Large-Scale Convex Optimization

Silvan Stadelmann silvasta@ethz.ch 20. Juni 2025

1 Introduction

Large Scale Problem of dimension n but iterations $\ll n$ desired
Convex One of the only problem classes that are "solvable"
Optimization with decision variable x, objective function f and feasible set $\mathcal{C} = \{\xi \in \mathbb{R}^n : g(\xi) \leq 0,\ h(\xi) = 0\}$
$x^\star$ **local minimum** if $\exists\, \epsilon > 0$ s.t. $f(x^\star) \leq f(x)$, $\forall x \in \mathcal{C} \cap B_\epsilon(x^\star)$, $B_\epsilon(x^\star) := \{\xi \in \mathbb{R}^n : |\xi - x^\star| < \epsilon\}$
Proposition 1. f (lower-semi-)continuous, $f(x) \to \infty$ for $|x| \to \infty$, $\mathcal{C}$ closed $\Rightarrow \exists$ of OP with: $\min_{x\in\mathcal{C}} f(x)$ and $x^\star \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$
Definition 1 (Lipschitz continuity). $q : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz with constant L if: $|q(x) - q(y)| \leq L|x - y| \forall x, y \in \mathbb{R}^m$
Class of OP P with $\mathcal{C} = [0, 1]^n$ and f is l^∞-Lipschitz with L
Proposition 2. For any algorithm $\exists$ problem in P, s.t. achieving $|f(x_N)-f(x)| < \epsilon$ requires $N \geq (\lfloor \frac{L}{2\epsilon} \rfloor)^n - 1$
Definition 2. OP convex if, f and g_i convex functions, h affine.
Definition 3. $q : \mathbb{R}^n \to \mathbb{R}$ convex (affine) if $\forall\, x, y \in \mathbb{R}^n$

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

Proposition 3. If OP convex, local minimum == global minimum

2 Convex Optimization Problem

Definition 4 (Convex Set). A set $\mathcal{C}$ is convex if and only if $\theta x + (1-\theta)y \in \mathcal{C}, \forall\, x, y \in \mathcal{C}, \forall\, \theta \in [0, 1]$
(hyperplane || half-space) $\{x \in \mathbb{R}^n \mid a^\mathsf{T} x (=\|\leq) b\}$
polyhedra $\{x \in \mathbb{R}^n \mid A^{q\times n} x \preceq b^{q\times 1}, C^{r\times n} x = d^{r\times 1}\}$

Operations that preserve convexity (sets)

Intersection $\mathcal{C}_1, \mathcal{C}_2$ cv $\Rightarrow \mathcal{C}_1 \cap \mathcal{C}_2$ convex (cv)
Image under affine map $\mathcal{C} \subseteq \mathbb{R}^n$ cv $\Rightarrow \{Ax + b \mid x \in \mathcal{C}\}$ cv
Inverse IoaM $\mathcal{C} \subseteq \mathbb{R}^m$ cv $\Rightarrow \{x \in \mathbb{R}^n \mid Ax + b \in \mathcal{C}\}$ cv

Separating Hyperplane Theorem

Theorem 1. $\mathcal{C} \subseteq \mathbb{R}^n$ non-empty closed (cl) convex set, $y \notin \mathcal{C}$ $\to \exists\, a \neq 0, b \in \mathbb{R}$ s.t. $a^\mathsf{T} x + b < a^\mathsf{T} y + b, \forall x \in \mathcal{C}$
Corollary 1. $\mathcal{C}_{\text{cl,cv}}$: intersection of cl half-spaces that contain $\mathcal{C}$

Support function

Idea represent any cl,cv set by its supporting hyperplanes

$$\sigma_{\mathcal{C}}(a) = \sup_{x\in\mathcal{C}} a^\mathsf{T} x \quad \text{if known, one can construct}$$
$$\mathcal{C} = \bigcap_{a\in\mathbb{R}^n} \{x \in \mathbb{R}^n \mid a^\mathsf{T} x - \sigma_c(a) \leq 0\}$$
$$= \{x \in \mathbb{R}^n \mid \sup_{a\in\mathbb{R}^n} a^\mathsf{T} x - \sigma_{\mathcal{C}}(a) \leq 0\}$$

Definition 5. $f : \mathbb{R}^n \to \mathbb{R}$ cv $\Leftrightarrow$ epigraph of f is cv set

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

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

Operations that preserve convexity (functions)

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

Check Convexity f is convex if it is composition of simple convex function with convexity preserving operations or if
$f : \mathbb{R}^n \to \mathbb{R}$ twice differentiable, $\partial^2 f/\partial x^2 \succeq 0\, \forall\, x \in \mathbb{R}^n$
$g : \mathbb{R} \to \mathbb{R}$ with $g(t) = f(x + tv)$ convex in $t\, \forall\, x, v \in \mathbb{R}^n$
Extended real numbers $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$
Indicator function $\psi_{\mathcal{C}}(x) := \begin{cases} +\infty & \text{if } x \notin \mathcal{C} \geq 0 \\ 0 & \text{if } x \in \mathcal{C} \end{cases}$
$\to$ 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 $f^\star(y) = \sup_{x\in\mathbb{R}^n} y^\mathsf{T} x - f(x)$

3 KKT and Lagrange Duality

Basic 2d Example for derivation: $\min_{x\in\mathbb{R}^2} f(x)$ s.t. $h(x) = 0$ $\to \nabla f(x^\star), \nabla h(x^\star)$ co-linear $\Leftrightarrow \exists\, \nu^\star \in \mathbb{R} : \nabla f(x^\star) + \nu^\star \nabla h(x^\star) = 0 \Leftrightarrow f(x) + \nu^\star h(x)$ is stationary at $x^\star$

Generalization for $n \to \infty$ and with constraints

$$\text{We consider } f^\star = \inf_{x\in\mathbb{R}^n} f(x) \text{ s.t. } h(x) = 0,\ g(x) \leq 0 \tag{1}$$

$$\text{Lagrange f. } \mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^\mathsf{T} g(x) + \nu^\mathsf{T} h(x) \tag{2}$$

Proposition 4 (Weak Duality). The dual function $d(\lambda, \nu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda, \nu)$ satisfies $d(\lambda, \nu) \leq f^\star, \forall \lambda \geq 0, \nu \in \mathbb{R}^{n_h}$
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 (1) is convex $\Rightarrow \exists \lambda \geq 0, \nu \in \mathbb{R}^{n_h}$ s.t. $d(\lambda, \nu) = f^\star$

KKT

Theorem 2 (KKT Conditions). Slater's condition holds and (1) is convex $\to x^\star \in \mathbb{R}^n$ is a minimizer of the primal (1) and $(\lambda^\star \geq 0, \nu^\star) \in \mathbb{R}^{n_g} \times \mathbb{R}^{n_h}$ is a maximizer of the dual $\Leftrightarrow$

$\nabla_x \mathcal{L}(x^\star, \lambda^\star, \nu^\star) = 0$	KKT-1 (Stationary Lagrangian)
$g(x^\star) \leq 0, h(x^\star) = 0$	KKT-2 (primal feasibility)
$\lambda^\star \leq 0, \nu^\star \in \mathbb{R}^{n_h}$	KKT-3 (dual feasibility)
$\lambda^{\star\mathsf{T}} g(x^\star) = 0 = \nu^{\star\mathsf{T}} h(x^\star)$	KKT-4 (compementary slackness)

In addition we have: $\sup_{\lambda\geq 0, \nu\in\mathbb{R}^{n_h}} q(\lambda, \nu) = \inf_{x\in\mathcal{C}} f(x)$
Remark Without Slater, KKT1-4 still implies $x^\star$ minimizes (1) and λ, ν maximizes dual, but the converse is no longer true. There can be primal-minimizer/dual-maximizer not satisfy KKT.

Subdifferential

For cv f we have $f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^\mathsf{T}(x - \bar{x}),\ \forall x, \bar{x} \in \mathbb{R}^n$
Definition 9. $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ cv, the subdifferential of f at $\bar{x}$ is: $\partial f(\bar{x}) := \{\lambda \in \mathbb{R}^n \mid f(x) \geq f(\bar{x}) + \lambda^\mathsf{T}(x - \bar{x}, \forall x \in \mathbb{R}^n)\}$
Proposition 6. f (like D9), $x^\star \in \operatorname{argmin}_x f(x) \Leftrightarrow 0 \in \partial f(x^\star)$
Proposition 7 (Relation to conjugate functions). For convex f with $\operatorname{epi}(f)$ closed: $y \in \partial f(x) \leftrightarrow x \in \delta f^\star(y)$

4 Convex Optimization Problems

Optimal value $f^\star = \inf\{f(x) \mid g_i(x) \leq 0, h_j = 0\}$
$f^\star = +\infty$ OP is infeasible, $f^\star = -\infty$ OP is unbound below

Feasibility Problem

Special case $f(x) = 0, \forall x \Leftrightarrow \min_s$ s.t. $g_i(x) \leq s, h_j(x) = 0$
Linear Programming minimize $c^\mathsf{T} x$ s.t. $Ax - b \geq 0,\ x \geq 0$
Step 1: $\mathcal{L}(x, \lambda_1, \lambda_2) = c^\mathsf{T} x - \lambda_1^\mathsf{T}(Ax - b) - \lambda_2^\mathsf{T} x,\ \lambda_i \geq 0$
Step 2: $\inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda_1, \lambda_2) = \begin{cases} \lambda_1^\mathsf{T} b & \text{if } c - A^\mathsf{T}\lambda_1 - \lambda_2 = 0 \\ -\infty & \text{if } c - A^\mathsf{T}\lambda_1 - \lambda_2 = 0 \end{cases}$
Step 3: Dual, maximize $b^\mathsf{T}\lambda$s.t.$c - A^\mathsf{T}\lambda \geq 0,\ \lambda \geq 0$ (again LP)
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.
Quadratic Programming convex if $P = P^\mathsf{T}$ positiv semi-definite minimize $\frac{1}{2}x^\mathsf{T} P x + q^\mathsf{T} x$ s.t. $Gx \leq h,\ Ax = b$

Second-Order Cone Program

minimize $f^\mathsf{T} x$ s.t. $|A_i x + b| \leq c_i^\mathsf{T} x + d_i, Fx = g$
Second-order cone $C_{n+1} = \{(x, t) \mid x \in \mathbb{R}^n, t \in \mathbb{R}, |x| \leq t\}$
Semi-Definite Programming with symmetric F_i, X, A_i
minimize $c^\mathsf{T} x$ s.t. $\sum_{i=1}^n x_i F_i + G \preceq 0, Ax = b$
Standard form minimize $\operatorname{tr}(CX)$ s.t. $X \geq 0, \operatorname{tr}(A_i X) = b_i$
$\operatorname{tr}(CX) = \sum_{i=1}^n \sum_{j=1}^n C_{ij} X_{ij}, C \in \mathbb{R}^{n\times n}, i = 1, .., m$
LP$\subset$ QP $\subset$ QCQP (Quadratically Constrained QP) $\subset$ SOCP $\subset$ SDP

5 Gradient methods - Part I

Definition 10 (smoothness). $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth (L-sm) if $\nabla f(x)$ satisfies $|\nabla f(x) - \nabla f(y)| \leq L|x - y|\ \forall x, y \in \mathbb{R}^n$
Taylor $\to f(y) \leq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{L}{2}|x - y|^2$
Definition 11 (strong convexity). $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex (μ-scv) if $f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{\mu}{2}|x - y|^2$

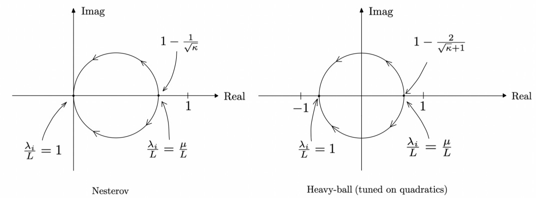
Gradient Descent

$x_{k+1} = x_k - T\nabla f(x_k)$ for $k = (k_0, \ldots, k_N)$ given x_0, T
Assume $f(x) = c_0 + b^\mathsf{T} x + \frac{1}{2}x^\mathsf{T} H x, H \succ 0 \Rightarrow Hx^\star = -b$
$x_{k+1} - x^\star = x_k - x^\star - T(b + Hx_k) = (I - TH)(x_k - x^\star)$
Convergence given by eigenvalues of $I - TH$, use $H = U\Lambda U^\mathsf{T}$
$x_N - x^\star = U(I - T\Lambda)^N U^\mathsf{T}(x_0 - x^\star) \to$ conv-rate $1 - T\lambda_i$
f: L-sm, μ-scv $\to \mu \leq \min \lambda_i$, $\max \lambda_i \leq L$, $\to$ conv-rate $\rho(T) =: \max_{\mu\leq h\leq L} |1 - Th| \to |x_N - x^\star| \leq \rho(T)^N |x_0 - x^\star|$
$T^\star = \frac{2}{L+\mu}$, with condition number $\kappa := \frac{L}{\mu}$ and $1 - \xi \leq e^{-\xi}$
$\rho(T^\star) = \frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1} = (1 - \frac{2}{\kappa+1}) \leq e^{-\frac{2}{\kappa+1}} \to$ algebraic complexity $N \geq \frac{\kappa+1}{2} \ln(\frac{|x_0 - x^\star|}{\epsilon})$ to achieve $|x_N - x^\star| \leq \epsilon$

Momentum-based methods

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

Nesterov accelerated gradient $T = 1, d = \frac{1}{\sqrt{k}+1}, \beta = \frac{\sqrt{k}-1}{\sqrt{k}+1}$
Heavy Ball tuned on quadratic $T = \frac{2\sqrt{k}}{\sqrt{k}+1}, d = \frac{1}{\sqrt{k}+1}, \beta = 0$



$C_{\text{Nesterov}}(1 - \frac{1}{\sqrt{\kappa}})^N \approx \frac{|q_N - q^\star|}{|q_0 - q^\star|} \approx C_{\text{HeavyBall}}(1 - \frac{2}{\sqrt{\kappa+1}})^N$
Theorem 3. f: L-sm, μ-scv $\to$ Nesterov's method satisfies:

$$|q_N - q^\star| \leq \sqrt{\kappa + 1}(1 - 1/\sqrt{\kappa})^{N/2}|q_0 - q^\star|$$
$$f(q_N) - f^\star \leq \frac{L + \mu}{2}(1 - 1/\sqrt{\kappa})^N |q_0 - q^\star|^2$$

Requires $N \geq 2\sqrt{\kappa} \ln(\frac{|q_0 - q^\star|}{\epsilon})$ to achieve $|x_N - x^\star| \leq \epsilon$
Theorem 4. For any first-order method $\exists f : \mathbb{R}^\infty \to \mathbb{R}$, μ-scv, L-sm, s.t. $|x_k - x^\star| \geq (1 - \frac{2}{\sqrt{\kappa+1}})^k |x_0 - x^\star| \forall k \geq 0$

Line search

Adaptive Methods

6 Gradient Methods - Part II

Definition 12. $\operatorname{prox}_{\mathcal{C}}(x) = \operatorname{argmin}_{y\in\mathcal{C}} \frac{1}{2}|x - y|^2$ with $\mathcal{C} \subset \mathbb{R}^n$
Lemma 1. cl, cv $\mathcal{C} \subset \mathbb{R}^n \to |\operatorname{prox}_{\mathcal{C}}(x) - \operatorname{prox}_{\mathcal{C}}(y)| \leq |x - y|$ $\leftarrow |\operatorname{prox}_{\mathcal{C}}(x) - \operatorname{prox}_{\mathcal{C}}(y)|^2 \leq (\operatorname{prox}_{\mathcal{C}}(x) - \operatorname{prox}_{\mathcal{C}}(y))^\mathsf{T}(x - y)$

Projected Gradient Descent

$x_{k+1} = \operatorname{prox}_{\mathcal{C}}(x_k - T\nabla f(x_k))$, for $x_0, k_{0..N}, T \in (0, 2/L)$
Proposition 9. f: L-sm, μ-scv $\to$ projected GD with $T = \frac{2}{L+\mu}$ satisfies $|x_N - x^\star| \leq |x_0 - x^\star|(1 - \frac{2}{\kappa+1})^N$ (κ still $\frac{L}{\mu}$)
Lemma 2. $f : \mathbb{R}^n \to \mathbb{R}$, L-sm, **cv** $\to \tilde{f}$ strongly-cv $\tilde{f}(x) = f(x) + \frac{\mu}{2}|x - x_0|^2$ and $|\tilde{x}^\star - x_0| \leq |x^\star - x_0|$ and $f(x) - f(x^\star) \leq \tilde{f}(x) - \tilde{f}(\tilde{x}^\star) + \frac{\mu}{2}|x^\star - x_0|^2, \mu > 0$
$\to$ from here one can apply GD or Nesterov, which results in: $f(x_N) - f(x_0) \leq \epsilon$ after $N \sim L|x^\star - x_0|^2/\epsilon$ iterations
Proposition 10 (Subgradient Method). cl, cv $\mathcal{C}$ contained in ball of radius R, $x_{0..N-1}$ satisfy $f(\frac{1}{N}\sum_{k=0}^{N-1}) - f(x^\star) \leq \frac{RL_f}{\sqrt{N}}$ under $x_{k+1} = \operatorname{prox}_{\mathcal{C}}(x_k - Tg_k)\ g_k \in \partial f(x_k), T = \frac{R}{L_f\sqrt{N}}$

Assumptions on f	Method	$N : f(x_N) - f(x^\star) \leqslant \varepsilon$	Optimal
μ-strongly convex	gradient descent	$N \sim \kappa \ln(1/\varepsilon)$	No
L-smooth	Nesterov	$N \sim \sqrt{\kappa} \ln(1/\varepsilon)$	Yes
L-smooth	gradient descent	$N \sim 1/\varepsilon$	No
	Nesterov (varying stepsize)	$N \sim 1/\sqrt{\varepsilon}$	Yes
L_f-Lipschitz, compact set	subgradient method	$N \sim 1/\varepsilon^2$	Yes