Large-Scale Convex Optimization

Stadelmann Silvan silvasta@ethz.ch

13. Juni 2025

1 T2 - Convex Sets and Convex Functions

Convex Sets

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

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

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

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

Convex Functions

Definition

Methods for establishing convexity

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

Log-Sum-Exp

$$f(x) = log(e^x_1 + ... + e^x_n)$$

differentiable approximation of max(x)

How to check convexity? $\rightarrow$second-order condition $\nabla^2 f >= 0$

Nonnegative Weighted Sum

$\alpha(f_1 + f_2)$ convex if f_1, f_2 convex, $\alpha > 0$

$\rightarrow$If f1 , . . . , fm are convex and w1 , . . . , wm ≥ 0, then w1 f1 + . . . + wm fm is convex

Composition with Affine Function

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

Examples

Log barrier for linear inequalities →transforms constrained problem in unconstrained

Norm Function

Composition

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

2 KKT and Lagrange Duality

Example

$$\min_{x \in \mathbb{R}^{\in}} f(x) \text{ s.t. } h(x) = 0$$

Generalization

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

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

$\rightarrow$ the corresponding Lagrange function is then:

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

cond...
Proposition 1 (Weak Duality)**.** The dual function

$$d(\lambda, \nu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \nu)$$

satisfies $d(\lambda, \nu) \leq f^\star, \ \forall \lambda \geq 0, \ \nu \in \mathbb{R}^n$

proof short
Definition 2 (Constraint qualification)**.** Let $\mathcal{C}$ be Convex, Slaters Condition is satisfied if $\exists \lambda \geq 0, vRn$ s.t. $d(l, v) = f*$
Proposition 2 (Strong Duality)**.** If Slater's condition holds and (1 TODO) is convex then $\exists \lambda \geq 0, \ \nu \in \mathbb{R}^n$ such that $d(\lambda, \nu) = f^\star$

proof extended, important graphic

KKT

Theorem 1 (KKT Conditions)**.** Let (1,TODO) be convex and Slaters condition hold. Then $x^\star \in \mathbb{R}^n$ is a minimizer of the primal (1t) and $\lambda^\star \geq 0, \ \nu^\star$ maximizer of the dual if and only if:

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

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

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

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

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

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

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

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

INF=SUP

Remark Without Slater, KKT 1 to 4 still implies $x^\star$ minimizes (1t) and (λ,ν) maximizes the dual, but the convergence is no longer true!

FORCE BALLANCE

What if f, g not differentiable?

Example $\inf_{x \in \mathbb{R}^n} |Ax - b|^2 + |x|_1$

where (l_1)-norm not differentiable at 0

Subdifferential

for convex f...
Definition 3 (name of the definition). $f\mathbb{R}^n \rightarrow \mathbb{R}$ convex. The subdifferential of f at $\bar{x}$ is: $\delta f(\bar{x}) := \{\lambda \in \mathbb{R}^n \mid f...\}$
Proposition 3. $f\mathbb{R}^n \rightarrow \mathbb{R}$ convex. $x^\star \in argmin...$
Proposition 4. f convex, $epi(f)$ closed $y \in df(x) arrow x \in \delta f^\star(y)$

EXAMPLE

Recitation 3

Information ML

Hard Margin SVM

Use hyperplane and support vectors for data classification.

SVM

Find the Maximum-Margin Hyperplane

Solve the Optimization Problem

- Introduce Lagrange multiplier $\alpha_i \geq 0$ for $i = 1, 2, \ldots, N$

  ...
- ...
- Solve $\alpha^\star$ by Strong Duality
- Obtain $w^\star$ and $b^\star$ using KKT

Soft Margin SVM

- Introduce some *slackness* ξ
- Point 2

Kernel Methods: Break the linearity

Introduce Nonlinear feature map $\phi(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$

Kernel $K(x_i, x_j) : \mathbb{R}^n X \mathbb{R}^m \rightarrow \mathbb{R}$

3 Convex Optimization Problem

minimize $f(x), g \leq 0, h(x) = 0$

1) Feasibility Problem

minimmize s

s.t. $g_i(x) \leq s \quad \forall i, \ldots, n_g, h(x) = 0$

2) Linear Programming

minimmize $c^\mathsf{T} x$ s.t. $Ax - b \geq 0$ and $x \geq 0$

$\rightarrow$ derive dual:

Step 1: $\mathcal{L}(x, \lambda_1, \lambda_1) = c^\mathsf{T} x - \lambda_1^\mathsf{T}(Ax - b) - \lambda_2^\mathsf{T} x$

Step 2: $\inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda_1, \lambda_1) = \{-\infty else, \lambda_1^\mathsf{T} b, if\ A^\mathsf{T}\lambda_1 + \lambda_2 = c$

Step 3: Dual Problem

$\sup_{\lambda_1 \geq 0} \lambda_1 b$ s.t. $0 \leq c - A^\mathsf{T}\lambda_2$

- dual as a linear program
- Skech, polyhedron, c-vector normal gives 'Levelsets' and optimal solution in or trough a corner (if exists)

Proposition 5. 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.

Example Shortest Path

Analogie with Fluid

Soltuion greater 0, not optimal edges = 0

3) Quadratic Programming

minimmize $\frac{1}{2}x^{\mathsf{T}}Px + q^Tx$

s.t. $Gx \leq h,\ Ax = b$

$\rightarrow$ if $P = P^{\mathsf{T}}$ is positive semi-definite then the problem is convex.

Example [optimal control] (basis for mpc)

Second-order cone program (SOCP)

minimmize f^Tx

s.t. $|A_ix + b| \leq c_i^Tx + d_i,\ Fx = g$

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
- $x \geq 0$, no short position

CALCULATIONS

Semidefinite programming (SDP)

minimmize $c^{\mathsf{T}}x$

s.t. $x_1F_1, \cdots + x_nF_n <= 0$ and $Ax - b = b$

$\rightarrow$ the 'standard' form

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

Diagramm

Recitation 4

Geometric Programming

Motivation

- Summary Change of variables, transformation of objectives and constraints

$\rightarrow$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

Sum of Squares

- Polynomial Optimization

$\rightarrow f, g_i, h_i$ polynomials

General case intractable

- Nonnegative polynomials

Small adaption with γ

find largest γsuch that $f(x) - \gamma$ nonnegative, NP Hard

$\rightarrow$ 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 $\leq d$
Theorem 2 (SOS). $p(x)$ is an SOS if and only if $\exists\, Q$ such that $Q >= 0$ and $p(x) = z(x)^\mathsf{T} Q z(x)$

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

4 Gradient methods - Part I

Definition 4 (smoothness). The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L-smooth if $\nabla f(x)$ satisfies

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

This result (with Taylors'Theorem) in:

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

Definition 5 (strong convexity). The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ strongly convex it it satisfies

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

Gradient Descent

PSEUDOCODE

$x = x_0 \in \mathbb{R}^n$

for x in range N …

HERLEITUNG

Optimal Step Size

$\mu \leq h \leq L$

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

GRAFIK

$$\rho(T^\star) = |XXX|$$

therefore with stepsize $T^\star$

$|x_N - x^\star| \leq \epsilon$

Momentum-based methods

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

Discretization of $\dot{q} = p, \dot{p} = -2d...$

Spring damper analogy

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

Nesterovs accelerated gradient methods

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

Heavy Ball (tuned quadratics)

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 $c = ((\lambda - L)/2, 0), r = \lambda + L$

Theorem 3 (NOT Nesterovs). $f\mu$ strongly convex, L smooth Nesterovs Method satisfies

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

proof with H Function

Recital 5 - More on Gradient Descent

Proberties of Smooth Functions

- L-smoothnes:

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

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

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

- 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 $1/T^2$

Proberties of Strongly-Convex Functions

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

...this implies

Smooth and Strongly-Convex

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

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

Gradient Descent - Part II

Projected gradient descent(smooth,strongly convex f)
Definition 6. $prox_{\mathcal{C}}(x) = argmin 1/2|x - y|^2$

C closed convex

CAUCHY SCHWARZ

This implies: $|prox_{\mathcal{C}}(x) - prox_{\mathcal{C}}(y)| \leq |x - y|$

Proof Other Information. TODO □

Algorithm
Proposition 6. satisfaction of GD

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

□

– norm ball

- probability simplex

when are projetions computationally cheap?

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 7 (Subgradient Method)**.** Closed, convex set $\mathcal{C}$ contained in ball of $r = R$

Consider update rule: $x_{k+1} = prox_{\mathcal{C}}(x_k - Tg_k), \ldots$ then $x_0, ...$

Proof. NOT SHOWED

□

TABLE

GRAPH with rates, IMPORTANTe

Recitation 7

- SGD vs GD

- N is large, GD to costly

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