(y^\top Q^{-1} y\right)} \geqslant \frac{4\lambda_1(Q) \cdot \lambda_n(Q)}{\left(\lambda_1(Q) + \lambda_n(Q)\right)^2}$$

As a result, we have

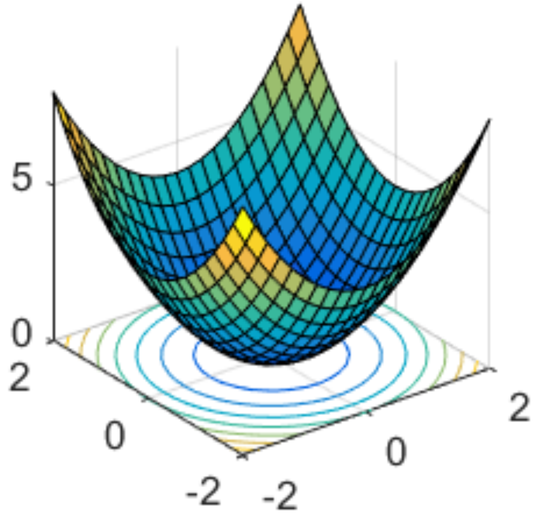
$$f(x_{t+1}) \leq \left(1 - \frac{4 \cdot \lambda_1(Q) \cdot \lambda_n(Q)}{\left(\lambda_1(Q) + \lambda_n(Q)\right)^2}\right) \cdot f\left(x_t\right) = \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)}\right)^2 f\left(x_t\right)$$

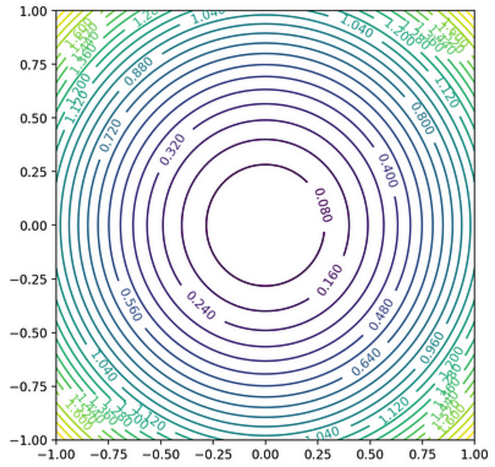
Let’s go beyond quadratic problems:

Strongly-convex and *smooth* problems

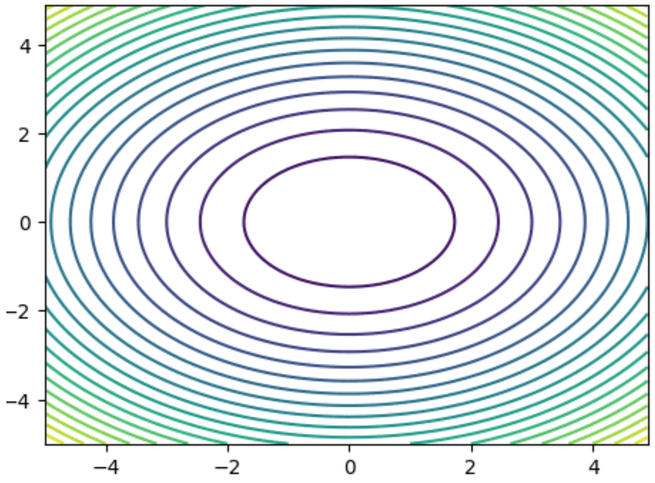
Why Strong Convexity and Smoothness?

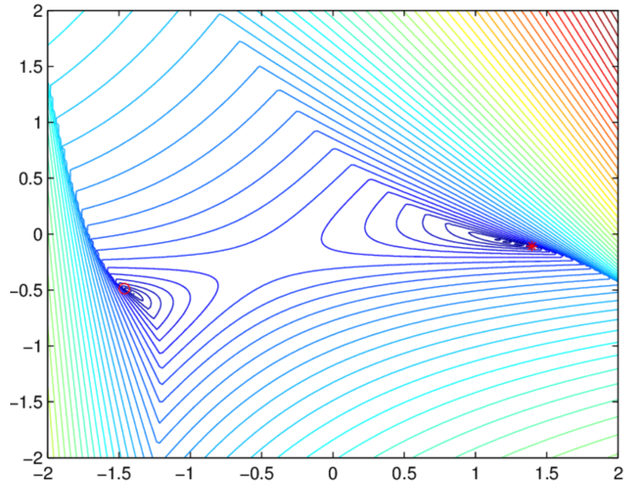
Strong convexity: GD can always attain sufficient per-step improvement





Smoothness: Gradient serves as a useful direction for improvement





Strict Convexity vs Strong Convexity

Definition: A function $f : X \to \mathbb{R}$ is called strictly convex if its domain X is a convex set and for any $x, y \in X$ with $x \neq y$ and any $\alpha \in [0,1]$, we have

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

Intuition: “The segment lies strictly above the function”

Definition: A *continuously differentiable* function $f : X \to \mathbb{R}$ is called μ-strongly convex if its domain X is a convex set and there exists some $\mu > 0$ such that for any $x, y \in X$

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \mu \|x - y\|^2$$

Intuition: 1-dimensional case

An Alternative Definition of Strong Convexity

Theorem 1: Let $f : X \to \mathbb{R}$ be a twice *continuously differentiable* function. Then, the following are equivalent characterization of strong convexity.

(1) There exists some $\mu > 0$ such that for any $x, y \in X$,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2$$

(2) There exists some $\mu > 0$ such that for any $x \in X$,

$$\nabla^2 f(x) - \mu I \succ 0$$

Intuition: Taylor expansion

(Proof: HW1 Problem)

Connecting Strict Convexity and Strong Convexity

Theorem 2: Let $f : X \to \mathbb{R}$ be a *continuously differentiable* function with an open convex domain X. If f is strongly convex, then f is also strictly convex.

Proof: Define $h(t) := f(x + t(y - x)), t \in \mathbb{R}$

Step 1: Consider $t, t' \in [0,1]$ such that $t < t'$

$$\underbrace{\big(\nabla f(x + t'(y - x)) - \nabla f(x + t(y - x))\big)^{\top}\big((t' - t)(y - x)\big)}_{= \left(\frac{dh(t')}{dt} - \frac{dh(t)}{dt}\right)(t' - t)} \geq \alpha(t' - t)^2\|y - x\|^2 > 0$$

Step 2: By Step 1, we know $\frac{dh}{dt}$ is strictly increasing. As a result,

$$\frac{h(t) - h(0)}{t} = \frac{1}{t}\int_0^t \frac{dh(s)}{ds}ds < \frac{1}{1 - t}\int_t^1 \frac{dh(s)}{ds}ds = \frac{h(1) - h(t)}{1 - t} \qquad \text{(Why?)}$$

Step 3: Hence, we have $t \cdot h(1) + (1 - t)h(0) > h(t)$

Lipschitz Smoothness and L-Smoothness

Definition: $f : \mathbb{R}^n \to \mathbb{R}^m$ is called Lipschitz continuous if there exists $L < \infty$ such that for all $x, y \in \mathbb{R}^n$

$$\|f(x) - f(y)\| \leq L\|x - y\|$$

Definition: $f : \mathbb{R}^n \to \mathbb{R}$ is called L-smooth if it has *Lipschitz continuous gradients*, i.e., there exists $L < \infty$ such that for all $x, y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Theorem 3: Let $f : X \to \mathbb{R}$ be *twice differentiable.* Then, f is L-smooth if and only if

$$\nabla^2 f(x) \leq LI$$

Equivalent Characterization of L-Smoothness for Convex Functions

Theorem 4: Let $f : X \to \mathbb{R}$ be a convex and differentiable function. Then, the following are equivalent characterization of L-smoothness:

(1) $f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2$, for all $x, y \in X$

(2) $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$, for all $x, y \in X$

(3) $(\nabla f(x) - \nabla f(y))^\top (y - x) \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2$, for all $x, y \in X$

(For the details, please see Chapter 5.1.2 of Amir Beck's textbook)

In the next few slides, we focus on GD for

μ-strongly convex and *L-smooth* objective functions

Convergence of GD for μ-Strongly Convex and Smooth Functions

Theorem (Convergence of GD under strong convexity and smoothness):

Let f be μ-strongly convex and L-smooth. Under GD with constant step sizes $\eta = 2/(\mu + L)$, we have

$$\| x_t - x^* \| \leqslant \left(\frac{\kappa - 1}{\kappa + 1} \right)^t \cdot \| x_0 - x^* \|^2$$

($\kappa := L/\mu$ is called the condition number)

Comparison:

- Step size: $\dfrac{2}{\mu + L}$ vs $\dfrac{2}{\lambda_1(Q) + \lambda_n(Q)}$
- Contraction: $\dfrac{\kappa - 1}{\kappa + 1}$ vs $\dfrac{1 - C(Q)}{1 + C(Q)}$

Proof: Convergence of GD for μ-Strongly Convex and Smooth Functions

Step 1: Let's rewrite

$$\nabla f(x_t) = \nabla f\left(x_t\right) - \nabla f\underbrace{(x^*)}_{x_t + 1 \cdot (x^* - x_t)} = \left(\int_0^1 \nabla^2 f\left(x_t + s \cdot \left(x^* - x_t\right)\right) \cdot ds \right) \left(x_t - x^*\right)$$

Step 2:

$$\|x_{t+1} - x^*\| = \|x_t - x^* - \eta \cdot \nabla f(x_t)\|$$

$$= \left\| \left(I - \eta \cdot \int_0^1 \nabla^2 f\left(x_t + s\left(x^* - x_t\right)\right) ds \right)(x_t - x^*) \right\|$$

$$= \underbrace{\sup_{0 \le s = 1} \left\| I - \eta \cdot \int_0^1 \nabla^2 f\left(x_t + s \cdot \left(x^* - x_t\right)\right) ds \right\|}_{\le |1 - \eta L|} \cdot \left\| x_t - x^* \right\|$$

Next Question: Do we still get “linear convergence” while relaxing the strong convexity condition?

Polyak-Łojasiewicz (PL) Condition in Non-Convex Optimization

Question: When can GD succeed under non-convex objective functions?

Polyak-Łojasiewicz Condition

Gradient norm Sub-optimality gap

$$\|\nabla f(\theta)\|^2 \geq 2\mu \cdot (f(\theta^*) - f(\theta)) \quad \text{for some } \mu > 0$$

(aka “gradient dominance”)

Boris Polyak

Stanisław Łojasiewicz

Interpretation:

- PL ensures that gradient grows fast as it moves away from the optimum
- PL ensures that every stationary point is a global optimum

Convergence of GD Under PL Condition

Theorem (Convergence of GD under PL and smoothness):

Let f satisfies PL condition and is L-smooth. Under GD with constant step sizes $\eta = 1/L$, we have

$$f\left(x_t\right) - f(x^*) \leqslant \left(1 - \frac{\mu}{L}\right)^t \cdot \left(f\left(x_0\right) - f(x^*)\right)$$

Proof: $f(x_{t+1}) - f(x^*)$

$$\leqslant f(x_t) - f(x^*) - \frac{1}{2L}\|\nabla f(x_t)\|^2 \qquad \text{..... (} \qquad \text{)}$$

$$\leqslant f(x_t) - f(x^*) - \frac{\mu}{L} \cdot (f(x_t) - f(x^*)) \qquad \text{..... (} \qquad \text{)}$$

$$= (1 - \frac{\mu}{L}) \cdot (f(x_t) - f(x^*)) \qquad \text{..... (} \qquad \text{)}$$

Question: Any known problem that satisfies a PL-like condition?

Example 1: Non-Uniform PL in Reinforcement Learning

Non-Uniform PL Condition (Mei et al., ICML 2020)

Gradient norm — Non-uniformity — Sub-optimality gap

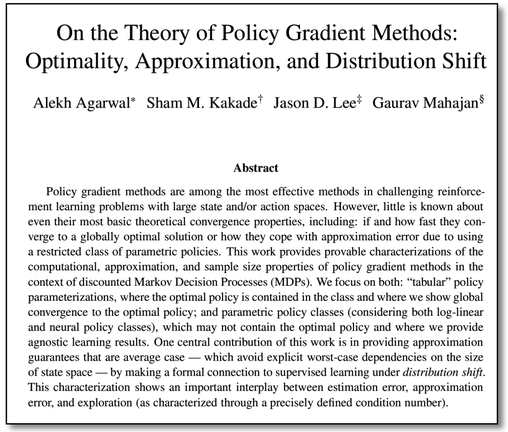
$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\rho^{\pi^*} / d_\mu^{\pi_\theta} \right\|_\infty} \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right] .$$

Nuance:

- Gradient could be extremely small if π_θ is far from an optimal one
- This “non-uniformity” results in complicated convergence analysis

Recent Breakthrough on Policy Gradient Theory in RL

(Agarwal et al., 2019)



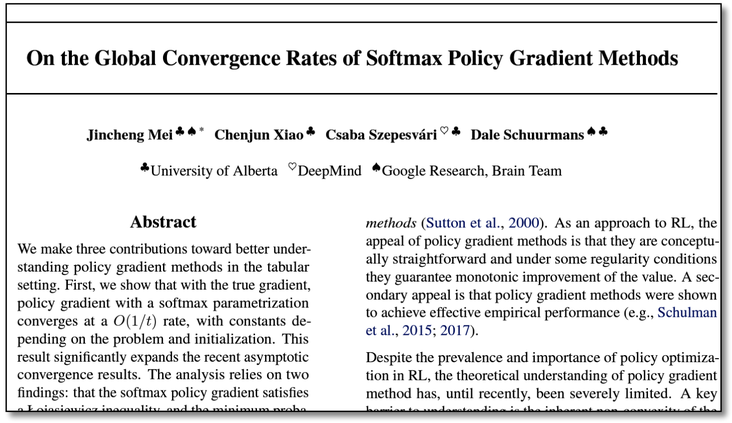
On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift

Alekh Agarwal* Sham M. Kakade† Jason D. Lee‡ Gaurav Mahajan§

Abstract

Policy gradient methods are among the most effective methods in challenging reinforcement learning problems with large state and/or action spaces. However, little is known about even their most basic theoretical convergence properties, including: if and how fast they converge to a globally optimal solution or how they cope with approximation error due to using a restricted class of parametric policies. This work provides provable characterizations of the computational, approximation, and sample size properties of policy gradient methods in the context of discounted Markov Decision Processes (MDPs). We focus on both: "tabular" policy parameterizations, where the optimal policy is contained in the class and where we show global convergence to the optimal policy; and parametric policy classes (considering both log-linear and neural policy classes), which may not contain the optimal policy and where we provide agnostic learning results. One central contribution of this work is in providing approximation guarantees that are average case — which avoid explicit worst-case dependencies on the size of state space — by making a formal connection to supervised learning under *distribution shift*. This characterization shows an important interplay between estimation error, approximation error, and exploration (as characterized through a precisely defined condition number).

(Mei et al., 2020)



On the Global Convergence Rates of Softmax Policy Gradient Methods

Jincheng Mei ♣♠* **Chenjun Xiao** ♣ **Csaba Szepesvári** ♡♣ **Dale Schuurmans** ♠♣

♣University of Alberta ♡DeepMind ♠Google Research, Brain Team

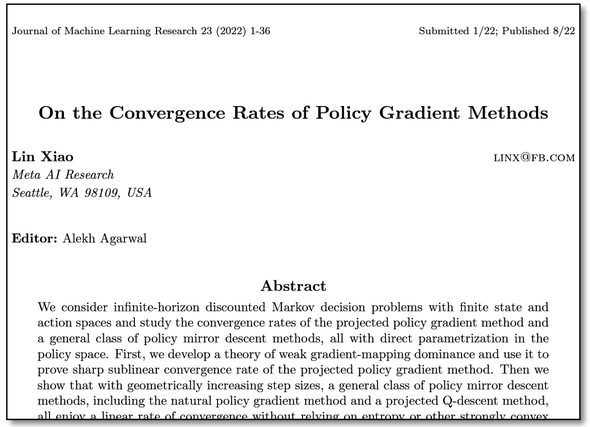
Abstract

We make three contributions toward better understanding policy gradient methods in the tabular setting. First, we show that with the true gradient, policy gradient with a softmax parametrization converges at a $O(1/t)$ rate, with constants depending on the problem and initialization. This result significantly expands the recent asymptotic convergence results. The analysis relies on two findings: that the softmax policy gradient satisfies

methods (Sutton et al., 2000). As an approach to RL, the appeal of policy gradient methods is that they are conceptually straightforward and under some regularity conditions they guarantee monotonic improvement of the value. A secondary appeal is that policy gradient methods were shown to achieve effective empirical performance (e.g., Schulman et al., 2015; 2017).

Despite the prevalence and importance of policy optimization in RL, the theoretical understanding of policy gradient method has, until recently, been severely limited. A key

(Xiao, 2022)



Journal of Machine Learning Research 23 (2022) 1-36 Submitted 1/22; Published 8/22

On the Convergence Rates of Policy Gradient Methods

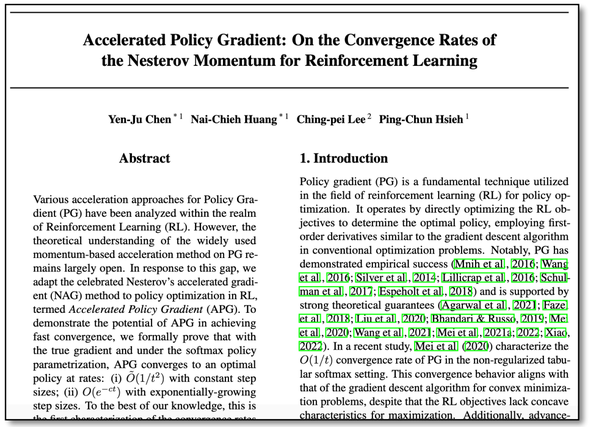
Lin Xiao LINX@FB.COM
Meta AI Research
Seattle, WA 98109, USA

Editor: Alekh Agarwal

Abstract

We consider infinite-horizon discounted Markov decision problems with finite state and action spaces and study the convergence rates of the projected policy gradient method and a general class of policy mirror descent methods, all with direct parametrization in the policy space. First, we develop a theory of weak gradient-mapping dominance and use it to prove sharp sublinear convergence rate of the projected policy gradient method. Then we show that with geometrically increasing step sizes, a general class of policy mirror descent methods, including the natural policy gradient method and a projected Q-descent method, all enjoy a linear rate of convergence without relying on entropy or other strongly convex

(Chen et al., 2024)



Accelerated Policy Gradient: On the Convergence Rates of the Nesterov Momentum for Reinforcement Learning

Yen-Ju Chen [*1] **Nai-Chieh Huang** [*1] **Ching-pei Lee** [2] **Ping-Chun Hsieh** [1]

Abstract

Various acceleration approaches for Policy Gradient (PG) have been analyzed within the realm of Reinforcement Learning (RL). However, the theoretical understanding of the widely used momentum-based acceleration method on PG remains largely open. In response to this gap, we adapt the celebrated Nesterov's accelerated gradient (NAG) method to policy optimization in RL, termed *Accelerated Policy Gradient* (APG). To demonstrate the potential of APG in achieving fast convergence, we formally prove that with the true gradient and under the softmax policy parametrization, APG converges to an optimal policy at rates: (i) $\tilde{O}(1/t^2)$ with constant step sizes; (ii) $O(e^{-ct})$ with exponentially-growing step sizes. To the best of our knowledge, this is

1. Introduction

Policy gradient (PG) is a fundamental technique utilized in the field of reinforcement learning (RL) for policy optimization. It operates by directly optimizing the RL objectives to determine the optimal policy, employing first-order derivatives similar to the gradient descent algorithm in conventional optimization problems. Notably, PG has demonstrated empirical success (Mnih et al., 2016; Wang et al., 2016; Silver et al., 2014; Lillicrap et al., 2016; Schulman et al., 2017; Espeholt et al., 2018) and is supported by strong theoretical guarantees (Agarwal et al., 2021; Fazel et al., 2018; Liu et al., 2020; Bhandari & Russo, 2019; Mei et al., 2020; Wang et al., 2021; Mei et al., 2021a; 2022; Xiao, 2022). In a recent study, Mei et al. (2020) characterize the $O(1/t)$ convergence rate of PG in the non-regularized tabular softmax setting. This convergence behavior aligns with that of the gradient descent algorithm for convex minimization problems, despite that the RL objectives lack concave characteristics for maximization. Additionally, advance-

COLT 2019

Asymptotic convergence to optimum under PG

ICML 2020

The first convergence rate of $O(1/t)$ under PG

JMLR 2022

Similar rates for a large class of PG

ICML 2024

Our recent result: $\tilde{O}(1/t^2)$ under Accelerated PG

Example 2: Overparametrized Linear Regression

Linear regression: Given N data samples $\{a_i \in \mathbb{R}^n, y_i \in \mathbb{R}\}$, find a linear model by minimizing

$$f(x) = \frac{1}{2}\sum_{i=1}^{N}(a_i^\top x - y_i)^2$$

Overparameterization: Model dimension > sample size

(This regime occurs frequently in deep learning)

Remark: This is a convex but not strongly convex problem (why?)

$$\nabla f(x) = \frac{1}{2}\sum_{i=1}^{N} a_i a_i^\top$$

Question: Does $f(x)$ satisfy the PL condition?

Example 2: Overparametrized Linear Regression

Let's show that $\|\nabla f(x)\|_2^2 \geq 2\lambda_{\min}\left(\boldsymbol{A}\boldsymbol{A}^\top\right) f(x)$

Notation: $\boldsymbol{A} = [a_1, \cdots, a_N]^\top$

$$\|\nabla f(x)\|_2^2 = (\boldsymbol{A}x - y)^\top \boldsymbol{A}\boldsymbol{A}^\top (\boldsymbol{A}x - y) \quad \text{..... ( )}$$

$$\geq \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)\|\boldsymbol{A}x - y\|^2 \quad \text{..... ( )}$$

$$= 2\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top) f(x) \quad \text{..... ( )}$$