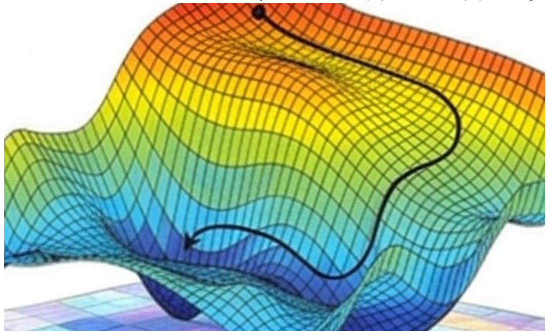
Large-Scale Convex Optimization

Silvan Stadelmann silvasta@ethz.ch 4. Juli 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 of objective function f with decision variable x considering feasible set $\mathcal{C} = \{\xi \in \mathbb{R}^n : g(\xi) \le 0,\ h(\xi) = 0\}$



Local minimum $x^\star$ if $\exists \epsilon > 0$ s.t. $f(x^\star) \le f(x)$, $\forall x \in \mathcal{C} \cap B_\epsilon(x^\star), B_\epsilon(x^\star) := \{\xi \in \mathbb{R}^n : |\xi - x^\star| < \epsilon\}$
Proposition 1. f (lower-semi-)continuous, radially unbounded, $\mathcal{C}$ closed $\Rightarrow \exists$ $\min_{x\in\mathcal{C}} f(x)$ 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: $|q(x) - q(y)| \le L|x - y| \forall x, y \in \mathbb{R}^m$

f is **Lipschitz** (Lip) with constant $L \Leftrightarrow |\nabla f(x)|_2 \le L$

OP class $\mathcal{P}$ with $\mathcal{C} = [0,1]^n$, f is l^∞-Lipschitz with constant L
Proposition 2. For any algorithm $\exists$ problem in $\mathcal{P}$, s.t. achieving $|f(x_N) - f(x)| < \epsilon$ requires $N \ge (\lfloor \frac{L}{2\epsilon} \rfloor)^n - 1$
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 $\forall\, x, y \in \mathbb{R}^n$

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

Proposition 3. OP convex $\Rightarrow$ local minimum $=$ global minimum

2 Convex Optimization Problem

Definition 4 (Convex Set). A set $\mathcal{C}$ is convex if and only if

$$\theta x + (1-\theta) y \in \mathcal{C} \forall\, x, y \in \mathcal{C}, \quad \forall\, \theta \in [0,1]$$

(hyperplane || half-space) $\{x \in \mathbb{R}^n \mid a^\mathsf{T} x (=||\le) b\}$
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, $y \notin \mathcal{C}$ $\to \exists\, a \ne 0, b \in \mathbb{R}$ 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

$$\sigma_\mathcal{C}(a) = \sup_{x\in\mathcal{C}} a^\mathsf{T} x \quad \text{if known, one can construct}$$

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

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

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

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

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

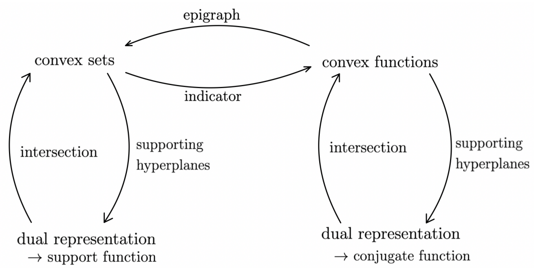
Operations that preserve convexity (functions)

- 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 $t \,\forall\, x, v \in \mathbb{R}^n$
$\to f$ convex (restriction to a line)
Extended real numbers $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$

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

$\to$ 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 $f^\star(y) = \sup_{x\in\mathbb{R}^n} y^T x - f(x)$
Concave $\nabla_x^2 f^\star \prec 0 \Rightarrow$ maximizer of sup satisfies $\nabla_x f^\star = 0$
Theorem 2 (Conjugate of Conjugate). $f : \mathbb{R}^n \to \bar{\mathbb{R}}$
(i) f proper, cv, epi(f) closed $\Rightarrow f^{\star\star} = f$
(ii) $f(x) \ge f^{\star\star}(x), \forall x \in \mathbb{R}^n$



3 KKT and Lagrange Duality

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

Lagrange $\mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^\mathsf{T} g(x) + \nu^\mathsf{T} h(x)$
Dual Function $d(\lambda, \nu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda, \nu)$

Proposition 4 (Weak Duality). $d(\lambda, \nu) \le f^\star, \forall \lambda \ge 0, \nu \in \mathbb{R}^h$
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 $\Rightarrow \exists \lambda \ge 0, \nu \in \mathbb{R}^{n_h}$ s.t. $d(\lambda, \nu) = f^\star$

KKT (Karush-Kuhn-Tucker) Conditions

Theorem 3 (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 $(\lambda^\star \ge 0, \nu^\star) \in \mathbb{R}^{n_g} \times \mathbb{R}^{n_h}$ is a maximizer of the dual $\Leftrightarrow$

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

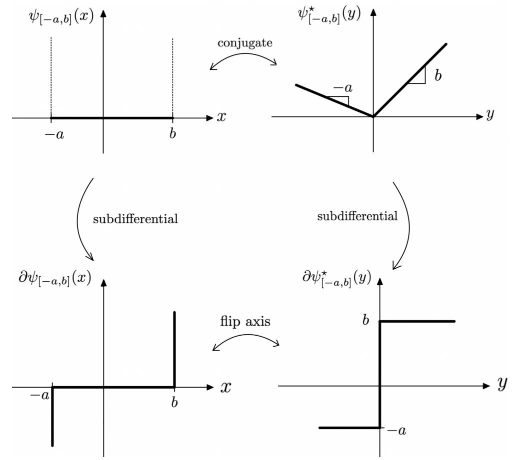
In addition we have: $\sup_{\lambda\ge 0, \nu\in\mathbb{R}^{n_h}} q(\lambda, \nu) = \inf_{x\in\mathcal{C}} f(x)$
Remark Without Slater, KKT1-4 still implies $x^\star$ minimizes (1) and λ, ν maximizes dual, but the converse is no longer true. There can be primal-minimizer/dual-maximizer not satisfy KKT.

Subdifferential

For cv f we have $f(x) \ge f(\bar{x}) + \nabla f(\bar{x})^\mathsf{T}(x - \bar{x}),\ \forall x, \bar{x} \in \mathbb{R}^n$
Definition 9. $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ cv, the subdifferential of f at $\bar{x}$ is:

$$\partial f(\bar{x}) := \{\lambda \in \mathbb{R}^n \mid f(x) \ge f(\bar{x}) + \lambda^\mathsf{T}(x - \bar{x}), \forall x \in \mathbb{R}^n\}$$

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 epi(f) closed: $y \in \partial f(x) \leftrightarrow x \in \partial f^\star(y)$



4 Convex Optimization Problems

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

Feasibility Problem

Special case $f(x) = 0, \forall x \Leftrightarrow \min_s$ s.t. $g_i(x) \le s, h_j(x) = 0$
Linear Programming minimize $c^\mathsf{T} x$ s.t. $Ax - b \ge 0,\ x \ge 0$
Step 1: $\mathcal{L}(x, \lambda_1, \lambda_2) = c^\mathsf{T} x - \lambda_1^\mathsf{T}(Ax - b) - \lambda_2^\mathsf{T} x,\ \lambda_i \ge 0$
Step 2: $\inf_{x\in\mathbb{R}^n} \mathcal{L} = \lambda_1^\mathsf{T} b$, if $c - A^\mathsf{T}\lambda_1 - \lambda_2 = 0$, else $-\infty$
Step 3: Dual, maximize $b^\mathsf{T}\lambda$ s.t. $c - A^\mathsf{T}\lambda \ge 0, \lambda \ge 0$ (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. $Gx \le h,\ Ax = b$
Second-Order Cone Program
minimize $f^\mathsf{T} x$ s.t. $|A_i x + b| \le c_i^\mathsf{T} x + d_i, Fx = g$
Second-order cone $C_{n+1} = \{(x,t) \mid x \in \mathbb{R}^n, t \in \mathbb{R}, |x| \le t\}$
$|A_i x + b| \le c_i^\mathsf{T} x + d_i \Leftrightarrow (A_i x + b, c_i^\mathsf{T} x + d_i) \in C_{n+1}$
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. $X \ge 0, \operatorname{tr}(A_i X) = b_i$
$\operatorname{tr}(CX) = \sum_{i=1}^n \sum_{j=1}^n C_{ij} X_{ij}, C \in \mathbb{R}^{n\times n}, i = 1,..,m$
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 $|\nabla f(x) - \nabla f(y)| \le L|x - y| \,\forall x, y \in \mathbb{R}^n$
Taylor $\to f(y) \le f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{L}{2}|x - y|^2$
Definition 11 (strong convexity). $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex **(μ-scv)** if $f(y) \ge f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{\mu}{2}|x - y|^2$
How to find μ/L, Spectra of Hessian $\nabla^2 f$, min/max eigenvalue

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 $\to \mu \le \min\lambda_i$, $\max\lambda_i \le L$, $\to$ conv-rate $\rho(T)$ $=: \max_{\mu\le h\le L} |1 - Th| \to |x_N - x^\star| \le \rho(T)^N |x_0 - x^\star|$
$T^\star = \frac{2}{L+\mu}$, with **condition number** $\kappa := \frac{L}{\mu}$ and $1 - \xi \le e^{-\xi}$
$\rho(T^\star) = \frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1} = (1 - \frac{2}{\kappa+1}) \le e^{-\frac{2}{\kappa+1}} \to$ algebraic complexity $N \ge \frac{\kappa+1}{2} \ln(\frac{|x_0 - x^\star|}{\epsilon})$ to achieve $|x_N - x^\star| \le \epsilon$

Momentum-based methods

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

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 4. f: L-sm, μ-scv $\to$ Nesterov's method satisfies:

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

Requires $N \ge 2\sqrt{\kappa} \ln(\frac{|q_0 - q^\star|}{\epsilon})$ to achieve $|x_N - x^\star| \le \epsilon$
Theorem 5. For any first-order method $\exists f : \mathbb{R}^\infty \to \mathbb{R}$, μ-scv, L-sm, s.t. $|x_k - x^\star| \ge (1 - \frac{2}{\sqrt{\kappa+1}})^k |x_0 - x^\star| \forall k \ge 0$
Line search optimal step $\nu_t^* = \operatorname{argmin}_{\nu\in\mathbb{R}} f(x_t - \nu\nabla f(x_t))$

6 Gradient Methods - Part II

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

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

f	Method	$N \sim$	Optimal
μ-sc, L-sm	GD NAG	$\kappa \ln(1/\epsilon)$ $\sqrt{\kappa}\ln(1/\epsilon)$	No Yes
L-smooth	GD NAG varying T	$1/\epsilon$ $1/\sqrt{\epsilon}$	No Yes
L-Lip, cmp set	Subgradient	$1/\epsilon^2$	Yes

Example project a on l_1-B: $\min_{x\in\mathbb{R}^n} \frac{1}{2}|x - a|_2^2$ s.t.$|x|_1 \le 1$
$\mathcal{L}(x, \lambda) = \frac{1}{2}|x - a|_2^2 + \lambda|x|_1 - \lambda$
$= \sum_{k=1}^n (\frac{1}{2}(x_k - a_k)^2 + \lambda|x|_1) - \lambda$
$d_k(\lambda) = \inf_{x_k\in\mathbb{R}} \phi(x_k) = \inf_{x_k\in\mathbb{R}}(\frac{1}{2}(x_k - a_k)^2 + \lambda|x|_1)$
$\partial\phi(x_k) = x_k - a_k + \lambda s, s(x_k = 0 : [1-1], else : sign(x_k))$
$d_k(\lambda) = \{\lambda \le |a_k| : \frac{1}{2}\lambda^2 + \lambda|a_k|, else : \frac{1}{2}a_k^2\}$,
$\nabla d_k(\lambda) = \max\{|a_k| - \lambda, 0\}$
Dual: max $d(\lambda)$ s.t. $\lambda \ge 0$ with $d(\lambda) = \sum_n^{k=1} d_k(\lambda) - \lambda$
$x_k = \{\lambda \ge |a_k| : 0; |ak| > 0 : a_k - \lambda, |ak| < 0 : a_k + \lambda\}$

7 Stochastic Gradient Descent (SGD)

Stochastic Gradient $(\tilde{y}_i - \phi(\tilde{x}_i;\theta))^\mathsf{T} \frac{\partial\phi}{\partial\theta}\big|_{\tilde{x}_i;\theta}, i \sim \{1,...,m\}$

Problem formulation: $\min_{x\in\mathbb{R}^n} F(x) = \min_{x\in\mathbb{R}^n} \mathbb{E}[f(x,\xi)]$

$\mathbb{E}_\xi[f(x,\xi)] = \begin{cases} \int_{\mathbb{R}^q} f(x,\bar{\xi})p_\xi(\bar{\xi})d\bar{\xi} & \text{continuous Random V} \\ \sum_{\bar{\xi}} f(x,\bar{\xi})p_\xi(\bar{\xi}) & \text{discrete R Variable} \end{cases}$

$\text{Var}_\xi[g(x,\xi)] = \mathbb{E}_\xi[|g(x,\xi)|^2] - |\nabla F(x)|^2$ Step 1: $\xi_k \leftarrow$ generate realization of ξ

Step 2: $x_{k+1} = x_k - T_k g(x_k,\xi_k)$, step size T_k, SG $g(.)$ $\nabla_x f(x,\bar{\xi}),\ \bar{\xi} \sim p_\epsilon$ or $\frac{1}{n_{mb}}\sum_{i=1}^{n_{mb}} \nabla_x f(x,\bar{\xi}_i), \xi_i \sim p_\epsilon$

$\Rightarrow$ The iterate x_k is now a random variable!

Assumptions on $F(x)$ and $g(x,\xi)$

A1 $F(x)$ is bounded below, ensures $\exists \min_x F(x)$ for $F: L$-sm

A2 $\mathbb{E}_\xi[g(x,\xi)] = \nabla F(x), \forall x \in \mathbb{R}^n$, ensures SG unbiased.

A3 $\exists M, M_v \geq 0$ s.t. $\text{Var}_\xi[g(x,\xi)] \leq M + M_v|\nabla F(x)|^2$ $\forall x \in \mathbb{R}^n$, ensures that variance is bounded.

Proposition 11. F μ-scv L-sm, SGD const. $T < \frac{1}{L(M_v+1)}$

$\mathbb{E}[F(x_k)] - F(x^\star) \leq \frac{TLM}{2\mu} + (1-T\mu)^k(F(x_0) - F(x^\star))$

$T = \frac{ln(N)}{\mu N} \to N \sim \left(\frac{LM}{2\mu^2} + F(x_0) - F(x^\star)\right)/\epsilon$

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

$(1-T\mu)^N \leq e^{-T\mu N}$ this in EQ

The role of mini batches $M \to M/n_{mb}$, $M_v \to M_v/n_{mb}$

Same analysis holds, But run SGD with T/nmb to get same result... Advantage in computation if paralellization possible!

Non-(Strongly-)Convex Functions

Proposition 12. F, L-sm, SGD with $T \leq \frac{1}{L(1+M_v)}$ achieves

$\mathbb{E}[\sum_{k=0}^{N-1}|\nabla F(x_k)|^2] \leq NTLM + \frac{2}{T}(F(x_0) - F_{\text{inf}})$

$F_{\text{inf}} = \inf_{x\in\mathbb{R}^n} F(x)$

SGD Table (1 and 3 optimal, except for finite-sum minimization)

F	Criterion $\leq \epsilon$	$N \sim$	$T_k \sim$
μ-sc L-sm	$\mathbb{E}[F(\bar{x}_N)] - F(x_0)$	$1/\epsilon$	$1/k$
L-sm	$\mathbb{E}[\frac{1}{N}\sum_{k=0}^{N-1}|\nabla F(x_k)|^2]$	$1/\epsilon^2$	$1/\sqrt{k}$
Lip, cv	$\mathbb{E}[F(\bar{x}_N)] - F(x_0)$	$1/\epsilon^2$	$1/\sqrt{k}$

8 ADMM

Parallelization $\min_{x\in\mathbb{R}^n}\sum_{i=1}^m f_i(x_i)$ s.t. $x_1 = \cdots = x_m$

Dual ascent

Consider: $\min_{x\in\mathbb{R}^n} f(x)$ s.t. $Ax = b, A \in \mathbb{R}^{m\times n}$

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

$\inf_{x\in\mathbb{R}^n}\mathcal{L}(x,\lambda) = \underbrace{-\sup_{x\in\mathbb{R}^n}\{(-\lambda^\mathsf{T}A)x - f(x)\}}_{-f^\star(-A^\mathsf{T}\lambda)} - \lambda^\mathsf{T}b$

Dual can be stated as: $\sup_{\lambda\in\mathbb{R}^m} \underbrace{-f^\star(-A^\mathsf{T}\lambda) - \lambda^\mathsf{T}b}_{:=d(\lambda)}$

Subgradient of d given by: $\partial d(\lambda) = A\partial f^\star(-A^\mathsf{T}\lambda) - b$

Recall $v \in \partial f^\star(u) \Leftrightarrow u \in \partial f(v)$ which means that the optimizer in $-\sup_{x\in\mathbb{R}^n}\{(-\lambda^\mathsf{T}A)x - f(x)\}$

satisfies: $-A^\mathsf{T}\lambda \in \partial f(x^\star) \Leftrightarrow x^\star \in \partial f^\star(-A^\mathsf{T}\lambda)$

As a Result, the subgradient $\partial d(\lambda)$ can be expressed via

$$\partial d(\lambda) = Ax - b, \text{where } x \in \text{argmin}_{\hat{x}\in\mathbb{R}^n}\{f(\hat{x}) + \hat{x}^\mathsf{T}A^\mathsf{T}\lambda\}$$

Dual Subgradient Method

$$x_k \in \text{argmin}_{\hat{x}\in\mathbb{R}^n}\{f(\hat{x}) + \hat{x}^\mathsf{T}A^\mathsf{T}\lambda_k\}$$
$$\lambda_{k+1} = \lambda_k + T_k(Ax_k - b), \quad T_k > 0$$

Example $f(x) = \sum_{i=1}^m f_i(x_i)$ with $Ax = b$

$x = (x_1,\ldots,x_n)$ and $A = [A_1,\ldots,A_m]$

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)

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

Derive ADMM

$$\text{Idea:} \quad \min_{x\in\mathbb{R}^n} f(x) + \frac{\rho}{2}|Ax-b|^2 \quad \text{s.t. } Ax = b \text{ with } \rho > 0$$

Leads to this Augmented Lagrangian

$$x_k = \text{argmin}_{x\in\mathbb{R}^n} f(x) + \lambda_k^\mathsf{T}Ax + \frac{\rho}{2}|Ax-b|^2$$
$$\lambda_{k+1} = \lambda_k + T_k(Ax_k - b) \quad (\text{typically } T_k = \rho)$$

Advantage Improved convergence properties even if f non-scv

Disadvantage Loose of decomposability/parallelization due to augmentation with quadratic term.

This motivates ADMM which tries to combine the best of both worlds. (Well conditioned minimization and parallelization)

Consider: $\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} f(x) + g(z)$ s.t. $Ax + Bz = c$

Augmented Objective: $\min f(x) + g(z) + \frac{\rho}{2}|Ax + Bz - c|^2$

Augmented Lagrangian: objective + $\lambda^\mathsf{T}(Ax + Bz - c)$

Alternating Direction Method of Multipliers

$$x_k = \underset{x\in\mathbb{R}^n}{\text{argmin}}\mathcal{L}_p(x, z_{k-1}, \lambda_k)$$
$$z_k = \underset{z\in\mathbb{R}^m}{\text{argmin}}\mathcal{L}_p(x_k, z, \lambda_k)$$
$$\lambda_{k+1} = \lambda_k + \rho(Ax_k + Bz_k - c)$$

Examples

$\min_{x\in\mathbb{R}^n}|Ax-b|_1 \to \min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m}|z|_1$ s.t. $Ax - z = b$

LP $\min_{x\in\mathbb{R}^n} c^\mathsf{T}x$ s.t. $Ax = b, x \geq 0$

$x_0 \in \mathbb{R}^n$ particular solution to $Ax_0 = b$

$Q \in \mathbb{R}^{n\times(n-m)}$ whose columns span null space of A.

Any solution x to $Ax = b$ can be represented by $x = x_0 + Qw$

$\to \min_{w\in\mathbb{R}^{n-m}} c^\mathsf{T}Qw + c^\mathsf{T}x_0$ s.t. $x_0 + Qw \geq 0$

$\to \min_{w,z} c^\mathsf{T}Qw + c^\mathsf{T}x_0$ s.t. $x_0 + Qw - z = 0, z \geq 0$

$\to \min_{w,z} c^\mathsf{T}Qw + c^\mathsf{T}x_0 + \psi_Z(z)$ s.t. $x_0 + Qw - z = 0$

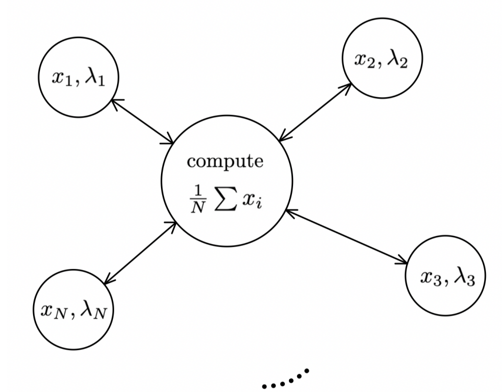
$(Z = \{z \in \mathbb{R}^n | z \geq 0\})$

UPDATE RULES

9 Distributed optimization with ADMM

Goal Solve s.t. each term can be handled by its own processor.

$$\min_{x_1..,x_N,z\in\mathbb{R}^n}\sum_{i=1}^N f_i(x_i) \quad \text{s.t.} \quad x_i = z \quad (f_i \text{ convex}) \quad (2)$$



(Distributed ADMM)

Global Consensus Problem

Step 1: Augmented Lagrangian to solve (2) with ADMM.

$$\mathcal{L}_p(x_i,..,\lambda_i) = \sum_{i=1}^N f_i(x_i) + \lambda_i^\mathsf{T}(x_i - z) + \frac{\rho}{2}|x_i - z|^2$$
$$= \sum_{i=1}^N f_i(x_i) + \frac{\rho}{2}|x_i - z + \frac{1}{\rho}\lambda_i|^2 - \frac{1}{2\rho}|\lambda_i|^2$$

Step 2: Formulate ADMM

$$x_i^{k+1} = \underset{x_i\in\mathbb{R}^n}{\text{argmin}} f_i(x_i) + \frac{\rho}{2}|x_i - z^k + \frac{1}{\rho}\lambda_i^k|^2$$
$$z^{k+1} = \underset{z\in\mathbb{R}^n}{\text{argmin}}\frac{\rho}{2}\sum_{i=1}^N|x_i^{k+1} - z + \frac{1}{\rho}\lambda_i^k|^2$$
$$= \frac{1}{N}\sum_{i=1}^N(x_i^{k+1} + \frac{1}{\rho}\lambda_i^k)$$
$$\lambda_i^{k+1} = \lambda_i^k + \rho(x_i^{k+1} - z^{k+1})$$

FURTHER REFORMULATIONS...

Sharing Problem

$$\min_{x_1,\ldots,x_N\in\mathbb{R}^n}\sum_{i=1}^N f_i(x_i) + g\left(\sum_{i=1}^N x_i\right) \quad (3)$$

$\to$ copy all the variables $x_i = z_i$

$\to$ formulate augmented Lagrangian

$\to$ state ADMM dynamics

Optimization over Graphs

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

$$\min_{x\in\mathbb{R}^n}\sum_{i\in V} f_i(x) \Rightarrow \min_{x_i\in|V|, z_i\in|E|}\sum_{i\in V}^N f_i(x_i)$$
$$\text{s.t. } x_i = z_{ij}, x_j = z_{ij} \quad \forall(i,j)\in E$$

Step 1: Augmented Lagrangian

Step 2: Form the Algorithm

ALGORITHM

$$x_i^{k+1} = \arg\min_{x_i\in\mathbb{R}^n} f_i(x_i) + \frac{d_i}{2}\left|x_i - \frac{1}{2}(x_i^k - \bar{x}_i^k) + \frac{1}{\rho}p_i^k\right|^2, \quad \text{(local update)}$$
$$\bar{x}_i^{k+1} = \frac{1}{d_i}\sum_{j\in\mathcal{N}_i} x_j^{k+1}, \quad \text{(communication with neighbours)}$$
$$p_i^{k+1} = p_i^k + \frac{\rho}{2}(x_i^{k+1} - \bar{x}_i^{k+1}), \quad \text{(local update)}$$

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 $m \ll n$ or $m \sim n$

Classic setting with outliers

l_2-norm: $\min_{x\in\mathbb{R}^n}|Ax - y|_2^2$ results in heavy impact for outliers

l_1-norm: $\min_{x\in\mathbb{R}^n}|Ax - y|$ can be reformulated and solved as

LP: $\min_{z\in\mathbb{R}^m} z^\mathsf{T}\mathbf{1}$ s.t. $-z \leq Ax - y \leq z, 0 \leq z$

Combined: $\phi_{\text{Hub}}(u) = \begin{cases} u^2 & \text{if } |u| \leq M \\ 2Mu - M^2 & \text{if } |u| > M \end{cases}$

Modern setting

$Ax = y$ infinite many solutions $\to$ add regulizer to find best

Tikhonov regression: $\min_{x\in\mathbb{R}^n}|Ax - y|_2^2 + \lambda|x|_2^2$

Least Absolute Shrinkage and Selection Operator

LASSO: $\min_{x\in\mathbb{R}^n}|Ax - y|_2^2 + \lambda|x|_1$

$\Leftrightarrow \min_{x\in\mathbb{R}^n}|Ax - y|_2^2$ s.t. $|x|_1 \leq c$ results in sparse solution

11 Classification

Setup Dataset with pairs of $(\tilde{x}_i, \tilde{y}_i),\ i = 1,...,N$ with data $\tilde{x}_i \in \mathbb{R}^n$ and class $\tilde{y}_i \in \{1,2,...,K\}$

Naive approach, linear regression

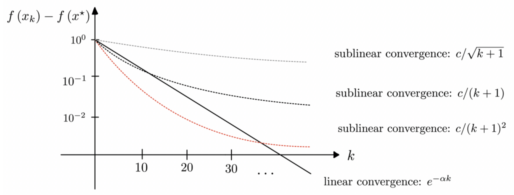
Slightly improved with probabilistic approach

Linear Discriminant Analysis

SVM aims to maximize margin of decision boundary

12 Adaptive decision-making (or Random)

:D



How to "Guess" the Gradient

Observation I: If $\mathbb{E}[uu^\top] = \mathrm{I}_d$, then $\mathbb{E}[uu^\top\nabla f(x)] = \nabla f(x)$

Observation II: $\frac{f(x+\lambda u) - f(x)}{\lambda} \to u^\top\nabla f(x)$

Zeroth-order gradient estimator: sample $u \sim \mathcal{N}(0, \mathrm{I}_d)$

$$g_1(x) = \frac{f(x+\lambda u) - f(x)}{\lambda}u,$$
$$g_2(x) = \frac{f(x+\lambda u) - f(x-\lambda u)}{2\lambda}u$$

Zeroth-order optimization: at each step, $u_t \sim \mathcal{N}(0, \mathrm{I}_d)$

$$x_{t+1} \leftarrow x_t - \eta\frac{f(x_t + \lambda u_t) - f(x_t - \lambda u_t)}{2\lambda}u_t$$