Fundamentals of Optimization for Machine Learning

Technische Universität München

January 3, 2025

Abstract

this is a lecture notes for the graduate level course: Fundamentals of Optimization for Machine Learning at TUM. The references are mainly from the professor's notes and the similar courses taught at EPFL and CMU.

Contents

1 convexity

1.1 Convex function

Definition 1.1 (convex). *a set $C \subset \mathbb{R}^d$ is called convex, if for any $x, y \in C$, the line-segment $\theta x + (1-\theta)y$ here $(0 \le \theta \le 1)$ also lies in C*

- linear: $\theta_1 x + \theta_2 y$ for $x, y \in C$
- conic: if we restrict $\theta_1, \theta_2 \ge 0$
- convex: if we restrict $\theta_1, \theta_2 \ge 0$ and $\theta_1 + \theta_2 = 1$

Theorem 1.2 (intersection). *let C_1, C_2 be convex sets. then, $C_1 \cap C_2$ is also convex.*

Proof. □

Definition 1.3 (convex hull). *let $x_1, x_2, ..., x_m \in \mathbb{R}^d$. their convex hull is*

$$co(x_1, ..., x_m) := \{\sum_i \theta_i x_i | \theta_i \ge 0, \sum_i \theta_i = 1\}$$

Example 1.4. *convex sets: examples*

- *Let $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$. Their convex hull is*

$$co(x_1, \ldots, x_m) := \left\{ \sum_i \theta_i x_i \mid \theta_i \ge 0,\ \sum_i \theta_i = 1 \right\}.$$

- *halfspace $\{x \mid a^T x \le b\}$.*
- *polyhedron $\{x \mid Ax \le b,\ Cx = d\}$.*
- *ellipsoid $\{x \mid (x - x_0)^T A (x - x_0) \le 1\}$, (A: semidefinite).*
- *probability simplex $\{x \mid x \ge 0, \sum_i x_i = 1\}$.*
- *Convex Cones. A convex set $K \subset \mathbb{R}^n$ is called a cone if for $x \in K$, the ray αx is in K for all $\alpha > 0$.*

more Examples:

- *nonneg orthant $\mathbb{R}^n_+$;*
- *Lorentz cone $\{(x, t) \in \mathbb{R}^n \times \mathbb{R}_+ \mid \|x\|_2 \le t\}$;*
- *PSD cone $S^n_+ := \{X \in \mathbb{R}^{n \times n} \mid X = X^T,\ Eig(X) \ge 0\}$.*

Exercise 1.5. *verify the following statements:*

- *Intersection of arbitrary collection of convex cones is a convex cone.*
- *Let $\{b_j\}_{j \in J}$ be vectors in $\mathbb{R}^n$. Then, $\{x \in \mathbb{R}^n \mid \langle x, b_j \rangle \le 0,\ j \in J\}$ is a convex cone (if J is finite, then this cone is polyhedral).*

- *A cone K is convex if and only if $K + K \subset K$.*
- *$\{(x,t) \in \mathbb{R}^n \times \mathbb{R}_+ \mid \|x\| \leq t\}$ is a cone for any norm $\|\cdot\|$.*
- *A real symmetric matrix A is called copositive if for every nonnegative vector x we have $x^T Ax \geq 0$. Verify that the set CP_n of $n \times n$ copositive matrices forms a convex cone.*
- *Spectrahedron: the set $S := \{x \in \mathbb{R}^n \mid x_1 A_1 + \cdots + x_n A_n \succeq 0\}$ is convex for symmetric matrices $A_1, \ldots, A_n \in \mathbb{R}^{m \times m}$. Additionally, observe that the spectrahedron is the inverse image of S_+^m under the affine map $A(x) = \sum_i x_i A_i$.*
- *The convex hull of $S = \{xx^T \mid x \in \mathbb{R}^n\}$ is S_+^n.*

Definition 1.6 (convex function)**.** *a function $f : \mathbb{R}^d \to \mathbb{R}$ is called convex if its domain $dom(f)$ is a convex set and for any $x, y \in dom(f)$ and $\theta \in [0,1]$ we have*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

Theorem 1.7 (Jensen inequality)**.** *let $f : I \to \mathbb{R}$ be continuous. then f is convex if and only if it is midpoint convex, i.e., if $f(\frac{x+y}{2}) \leq \frac{f(x)+f(y)}{2}$ for any $x, y \in I$*

Proof. □

Theorem 1.8. *there are three theorems for recognizing convex functions:*

- *if f is continuous and midpoint convex, then it is convex*

  Proof. □

- *if f is differentiable, then f is convex if and only if $domf$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in domf$*

  Proof. □

- *if f is twice differentiable, then f is convex if and only if $domf$ is convex and $\nabla^2 f(x) \succ 0$ at every $x \in domf$*

  Proof. □

Example 1.9. *the pointwise maximum of a family of convex functions is convex. that is, if $f(x,y)$ is a convex function of x for every y in an arbitrary 'index set' $\mathcal{Y}$, then*

$$f(x) := \max_{y \in \mathcal{Y}} f(x,y)$$

is a convex function of x (set $\mathcal{Y}$ is arbitrary)

Proof. □

Example 1.10. *let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. let $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. prove that $g(x) = f(Ax + b)$ is convex.*

Proof. □

Theorem 1.11. *let $\mathcal{Y}$ be a nonempty convex set. suppose $L(x,y)$ is convex in (x,y), then*

$$f(x) := \inf_{y \in \mathcal{Y}} L(x,y)$$

is a convex function of x, provided $f(x) > -\infty$

Proof. □

Example 1.12 (indicator function)**.** *Let $\mathbf{1}_{\mathcal{X}}$ be the indicator function for $\mathcal{X}$ defined as:*

$$\mathbf{1}_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases}$$

Exercise 1.13. *Verify $\mathbf{1}_{\mathcal{X}}(x)$ is convex if and only if $\mathcal{X}$ is convex.*

Proof. □

Remark. *Using $\mathbf{1}_{\mathcal{X}}(x)$ we can rewrite the constrained problem*

$$\min_x f(x), \quad x \in \mathcal{X},$$

as the following unconstrained problem

$$\min_x f(x) + \mathbf{1}_{\mathcal{X}}(x).$$

Definition 1.14 (norm)**.** *let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that satisfies:*

- *$f(x) \geq 0$, and $f(x) = 0$ if and only if $x = 0$*
- *$f(\lambda x) = |\lambda| f(x)$ for any $\lambda \in \mathbb{R}$*
- *$f(x + y) \leq f(x) + f(y)$*

such a function is called a norm, denoted by $\|\cdot\|$

Theorem 1.15. *norms are convex*

Proof. immediate from subadditivity and positive homogeneity □

Example 1.16. *Let $\mathcal{Y}$ be a convex set. Let $x \in \mathbb{R}^d$ be some point. The distance of x to the set $\mathcal{Y}$ is defined as*

$$dist(x, \mathcal{Y}) := \inf_{y \in \mathcal{Y}} \|x - y\|.$$

Because $\|x - y\|$ is jointly convex in (x, y), the function $dist(x, \mathcal{Y})$ is a convex function of x.

Example 1.17. *vector norms*

- *The Euclidean or ℓ_2-norm is $\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}$.*
- *Let $p \geq 1$; ℓ_p-norm is $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$.*
- *Verify that $\|x\|_p$ is indeed a norm.*
- *(ℓ_∞-norm): $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$.*
- *(Frobenius-norm): Let $A \in \mathbb{C}^{m \times n}$. The Frobenius norm of A is $\|A\|_F := \sqrt{\sum_{ij} |a_{ij}|^2}$; that is, $\|A\|_F = \sqrt{\mathrm{Tr}(A^* A)}$.*

Example 1.18. *Let $A \in \mathbb{R}^{m \times n}$, and let $\|\cdot\|$ be any vector norm. We define an induced matrix norm as*

$$\|A\| := \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Example 1.19. *Let A be any matrix. Its operator norm is*

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

It can be shown that $\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

- **Warning!** Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \notin \{1, 2, \infty\}$
- *Schatten p-norm*: ℓ_p-norm of vector of singular value.

Exercise 1.20. *Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ be singular values of a matrix $A \in \mathbb{R}^{m\times n}$. Prove that*

$$\|A\|_{(k)} := \sum_{i=1}^{k} \sigma_i(A),$$

is a norm; $1 \leq k \leq n$.

Exercise 1.21. *Prove that the following functions are convex.*

- $f(x, y) = \frac{x^2}{y}$ *for* $y > 0$ *on* $\mathbb{R} \times \mathbb{R}_{++}$
- $f^*(y) = \sup_{x\in dom f} \langle x, y\rangle - f(x)$
- $\operatorname{Tr} f(X)$, *where f is scalar convex, X Hermitian*
- $f(X) = -\log\det(X)$ *on positive definite matrices*
- $f(x) = \log(1 + e^x)$ – *logistic loss, on* $\mathbb{R}$
- $f(x) = \log\left(\sum_j e^{a_j^T x}\right)$ – *log-sum-exp on* $\mathbb{R}^d$
- $f(x) = \log \frac{1-x^a}{1-x}$ *for* $a \geq 5$ *on* $(0, 1)$
- $f(x) = \log \int_0^\infty t^{x-1}e^{-t}\, dt$ *on* $x > 0$

1.2 Fenchel conjugates

Definition 1.22. *the fenchel conjuate for function f is defined as: $f^*(y) := \sup_{x\in dom\ f} \langle x, y\rangle - f(x)$*

Remark. *f^* is convex, even if f is not. If f differentiable, then $f^*(\nabla f(x)) = \langle x, \nabla f(x)\rangle - f(x)$ (Fenchel-Legendre transform).*

Remark. *Fenchel-Young inequality: $f^*(u) + f(x) \geq \langle u, x\rangle$*

Fenchel transforms satisfy the beautiful *duality* property:

Theorem 1.23. *Let f be a closed convex function (i.e., epi $f = \{(x, t) \mid f(x) \leq t\}$ is a closed convex set; equivalently, f is lower semi-continuous). Then, $f^{**} = f$.*

this theorem is a special case of the following theorem:

Remark (biconjugate theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be a proper function, i.e., $f(x) > -\infty$ for all $x \in \mathbb{R}^n$ and $f(x) < +\infty$ for at least one $x \in \mathbb{R}^n$. Then, the biconjugate f^{**} of f is given by:*

$$f^{**}(x) = \operatorname{cl}\operatorname{conv}(f)(x),$$

where $\operatorname{cl}\operatorname{conv}(f)$ is the lower semicontinuous convex hull of f. clearly that for lower semicontinuous function, $f = \operatorname{cl}\operatorname{conv}(f)$

Exercise 1.24. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and twice continuously differentiable function. Let f^* be its Fenchel conjugate. Show that $\nabla f^*(\nabla f(x)) = x$*

solution. The Fenchel conjugate $f^*(y)$ is defined as:

$$f^*(y) = \sup_{x\in\mathbb{R}^n} \big(\langle y, x\rangle - f(x)\big),$$

where $\langle y, x\rangle$ denotes the inner product of y and x. let $g(x) \triangleq \langle x, y\rangle - f(x)$, for fixed y, we find x' maximizing $g(x)$ by first order optimality:

$$\nabla_x g(x) = \nabla_x[\langle x, y\rangle - f(x)] = 0 \Rightarrow y = \nabla f(x')$$

which says that the maximizer x' satisfies that $y = \nabla f(x')$, i.e., $f^*(y) = \langle y, x'\rangle - f(x')$, re-use first order optimality w.r.t. y:

$$\nabla_y f^*(y) = x' \Rightarrow \nabla f^*(\nabla f(x')) = x'$$

notice that x' is only related to y and y is arbitrary, so this equation holds for all $x \in \mathbb{R}^n$ □

Exercise 1.25. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and twice continuously differentiable function. Let f^* be its Fenchel conjugate. Show that $\nabla f^*(\nabla f(x)) = x$ and $\nabla^2 f(x) \succ 0$, Show that $\nabla^2 f^*(\nabla f(x)) = [\nabla^2 f(x)]^{-1}$ note both sids are $n \times n$ matrices*

solution. first $\nabla^2 f(x) \succ 0$ guarantees that $\nabla^2 f(x)$ is invertible. using the conclusion of last exercise, we have

$$\nabla[\nabla f^*(\nabla f(x))] = \nabla x \in \mathbb{R}^{n\times n}$$

by chain rule:

$$\nabla[\nabla f^*(\nabla f(x))] = \nabla^2 f^*(\nabla f(x)) \cdot \nabla[\nabla f(x)] = I^{n\times n} \Rightarrow \nabla^2 f^*(\nabla f(x)) = [\nabla^2 f(x)]^{-1}$$

□

Exercise 1.26. *Show that* $f^* = f \iff f = \frac{1}{2}\|\cdot\|_2^2$. *(I guess f is restricted to be a norm)*

solution. ($\Leftarrow$): $f^*(y) = \sup_{x\in dom}(\langle x, y\rangle - f(x))$, find the maximum of $\langle x, y\rangle - f(x)$:

$$\nabla_x[\langle x, y\rangle - \frac{1}{2}\|x\|^2] = 0 \to y = x$$

substitute x by y in the definition of Fenchel conjugate:

$$f^*(y) = \langle y, y\rangle - \frac{1}{2}\|y\|^2 = \frac{1}{2}\|y\|^2$$

($\Rightarrow$): The Fenchel conjugate of $f(x)$ is defined as:

$$f^*(y) = \sup_{x\in\mathbb{R}^n} \big(\langle y, x\rangle - f(x)\big).$$

If $f^* = f$, then:

$$f(x) = \sup_{y\in\mathbb{R}^n} \big(\langle x, y\rangle - f(y)\big).$$

For $f(x)$ to be self-conjugate, $f(x)$ must satisfy: 1. $f(x)$ is strictly convex; 2. $\nabla^2 f(x)$ is constant (explained below); 3. The function must be quadratic.

Why $\nabla^2 f(x)$ is constant: Since $f^* = f$, the second-order derivatives satisfy:

$$\nabla^2 f^*(\nabla f(x)) = [\nabla^2 f(x)]^{-1}.$$

If $\nabla^2 f(x)$ were not constant, this inverse relationship would imply a dependence on x, violating the self-conjugacy condition. Thus, $\nabla^2 f(x)$ must be constant.

A strictly convex function with constant Hessian is quadratic:

$$f(x) = \frac{1}{2}x^T Q x + c^T x + d,$$

where $Q \succ 0$ is a positive definite matrix.

For $f^* = f$: 1. $Q = Q^{-1}$ implies $Q = I$ (identity matrix); 2. The linear term $c^T x$ must vanish to ensure symmetry, so $c = 0$; 3. The constant term d can be ignored since it does not affect the conjugate.

Thus:

$$f(x) = \frac{1}{2}\|x\|_2^2.$$

□

Example 1.27. *$f(x) = ax + b$; then,*

$$f^*(z) = \sup_x zx - (ax + b) = \infty, \text{ if } (z - a) \neq 0.$$

Thus, $domf^* = \{a\}$, and $f^*(a) = -b$.

Example 1.28. *Let $a > 0$, and set $f(x) = -\sqrt{a^2 - x^2}$ if $|x| \le a$, and $+\infty$ otherwise. Then, $f^*(z) = a\sqrt{1 + z^2}$.*

Example 1.29. *$f(x) = \frac{1}{2}x^T Ax$, where $A \succ 0$. Then, $f^*(z) = \frac{1}{2}z^T A^{-1} z$.*

Exercise 1.30. *If $f(x) = \max(0, 1 - x)$, then $domf^*$ is $[-1, 0]$, and within this domain, $f^*(z) = z$.*

Definition 1.31 (support function)**.** $\sigma_C(x) = \sup_{z\in C} z^T x$

support function for the unit norm ball is called *dual norm*

Definition 1.32. *let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. its dual norm is*

$$\|u\|_* := \sup\{u^T x | \|x\| \leq 1\} = \sigma_{\|x\|\leq 1}(u)$$

Exercise 1.33.

- *verify that $\|u\|_*$ is a norm*
- *let $1/p + 1/q = 1$, where $p, q \geq 1$. show that $\|\cdot\|_q$ is dual to $\|\cdot\|_p$. in particular, the ℓ_2-norm is self-dual*
- *verify the generalized Hölder inequality $u^T x \leq \|u\|\|x\|_*$ using the definition of dual norms*

Example 1.34. *let $f(x) = \|x\|$. we have $f^*(z) = \delta_{\|\cdot\|^*\leq 1}(z)$. thus conjugate of a norm is the indicator of unit dual norm ball*

Let $\Gamma_0(\mathbb{R}^d)$ denote class of closed, convex functions on $\mathbb{R}^d$. The (Legendre)-Fenchel transform of $f \in \Gamma_0$ is defined as

$$\mathcal{L} : f \mapsto \sup_y \langle \cdot, y\rangle - f(y)$$

(so that $(Lf)(x) = f^*(x)$).

Theorem 1.35. *Let $\mathcal{T}$ be a transform that maps $\Gamma_0 \to \Gamma_0$ and satisfies: (i) $\mathcal{T}(Tf) = f$ (closure); and (ii) $f \leq g \Rightarrow Tf \geq Tg$. Then, $\mathcal{T}$ must "essentially" be the Fenchel transform. More precisely, there exists $c \in \mathbb{R}$, $v \in \mathbb{R}^d$ and $B \in GL_n(\mathbb{R})$ such that*

$$(Tf)(x) = (Lf)(Bx + v) + \langle v, x\rangle + c$$

Remark. *there are other forms of convexity:*

- ***Log-convex:*** *$\log f$ is convex; log-cvx $\Rightarrow$ cvx;*
- ***Log-concavity:*** *$\log f$ concave; not closed under addition!*
- ***Exponentially convex:*** *$[f(x_i + x_j)] \succeq 0$, for $x_1, \ldots, x_n$*
- ***Operator convex:*** *$f(\lambda X + (1-\lambda)Y) \preceq \lambda f(X) + (1-\lambda) f(Y)$*
- ***Quasiconvex:*** *$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}$*
- ***Pseudoconvex:*** *$\langle \nabla f(y), x - y\rangle \geq 0 \Rightarrow f(x) \geq f(y)$*
- ***Discrete convexity:*** *$f : \mathbb{Z}^n \to \mathbb{Z}$; "convexity + matroid theory."*
- ***Geodesic convexity:*** *$f(\gamma_{x,y}(t)) \leq (1-t)f(x) + tf(y)$ on a "geodesic space"*

1.3 Subdifferentials

Definition 1.36. *A vector $\mathbf{g} \in \mathbb{R}^n$ is called a subgradient at a point y, if for all $x \in \operatorname{dom} f$, it holds that*

$$f(x) \geq f(y) + \langle \mathbf{g}, x - y\rangle$$

Definition 1.37. *The set of all subgradients at y denoted by $\partial f(y)$. This set is called subdifferential of f at y*

If f is convex, $\partial f(x)$ is nice:

- If $x \in$ relative interior of dom f, then $\partial f(x)$ nonempty
- If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If $\partial f(x) = \{\mathbf{g}\}$, then f is differentiable and $\mathbf{g} = \nabla f(x)$

there are some basic facts:

- f is convex, differentiable: $\nabla f(y)$ the unique subgradient at y
- A vector $\mathbf{g}$ is a subgradient at a point y if and only if $f(y) + \langle \mathbf{g}, x - y\rangle$ is globally smaller than $f(x)$.
- Often one subgradient costs approx as much as $f(x)$
- Determining all subgradients at a given point — difficult.
- Subgradient calculus: great achievement in convex analysis

- Without convexity, things become wild (e.g., chain rule fails!)

Example 1.38. *$f(x) = \|x\|_2$. Then,*

$$\partial f(x) := \begin{cases} \|x\|_2^{-1} x & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Proof.

$$\|z\|_2 \geq \|x\|_2 + \langle g, z - x\rangle \|z\|_2 \geq \langle g, z\rangle \Rightarrow \|g\|_2 \leq 1.$$

□

Remark (Some basic calculus)**.**

- *If f and k are differentiable, we know that*
  - ***Addition:*** $\nabla(f + k)(x) = \nabla f(x) + \nabla k(x)$
  - ***Scaling:*** $\nabla(\alpha f(x)) = \alpha \nabla f(x)$
- ***Chain rule***
  If $f : \mathbb{R}^n \to \mathbb{R}^m$, and $k : \mathbb{R}^m \to \mathbb{R}^p$. Let $h : \mathbb{R}^n \to \mathbb{R}^p$ be the composition $h(x) = (k \circ f)(x) = k(f(x))$. Then,

$$Dh(x) = Dk(f(x))Df(x).$$

- *If $f : \mathbb{R}^n \to \mathbb{R}$ and $k : \mathbb{R} \to \mathbb{R}$, then using the fact that $\nabla h(x) = [Dh(x)]^T$, we obtain*

$$\nabla h(x) = k'(f(x))\nabla f(x).$$

- If f is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- Scaling $\alpha > 0$, $\partial(\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$
- Addition*: $\partial(f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)
- Chain rule*: Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax + b)$. Then,

$$\partial h(x) = A^T \partial f(Ax + b).$$

- Chain rule*: $h(x) = f \circ k$, where $k : X \to Y$ is differentiable.

$$\partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))$$

- Max function*: If $f(x) := \max_{1 \leq i \leq m} f_i(x)$, then

$$\partial f(x) = \text{conv} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

  convex hull over subdifferentials of "active" functions at x
- Conjugation: $z \in \partial f(x)$ if and only if $x \in \partial f^*(z)$

Exercise 1.39. *Is it true that convex functions have at least one point in their domain where the subdifferential is non-empty? Justify.*

solution. currently I cannot find a counter example. according to this theorem, my (temporary) answer is that the proposition is true.

Theorem 1.40 (Existence of subgradients (theorem 3.38 in this book))**.** *Consider a proper convex function $f : E \to \mathbb{R}$. Then the subdifferential $\partial f(x)$ is nonempty at every point $x \in \text{ri}(\text{dom} f)$.*

□

Exercise 1.41. *show that if g_1, g_2 are subgradients at a point y for a function $f : \mathbb{R}^d \to \mathbb{R}$, then any convex combination of g_1 and g_2 is also a subgradient at y.*

solution. For g_1 and g_2 to be subgradients at y, the following inequalities hold:

$$f(x) \geq f(y) + g_1^\top (x - y), \quad \forall x \in \mathbb{R}^d, \tag{1}$$

$$f(x) \geq f(y) + g_2^\top (x - y), \quad \forall x \in \mathbb{R}^d. \tag{2}$$

Let g_λ be an arbitrary convex combination of g_1 and g_2:

$$g_\lambda = \lambda g_1 + (1 - \lambda) g_2, \quad \text{where } \lambda \in [0, 1].$$

We need to show that g_λ satisfies the subgradient inequality:

$$f(x) \geq f(y) + g_\lambda^\top (x - y), \quad \forall x \in \mathbb{R}^d. \tag{3}$$

Substitute $g_\lambda = \lambda g_1 + (1 - \lambda) g_2$ into inequality (3):

$$g_\lambda^\top (x - y) = \big(\lambda g_1 + (1 - \lambda) g_2\big)^\top (x - y).$$

Expand:

$$g_\lambda^\top (x - y) = \lambda g_1^\top (x - y) + (1 - \lambda) g_2^\top (x - y). \tag{4}$$

From inequalities (1) and (2), we know:

$$f(x) \geq f(y) + g_1^\top (x - y), \quad f(x) \geq f(y) + g_2^\top (x - y).$$

Take a convex combination of these two inequalities with weights λ and $1 - \lambda$:

$$\lambda f(x) \geq \lambda f(y) + \lambda g_1^\top (x - y),$$

$$(1 - \lambda) f(x) \geq (1 - \lambda) f(y) + (1 - \lambda) g_2^\top (x - y).$$

Add the two inequalities:

$$\lambda f(x) + (1 - \lambda) f(x) \geq \lambda f(y) + (1 - \lambda) f(y) + \lambda g_1^\top (x - y) + (1 - \lambda) g_2^\top (x - y).$$

Simplify:

$$f(x) \geq f(y) + \big(\lambda g_1 + (1 - \lambda) g_2\big)^\top (x - y). \tag{5}$$

Thus, $f(x) \geq f(y) + g_\lambda^\top (x - y)$, proving that g_λ is a subgradient.

□

1.4 Computing subgradients

Example 1.42 (Subgradient for pointwise sup)**.**

$$f(x) := \sup_{y \in \mathcal{Y}} h(x, y)$$

Simple way to obtain some $g \in \partial f(x)$:

- *Pick any y^* for which $h(x, y^*) = f(x)$*
- *Pick any subgradient $g \in \partial h(x, y^*)$*
- *This $g \in \partial f(x)$*

$$\begin{array}{ll} h(z, y^*) & \geq h(x, y^*) + g^T (z - x) \\ h(z, y^*) & \geq f(x) + g^T (z - x) \\ f(z) & \geq h(z, y^*) \quad (\textit{because of sup}) \\ f(z) & \geq f(x) + g^T (z - x). \end{array}$$

Example 1.43. *Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And*

$$f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).$$

This f a max (in fact, over a finite number of terms)

- *Suppose $f(x) = a_k^T x + b_k$ for some index k*

- *Here* $f(x;y) = f_k(x) = a_k^T x + b_k$, *and* $\partial f_k(x) = \{\nabla f_k(x)\}$
- *Hence,* $a_k \in \partial f(x)$ *works!*

Example 1.44 (Subgradient of expectation)**.** *Suppose* $f = \mathbb{E}f(x,u)$*, where* f *is convex in* x *for each* u *(an r.v.)*

$$f(x) := \int f(x,u)p(u)\,du$$

- *For each* u *choose any* $g(x,u) \in \partial_x f(x,u)$
- *Then,* $g = \int g(x,u)p(u)\,du = \mathbb{E}g(x,u) \in \partial f(x)$

Example 1.45 (Subgradient of composition)**.** *Suppose* $h : \mathbb{R}^n \to \mathbb{R}$ *is convex and increasing; each* f_i *is convex*

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

We can find a vector $g \in \partial f(x)$ *as follows:*

- *For* $i = 1$ *to* n*, compute* $g_i \in \partial f_i(x)$
- *Compute* $u \in \partial h(f_1(x), \ldots, f_n(x))$
- *Set* $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$*; this* $g \in \partial f(x)$
- *Compare with* $\nabla f(x) = J\nabla h(x)$*, where* J *is the matrix of* $\nabla f_i(x)$

Exercise: *Verify* $g \in \partial f(x)$ *by showing* $f(z) \geq f(x) + g^T(z-x)$

Exercise 1.46. *Explain how to compute* ***one subgradient*** *for the following convex functions:*

1. $f(x) = \max_{u:Au\leq b} x^T u$
2. *Letting* $x = \left(x^{(1)}, \ldots, x^{(m)}\right) \in \mathbb{R}^{mn}$, $f(x) = \sum_{i=1}^m \|x^{(i)}\|_\infty$

Exercise 1.47. *For each of the convex functions below, explain how to calculate a subgradient at a given* x*.*

(a) $f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$

(b) $f(x) = \max_{i=1,\ldots,m} |a_i^T x + b_i|$

(c) $f(x) = \sup_{0\leq t\leq 1} p(t)$*, where* $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$

(d) $f(x) = x_{[1]} + \cdots + x_{[k]}$*, where* $x_{[i]}$ *denotes the* i*-th largest element of the vector* x

(e) $f(x) = \inf_{Ay\leq b} \|x - y\|^2$*, i.e., the square of the distance of* x *to the polyhedron defined by* $Ay \leq b$*. You may assume that the inequalities* $Ay \leq b$ *are strictly feasible.*

(f) $f(x) = \sup_{Ay\leq b} y^T x$*, i.e., the optimal value of an LP as a function of the cost vector. (You can assume that the polyhedron defined by* $Ay \leq b$ *is bounded.)*

Exercise 1.48 (Fenchel conjugates, Biconjugates and Infimal convolution)**.**

1. *Let* $\alpha > 0$*. Then show that* $(\alpha \cdot |\cdot|)^* = \ell_{[-\alpha,\alpha]}$*, where* $\ell_{[-\alpha,\alpha]}$ *is the indicator function for the interval* $[-\alpha, \alpha]$*.*

   solution. recall the definition of fenchel conjugate for function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$:

   $$f^*(y) = \sup_{x\in\mathbb{R}^n} \{\langle y, x\rangle - f(x)\}$$

   given $f(x) = \alpha|x|$, we have

   $$f^*(y) = \sup_{x\in\mathbb{R}} \{yx - \alpha|x|\}$$

   - for $x \geq 0$, $|x| = x$, and $yx - \alpha|x| = yx - \alpha x = x(y-\alpha)$. the supremum occurs when $y + \alpha \leq 0$, giving $\sup = +\infty$ if $y \leq \alpha$.
   - for $x \leq 0$, $|x| = -x$, we have $yx - \alpha|x| = yx + \alpha x = x(y+\alpha)$. the supremum occurs when $y + \alpha \leq 0$, giving $\sup = +\infty$ is $y \leq -\alpha$
   - for $|y| \leq \alpha$, we have $\sup yx - \alpha|x| = 0$

so the fenchel conjugate is therefore:

$$f^*(y) = \begin{cases} 0, & |y| \leq \alpha \\ +\infty, & \text{otherwise} \end{cases}$$

□

2. *Use the Biconjugate Theorem to show that the biconjugate of the indicator function of a convex set is the indicator function itself.*

solution. Let $\delta_C(x)$ be the indicator function of a convex set C, defined as:

$$\delta_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{otherwise.} \end{cases}$$

From the **Biconjugate Theorem**, the biconjugate of any proper, closed, convex function f equals the function itself:

$$f^{**}(x) = f(x).$$

Since the indicator function $\delta_C(x)$ is closed, convex, and proper, we directly conclude:

$$\delta_C^{**}(x) = \delta_C(x).$$

Thus, the biconjugate of the indicator function of a convex set is the indicator function itself. □

3. *Recall that the infimal convolution of two proper functions f, g on X is given by:*

$$(f \square g)(y) = \inf_x \{f(x) + g(y - x)\}.$$

(a) ***Fenchel Conjugate of the Infimal Convolution:*** *Let f, g be proper on X. Then show that:*

$$(f \square g)^* = f^* + g^*.$$

solution. Fenchel Conjugate of the Infimal Convolution
The infimal convolution of two proper functions f and g is defined as:

$$(f \square g)(y) = \inf_{x \in \mathbb{R}^n} \{f(x) + g(y - x)\}.$$

To compute the Fenchel conjugate, we use the definition:

$$(f \square g)^*(z) = \sup_{y \in \mathbb{R}^n} \{\langle z, y \rangle - (f \square g)(y)\}.$$

Substituting the definition of $(f \square g)(y)$:

$$(f \square g)^*(z) = \sup_{y \in \mathbb{R}^n} \left\{ \langle z, y \rangle - \inf_{x \in \mathbb{R}^n} \{f(x) + g(y - x)\} \right\} = \sup_{y \in \mathbb{R}^n} \sup_{x \in \mathbb{R}^n} \{\langle z, y \rangle - f(x) - g(y - x)\}.$$

$$(f \square g)^*(z) = \sup_{x \in \mathbb{R}^n} \sup_{y \in \mathbb{R}^n} \{\langle z, y \rangle - f(x) - g(y - x)\} = \sup_{x \in \mathbb{R}^n} \left\{ \sup_{y \in \mathbb{R}^n} \{\langle z, y - x \rangle - g(y - x)\} + \langle z, x \rangle - f(x) \right\}.$$

$$= \sup_{x \in \mathbb{R}^n} g^*(z) + \langle z, x \rangle - f(x) = g^*(z) + \sup_{x \in \mathbb{R}^n} \{\langle z, x \rangle - f(x)\} = g^*(z) + f^*(z)$$

$$(f \square g)^*(z) = f^*(z) + g^*(z).$$

Thus, we conclude:

$$(f \square g)^* = f^* + g^*.$$

□

(b) ***Fenchel Conjugate of the Sum:*** *Let f, g be convex, lower semicontinuous, and proper on X. Assume that* $(\operatorname{ri}\operatorname{dom} f) \cap (\operatorname{ri}\operatorname{dom} g) \neq \emptyset$. *Then show that:*

$$(f+g)^* = f^* \square g^*.$$

And the infimum in the definition of the infimal convolution is attained, i.e., a minimum exists.

solution. by definition of infimal convolution:

$$f^* \square g^*(z) = \inf_{x \in \mathbb{R}^n} \{f^*(x) + g^*(z-x)\} = \inf_{x \in \mathbb{R}^n} \left\{ \sup_{y \in \mathbb{R}^n} \{\langle x, y\rangle - f(y)\} + \sup_{y \in \mathbb{R}^n} \{\langle z-x, y\rangle - g(y)\} \right\}.$$

$$= \inf_{x \in \mathbb{R}^n} \left\{ \sup_{y \in \mathbb{R}^n} \{\langle x, y\rangle + \langle z-x, y\rangle - f(y) - g(y)\} \right\} = \inf_{x \in \mathbb{R}^n} \left\{ \sup_{y \in \mathbb{R}^n} \{\langle z, y\rangle - f(y) - g(y)\} \right\} = \inf_{x \in \mathbb{R}^n} \{f(z) + g(z)\}^*$$

Therefore:

$$(f+g)^* = f^* \square g^*.$$

The condition that the infimum is attained ensures that the result holds rigorously.

□

2 optimality conditions

Definition 2.1. *A point $x^* \in \mathcal{X}$ is locally optimal if $f(x^*) \leq f(x)$ for all x in a neighborhood of x^*. Global if $f(x^*) \leq f(x)$ for all $x \in \mathcal{X}$.*

Theorem 2.2. *For convex f, locally optimal point also global.*

Proof.

- Let x^* be a local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.
- We need to show that $f(y) \geq f(x^*) = p^*$.
- $\mathcal{X}$ is convex, so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is convex, and $x^*, y \in \operatorname{dom} f$, we have

$$f(x_\theta) \leq \theta f(y) + (1-\theta) f(x^*)$$
$$f(x_\theta) - f(x^*) \leq \theta(f(y) - f(x^*)).$$

- Since x^* is a local minimizer, for small enough $\theta > 0$, lhs ≥ 0.
- So rhs is also nonnegative, proving $f(y) \geq f(x^*)$ as desired.

□

Definition 2.3. *The set of optimal solutions $\mathcal{X}^*$ may be empty. If $\mathcal{X} = \emptyset$, i.e., no feasible solutions, then $\mathcal{X}^* = \emptyset$. When only inf not min, e.g., $\inf e^x$ as $x \to -\infty$ in general, we should worry about the question "Is $\mathcal{X}^* = \emptyset$?"*

Exercise 2.4. *Verify that $\mathcal{X}^*$ is always a convex set.*

2.1 First-order conditions: unconstrained

Theorem 2.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.*

Proof. Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \to 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d\rangle.$$

Similarly, using $-d$ it follows that $\langle \nabla f(x^*), d\rangle \leq 0$, so $\langle \nabla f(x^*), d\rangle = 0$ *must* hold. Since d is arbitrary, $\nabla f(x^*) = 0$. □

Exercise 2.6. *Prove that if f is convex, then $\nabla f(x^*) = 0$ is actually sufficient for global optimality! For general f this is not true. (This property is what makes convex optimization special!)*

Exercise 2.7. *If $\nabla f(x^*) = 0$ implies x^* is global opt, must f be convex?*

2.2 First-order conditions: constrained

Theorem 2.8. *For convex f, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$. Thus, x^* is optimal if and only if*

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \quad \text{for all } y \in \mathcal{X}.$$

($\mathcal{X}$ is convex). If $\mathcal{X} = \mathbb{R}^n$, this reduces to $\nabla f(x^*) = 0$. If $\nabla f(x^*) \neq 0$, it defines supporting hyperplane to $\mathcal{X}$ at x^*

Proof. Let f be continuously differentiable, possibly nonconvex. Suppose $\exists y \in \mathcal{X}$ such that $\langle \nabla f(x^*), y - x^* \rangle < 0$. Using mean-value theorem of calculus, $\exists \xi \in [0, 1]$ such that

$$f(x^* + t(y - x^*)) = f(x^*) + \langle \nabla f(x^* + \xi t(y - x^*)), t(y - x^*) \rangle$$

(we applied MVT to $g(t) := f(x^* + t(y - x^*))$). For sufficiently small t, since ∇f continuous, by assumption on y,

$$\langle \nabla f(x^* + \xi t(y - x^*)), y - x^* \rangle < 0$$

This in turn implies that $f(x^* + t(y - x^*)) < f(x^*)$. Since $\mathcal{X}$ is convex, $x^* + t(y - x^*) \in \mathcal{X}$ is also feasible. Contradiction to local optimality of x^*

□

2.3 Optimality without differentiability

Theorem 2.9 (Fermat's rule)**.** *let $f : \mathbb{R}^n \to (-\infty, \infty]$. then,*

$$Argmin f = zer(\partial f) := \{x \in \mathbb{R}^n | 0 \in \partial f(x)\}$$

Proof. $x \in$ Argmin f implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$. equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle, \forall y, \Leftrightarrow 0 \in \partial f(x)$ □

Example 2.10 (nonsmooth problem)**.**

$$\min_x f(x) \quad s.t. \quad x \in \mathcal{X}$$

$$\min_x f(x) + \mathbf{1}_{\mathcal{X}}(x)$$

by the theorem, we know that minimizing x must satisfy: $0 \in \partial(f + \mathbf{1})(x)$. Assuming

$$ri(dom f) \cap ri(dom \mathcal{X}) \neq \emptyset, 0 \in \partial f(x) + \partial 1_{\mathcal{X}}(x)$$

recall $g \in \partial_{\mathbf{1}_{\mathcal{X}}}(x)$ iff $\mathbf{1}_{\mathcal{X}}(y) \geq 1_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y. so $g \in \partial_{\mathbf{1}_{\mathcal{X}}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle$, $\forall y \in \mathcal{X}$. we define the subdifferential of the indicator $\mathbf{1}_{\mathcal{X}}(x)$, aka normal cone:

$$\mathcal{N}_{\mathcal{X}}(x) := \{g \in \mathbb{R}^n | 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X}\}$$

if f is differentiable, we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$.
$-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*) \Leftrightarrow \langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in \mathcal{X}$.

2.4 Optimality conditions via KKT

consider the problem:

$$\min f(x), \quad \text{subject to} \quad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$.
- Can we simplify this using Lagrangian?
- Let $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)$
- Let $d^* := \sup_\lambda g(\lambda)$.

Assume strong duality and that p^*, d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f(x^*) = p^*.$$

Thus, equalities hold in the above chain, and

$$x^* \in \arg\min_x \mathcal{L}(x, \lambda^*).$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)\big|_{x=x^*} = \nabla f(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f(x^*)$, we also have

$$\sum_i \lambda_i^* f_i(x^*) = 0.$$

But $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$, so *complementary slackness*

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m.$$

Definition 2.11 (**Karush-Kuhn-Tucker Conditions (KKT)**).

$$f_i(x^*) \leq 0, \quad i = 1, \ldots, m \quad \textit{(primal feasibility)}$$

$$\lambda_i^* \geq 0, \quad i = 1, \ldots, m \quad \textit{(dual feasibility)}$$

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m \quad \textit{(complementary slackness)}$$

$$\nabla_x \mathcal{L}(x, \lambda^*)\big|_{x=x^*} = 0 \quad \textit{(Lagrangian stationarity)}$$

- Thus, if strong duality holds, and (x^*, λ^*) exists, then KKT conditions are necessary for pair (x^*, λ^*) to be optimal.
- If problem is convex, then KKT also sufficient.

Exercise 2.12. *Prove the above sufficiency of KKT.*
Hint: Use that $\mathcal{L}(x, \lambda^*)$ is convex, and conclude from KKT conditions that $g(\lambda^*) = f_0(x^*)$, so that (x^*, λ^*) optimal primal-dual pair.

Example 2.13.

$$\min_x \quad \frac{1}{2}\|x - y\|^2, \quad s.t. \quad x^T 1 = 1, \quad x \geq 0.$$

KKT Conditions:

$$\mathcal{L}(x, \lambda, \nu) = \frac{1}{2}\|x - y\|^2 - \sum_i \lambda_i x_i + \nu(x^T 1 - 1)$$

$$\frac{\partial \mathcal{L}}{\partial x_i} = x_i - y_i - \lambda_i + \nu = 0$$

$$\lambda_i x_i = 0$$

$$\lambda_i \geq 0$$

$$x^T 1 = 1, \quad x \geq 0$$

Challenge A. *Solve the above conditions in $O(n \log n)$ time.*
Challenge A+. *Solve the above conditions in $O(n)$ time.*

2.5 Nonsmooth KKT: via subdifferentials

3 nonconvex optimality, stationarity

3.1 tractable nonconvex problems

4 first-order methods

4.1 Gradient Descent

Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$. Consider updating x using

$$x(\eta) = x + \eta d,$$

where **direction** $d \in \mathbb{R}^n$ is obtuse to $\nabla f(x)$, i.e., $\langle \nabla f(x), d \rangle < 0$. and η is the stepsize.

Using the Taylor expansion:

$$f(x(\eta)) = f(x) + \eta \langle \nabla f(x), d \rangle + o(\eta),$$

where $\langle \nabla f(x), d \rangle$ dominates $o(\eta)$ for small η. Since d is obtuse to $\nabla f(x)$, this implies $f(x(\eta)) < f(x)$.

The iterative update is given by:

$$x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots$$

where stepsize $\eta_k \geq 0$ usually ensures $f(x^{k+1}) < f(x^k)$, and descent direction d^k satisfies $\langle \nabla f(x^k), d^k \rangle < 0$.

Numerous ways exist to select η_k and d^k, with many methods seeking monotonic descent:

$$f(x^{k+1}) < f(x^k).$$

4.1.1 Descent direction

Scaled gradient: $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$

- Newton's method: $(D^k = [\nabla^2 f(x^k)]^{-1})$
- Quasi-Newton: $D^k \approx [\nabla^2 f(x^k)]^{-1}$
- Steepest descent: $D^k = I$
- Diagonally scaled: D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \right)^{-1}$
- Discretized Newton: $D^k = [H(x^k)]^{-1}$, H is the Hessian via finite-difference method: $\nabla^2 f(x) \simeq \frac{\nabla f(x+h) - \nabla f(x)}{h}$.
- ...

Exercise 4.1. *Verify that $\langle \nabla f(x^k), d^k \rangle < 0$ for above choices*

4.1.2 Stepsize selection

- Constant: $\eta_k = 1/L$ (for suitable value of L)
- Diminishing: $\eta_k \to 0$ but $\sum_k \eta_k = \infty$.

Exercise 4.2. *Prove that the latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.*

Sketch: Say, $x^k \to \bar{x}$; then for sufficiently large m and n, $(m > n)$

$$x^m \approx x^n \approx \bar{x}, x^m \approx x^n - \left(\sum_{k=n}^{m-1} \eta_k \right) \nabla f(\bar{x}).$$

The sum can be made arbitrarily large, contradicting nonstationarity of $\bar{x}$

- Exact: $\eta_k := \arg\min_{\eta \geq 0} f(x^k + \eta d^k)$
- Limited min: $\eta_k = \arg\min_{0 \leq \eta \leq s} f(x^k + \eta d^k)$

- Armijo-rule. Given fixed scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\eta_k = \beta^{m_k} s,$$

where we try $\beta^m s$ for $m = 0, 1, \ldots$ until sufficient descent

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k), d^k \rangle$$

If $\langle \nabla f(x^k), d^k \rangle < 0$, stepsize guaranteed to exist

Usually, σ small $\in [10^{-5}, 0.1]$, while β from 1/2 to 1/10 depending on how confident we are about initial stepsize s.

Exercise 4.3.

- *Let D be the $(n-1) \times n$ differencing matrix* $D = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \\ & & & -1 \quad 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n}$,
- $f(x) = \frac{1}{2}\|D^T x - b\|_2^2 = \frac{1}{2}(\|D^T x\|_2^2 + \|b\|_2^2 - 2\langle D^T x, b\rangle)$
- *Notice that* $\nabla f(x) = D(D^T x - b)$
- *Try different choices of b, and different initial vectors x_0*
- *Exercise: Experiment to see how large n must be before gradient method starts outperforming CVXPY*
- *Exercise : Minimize $f(x)$ for large n; e.g., $n = 10^6, n = 10^7$*
- *Exercise : Repeat same exercise with constraints: $x_i \in [-1, 1]$.*

4.1.3 Gradient descent – convergence

Definition 4.4 (Lipschitz continuous gradient)**.** *we say f has Lipschitz continuous gradient, $f \in C_L^1$ if $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$*

Exercise 4.5. *show that if $f \in C_L^1$ is twice differentiable, then $\|\nabla^2 f(x)\|_2 \leq L$*

Lemma 4.6. *Let $f \in \mathcal{C}_L^1$. Then,*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|_2^2.$$

Proof. By Taylor's theorem, for $z_t = y + t(x - y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t), x - y\rangle dt.$$

Adding and subtracting $\langle \nabla f(y), x - y\rangle$ we obtain

$$|f(x) - f(y) - \langle \nabla f(y), x - y\rangle| = \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(y), x - y\rangle dt \right|.$$

$$\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(y), x - y\rangle| dt$$

$$\leq \int_0^1 \|\nabla f(z_t) - \nabla f(y)\|_2 \|x - y\|_2 dt$$

$$\leq L \int_0^1 t\|x - y\|_2^2 dt$$

$$= \frac{L}{2}\|x - y\|_2^2.$$

Thus, $f(x)$ is bounded above and below with quadratic functions.

□

Corollary 4.7. *If $f \in \mathcal{C}_L^1$, and $0 < \eta_k < 2/L$, then $f(x^{k+1}) < f(x^k)$.*

Proof.

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2$$

$$= f(x^k) - \eta_k \|\nabla f(x^k)\|_2^2 + \frac{\eta_k^2 L}{2}\|\nabla f(x^k)\|_2^2$$

$$= f(x^k) - \eta_k \left(1 - \frac{\eta_k}{2} L\right) \|\nabla f(x^k)\|_2^2.$$

If $\eta_k < 2/L$, we have descent. Minimize over η_k to get the best bound, giving $\eta_k = 1/L$

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|_2^2$$

.

□

Theorem 4.8. *Let $f \in \mathcal{C}_L^1$. $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$.*

Proof. directly get from next theorem. □

Theorem 4.9. *Let $f \in \mathcal{C}_L^1$. $\min_{1 \le k \le T} \|\nabla f(x^k)\|_2 = O(1/T)$.*

Proof. We showed that from the last corollary:

$$f(x^k) - f(x^{k+1}) \ge \frac{1}{2L}\|\nabla f(x^k)\|_2^2,$$

Sum up the above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{1}{2L}\sum_{k=0}^{T} \|\nabla f(x^k)\|_2^2 \le f(x^0) - f(x^{T+1}) \le f(x^0) - f^*.$$

We assume $f^* > -\infty$, so the right-hand side is some fixed positive constant. Thus, as $k \to \infty$, the left-hand side must converge; thus

$$\|\nabla f(x^k)\|_2 \to 0 \quad \text{as } k \to \infty.$$

$$\min_{0 \le k \le T} \|\nabla f(x^k)\|_2^2 \le \frac{1}{T+1}\sum_{k=0}^{T} \|\nabla f(x^k)\|_2^2,$$

so $O\left(\frac{1}{\epsilon}\right)$ for $\|\nabla f\|^2 \le \epsilon$.

Notice, we did not require f to be convex...

□

Corollary 4.10. *If f is a convex function $\in \mathcal{C}_L^1$, then*

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \le \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Proof. □

Definition 4.11 (Strong convexity)**.** *f is said to be strong convex, denoted $f \in \mathcal{S}_{L,\mu}^1$ if*

$$f(x) \ge f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2$$

next we consider the convergence rate, the below theorem is useful in following proof.

Theorem 4.12. *Suppose $f \in \mathcal{S}^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$,*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2.$$

Proof. using corollary 5.10 □

Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - \eta \nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\eta \langle \nabla f(x^k), x^k - x^* \rangle + \eta^2 \|\nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\eta \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \eta^2 \|\nabla f(x^k)\|_2^2 \\ &\leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right) r_k^2 + \eta \left(\eta - \frac{2}{\mu + L}\right) \|\nabla f(x^k)\|_2^2 \end{aligned}$$

where we used **Thm 5.12** with $\nabla f(x^*) = 0$ for the last inequality. Now assume the stepsize $\eta < \frac{2}{\mu + L}$, so that we obtain:

$$r_{k+1}^2 \leq \theta r_k^2 \implies r_k^2 \leq \theta^k r_0^2, \quad \theta = \left(1 - \frac{2\eta\mu L}{\mu + L}\right)$$

we summarize the discussion by giving the following theorem

Theorem 4.13. *If $f \in \mathcal{S}^1_{L,\mu}$, $0 < \eta < 2/(L + \mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies*

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2^2.$$

Moreover, if $\eta = 2/(L + \mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2} \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa := L/\mu$ is the condition number.

Proof. the first inequality is proved above. the second inequality involving $f(x^k) - f^*$ is by Lipschitz continuity (**Exercise 5.5**) and $\nabla f(x^*) = 0$:

$$f(x^k) - f(x^*) = \nabla f(x^*)(x^k - x^*) + \frac{\mu}{2} \nabla^2 f(x) \|x^k - x^*\|_2^2 \leq 0 + \frac{L}{2} \|x^k - x^*\|_2^2$$

$$\leq \frac{L}{2} \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2^2 = \frac{L}{2} \left(1 - \frac{4\mu L}{(\mu + L)^2}\right)^k \|x^0 - x^*\|_2^2 = \frac{L}{2} \left(\frac{\mu - L}{\mu + L}\right)^{2k} \|x^0 - x^*\|_2^2$$

□

we now do not use strong convexity and want to prove the well-known $O(1/T)$ rate
⋆ Let $\eta_k = 1/L$
⋆ Shorthand notation $g^k = \nabla f(x^k), g^* = \nabla f(x^*)$
⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma 4.14. *Distance to min shrinks monotonically; $r_{k+1} \leq r_k$*

Proof. Descent lemma implies that:

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L} \|g^k\|_2^2.$$

Consider,

$$r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2.$$

$$r_{k+1}^2 = r_k^2 + \eta_k^2 \|g^k\|_2^2 - 2\eta_k \langle g^k, x^k - x^* \rangle$$

$$= r_k^2 + \eta_k^2 \|g^k\|_2^2 - 2\eta_k \langle g^k - g^*, x^k - x^* \rangle \quad \text{as } g^* = 0$$

$$\leq r_k^2 + \eta_k^2 \|g^k\|_2^2 - \frac{2\eta_k}{L} \|g^k - g^*\|_2^2 \quad (\textit{Exercise : Why?})$$

$$= r_k^2 - \eta_k \left(\frac{2}{L} - \eta_k \right) \|g^k\|_2^2.$$

Since $\eta_k < \frac{2}{L}$, it follows that $r_{k+1} \leq r_k$. □

Lemma 4.15. *Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1 - \beta_k)$.*

Proof.

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\leq} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is,

$$\|g^k\|_2 \geq \Delta_k / r_k.$$

In particular, since $r_k \leq r_0$, we have

$$\|g^k\|_2 \geq \Delta_k / r_0.$$

Now we have a bound on the gradient norm... Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides:

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k \left(1 - \frac{\Delta_k}{2Lr_0^2} \right) = \Delta_k(1 - \beta_k).$$

□

Theorem 4.16. *Let $f \in \mathcal{C}_L^1$ be convex; let $\{x^k\}$ be generated as above, with $\eta_k = 1/L$. Then, $f(x^{T+1}) - f(x^*) = O(1/T)$.*

Proof. **we want to bound:** $f(x^{T+1}) - f(x^*)$. from the above lemma:

$$\Delta_{k+1} \leq \Delta_k(1 - \beta_k) \implies \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k}(1 + \beta_k) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}.$$

- Sum both sides over $k = 0, \ldots, T$ (telescoping) to obtain:

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}.$$

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

- Rearrange to conclude:

$$f(x^T) - f^* \leq \frac{2L\Delta_0 r_0^2}{2Lr_0^2 + T\Delta_0}.$$

- Use descent lemma to bound $\Delta_0 \leq (L/2)\|x^0 - x^*\|_2^2$; simplify:

$$f(x^T) - f(x^*) \leq \frac{2L\Delta_0 \|x^0 - x^*\|_2^2}{T+4} = O(1/T).$$

□

4.2 Subgradient method

Does "gradient descent" still work if $f(x)$ is not differentiable: $f \in C^0_G$?

$$x^{k+1} = x^k - \eta_k g^k$$

where $g^k \in \partial f(x^k)$ is any subgradient.

Stepsize $\eta_k > 0$ must be chosen

Method generates sequence $\{x^k\}_{k\geq 0}$:

- Does this sequence converge to an optimal solution x^*?
- If yes, then how fast?
- What if we have constraints: $x \in \mathcal{C}$?

Example 4.17 (lasso regression)**.**

$$\min \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

$$x^{k+1} = x^k - \eta_k \left(A^T(Ax^k - b) + \lambda\, sgn(x^k)\right)$$

4.2.1 Subgradient method – stepsizes

- **Constant**: Set $\eta_k = \eta > 0$, for $k \geq 0$.
- **Normalized**: $\eta_k = \eta/\|\nabla f^k\|_2 \quad (\|x^{k+1} - x^k\|_2 = \eta)$.
- **Square summable**:
$$\sum_k \eta_k^2 < \infty, \quad \sum_k \eta_k = \infty.$$
- **Diminishing**:
$$\lim_{k\to\infty} \eta_k = 0, \quad \sum_k \eta_k = \infty.$$
- **Adaptive stepsizes**: (not covered).
- **Not a descent method!**
Could use best f^k so far: $f^k_{\min} := \min_{0\leq i\leq k} f^i$

4.2.2 Subgradient method – convergence

Theorem 4.18. *Assume that f is convex, and*

1. *min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$.*
2. *$\mathrm{dom}(f)$ is bounded: $\|x^0 - x^*\|_2 \leq R$,*
3. *subgradients is bounded $\|g\|_2 \leq G$ for all $g \in \partial f$.*

Then for a fixed step size k, let $f^k_{min} := \min_{i=0,1,\ldots,k} f^i$, the subgradient method satisfies:

$$f^k_{min} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_k^2}{2\sum_{t=1}^k \eta_k}$$

proof idea: Bound the distance between each iterate and the optimal solution using the update rule $x^{(t)} = x^{(k-1)} - \eta_k g^{(k-1)}$.

Proof. We know that $\eta_k g^{(k-1)\top}(x^{k-1} - x^*) \geq f(x^{k-1}) - f^*$, then:

$$\begin{aligned}\|x^k - x^*\|_2^2 &= \|x^{k-1} - \eta_k g^{k-1} - x^*\|_2^2 \\ &= \|x^{k-1} - x^*\|_2^2 - 2\eta_k (g^{k-1})^\top (x^{k-1} - x^*) + \eta_k^2 \|g^{k-1}\|_2^2 \\ &\leq \|x^{k-1} - x^*\|_2^2 - 2\eta_k (f(x^{k-1}) - f(x^*)) + \eta_k^2 \|g^{k-1}\|_2^2.\end{aligned}$$

Iterating, we have:

$$\|x^k - x^*\|_2^2 \leq \|x^0 - x^*\|_2^2 - 2\sum_{i=1}^{k} \eta_i (f(x^{i-1}) - f(x^*)) + \eta_i^2 \sum_{i=1}^{k} \|g^{i-1}\|_2^2.$$

by assumption we have $\|g\|_2 \leq G$ so $\|g^{i-1}\|_2^2 \leq G^2$ for all i, and $\|x^k - x^*\|_2^2 \geq 0$, also define $R \equiv \|x^0 - x^*\|_2$:

$$0 \leq R^2 - 2\sum_{i=1}^{k} \eta_i (f(x^{i-1}) - f(x^*)) + G^2 \sum_{i=1}^{k} \eta_i^2.$$

Resulting in the *basic equation* from which we can read off all our convergence results:

$$f(x_{\text{best}}^k) - f(x^*) \leq \frac{R^2 + G^2 \sum_{i=1}^k \eta_i^2}{2\sum_{i=1}^k \eta_i}.$$

□

Suppose we want $f_{\min}^k - f^* \leq \epsilon$, how big should k be? Optimize the bound for η_t: want

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2\sum_{t=1}^k \eta_t} \leq \epsilon.$$

For fixed k: best possible stepsize is constant η

$$\frac{R^2 + G^2 k\eta^2}{2k\eta} \leq \epsilon \implies \eta = \frac{R}{G\sqrt{k}}.$$

Then, after k steps $f_{\min}^k - f^* \leq \frac{RG}{\sqrt{k}} \leq \epsilon \Rightarrow k \geq (\frac{RG}{\epsilon})^2$.
For accuracy ϵ, we need at least $(RG/\epsilon)^2 = O(1/\epsilon^2)$ steps
(quite slow but already hits the lower bound!) Then the convergence rate is $O(1/\epsilon^2)$, which is much slower than gradient descent $O(1/\epsilon)$. We can't improve this convergence rate by varying step size or choosing unbalanced allocations.

Exercise 4.19. *analyze* $\lim_{k\to\infty} f_{min}^k - f^*$ *for the different choices of stepsize that we mentioned.*

4.2.3 Projected subgradient method

we add a constraint on the domain:

$$\min f(x) \quad s.t. \quad x \in \mathcal{C}$$

the previously gradient descent: $x^{t+1} = x^t - \eta_t g^t$ could be infeasible!

Definition 4.20 (Projection to the closest feasible point)**.**

$$P_C(x) = \arg\min_{y \in C} \|x - y\|^2$$

(Assume C is closed and convex, then projection is unique.)

the update algorithm is

$$x^{k+1} = P_C(x^k - \eta_k g^k)$$

where $g^k \in \partial f(x^k)$ is any subgradient. it is great as long as projection is "easy".
we ask the Same questions as before:

- **Does it converge? For which stepsizes? How fast?**

the following property of projection is important in the following convergence analysis.

Theorem 4.21.

$$\|P_C(x_1) - P_C(x_2)\|_2 \leq \|x_1 - x_2\|_2$$

(aka: non-expansivity of projections)
which will follow from the more general result:

$$\|P_C(x_1) - P_C(x_2)\|_2^2 \leq \langle P_C(x_1) - P_C(x_2), x_1 - x_2\rangle$$

(aka: firm non-expansivity of projections)

Proof. we first prove firm non-expansivity of projections:
Let C be a closed, convex set. From first-order optimality conditions

$$\langle \nabla f(x^*), x - x^*\rangle \geq 0, \quad \forall x \in C.$$

Thus, noting $\nabla f(x^*) = x^* - y$ and $x^* = P_C(y)$, we get

$$\langle y - P_C(y), x - P_C(y)\rangle \leq 0, \quad \forall x \in C.$$

Using the above inequality, for two points x_1, x_2 we obtain

$$\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1)\rangle \leq 0,$$

$$\langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2)\rangle \leq 0,$$

$$\langle P_C(x_1) - P_C(x_2), x_2 - x_1 + P_C(x_1) - P_C(x_2)\rangle \leq 0.$$

Combining these results:

$$\|P_C(x_1) - P_C(x_2)\|_2^2 \leq \langle P_C(x_1) - P_C(x_2), x_1 - x_2\rangle.$$

□

Exercise 4.22. *Prove that NE follows from this FNE property.*

Same as the above assumption: Assume that f is convex, and

1. min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$.
2. $\mathrm{dom}(f)$ is bounded: $\|x^0 - x^*\|_2 \leq R$,
3. subgradients is bounded $\|g\|_2 \leq G$ for all $g \in \partial f$.

let $z^{t+1} = x^t - \eta_t g^t$, then $x^{t+1} = P_C(z^{t+1})$. we care about $\|x^{t+1} - x^*\|$.

Actually we only need to use the nonexpansiveness of projection:

$$\|x^{t+1} - x^*\|_2^2 = \|P_C(x^t - \eta_t g^t) - P_C(x^*)\|_2^2 \leq \|x^t - \eta_t g^t - x^*\|_2^2$$

we can reach the same convergence results as in unconstrained case.

Example 4.23 (examples of simple projections)**.**

- ***Nonnegativity****:* $x \geq 0$, $P_C(z) = [z]_+$.
- ℓ_∞***-ball****:* $\|x\|_\infty \leq 1$
  - *Projection:* $\min \|x - z\|^2$ *s.t.* $x \leq 1$ *and* $x \geq -1$
  - $P_{\|x\|_\infty \leq 1}(z) = y$ *where* $y_i = sgn(z_i)\min\{|z_i|, 1\}$
- ***Linear equality constraints****:* $Ax = b$ $\quad$ *(*$A \in \mathbb{R}^{n\times m}$ *has rank* n*)*

$$P_C(x) = z - A^T(AA^T)^{-1}(Az - b)$$

$$= (I - A^T(AA^T)^{-1}A)z + A^T(AA^T)^{-1}b$$

- ***Simplex****:* $x^\top 1 = 1$ *and* $x \geq 0$
  - *Doable in* $O(n)$ *time; similarly for* ℓ_1*-norm ball.*

5 Constrained Optimization via Frank-Wolfe methods

recall the projected GD method:

$$x^{k+1} = P_{\mathcal{M}}(x^k - \frac{1}{L}\nabla f(x^k)) = \arg\min_{x\in\mathcal{M}} \left(f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|^2 \right)$$

this form is actually equivalent to the original projected method:

$$x^{k+1} = P_{\mathcal{M}} \left(x^k - \frac{1}{L}\nabla f(x^k) \right)$$

where $P_{\mathcal{M}}(x) \triangleq \arg\min_{y\in\mathcal{M}} \|x - y\|^2$. Then we have

$$x^{k+1} = P_{\mathcal{M}} \left(x^k - \frac{1}{L}\nabla f(x^k) \right) = \arg\min_{x\in\mathcal{M}} \left(\|x - (x^k - \frac{1}{L}\nabla f(x^k))\|^2 \right)$$

$$= \arg\min_{x\in\mathcal{M}} \left(\|x - x^k\|^2 + \frac{1}{L^2}\|\nabla f(x^k)\|^2 + \frac{2}{L}\langle \nabla f(x^k), x - x^k \rangle \right) = \arg\min_{x\in\mathcal{M}} \left(\|x - x^k\|^2 + \frac{2}{L}\langle \nabla f(x^k), x - x^k \rangle \right)$$

note that $f(x^k)$ can be omitted in the first equation, and after multiplying a parameter $\frac{2}{L}$ we get the second form of projection.

5.1 Frank-Wolfe method

Instead of using a quadratic expansion as shown above for the projected GD method, it uses a local linear expansion of f.

$$s^k \in \arg\min_{x\in\mathcal{M}} f(x^k) + \langle \nabla f(x^k), x - x^k \rangle, \quad x^{k+1} = x^k + \eta(s^k - x^k), \quad \eta \in [0, 1]$$

in which we omit the quadratic term in the first equation. we write the FW method in a more standard way:

1. Start with some guess $x^0 \in \mathcal{M}$.
2. Form linear approximation of f at x^k:

   $$\phi_f(x, x^k) := f(x^k) + \langle \nabla f(x^k), x - x^k \rangle = \langle \nabla f(x^k), x \rangle = \nabla f(x^k)^T x$$

   here we omit all term irrelevant to x
3. Find $s^k \in \arg\min_{x\in\mathcal{M}} \phi_f(x, x^k)$.
4. Take convex combination of x^k and s^k:

   $$x^{k+1} = x^k + \eta_k(s^k - x^k), \quad \eta_k \in [0, 1]$$

5. Repeat above procedure until $f(x^k) \le f(x^*) + \epsilon$.

5.2 norm constraints

Now, we look at how the Frank-Wolfe method deals with optimization problems where the constraint is on the norm of the solution, i.e., what happens when

$$C = \{x : \|x\| \le t\}$$

for a norm $\|\cdot\|$. By definition of the problem, we have

$$s \in \arg\min_{\|s\|\le t} \nabla f(x^{(k-1)})^T s.$$

We see that:

$$\min_{\|s\|\le t} \nabla f(x^{(k-1)})^T s = -\max_{\|s\|\le t} \nabla f(x^{(k-1)})^T s = -t \cdot \max_{\|z\|\le 1} \nabla f(x^{(k-1)})^T z.$$

Therefore:

$$\arg\min_{\|s\|\le t} \nabla f(x^{(k-1)})^T s = -t \cdot \arg\max_{\|z\|\le 1} \nabla f(x^{(k-1)})^T z,$$

which is equal to the subdifferential of the dual norm. The dual norm is written as:

$$\|z\|_* = \max_{\|x\| \leq 1} x^T z.$$

Hence, following the rule of the subgradients of the max function, we have:

$$s = -t \cdot \left(\arg\max_{\|s\| \leq t} \nabla f(x^{(k-1)})^T s \right) = -t \cdot \delta \|\nabla f(x^{(k-1)})\|_*,$$

where $\|\cdot\|_*$ denotes the corresponding dual norm. That is, if we know how to compute the subgradients of the dual norm, then we can easily perform the Frank-Wolfe steps. Since we have a closed-form update now, this can often be much cheaper or simpler than projection onto $C = \{x : \|x\| \leq t\}$. We now look at some examples of norm-based constraints and how to derive the Frank-Wolfe method in these special cases.

Example 5.1 (ℓ_1-norm regularization)**.** *we are solving the optimization problem:*

$$\min_{\|x\|_1 \leq \gamma} f(x)$$

note that the specialty is $\mathcal{M} \triangleq \{x : \|x\|_1 \leq \gamma\}$*. so we only focus on step 3:*

1. *choose* $i_k \in \arg\max_{1 \leq i \leq d} |\nabla_i f(x^k)|$
2. *set* $s_i^k = -\gamma \ sign(\nabla_{i_k} f(x^k)) e_{i_k}$ *for* $i = i_k$ *and* 0 *otherwise.*

The dual norm of the ℓ_1 *norm is the* ℓ_∞ *norm. So, we have:*

$$s^{(k-1)} \in -t\delta \|\nabla f(x^{(k-1)})\|_\infty.$$

The Frank-Wolfe update requires the subgradient of the ℓ_∞ *norm. If* $\nabla f(x^{(k-1)})$ *has its component-wise maximum at* i*, then the subgradient is the standard basis vector* e_i*. This can be written as:*

$$i_{k-1} \in \arg\max_{i=1,\dots,p} |\nabla_i f(x^{(k-1)})|$$

$$x^{(k)} = (1 - \gamma_k) x^{(k-1)} - \gamma_k t \cdot sign(\nabla_{i_{k-1}} f(x^{(k-1)})) \cdot e_{i_{k-1}}.$$

This looks like greedy coordinate descent since coordinate descent goes through the vector cyclically, whereas Frank-Wolfe here picks the largest component at each iteration. Note that this update is simpler than projecting onto a ℓ_1*-ball, although they both have the same computational complexity, i.e.,* $O(d)$*,* d *is the dimension of* f*.*

Example 5.2 (ℓ_p-norm regularization)**.** *More generally, for the* ℓ_p*-regularized problem:*

$$\min_x f(x) \quad \textit{subject to} \quad \|x\|_p \leq t,$$

for $1 \leq p \leq \infty$*, we have:*

$$s^{(k-1)} \in -t\delta \|\nabla f(x^{(k-1)})\|_q,$$

where p, q *are such that* $\|\cdot\|_p$ *is the dual of* $\|\cdot\|_q$*. Recall that this is true if and only if* $1/p + 1/q = 1$*. It is interesting to note that the subgradient of a given dual norm can be computed efficiently using the following function, where* α *is a scaling factor and the rest is the scaled form of the subgradient of max over* ℓ_q *norm and recall the subgradient of the* ℓ_q*-norm can be written as:* $g_i \propto sign(z_i) \cdot |z_i|^{q-1}$*. :*

$$s_i^{(k-1)} = -\alpha \cdot sign(\nabla_i f(x^{(k-1)})) \cdot |\nabla_i f(x^{(k-1)})|^{q-1}, \quad i = 1, \cdots, n,$$

where α *is such that the constraint is satisfied, i.e.,* $\|s^{(k-1)}\|_q = t$*.*

This is followed by the main update with a convex combination (γ_k) *of* $(s^{(k-1)}, x^{(k-1)})$*. Note that these update rules are a lot simpler than projection onto the* ℓ_p*-ball for any general* p*, since there exists no general projection rule. Aside from special cases* $(p = 1, 2, \infty)$*, these projections cannot be computed directly; the projection step must be treated as an optimization.*

Example 5.3 (Trace-norm regularization)**.** *Here we discuss an example of Frank-Wolfe on a matrix-valued problem, where you have a much cheaper linear optimization oracle vs a projection (very big difference comparatively). The trace-regularized problem takes the following form:*

$$\min_X f(X) \quad \textit{subject to } \|X\|_{tr} \leq t$$

Recall that the trace norm is the sum of the singular values of X. Now, to derive our update S we need to (1) compute the dual norm, (2) find the subgradients of the resultant dual norm. We first note that the dual of the trace norm is the operator norm (largest singular value of X), and obtain the following general update:

$$S^{(k-1)} \in -t\partial \left\| \nabla f(X^{(k-1)}) \right\|_{op}$$

We now claim that the update $S^{(k-1)}$ can be explicitly written as:

$$S^{(k-1)} = -tuv^T,$$

where u, v are the leading left and right singular vectors of $\nabla f(X^{(k-1)})$ (recall proof from previous lectures). Hence, this means that:

$$\partial \left\| \nabla f(X^{(k-1)}) \right\|_{op} = uv^T.$$

Further, note that we can compute u, v using the power method on $\nabla f(X^{(k-1)})$, which is very cheap if the matrix is sparse.

We next consider the alternative of projection onto the norm ball. This would require computing the full SVD, which is more complex and expensive than Frank-Wolfe.

5.3 Stepsize selection

- **Oblivious:** Set $\eta_k = 2/(k+2), \quad$ for $k \geq 0$
  (Not a descent method)
- **Exact line-search:**

$$\eta_k \in \arg\min_{\eta \in [0,1]} f(x^k + \eta(s^k - x^k))$$

  Exercise 5.4. $f(x) = \frac{1}{2}\|Ax - b\|^2$

$$\implies \eta_k = clip_{[0,1]}\left(\frac{\langle Ax^k - As^k, Ax^k - b \rangle}{\|Ax^k - As^k\|^2} \right)$$

- **Approx. line-search:**

$$\eta_k = \text{clip}_{[0,1]}\left(\frac{\langle \nabla f(x^k), x^k - s^k \rangle}{LR^2} \right)$$

- **Fully corrective*:** Solve

$$x^{k+1} \in \arg\min_{x \in \text{conv}\{s^0, s^1, \ldots, s^k\}} f(x)$$

5.4 Convergence analysis

Theorem 5.5. *Let $f \in \mathcal{C}_L^1$ be convex; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,*

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}.$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$:

$$\begin{aligned}
& f(x^{k+1}) - f^* \\
& \leq f(x^k) - f^* + \eta_k \langle \nabla f(x^k), s^k - x^k \rangle + \frac{1}{2}\eta_k^2 L \|s^k - x^k\|^2 \\
& \leq f(x^k) - f^* + \eta_k \langle \nabla f(x^k), s^k - x^k \rangle + \frac{1}{2}\eta_k^2 LR^2 \\
& \leq f(x^k) - f^* + \eta_k \langle \nabla f(x^k), x^* - x^k \rangle + \frac{1}{2}\eta_k^2 LR^2 \\
& \leq f(x^k) - f^* + \eta_k (f(x^*) - f(x^k)) + \frac{1}{2}\eta_k^2 LR^2 \\
& = (1 - \eta_k)(f(x^k) - f(x^*)) + \frac{1}{2}\eta_k^2 LR^2.
\end{aligned}$$

Verify: Inductively, this leads to:

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}.$$

□

5.5 Invariance under affine transforms

One of the important properties that are not shared with the projected gradient method is affine invariance. For nonsingular matrix A, define $x = Ax'$ and $F(x') = f(x) = f(Ax')$. Consider Frank-Wolfe on F and $x \in C, x = Ax' \iff x' \in A^{-1}C$, we have:

$$s' = \arg\min_{z \in A^{-1}C} \nabla F(x')^T z \tag{1}$$

$$(x')^+ = (1-\gamma)x' + \gamma s'$$

By multiplying A on both sides:

$$A(x')^+ = (1-\gamma)Ax' + \gamma As'$$

and then by applying $\nabla F(x') = A^T \nabla f(Ax')$,

$$As' = A \cdot \arg\min_{z \in A^{-1}C} \nabla F(x')^T z$$

$$= A \cdot \arg\min_{Az \in C} \nabla f(Ax')^T Az$$

$$= AA^{-1} \arg\min_{w \in C} \nabla f(Ax')^T w$$

$$= \arg\min_{w \in C} \nabla f(Ax')^T w \tag{2}$$

which produces the same Frank-Wolfe update as that from f by comparing (1) and (2).

5.6 Curvature constant

5.7 Stopping criterion

- **(Optimality condition)** Recall from Lecture 4:

$$\langle \nabla f(x^*), x^* - x \rangle \leq 0, \quad \forall x \in \mathcal{M}.$$

- **(Definition)** Frank-Wolfe gap / directional derivative:

$$G_{\text{FW}}(x^k) = \max_{x \in \mathcal{M}} \langle \nabla f(x^k), x^k - x \rangle = \langle \nabla f(x^k), x^k - s^k \rangle.$$

  - $G_{\text{FW}}(x) \geq 0$ for all $x \in \mathcal{M}$.
  - $G_{\text{FW}}(x) = 0$ iff $x = x^*$.

  we can show that these are upper bounds on $f(x^k) - f^*$:

  Exercise 5.6. *If f is convex, then:*

$$f(x^k) - f^* \leq G_{FW}(x^k).$$

Proof. By the first-order convexity condition:

$$f(s) \geq f(x^{(k)}) + \nabla f(x^{(k)})^T (s - x^{(k)}).$$

Minimizing both sides over all $s \in \mathcal{C}$ yields:

$$f^* \geq f(x^{(k)}) + \min_{s \in \mathcal{C}} \nabla f(x^{(k)})^T (s - x^{(k)}),$$

$$f^* \geq f(x^{(k)}) + \nabla f(x^{(k)})^T (s^{(k)} - x^{(k)}).$$

Which can then be re-written as:

$$f^* \geq f(x^{(k)}) + \nabla f(x^{(k)})^T (s^{(k)} - x^{(k)}),$$
$$-\nabla f(x^{(k)})^T (s^{(k)} - x^{(k)}) \geq f(x^{(k)}) - f^*,$$
$$\nabla f(x^{(k)})^T (x^{(k)} - s^{(k)}) \geq f(x^{(k)}) - f^*.$$

□

5.8 further topics

5.8.1 Speeding up: Frank-Wolfe with Away Steps Algorithm

we notice that FW method suffers from the zigzagging phenomenon. To deal with the problem, an idea is to incorporate steps that go away from an extreme point.
Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$, where $\mathcal{A}$ is a finite set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the active set $S^k = \{x^0\}$.
2. Find FW direction:
$$s^k \in \arg\min_{x \in \mathcal{M}} \langle \nabla f(x^k), x \rangle.$$
3. Return if (according to the stopping criterion):
$$G_{\text{FW}}(x^k) = \langle \nabla f(x^k), x^k - s^k \rangle \leq \epsilon.$$
4. Find away direction:
$$v^k \in \arg\max_{x \in S^k} \langle \nabla f(x^k), x \rangle.$$
5. If:
$$\langle -\nabla f(x^k), s^k - x^k \rangle \geq \langle -\nabla f(x^k), x^k - v^k \rangle, \tag{1}$$
take a FW step:
$$x^{k+1} = x^k + \eta_k (s^k - x^k),$$
where:
$$\eta_k \in \arg\min_{\eta \in [0,1]} f(x^k + \eta(s^k - x^k)).$$
Remark. *we note that both sides of (1) are nonnegative, then the value of inner product (projection) can reflect which direction is better: along the smaller direction, the decent of gradient is steeper, i.e., the direction is better.*
6. Otherwise, take an away step:
$$x^{k+1} = x^k + \eta_k (x^k - v^k),$$
where:
$$\eta_k \in \arg\min_{\eta \in [0, \eta_{\max}]} f(x^k + \eta(x^k - v^k)).$$
7. Update S^{k+1} and repeat the procedure.

we need to check the convergence of this algorithm:
Assumptions:

- $f \in \mathcal{S}^1_{L,\mu}$,
- $\mathcal{M} = \text{conv}(\mathcal{A})$, where $\mathcal{A}$ is a finite set of vectors,
- (Diameter) $R = \max_{x,y \in \mathcal{M}} \|x - y\|$,
- (Facial dist.) $\Phi = \min_{F \in \text{faces}(\text{conv}(\mathcal{A}))} \text{dist}(F, \text{conv}(\mathcal{A} \setminus F))$.

Then,
$$f(x^k) - f^* \leq \left(1 - \frac{\mu \Phi^2}{L \cdot 4R^2}\right)^{k/2} (f(x^0) - f^*).$$

$\frac{R^2}{\Phi^2}$ can be interpreted as the *condition number* of $\mathcal{M}$.

5.8.2 FW with subgradients

5.8.3 Nonconvex FW

5.8.4 Stochastic FW

6 BCD, AltMin, Product Space trick

6.1 Coordinate descent

solving the problem:

$$\min f(x) = \sum_{i=1}^{n} f_i(x), \quad x \in \mathbb{R}^d$$

the idea of coordinate descent is to go through $x_1, ..., x_d$ one by one

- For $k = 0, 1, \ldots$
  - Pick an index i from $\{1, \ldots, d\}$
  - Optimize the i-th coordinate

$$x_i^{k+1} \leftarrow \underset{\xi \in \mathbb{R}}{\operatorname{argmin}} f(\underbrace{x_1^{k+1}, \ldots, x_{i-1}^{k+1}}_{\text{done}}, \underbrace{\xi}_{\text{current}}, \underbrace{x_{i+1}^{k}, \ldots, x_d^{k}}_{\textit{todo}})$$

- Decide when/how to stop; *return* x^k

Remark.

- *One of the simplest optimization methods*
- *Old idea: Gauss-Seidel, Jacobi methods for linear systems!*
- *Can be "slow", but sometimes very competitive*
- *Gradient, subgradient, incremental methods also "slow"*
- *But incremental, stochastic gradient methods are scalable*
- *Renewed interest in CD was driven by ML*
- *Notice: in general CD is "derivative free"*

Exercise 6.1 (CD for least squares)**.**

$$\min_x \|Ax - b\|_2^2$$

Obtain an update for j-th coordinate.

$$x_j \leftarrow \frac{\sum_{i=1}^m a_{ij} \left(b_i - \sum_{l \neq j} a_{il} x_l\right)}{\sum_{i=1}^m a_{ij}^2}$$

Remark. ***Advantages***

- *Each iteration usually cheap (single variable optimization)*
- *No extra storage vectors needed*
- ***No stepsize tuning***
- *No other pesky parameters (usually) that must be tuned*
- *Simple to implement*
- *Can work well for large-scale problems*

Disadvantages

- *Tricky if single variable optimization is hard*
- *Convergence theory can be complicated*
- *Can slow down near optimum*
- *Nonsmooth case more tricky*
- ***Explore:*** *not easy to use for deep learning...*

6.2 Block coordinate descent

$$\min f(x) := f(x_1, \ldots, x_m)$$
$$x \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m.$$

Gauss-Seidel update

$$x_i^{k+1} \leftarrow \operatorname*{argmin}_{\xi \in \mathcal{X}_i} f(\underbrace{x_1^{k+1}, \ldots, x_{i-1}^{k+1}}_{\text{done}}, \underbrace{\xi}_{\text{current}}, \underbrace{x_{i+1}^{k}, \ldots, x_m^k}_{\text{todo}})$$

Jacobi update (easy to parallelize)

$$x_i^{k+1} \leftarrow \operatorname*{argmin}_{\xi \in \mathcal{X}_i} f(\underbrace{x_1^{k}, \ldots, x_{i-1}^{k}}_{\text{done}}, \underbrace{\xi}_{\text{current}}, \underbrace{x_{i}^{k}, \ldots, x_m^k}_{\text{todo}})$$

Convergence of BCD:

Theorem 6.2. *Let f be C^1 over $\mathcal{X} := \prod_{i=1}^m \mathcal{X}_i$. Assume for each block i and $x \in \mathcal{X}$, the minimum*

$$\min_{\xi \in \mathcal{X}_i} f(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_m)$$

is uniquely attained. Then, every limit point of the sequence $\{x^k\}$ generated by BCD, is a stationary point of f.

Corollary 6.3. *If f is in addition convex, then every limit point of the BCD sequence $\{x^k\}$ is a global minimum.*

- **Unique solutions** of subproblems not always possible
- Above result is only **asymptotic** (holds in the limit)
- **Warning!** BCD may cycle indefinitely without converging, if number blocks > 2 and objective nonconvex.

Two Blocks - BCD

$$\text{minimize } f(x) = f(x_1, x_2) \quad x \in \mathcal{X}_1 \times \mathcal{X}_2.$$

Theorem 6.4. *(Grippo & Sciandrone (2000)). Let f be continuously differentiable. Let $\mathcal{X}_1, \mathcal{X}_2$ be closed and convex. Assume both BCD subproblems have solutions and the sequence $\{x^k\}$ has limit points. Then, every limit point of $\{x^k\}$ is stationary.*

- *No need of **unique solutions** to subproblems*
- *BCD for 2 blocks is also called: **Alternating Minimization***

6.3 CD-projection onto convex sets

$$\min \quad \frac{1}{2}\|x - y\|_2^2 \quad \text{s.t.} \quad x \in C_1 \cap C_2 \cap \cdots \cap C_m.$$

quick recap of proximal operators: let $\mathbf{1}_{\mathcal{X}}$ be the indicator function for closed, cvx $\mathcal{X}$. recall orthogonal projection $P_{\mathcal{X}}(y)$

$$P_{\mathcal{X}}(y) := \arg\min \frac{1}{2}\|x-y\|_2^2 \quad s.t. \quad x \in \mathcal{X}$$

rewrite orthogonal projection $P_{\mathcal{X}}(y)$ as

$$P_{\mathcal{X}}(y) := \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \mathbf{1}_{\mathcal{X}}(x)$$

proximity: replace $\mathbf{1}_{\mathcal{X}}$ by some convex function:

$$prox_r(y) := \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + r(x)$$

the operator $prog_r : \mathbb{R}^n \to \mathbb{R}^n$ is called proximal operator. $Prox_r$ has several nice properties:

- the solution is unique: $0 \in x - y + \partial r(x) \to y \in (Id + \partial r)(x) \to x = (Id + \partial r)^{-1} y \to x = prox_r(y)$
- non-expansivity and firm non-expansivity

Solution 1: Rewrite using indicator functions

$$\min \quad \frac{1}{2}\|x-y\|_2^2 + \sum_{i=1}^{m} \delta_{C_i}(x).$$

- Original problem over $\mathcal{H} = \mathbb{R}^n$
- Suppose we have $\sum_{i=1}^n f_i(x)$
- Introduce n new variables $(x_1, \ldots, x_n)$
- Now problem is over domain $\mathcal{H}^n := \chi_{i=1}^n \mathcal{H}$
- New constraint: $x_1 = x_2 = \cdots = x_n$

$$\min_{(x_1,\ldots,x_n)} \quad \sum_i f_i(x_i) \quad \text{s.t.} \quad x_1 = x_2 = \cdots = x_n. \Leftrightarrow \min_x f(x) + \mathbf{1}_{\mathcal{B}}(x)$$

where $x \in \mathcal{H}^n$ and $\mathcal{B} = \{z \in \mathcal{H}^n \mid z = (x, x, \ldots, x)\}$.

- Let $\mathbf{y} = (y_1, \ldots, y_n)$
- $\text{prox}_f(\mathbf{y}) = (\text{prox}_{f_1}(y_1), \ldots, \text{prox}_{f_n}(y_n))$
- $\text{prox}_{\mathcal{B}} \equiv \Pi_{\mathcal{B}}(\mathbf{y})$ can be solved as follows:

$$\min_{\mathbf{z}\in\mathcal{B}} \frac{1}{2}\|\mathbf{z}-\mathbf{y}\|_2^2 \Leftrightarrow \min_{x\in\mathcal{H}} \sum_i \frac{1}{2}\|x-y_i\|_2^2 \Leftrightarrow x = \frac{1}{n}\sum_i y_i$$

Solution 2: Take dual of the above formulation

$$\min \frac{1}{2}\|x-y\|_2^2 + f(x) + h(x)$$

$$L(x,z,w,\nu,\mu) := \frac{1}{2}\|x-y\|_2^2 + f(z) + h(w) + \nu^T(x-z) + \mu^T(x-w)$$

$$g(\nu,\mu) := \inf_{x,z,w} L(x,z,w,\nu,\mu)$$

$$x - y + \nu + \mu = 0 \implies x = y - \nu - \mu$$

$$g(\nu,\mu) = -\frac{1}{2}\|\nu+\mu\|_2^2 + (\nu+\mu)^T y - f^*(\nu) - h^*(\mu)$$

Dual as minimization problem:

$$\min k(\nu, \mu) := \frac{1}{2}\|\nu + \mu - y\|_2^2 + f^*(\nu) + h^*(\mu)$$

Apply CD to $k(\nu, \mu) = \frac{1}{2}\|\nu + \mu - y\|_2^2 + f^*(\nu) + h^*(\mu)$

$$\nu_{k+1} = \arg\min_{\nu} k(\nu, \mu_k), \quad \mu_{k+1} = \arg\min_{\mu} k(\nu_{k+1}, \mu)$$

- $0 \in \nu + \mu_k - y + \partial f^*(\nu)$
- $0 \in \nu_{k+1} + \mu - y + \partial h^*(\mu)$
- $y - \mu_k \in \nu + \partial f^*(\nu) \implies \nu = \text{prox}_{f^*}(y - \mu_k)$
- Similarly, $\mu = y - \nu_{k+1} - \text{prox}_h(y - \nu_{k+1})$

$$\boxed{\nu_{k+1} \leftarrow y - \mu_k - \text{prox}_f(y - \mu_k), \quad \mu_{k+1} \leftarrow y - \nu_{k+1} - \text{prox}_h(y - \nu_{k+1})}$$

- Simplify, and use Lagrangian stationarity to obtain primal

$$x = y - \nu - \mu \implies y - \mu = x + \nu$$

- Thus, the CD iteration may be rewritten as
  1. $t_k \leftarrow \text{prox}_f(x_k + \nu_k)$
  2. $\nu_{k+1} \leftarrow x_k + \nu_k - t_k$
  3. $x_{k+1} \leftarrow \text{prox}_h(\mu_k + t_k) \quad (\Leftarrow \text{prox}_h(y - \nu_{k+1}) = \mu_{k+1} - y - \nu_{k+1} = x_{k+1})$
  4. $\mu_{k+1} \leftarrow \mu_k + t_k - x_{k+1}$
- This is the proximal-Dykstra method!

6.4 CD – nonsmooth case

6.5 CD – iteration complexity

6.6 Randomized BCD

7 Optimization in Deep Learning

8 Intro to bilevel optimization