Large-Scale Convex Optimization

Stadelmann Silvan silvasta@ethz.ch

June 18, 2025

1 Introduction

Large Scale Problem of dimension n but iterations $\ll n$ desired
Convex One of the only problem classes that are "solvable"
Mathematical Optimization

$$\begin{aligned} &\text{minimize} f(x) \\ &\text{s.t.} g_i(x) \le 0, \quad i = 1, \dots, n_g \\ &h_i(x) = 0, \quad i = 1, \dots, n_h \end{aligned} \tag{1}$$

- $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ decision variable (most of our algorithms also work for $n \to \infty$)
- f objectivce function
- $\mathcal{C} = \{\xi \in \mathbb{R}^n : g(\xi) \le 0,\ h(\xi) = 0\}$ fesabile set

1.1 Important Definitions

- $x^\star$ is a *global minimum* if $f(x^*) \le f(x)$
- $x^\star$ is a *local minimum* if there exists $\epsilon > 0$ s.t.

$$f(x^\star) \le f(x) \quad \forall x \in C \cap B_\epsilon(x^\star)$$

$B_\epsilon(x^\star) := \{\xi \in \mathbb{R}^n : |\xi - x^\star| < \epsilon\}$ open ball, center $x^\star$, radius ϵ

1.2 Existance of minimum

1.2.1 Counter examples

a) unbounded level sets, f.e. $1/x$
b) C open f.e. $(0, 1)$ but minimum at f.e. 0
c) f not l.s.c. (lower semi-continuous)

Proposition 1. f (lower-semi-)continuous, $f(x) \to \infty$ for $|x| \to \infty$, $\mathcal{C}$ closed $\Rightarrow \exists$ minimizer of (4) described by: $\min_{x\in\mathcal{C}} f(x)$ and $\operatorname{argmin}_{x\in\mathcal{C}} f(x)$

1.2.2 Examples

x:
- assets in a portfolio
- control inputs
- schedule assignment
- resource allocation

$\mathcal{C}$:
- all possible trade assets
- actuation limits

f:
- cost (negative returns)
- deviaton from target
- waiting times / delas
- risk (a certain resource fails)

1.2.3 First Order Algorithmus

Initialize x_0
for k = 0,...,#iterations -1
$(f(x_k), \nabla f(x_k))$ <- call first-order oracle
Determine x_{k+1} based on $..f..$
end

Definition 1 (Lipschitz continuity)**.** ... $q : R^m \to R^n$ ... if

$$|q(x) - q(y)| \le L|x - y| \forall x, y \in R^m$$

...definition P...

Proposition 2. For any algorithm, there exists a problem in P, such that achieving $|f(xN) - f(x)| <$ requires

$$N \ge (upper(L/2\epsilon))^n - 1$$

Example
(for L=1, ϵ = 0.0005, n=27, N larger than #atoms in universe)

Proof. **Idea** Construct f where $(f(x_0) = 0, \nabla f(x_0) = 0)$, $(f(x_1) = 0, \nabla f(x_1) = 0)$, ... but the actual $min_{x\in C} f(x)$ is small.
Grid(x1,x2)
raster 1/3, 9 boxes in (1,1), for $N \le 7$ (8 steps) one grid cell is not visited
Hence $f(x_i) = 0, i \in [0, 7]$ but $f(x^*) = -L/6$
Generalization
- Partition unit cube into s^n small boxes with side length $1/s$ and $min_x inC = -L/2s$ - therefore $f(x_i) - f(x_star) \ge L/2s$ for $i = 0 \dots s^n - 2$ - roughly ... - therefore $N = \dots$ □

Definition 2. The optimization problem 4 is convex if f and g_i are convex functions, $i = 1, \dots, n_g$, and h is affine.

Definition 3. Function $q : \mathbb{R}^n \to \mathbb{R}$ is convex (affine) if for any $x, y \in \mathbb{R}^n$

$$q(\theta x + (1-\theta) y) \le \theta q(x) + (1-\theta) q(y) \quad \forall \theta \in [0, 1]$$

1.2.4 Software Frameworks

- CVX Python - Yalmip

Proposition 3. $x^\star$ local minimum of (4), if (4) convex, then $x^\star$ global minimum of (4)

Proof. Counter example, $\exists\, y \ne x^\star \in C$ such that $f(y) \le f(x^\star)$ □

1.3 Recitation

LOOK AT SLIDES or FIND r1.md

2 Convex sets and convex functions

Definition 4 (Convex Set)**.** A set $\mathcal{C}$ is convex if and only if $\forall x, y \in \mathcal{C}$ and $\forall \theta \in [0, 1]$: $\quad \theta x + (1-\theta) y \in \mathcal{C}$.

Examples of convex sets:
- hyperplane $\{x \in \mathbb{R}^n \mid a^\mathsf{T} x = b\}$
- half-space $\{x \in \mathbb{R}^n \mid a^\mathsf{T} x \le b\}$
- polyhedron $\{x \in \mathbb{R}^n \mid Ax \preceq b,\ Cx = d\}$ $A \in \mathbb{R}^{q\times n},\ C \in \mathbb{R}^{r\times n},\ b \in \mathbb{R}^q,\ d \in \mathbb{R}^r$
- ...more...

2.1 Operations that preserve convexity (sets)

- **Intersection** $\mathcal{C}_1, \mathcal{C}_2$ convex $\Rightarrow \mathcal{C}_1 \cap \mathcal{C}_2$ convex
- **Image under affine map** $\mathcal{C} \subseteq \mathbb{R}^n$ convex $\Rightarrow \{Ax + b \mid x \in \mathcal{C}\}$ convex
- inverse image of an affine map: ...

2.2 Separating Hyperplane Theorem

Theorem 1. $\mathcal{C} \subseteq \mathbb{R}^n$ non-empty closed convex set, $y \notin \mathcal{C} \to \exists\, a \ne 0, b \in \mathbb{R}$ s. t. $a^\mathsf{T} x + b < a^\mathsf{T} y + b, \forall x \in \mathcal{C}$

Proof. **Claim** $\exists\, \hat{x} \in C$ s.t. $|\hat{x} - y| \le |x - y| \quad \forall x \in \mathcal{C}$
Proof of claim $|x - y|$ has bounded level sets, $\mathcal{C}$ is non-empty and closed $\Rightarrow \exists\, \hat{x} \in \operatorname{argmin}_{x\in\mathcal{C}} |x - y|$
Hyperplane, we choose $a := y - \hat{x}, b := -a^\mathsf{T}\hat{x} = -(y - \hat{x})^\mathsf{T}\hat{x}$
As a result, $a^\mathsf{T} x + b = (y - \hat{x})^\mathsf{T}(x - \hat{x})$ and therefore $a^\mathsf{T} y + b = |y - \hat{x})|^2 > 0$. The following claim shows that the hyperplane $a^\mathsf{T} y + b$ seperates $\mathcal{C}$ and y.
Claim $a^\mathsf{T} y + b \le 0 \quad \forall x \in \mathcal{C}$
Proof of claim Assume not. $\to \exists\, x \in \mathcal{C}$ s.t. $(y - \hat{x})^\mathsf{T}(x - \hat{x}) > 0$
PARAMETRIZE θ
Contradiction $\hat{x}$ nearest point to y
(Details in Lecture notes) □

Corollary 1. A closed convex set $\mathcal{C} = \mathbb{R}^n$ is the intersection of the closed half-spaces that contain $\mathcal{C}$.

Proof. $\mathcal{S}$ intersection of closed half-spaces that contain $\mathcal{C}$
1) $\mathcal{C} \subseteq \mathcal{S} : x \in \mathcal{C} \Rightarrow x$ is contained in every half-spaces that contains $\mathcal{C} \Rightarrow x$ is also contained in the intersections of half-spaces that contains $\mathcal{C}$ $\Rightarrow x \in \mathcal{S}$
2) $\mathcal{S} \subseteq \mathcal{C}$: Assume not $\to \exists\, \hat{x} \in \mathcal{S}$ with $\hat{x} \notin \mathcal{C}$.
By the Seperating Hyperplane Theorem there exists a hyperplane that seperates $\hat{x}$ from $\mathcal{C}$. That means there exists a closed half-space that contains $\mathcal{C}$ but not $\hat{x}$, hence $\hat{x} \notin \mathcal{C}$, contradiction. □

2.3 Support function

Idea represent any closed convex set by its supporting hyperplanes
Support Function: $\sigma_\mathcal{C}(a) = \sup_{x\in\mathcal{C}} a^T x$

CALCULATION EXAMPLE
If we know the $\sigma_\mathcal{C}(a)$, we arrive at at

$$\mathcal{C} = \bigcap_{a\in\mathbb{R}^n} \{x \in \mathbb{R}^n \mid a^\mathsf{T} x - \sigma_c(a) \le 0\}$$

$$= \{x \in \mathbb{R}^n \mid \sup_{a\in\mathbb{R}^n} a^\mathsf{T} x - \sigma_\mathcal{C}(a) \le 0\}$$

Definition 5. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph is a convex set, where

$$\text{epi}(f) := \{(x, t) \in \mathbb{R}^{n+1} | f(x) \le t\}$$

$\to$ this provides a link between convex sets and functions

2.4 Operations that preserve convexity (functions)

- the pointwise maximum of convex functions is convex
- the sum of convex functions is convex
- $f(Ax + b)$ is convex if f is convex

2.4.1 How to check if f is convex?

- if $f : \mathbb{R}^n \to \mathbb{R}$ twice differentiable, $\partial^2 f / \partial x^2 \succeq 0\ \forall\, x \in \mathbb{R}^n$
- if $g : \mathbb{R} \to \mathbb{R}$ with $g(t) = f(x + tv)$ convex in $t\ \forall\, x, v \in \mathbb{R}^n$, then f is convex
- composition of simple convex function with convexity preserving operations

Extended real numbers $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$

Indicator function $\psi_\mathcal{C}(x) := \begin{cases} +\infty & \text{if } x \notin \mathcal{C} \ge 0 \\ 0 & \text{if } x \in \mathcal{C} \end{cases}$

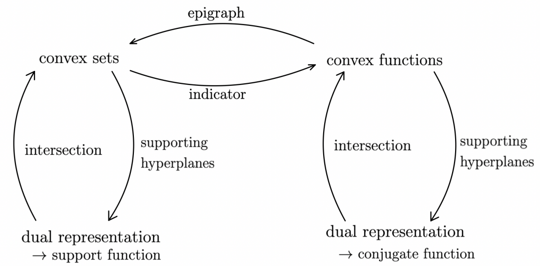
→ this provides another link between convex sets and functions
We can write $\min_{x\in\mathcal{C}} f(x)$ as $\min_{x\in\mathbb{R}^n} f(x) + \psi_{\mathcal{C}}(x)$

Definition 6 (3). $f:\mathbb{R}^n \to \bar{\mathbb{R}}$ is called proper if f is bounded below and if $\exists\, x \in \mathbb{R}^n$ s. t. $f(x) < \infty$

Definition 7 (Legendre Transformation). The conjugate function of $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is defined as $f^\star(y) = \sup_{x\in\mathbb{R}^n} y^\mathsf{T} x - f(x)$

IMAGE F-STAR

2.5 Summary of Concepts



QUESTION
Theorem 2

2.6 Recitation

2.6.1 Convex Sets

A set $\mathcal{C}$ is convex if and only if for all $x, y \in \mathcal{C}$ and $\theta \in [0,1]$:

$$\theta x + (1-\theta)y \in \mathcal{C}$$

2.6.2 Convex Cone

conic combination
Given $x_1, ..., x_n$
any point of the form:
$\theta_1 x_1, ..., \theta_n x_n$
$\theta_i \geq 0$
convex cone
XXX

2.6.3 Positive Semidefinite Cone

Notation
$\mathbb{S}^n$ set of symetric nxn matrices
$\mathbb{S}^n_+$ HHH
$\mathbb{S}^n_{++}$ HHH not convex cone
Example
Sylvester Condition

2.6.4 Convex Functions

Definition

2.6.5 Methods for establishing convexity

1. Verify from definition
2. Second order condition
3. Operations that preserve convexity

2.6.6 Log-Sum-Exp

$$f(x) = log(e_1^x + ... + e_n^x)$$

differentiable approximation of max(x)
How to check convexity?
Second-order condition $\nabla^2 f \geq 0$

2.6.7 Nonnegative Weighted Sum

$\alpha(f_1 + f_2)$ convex if f_1, f_2 convex, $\alpha > 0$
$f_1, \dots, f_m$ convex, $w_1, \dots, w_m \geq 0 \Rightarrow w_1 f_1 + \dots + w_m f_m$ convex

2.6.8 Composition with Affine Function

$$g(x) = f(Ax + b)$$

Examples
Log barrier for linear inequalities → transforms constrained problem in unconstrained
Norm Function

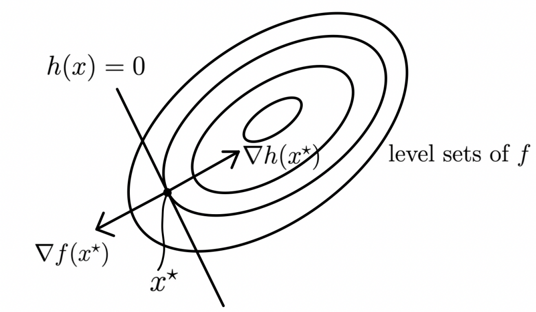
2.6.9 Composition

$$f(x) = h(g(x))$$

3 KKT and Lagrange Duality

3.1 Example

Optimization problem: $\min_{x\in\mathbb{R}^2} f(x)$ s.t. $h(x) = 0$



We note the following: $\nabla f(x^\star)$ and $\nabla h(x^\star)$ are colinear
$\Leftrightarrow \exists\, \nu^\star \in \mathbb{R} : \nabla f(x^\star) + \nu^\star \nabla h(x^\star) = 0$
$f(x) + \nu^\star h(x)$ is stationary at $x^\star$, where $\nu^\star$ can be interpreted as cost of violationg constraint

3.2 Generalization

Generalization to $n \geq 2$ and presence of inequality constraints

$$f^\star = \underset{x\in\mathbb{R}^n}{inf}\, f(x) \text{ s.t. } h(x) = 0,\ g(x) \leq 0 \qquad (2)$$

with corresponding Lagrange function

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^\mathsf{T} g(x) + \nu^\mathsf{T} h(x) \qquad (3)$$

where $\lambda_i \geq 0, \nu_i \in \mathbb{R}$ are the dual variables or multipliers that can be interpreted as cost for violationg constraints.

Proposition 4 (Weak Duality). The dual function $d(\lambda,\nu) = \underset{x\in\mathbb{R}^n}{inf}\, \mathcal{L}(x,\lambda,\nu)$ satisfies
$d(\lambda,\nu) \leq f^\star,\ \forall \lambda \geq 0,\ \nu \in \mathbb{R}^{n_h}$

Proof. SHORT □

Definition 8 (Constraint qualification). $\mathcal{C}$ convex, Slaters Condition holds if $\exists\, \hat{x} \in \mathbb{R}^n$ s.t. $h(\hat{x}) = 0$ and $g(\hat{x}) < 0$

Proposition 5 (Strong Duality). If Slater's condition holds and (2) is convex then $\exists \lambda \geq 0, \nu \in \mathbb{R}^{n_h}$ s.t. $d(\lambda,\nu) = f^\star$

Proof. EXTENDED GRAPHIC □

3.3 KKT

Theorem 2 (KKT Conditions). Slater's condition holds and (2) is convex. Then $x^\star \in \mathbb{R}^n$ is a minimizer of the primal (2) and $(\lambda^\star \geq 0,\ \nu^\star) \in \mathbb{R}^{n_g} \times \mathbb{R}^{n_h}$ is a maximizer of the dual if and only if:

$$KKT-1\ (Stationary\ Lagrangian)$$

$$\nabla_x \mathcal{L}(x^\star, \lambda^\star, \nu^\star) = 0$$

$$KKT-2\ (primal\ feasibility)$$

$$g(x^\star) \leq 0, h(x^\star) = 0$$

$$KKT-3\ (dual\ feasibility)$$

$$\lambda^\star \leq 0, \nu^\star \in \mathbb{R}^{n_h}$$

$$KKT-4\ (complementary\ slackness)$$

$${\lambda^\star}^\mathsf{T} g(x^\star) = 0,\ {\nu^\star}^\mathsf{T} h(x^\star) = 0$$

In addition we have: $INF = SUP$

QUESTION Proof?

Remark Without Slater, KKT 1 to 4 still implies $x^\star$ minimizes (2) and (λ,ν) maximizes the dual, but the converse is no longer true, there can be primal/dual minimizer maximizer that do not satisfy KKT1-4
FORCE BALLANCE

3.4 What if f, g not differentiable?

Example $\underset{x\in\mathbb{R}^n}{inf} |Ax - b|^2 + |x|_1$
where (l_1)-norm not differentiable at 0

3.5 Subdifferential

for convex f...

Definition 9. $f:\mathbb{R}^n \to \bar{\mathbb{R}}$ convex, the subdifferential of f at $\bar{x}$ is: $\partial f(\bar{x}) := \{\lambda \in \mathbb{R}^n \mid f...\}$

Proposition 6. $f : \mathbb{R}^n \to \mathbb{R}$ convex. $x^\star \in argmin...$

Proposition 7 (Relation to conjugate functions). f convex, $epi(f)$ closed: $y \in \partial f(x) \leftrightarrow x \in \delta f^\star(y)$

3.6 Recitation 3

3.6.1 Information ML

3.6.2 Hard Margin SVM

Use hyperplane and support vectors for data classification.

3.6.3 SVM

Find the Maximum-Margin Hyperplane

3.6.4 Solve the Optimization Problem

- Introduce Lagrange multiplier $\alpha_i \geq 0$ for $i = 1, 2, \dots, N$
  ...
- ...
- Solve $\alpha^\star$ by Strong Duality
- Obtain $w^\star$ and $b^\star$ using KKT

3.6.5 Soft Margin SVM

- Introduce some *slackness* ξ
- Point 2

3.6.6 Kernel Methods: Break the linearity

Introduce Nonlinear feature map $\phi(x) : \mathbb{R}^n \to \mathbb{R}^m$
Kernel $K(x_i, x_j) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$

4 Convex Optimization Problem

Recall general optimization Problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{s.t.}\, g_i(x) \le 0, \quad & i=1,\dots,n_g \\ h_i(x)=0, \quad & i=1,\dots,n_h \end{aligned} \tag{4}$$

OPTIMAL VALUE

4.1 Feasibility Problem

$$\begin{aligned} \text{minimize} \quad & s \\ \text{s.t.}\, g_i(x) \le s, \quad & i=1,\dots,n_g \\ h_i(x)=0, \quad & i=1,\dots,n_h \end{aligned} \tag{5}$$

4.2 Linear Programming

$$\text{minimize } c^\top x \quad \text{s.t. } Ax-b\ge 0,\ x\ge 0 \tag{6}$$

Derive dual problem:

Step 1: $\mathcal{L}(x,\lambda_1,\lambda_2) = c^\top x - \lambda_1^\top(Ax-b) - \lambda_2^\top x,\ \lambda_i \ge 0$

Step 2: $\underset{x\in\mathbb{R}^n}{inf}\ \mathcal{L}(x,\lambda_1,\lambda_2) = \begin{cases}\lambda_1^\top b & \text{if } c - A^\top\lambda_1-\lambda_2 = 0\\ -\infty & \text{if } c-A^\top\lambda_1-\lambda_2=0\end{cases}$

Step 3: Dual Problem (again linear programm)

$$\text{maximize } b^\top\lambda \quad \text{s.t. } c-A^\top\lambda\ge 0,\ \lambda\ge 0 \tag{7}$$

4.2.1 Skech

- Polyhedron
- c-vector normal gives 'Levelsets'
- Optimal solution in or trough a corner (if exists)

Proposition 8. The optimal solution of a linear program (if it exists) lies always on the boundary of the feasible set and there exists an optimal solution that is a vertex of the feasible set.

4.2.2 Shortest Path

Analogie with Fluid
Soltuion greater 0, not optimal edges = 0

4.3 Quadratic Programming

minimmize $\frac{1}{2}x^\top Px + q^\top x$ s.t. $Gx\le h,\ Ax=b$

If $P=P^\top$ is positive semi-definite then the problem is convex.

Example [optimal control] (basis for mpc)

4.3.1 Second-order cone program (SOCP)

minimmize $f^\top x$
s.t. $|A_ix+b|\le c_i^\top x + d_i,\ Fx=g$
Cone: Cn+1=

Example [Markovitz portfolio optimization:]

- n number of assets/stocks
- x_i relative value of asset i
- p_i price change of stock i
- p^Tx overall return

Constraints

- $x^T\mathbf{1}=B$, total amount
- $x\ge 0$, no short position

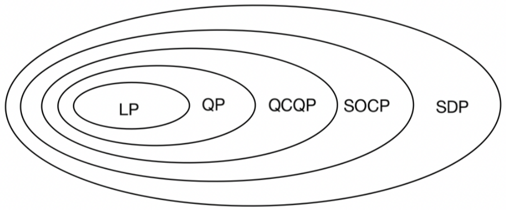
CALCULATIONS

4.4 Semidefinite programming (SDP)

minimmize $c^\top x$
s.t. $x_1F_1,\dots+x_nF_n <= 0$ and $Ax-b=b$
→ the 'standard' form

$$\min_{x\in\mathbb{R}^{n*n}} tr(CX)$$

Problem Classes



4.5 Recitation 4

4.5.1 Geometric Programming

Motivation

- Summary Change of variables, transformation of objectives and constraints

→convex problem in standard form

- Monomial function
- Posynomial function
- Problem formulation
- Example
- Technique

Variable transformation $y_i = \log x_i$ on objective and constraints.

- Transformation

4.5.2 Sum of Squares

- Polynomial Optimization

→ f, g_i, h_i polynomials
General case intractable

- Nonnegative polynomials

Small adaption with γ
find largest γsuch that $f(x)-\gamma$ nonnegative, NP Hard
→ chose γvery high, results in sum of squares

Definition A polynomial $f(x)$ is a sum of squares (SOS), if it can be written as

$$f(x)=\sum_i g_i^2(x) \quad g_i\text{: polynomial}$$

- Verification

$z(x)$ as vector that contains all polynomials of degree $\le d$

Theorem 3 (SOS). $p(x)$ is an SOS if and only if $\exists Q$ such that $Q >= 0$ and $p(x)=z(x)^\top Qz(x)$

Proof
Example

SOS for Lyapunov Stability Analysis

Dynamic
$\dot x_1 = -x_1^3 + x_2$
$\dot x_2 = -x_1 - x_2$
Equilibrium
$x=(x_1,x_2)=(0,0)$
$V(x) = ax1^2 + bx2^2$ vdot = dVf(x) = [2ax1,2bx2]*dynvec
verify vx>0,-vdot>0

5 Gradient methods - Part I

Definition 10 (smoothness). The function $f:\mathbb{R}^n\to\mathbb{R}$ is L-smooth if $\nabla f(x)$ satisfies

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

This result (with Taylors'Theorem) in:

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

Definition 11 (strong convexity). The function $f:\mathbb{R}^n\to\mathbb{R}$ is μ-strongly convex if it satisfies

$$f(y)\ge f(x)+\nabla f(x)^\top(y-x)+\frac{\mu}{2}|x-y|^2 \quad \forall x,y\in\mathbb{R}$$

5.1 Gradient Descent

Given x_0 and stepsize T > 0

$x_{k+1}=x_k - T\nabla f(x_k)$ for $k=(k_0,\dots,k_N)$

HERLEITUNG

Optimal Step Size

$\mu\le h\le L$

$$T^\star = \frac{2}{L+\mu}$$

GRAFIK

Convergence rate

$$\rho(T^\star)=|1-\frac{2L}{L+\mu}|=\frac{L-\mu}{L+\mu}$$

therefore with stepsize $T^\star$
$|x_N-x^\star|\le\epsilon$ if $N\ge\frac{\kappa+1}{2}\ln(\frac{|x_0-x^\star|}{\epsilon})$

5.2 Momentum-based methods

$$\begin{aligned} q_{k+1} &= q_k + T_{p_{k+1}} \\ p_{k+1} &= (1-2dT)p_k - T\nabla f(q_k+\beta p_k)/L\end{aligned}\tag{8}$$

SPRING DAMPER ANALOGY
Nesterovs accelerated gradient methods

- for $T=1, d=\frac{1}{\sqrt{k}+1}, \beta=\frac{\sqrt{k}-1}{\sqrt{k}+1}$

Heavy Ball (tuned quadratics)

- for $T=\frac{2\sqrt{k}}{\sqrt{k}+1}, d=\frac{1}{\sqrt{k}+1}, \beta=0$

What is the convergence rate?

EXAMPLE DIAGONALIZATION
EIGENVALUE analysis
ROOT Locus

- Nesterov on circle $c=(r/0), r=\lambda_i/L=\mu/L$
- Heavy ball circle $c=((\lambda-L)/2,0), r=\lambda+L$

TODO

Theorem 4 (NOT Nesterovs). $f\mu$ strongly convex, L smooth Nesterovs Method satisfies

$$|x_N-x^\star|\le(1-\frac{2}{\sqrt{k}+1})|x_0-x^\star|\forall k\ge 0$$

proof with H Function

5.3 Recital 5 - More on Gradient Descent

5.3.1 Proberties of Smooth Functions

- L-smoothnes:

$$f(y)\le f(x)+\nabla f(x)^\top(y-x)+\frac{L}{2}|x-y|^2 \quad \forall x,y\in$$

5.3.2 Gradien Descent

- Smooth and Convex

xstar argmin f
f is also L-smooth
select $\eta=\frac{1}{2L}$

summing up
- sufficient decrease
- this results in

$$f(x_T) \leq f(x_{T-1}) \leq \dots \leq f(x_1) \leq f(x_0)$$

- As a result, with stepsite $\eta = \frac{1}{2L}$, GD Converges with

$$f(x_T) - minf(x) \leq \frac{2L}{T}||x_0 - x^\star||^2$$

- can do better, nestrov $1/T^2$

5.3.3 Proberties of Strongly-Convex Functions

- μ-strong-convexity: $f(y) \geq ..\mu/2..$
...this implies

5.3.4 Smooth and Strongly-Convex

$\eta = \frac{1}{lL}$ converges with ...

5.3.5 Stepsize

- guess if dont know L
- start with η ca ϵ
- doulbe η until checkable condition does not hold

Line search
- 1-dimensional programming
- find η with optimization for every step
- can result in stepsive $\geq 1/L$ as it is for normal GD

Line search for Heavy Ball Method
- works also for quadratics
- conjugate GD, orthogonalize?

Adaptive Methods
- normalized GD
- AdaGrad-Norm (Adaptive Gradien estimation)
- AdaM (Adaptive Momentum estimation)
- AdamW

6 Gradient Descent - Part II

Projected gradient descent(smooth,strongly convex f)

Definition 12. $prox_{\mathcal{C}}(x) = argmin 1/2|x-y|^2$
C closed convex
CAUCHY SCHWARZ
This implies: $|prox_{\mathcal{C}}(x) - prox_{\mathcal{C}}(y)| \leq |x-y|$

Proof Other Information. TODO □

Algorithm

Proposition 9. satisfaction of GD

Proof. Restricted on quadtratic functions: $\frac{1}{2}x^T Hx + b^{\mathsf{T}}x + c$ □

when are projetions computationally cheap?
– norm ball
- probability simplex

What if f is not strongly convex? ($\mu = 0$)
$\rightarrow$idea: apply small amount of regulairzation
$f : \mathbb{R}^n \rightarrow \mathbb{R}$ L-smooth, convex

$$\hat{f}(x) = f(x) + \frac{\mu}{2}|x - x_0|^2$$

and
XXX (IEQ 1 2)
are satisfied $\forall x_0 \, in \mathbb{R}^n, \mu > 0,$ where...x(hat)star argmin f(hat)

Proof. XXX □

hence we can apply GD or Nesterov
calc
For Nesterov: ... $e^{sqrt\frac{\mu}{L+\mu}}$...
sqrt ESSENCE of morning
chose .. $\frac{2\ln(N)}{N}$
BOX Hence if f smooth and (not strongly) convex we need aproximately N tilde $L|x^\star - x_0|^2/epsilon$ iterations to reach $f(x_N) - f(x_0) \leq \epsilon$

What if f is non-smooth?
i.e. L_f Lipschitz but nor neccessairly differentiable
Example $f(x) = |x|$
Leads to osciliations with $\nabla f = \{+1 \mid -1\}$

Proposition 10 (Subgradient Method)**.** Closed, convex set $\mathcal{C}$ contained in ball of $r = R$
Consider update rule: $x_{k+1} = prox_{\mathcal{C}}(x_k - Tg_k), \dots$ then $x_0, \dots$

Proof. NOT SHOWED □

TABLE
GRAPH with rates, IMPORTANTe

6.1 Recitation

TODO

7 Stochastic gradient descent

MOTIVATION EXAMPLE
- Regression: $tily = \phi(tilx_i\theta) + \epsilon$
ϕfunction approximation with parameter θ
- Data points; $(x_1, y_1) \dots m$
- Minimize: SOME LS
- Gradient: $-\frac{1}{m}\sum_{i=1}^m (y_i - \phi(x_i, \theta))DTF$
$\rightarrow$ computationally intractable if m is large

Goal Obtain approximated solution quickly
$\Rightarrow$ Compute Stochastic gradient
$-(y_i - \phi(x_i, \theta))DTF, \; i \in Unif?(\{1, ..., m\})$
$\Rightarrow$ the gradient is **unbiased**
More generally we consider
$\min_{x \in \mathbb{R}^n} F(x) = \min_{x \in \mathbb{R}^n} \mathbb{E}[f(x, \xi)]$
Where ξ is a continuous or discrere Random Variable.
ALGORITHM Stochastic gradient descent
Step 1: $\xi_k \leftarrow$ generate realization of ξ
Step 2: $x_{k+1} = x_k - T_k g(x_k, \xi_k)$ with T_k step sice
Stochastic gradient $g(x_k, \xi_k)$ examples:
$\nabla_x f(x\bar{\xi}), \; \bar{\xi} \; til \; p_\epsilon$ or somesume
$\Rightarrow$ The iterate x_k is now a random variable!

7.1 Assumptions on $F(x)$ and $g(x_k, \xi_k)$

A1
A2
A3

Proposition 11. F is μ-strongly convex and L-smooth with stepsize

$$0 < T < \frac{1}{L(M_v + 1)}$$

satisfies

$$\mathbb{E}[F(x_k)] - F(x^\star) \leq XXX$$

With T = $\frac{ln(N)}{\mu N}$ we require about

$$N\,()/\epsilon$$

iterations to ensure $\mathbb{E}[F(x_k)] - F(x^\star) \leq XXX \leq \epsilon$

Proof with most important SGD-EQ. XXX

$$\mathbb{E}[F(x_{k+1}) \mid x_k] \leq F(x_k) - T|\nabla F(x_k)|^2 + XXX$$

(1)
Strong convexity implies:

$$F(x) \leq F(x^\star) + \frac{1}{2\mu}|\nabla F(x)| \quad \forall x \in \mathbb{R}^n$$

from there we can conclude:
XXX □

Remarks
$(1 - T\mu)^N \leq e^{-T\mu N}$ this in EQ
- $T_k = \frac{\ln(N)}{N}$ then E[XXX]<=
- $\sum_{k=0}^\infty T_k = \infty, \; \sum_{k=0}^\infty T_k^2 \leq \infty$
- $T_k = \frac{\beta}{\gamma + k}$

The role of mini batches
same analysis applies $M \rightarrow M/n_{mb}, M_v \rightarrow M_v/n_{mb}$
EQ
But we can also run SGD with step T/nmb and get same result
Advantage in computation if paralellization possible

Can we do non-(strongly-)convex functions?

Proposition 12. F L-smooth, then SGD with stepize $0 < T \leq \frac{1}{L(1+M_v)}$ achieves
$E[\frac{1}{N}SUM] \leq TLM + \frac{2(F(x_0) - F_{inf})}{TN}$
$F_{inf} = \underset{x \in \mathbb{R}^n}{inf} F(x)$

Proof. similat to previous proposition, from (1) we infer:

$$E[F(x_{k+1})] - E[F(x_{k+1})] \leq -\frac{T}{2}E[X]XX$$

XXX
SUM □

7.2 Table

7.3 Recitation

- SGD vs GD
- N is large, GD to costly

7.4 Methods to imporve SGD

- Mini Batch
- Momentum, moving average of gradients
- Control Variates
- Variance Reduction Techniques
- SAGA stochastic avarageing gradient
- Stochastic Variance Reduced Gradient (SVRG)
- Summary

Explanations on Code

8 Alternating Direction Method of Multipliers (ADMM)

Motivation
Last week:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^m f_i(x) \tag{9}$$

Today: exploit parallesization

$$\min_{x_1,\dots,x_m} \sum_{i=1}^{m} f_i(x_i) \text{ s.t. } x=(x_1,\dots,x_m) \quad (10)$$

8.1 Dual ascent

Start with:

$$\min_{x\in\mathbb{R}^n} f_i(x) \text{ s.t. } Ax=b \quad (11)$$

Derive dual:

$$\mathcal{L}(x,\lambda)=f(x)+\lambda^{\mathsf{T}}(Ax-b)$$

$$\underset{x\in\mathbb{R}^n}{inf}\,\mathcal{L}(x,\lambda)=-\underset{x\in\mathbb{R}^n}{sup}\,\{(-\lambda^{\mathsf{T}}A)x-f(x)\}-\lambda^{\mathsf{T}}b$$

fstar
d(lambda)
The subgradient is given by:

$$\partial d(\lambda)=A\partial f^\star(-A^{\mathsf{T}}\lambda)-b$$

optimizer satisfies...
BOX
Two results in dual subgradient ascent
$\lambda_{k+1}=\lambda_k+T_k(Ax_k-b)$, $x_k\in A$

8.2 Example 1

Starting from (9) and with $Ax=0$ s.t. $x_1-x_2=x_2-x_3=\dots=x_m-x_1=0$
BLACKBOARD

$$x_k\in\underset{x_1,\dots,x_m\in\mathbb{R}^n}{\text{argmin}}\,(\sum_{i=1}^{m}f_i(x))+\lambda_1(x1-x2)+\lambda_2(x2-x3)+\dots$$

With that the subgradient becomes

$$x_{k_i}\in\underset{\hat{x_i}\in\mathbb{R}^n}{\text{argmin}}\{f_i(\hat{x_i})-\lambda_{k_{i-1}}^{\mathsf{T}}\hat{x_i}+\lambda_{k_i}^{\mathsf{T}}\hat{x_i}\}$$

for $i=2,3,\dots,m-1$ in parallel
$\lambda_{k+1,i}=\lambda_{k,i}+T_k(x_{k_i}-x_{k_{i+1}})$

8.3 Real life examples

Video Quadcopter
- Not attached Pendulum
- Nonconvex OP
- Trajectory offline computed
- Track it with time-variying LQR feedback controller

Video Robotarm
- Table tennis
- Very flexibel arm

Dynamic control of magnetic navigation
- Balance stick on 4 magnets
- Precise control of fields

8.4 Example 2

$f(x=)\sum_{i=1}^m f_i(x_i)$ with $Ax=b$
$x=(x_1,\dots,x_n)$ and $A=[A_1,\dots,A_m]$
Dual subgradient becomes
$x_{k_i}\in\underset{\hat{x_i}}{\text{argmin}}\{f_i(\hat{x_i})+\lambda_k^{\mathsf{T}}A_i\hat{x_i}\}$ (local minimization)
$\lambda_{k+1}=\lambda_k+T_k(\sum_{i=1}^m A_i x_{k_i}-b)$ (broadcasting)
IMAGE

Proposition 13. f convex with closed epigraph, f is μ-strongly convex if and only if $f^\star$ is $1/\mu$-smooth.

From that we conclude
$d(\lambda)=-f^\star(-A^{\mathsf{T}}\lambda)-\lambda^{\mathsf{T}}$
f μ-strongly convex $\rightarrow$ $f^\star$ is $1/\mu$ smooth $\rightarrow$ $d(\lambda)$ is $\bar{\sigma}(AA^{\mathsf{T}})$ $1/\mu$-smooth
f is L-smooth $\rightarrow$ $f^\star$ is $1/L$ strongly convex $\rightarrow$ $d(\lambda)$ is $\bar{\sigma}(AA^{\mathsf{T}})$ $1/L$-smoothly convex
Problem f μ-strongly convex is hardly restricting condition

8.5 ADMM

$$\min_{x\in\mathbb{R}^n} f(x)+\frac{\rho}{2}|Ax-b|^2$$

s.t. $Ax=b$ with $\rho>0$

8.5.1 Augmented Lagrangian

$$x_k=A$$

$$\lambda_{k+1}=A$$

ADVANTAGE
DISADVANTAGE
SOLUTION

8.6 Alternating direction method of multipliers

CONSIDER f,g
form augmented objective
augmented Lagrangian
ADMM

$$x_k=\underset{x\in\mathbb{R}^n}{\text{argmin}}\mathcal{L}_p(x,z_{k-1},\lambda_k)$$

$$z_k=\underset{x\in\mathbb{R}^n}{\text{argmin}}\mathcal{L}_p(x_k,z,\lambda_k)$$

$$\lambda_{k+1}=\lambda_k+\rho(Ax_k+Bz_k-c)$$

EXAMPLE Images Low/High rank

8.7 Recitation

8.7.1 Recap

Optimization Problem
min f,g
Augmented Lagrangian
ADMM

$$x_k=\underset{x\in\mathbb{R}^n}{\text{argmin}}\mathcal{L}_p(x,z_{k-1},\lambda_k)$$

$$z_k=\underset{x\in\mathbb{R}^n}{\text{argmin}}\mathcal{L}_p(x_k,z,\lambda_k)$$

$$\lambda_{k+1}=\lambda_k+\rho(Ax_k+Bz_k-c)$$

Wecan also consider completing the square in the augmented Lagrangian as

$$\mathcal{L}_p(x,z,\lambda)=f(x)+g(z)+\frac{\rho}{2}|Ax+Bz-c+\frac{\lambda}{\rho}|^2-\frac{1}{2\rho}|\lambda|^2$$

and introduce new dual variable $\mu=\frac{\lambda}{\rho}$ to obtain a scaledversion of ADMM.
SYSTEM
x_k
z_k
ν_{k+1}

8.7.2 Contrained optimization via ADMM

min...
Solve with ADMM:
1. Transform
2. Apply ADMM

8.7.3 Solving QPs with ADMM

QP
1. Transform to ADMM form:
min f,g s.t. A..
2. Apply ADMM
x_k
z_k
ν_{k+1}
3. Simplify the minimization steps
x-minimiztion is again QP with constraints
- Lagrangian
- KKT $\nabla_x\mathcal{L}(x,\mu)=\dots=0\Leftrightarrow$ Matrix system

z-minimiztion
- since g indicator function ...
- ..projection step..

$z_k=\underset{\mathbb{R}_+^n}{\text{prox}}(x_k+\mu_k)$

9 Distributed optimization with ADMM

Motivation
- Slides - Distributed computation - Vanilla vs averaging

We start with:

$$\min_{x\in\mathbb{R}^n}\sum_{i=1}^{m}f_i(x)$$

Goal Solve problem such that each term can be handled by its own processor.
Reformulation:

$$\min_{x_1,\dots,x_N\in\mathbb{R}^n,z\in\mathbb{Z}^n}\sum_{i=1}^{N}f_i(x_i) \quad \text{s.t.} \quad x_i=z, \quad i=1 \quad (12)$$

and apply ADMM

9.1 Global consensus problem

Solve (12) with ADMM
Step 1: Form augmented Lagrangian
$\mathcal{L}_p()=SUM$
Step 2: Formulate ADMM

$$\lambda_i^{k+1}=\lambda_i^k+\rho(x_i^{k+1}-z^{k+1})$$

$\rho N z_{k+1}=\dots$
$\sum_{i=1}^N\lambda_i=\sum_{i=1}^N\{\lambda_i^k-\lambda_i^k\}=0$
therefore with $\lambda_i=0$ for $i=1,\dots,N$
$z^{k+1}=\dots$
this results in...
GRAFIK

9.2 Sharing Problem

$$\min_{x_1,\dots,x_N\in\mathbb{R}^n} \sum_{i=1}^N f_i(x_i) + g(\sum_{i=1}^N x_i)$$

Apply ADMM:
$x_i = z_i,\ i = 1,\dots,N$
Step 1: Form augmented Lagrangian

$$\mathcal{L}_p(x_1,\dots,x_N,z_1,\dots,z_\lambda,x_1,\dots,\lambda_N) = \sum_{i=1}^N f_i(x_i)+g(\sum_{i=1}^N z_i)+\rho\dots$$

Step 2: Formulate ADMM dynamics

$$x_i^{k+1} \tag{13}$$

$$z_i^{k+1} \tag{14}$$

$$\lambda_i^{k+1} \tag{15}$$

Simplify (14) with $a_i = \dots$
stationary contidions for (14) $O \in$
...greatly simplified by introducing averages $\bar{z}^{k+1}, \bar{a}$
Then we arrive at N stationary contidions...
R
NR
O
$\bar{z}^{k+1}$
z^{k+1}
λ_i^{k+1}
all λ_i^{k+1} equal
FINAL DYNAMICS

$$x_i^{k+1}$$

$$\bar{z}^{k+1}$$

$$\lambda^{k+1}$$

Priciples: (not shown)

9.2.1 Dual of Sharing Problem

derivations (not shown)
sup

9.3 Optimization over Graphs

$g = (V, E)$ undirected graph with vertices V and edges E
Solve

$$\min_{x\in\mathbb{R}^n} \sum_{i\in V} f_i(x)$$

where each vertex has local data and we would like to fit a model with shared parameters
GRAFIK
Idea Reformulation with constraints

$$\min_{x_1,\dots,x_{|V|},z_1,\dots,z_{|V|}} \sum_{i\in V}^N f_i(x_i) \text{ s.t. } x_i = z_{ij}, x_j = z_{ij}$$

Step 1: Form augmented Lagrangian
$\mathcal{L}_p() = SUM + SUMSUM$
Step 2: Formulate ADMM
DERIVATIONS
FINAL RESULTS
$x_i^{k+1} \in \operatorname{argmin}_{x\in\mathbb{R}^n} f_i(x_i) + SUM$
$\bar{x}_i^{k+1}$
p_i^{k+1}

9.4 Recitation

QUIZ Qestions
2023 1b) conjugate function
2020 2a) Hyperplane with dual-minimization

10 Signal denoising and regression

Linear equation $y = Ax, y \in \mathbb{R}^n, x \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n}$
- classic setting $m \gg n$
- modern setting $m \ll n$

10.1 Classic setting with outliers

$$\min_{x\in\mathbb{R}^n} |Ax - y|_2^2$$

uses l_2-norm to penaliize large residuals
GRAFIK
but as a result, outliers have a lot of weight
Grafik
Weight of outliers can be reduced with l_1-norm

$$\min_{x\in\mathbb{R}^n} |Ax - y|_2^2 \tag{16}$$

Rewrite (16) as convex program
min,sum,zi
s.t.
AGAIN REformulate
-> linear program
For best of both worlds:

$$\forall (i,j) \in E \quad \phi_{\text{Hub}}(u) = \begin{cases} u^2 & \text{if } |u| \geq M \\ 2Mu - M^2 & \text{if } |u| > M \end{cases}$$

resulting OP:
min,sum,fub,()

10.2 Modern setting

- $Ax = y$ has infinetly many solutions
- Which one is the best?
- add regulizer

Tikhonov regulizer: $\min_{x\in\mathbb{R}^n} |Ax - y|_2^2 + \lambda|x|_2^2$
Least Absolute Shrinkage and Selecttion Operator:
$\min_{x\in\mathbb{R}^n} |Ax - y|_2^2 + \lambda|x|_1$
is equalent to
$\min_{x\in\mathbb{R}^n} |Ax - y|_2^2$ s.t. $|x|_1 \leq c$

10.2.1 Example

Audio signal, $f_1 = 102$ Hz, $f_2 = 305$ Hz

$$\tilde{x}(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t) + n(t) \tag{17}$$

Signal evaluated at 100 randomly selected points $t_i \in [0, 1]$
MATLAB script
-> how choose λ?
Projection on l_1 Ball
Approach to solve $\min_{x\in\mathbb{R}^n} \frac{1}{2}|Ax - y|_2^2$ s.t. $|x|_1 \leq c$ with projection:

$$\operatorname*{argmin}_{x\in\mathbb{R}^n} \frac{1}{2}|x - y|_2^2$$

results in Lagrange function:

$$\begin{aligned}\mathcal{L}(x,\lambda) &= \frac{1}{2}|x - y|_2^2 - \lambda(c - |x|_1) \\ &= (\sum_{i=1}^n (x_i - y_i)^2 + \lambda|x_i|) - \lambda c, \lambda \geq 0\end{aligned}$$

where we set $l_i(x,\lambda) = \frac{1}{2}(x_i - y_i)^2 + \lambda|x_i|$
Figure of $\partial_x l_i$ with respect to x_i
IMAGE
Result:

$$\begin{aligned}x_i = \{\operatorname*{argmin}_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda)\}_i &= \{t1\} \\ &= \{t1\} \\ &= \{t1\}\end{aligned}$$

-> how choose λ?
if ... then ..
Example 3 Image denoising
Example 4 Face recognition

10.3 Recitation

11 Classification

$\tilde{y}(\tilde{x})$ takes values in discrete categories
Setup: dataset of $(\tilde{x}_i, \tilde{y}_i),\ i = 1, \dots, N$ with $\tilde{x}_i \in \mathbb{R}^n,\ \tilde{y}_i \in \{1, 2, \dots, K\}$
Naive Approach
Classify with
$f^{\text{naive}}(\tilde{x}) =$

11.1 Recitation

12 Adaptive decision-making

12.1 Recitation