Recap

Pontus Giselsson
Outline

• Convex analysis
• Composite optimization and duality
• Solving composite optimization problems – Algorithms
Convex Analysis
Convex sets

- A set $C$ is convex if for every $x, y \in C$ and $\theta \in [0, 1]$:

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

- “Every line segment that connect any two points in $C$ is in $C$”

- Will assume that all sets are nonempty and closed
Separating hyperplane theorem

- Suppose that $R, S \subseteq \mathbb{R}^n$ are two non-intersecting convex sets
- Then there exists hyperplane with $S$ and $R$ in opposite halves

\[ s^T x = r \]

Example

Counter-example

$R$ nonconvex

- Mathematical formulation: There exists $s \neq 0$ and $r$ such that

\[ s^T x \leq r \quad \text{for all } x \in R \]
\[ s^T x \geq r \quad \text{for all } x \in S \]

- The hyperplane \( \{ x : s^T x = r \} \) is called *separating hyperplane*
A strictly separating hyperplane theorem

• Suppose that $R, S \subseteq \mathbb{R}^n$ are non-intersecting closed and convex sets and that one of them is compact (closed and bounded)
• Then there exists hyperplane with strict separation

$$s^T x = r$$

Example

$R = \{(x, y) : y \geq x^{-1}, x > 0\}$

$S = \{(x, y) : y \leq 0\}$

Counter example $R, S$ not compact

• Mathematical formulation: There exists $s \neq 0$ and $r$ such that

$$s^T x < r$$ for all $x \in R$$

$$s^T x > r$$ for all $x \in S$$
Consequence – $S$ is intersection of halfspaces

A closed convex set $S$ is the intersection of all halfspaces that contain it.

Proof:

- Let $H$ be the intersection of all halfspaces containing $S$.
- $\Rightarrow$: Obviously $x \in S \Rightarrow x \in H$.
- $\Leftarrow$: Assume $x \not\in S$, since $S$ closed and convex and $x$ compact (a point), there exists a strictly separating hyperplane, i.e., $x \not\in H$. 

![Diagram](image-url)
Supporting hyperplanes

- Supporting hyperplanes touch set and have full set on one side:

- We call the halfspace that contains the set *supporting halfspace*
- $s$ is called *normal vector* to $S$ at $x$
- Definition: Hyperplane $\{y : s^T y = r\}$ supports $S$ at $x \in \text{bd } S$ if
  \[ s^T y \leq r \text{ for all } y \in S \quad \text{and} \quad s^T x = r \]
Supporting hyperplane theorem

Let $S$ be a nonempty convex set and let $x \in \text{bd}(S)$. Then there exists a supporting hyperplane to $S$ at $x$.

- Does not exist for all point on boundary for nonconvex sets
- Many supporting hyperplanes exist for points of nonsmoothness
Connection to duality and subgradients

Supporting hyperplanes are at the core of convex analysis:

- Subgradients define supporting hyperplanes to $\text{epi } f$
- Conjugate functions define supporting hyperplanes to $\text{epi } f$
- Duality is based on subgradients, hence supporting hyperplanes:
  - Consider $\min_x (f(x) + g(x))$ and primal solution $x^*$
  - Dual problem $\min_\mu (f^*(\mu) + g^*(-\mu))$ solution $\mu^*$ satisfies
    \[
    \mu^* \in \partial f(x^*) \quad \quad -\mu^* \in \partial g(x^*)
    \]
    i..e, dual problem finds subgradients at optimal point

1 When solving $\min_x (f(Lx) + g(x))$ dual problem finds $\mu$ such that $L^T \mu \in \partial(f \circ L)(x)$ and $-L^T \mu \in \partial g(x)$.
Convex functions

• Graph below line connecting any two pairs \((x, f(x))\) and \((y, f(y))\)

\begin{itemize}
  \item Function \(f : \mathbb{R}^n \rightarrow \mathbb{R}\) is convex if for all \(x, y \in \mathbb{R}^n\) and \(\theta \in [0, 1]\):
  \[
  f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)
  \]
  (in extended valued arithmetics)
  \item A function \(f\) is concave if \(-f\) is convex
\end{itemize}
Epigraphs and convexity

• Let \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \)

• Then \( f \) is convex if and only \( \text{epi} f \) is a convex set in \( \mathbb{R}^n \times \mathbb{R} \)

• \( f \) is called closed (lower semi-continuous) if \( \text{epi} f \) is closed set
First-order condition for convexity

- A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if
  \[ f(y) \geq f(x) + \nabla f(x)^T (y - x) \]
  for all $x, y \in \mathbb{R}^n$

- Function $f$ has for all $x \in \mathbb{R}^n$ an affine minorizer that:
  - has slope $s$ defined by $\nabla f$
  - coincides with function $f$ at $x$
  - is supporting hyperplane to epigraph of $f$
  - defines normal $(\nabla f(x), -1)$ to epigraph of $f$
Subdifferentials and subgradients

- Subgradients $s$ define affine minorizers to the function that:
  - coincide with $f$ at $x$
  - define normal vector $(s, -1)$ to epigraph of $f$
  - can be one of many affine minorizers at nondifferentiable points $x$

Subdifferential of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at $x$ is set of vectors $s$ satisfying

$$f(y) \geq f(x) + s^T(y - x) \quad \text{for all } y \in \mathbb{R}^n,$$

Notation:
- subdifferential: $\partial f : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ (power-set notation $2^{\mathbb{R}^n}$)
- subdifferential at $x$: $\partial f(x) = \{ s : (1) \text{ holds} \}$
- elements $s \in \partial f(x)$ are called subgradients of $f$ at $x$
Subgradient existence – Nonconvex example

• Function can be differentiable at $x$ but $\partial f(x) = \emptyset$

- $x_1$: $\partial f(x_1) = \{0\}$, $\nabla f(x_1) = 0$
- $x_2$: $\partial f(x_2) = \emptyset$, $\nabla f(x_2) = 0$
- $x_3$: $\partial f(x_3) = \emptyset$, $\nabla f(x_3) = 0$

• Gradient is a local concept, subdifferential is a global property
Existence for extended-valued convex functions

• Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex, then:
  1. Subgradients exist for all $x$ in relative interior of $\text{dom} f$
  2. Subgradients sometimes exist for $x$ on boundary of $\text{dom} f$
  3. No subgradient exists for $x$ outside $\text{dom} f$

• Examples for second case, boundary points of $\text{dom} f$:

\[-\sqrt{1-x^2} + \iota_{[-1,1]}(x)\]

\[x^2 + \iota_{[-2,2]}(x)\]

• No subgradient (affine minorizer) exists for left function at $x = 1$
Fermat’s rule

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then $x$ minimizes $f$ if and only if

$0 \in \partial f(x)$

• Proof: $x$ minimizes $f$ if and only if

$$f(y) \geq f(x) + 0^T(y - x) \quad \text{for all } y \in \mathbb{R}^n$$

which by definition of subdifferential is equivalent to $0 \in \partial f(x)$

• Example: several subgradients at solution, including $0$

![Graph showing Fermat's rule](image)
Fermat’s rule – Nonconvex example

• Fermat’s rule holds also for nonconvex functions
• Example:

\( \partial f(x_1) = 0 \) and \( \nabla f(x_1) = 0 \) (global minimum)
\( \partial f(x_2) = \emptyset \) and \( \nabla f(x_2) = 0 \) (local minimum)

• For nonconvex \( f \), we can typically only hope to find local minima
Subdifferential calculus rules

• Subdifferential of sum $\partial(f_1 + f_2)$
• Subdifferential of composition with matrix $\partial(g \circ L)$
Subdifferential of sum

If $f_1, f_2$ closed convex and $\text{relint dom } f_1 \cap \text{relint dom } f_2 \neq \emptyset$:

$$\partial (f_1 + f_2) = \partial f_1 + \partial f_2$$

• One direction always holds: if $x \in \text{dom } \partial f_1 \cap \text{dom } \partial f_2$:

$$\partial (f_1 + f_2)(x) \supseteq \partial f_1(x) + \partial f_2(x)$$

Proof: let $s_i \in \partial f_i(x)$, add subdifferential definitions:

$$f_1(y) + f_2(y) \geq f_1(x) + f_2(x) + (s_1 + s_2)^T(y - x)$$

i.e. $s_1 + s_2 \in \partial (f_1 + f_2)(x)$

• If $f_1$ and $f_2$ differentiable, we have (without convexity of $f$)

$$\nabla (f_1 + f_2) = \nabla f_1 + \nabla f_2$$
Subdifferential of composition

If $f$ closed convex and $\text{relint dom}(f \circ L) \neq \emptyset$:

$$\partial(f \circ L)(x) = L^T \partial f(Lx)$$

- One direction always holds: If $Lx \in \text{dom } f$, then

$$\partial(f \circ L)(x) \supseteq L^T \partial f(Lx)$$

Proof: let $s \in \partial f(Lx)$, then by definition of subgradient of $f$:

$$(f \circ L)(y) \geq (f \circ L)(x) + s^T (Ly - Lx) = (f \circ L)(x) + (L^T s)^T (y - x)$$

i.e., $L^T s \in \partial(f \circ L)(x)$

- If $f$ differentiable, we have chain rule (without convexity of $f$)

$$\nabla (f \circ L)(x) = L^T \nabla f(Lx)$$
A sufficient optimality condition

Let $f : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$, $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, and $L \in \mathbb{R}^{m \times n}$ then:

$$\text{minimize } f(Lx) + g(x)$$  \hspace{1cm} (1)

is solved by every $x \in \mathbb{R}^n$ that satisfies

$$0 \in L^T \partial f(Lx) + \partial g(x)$$  \hspace{1cm} (2)

- Subdifferential calculus inclusions say:

$$0 \in L^T \partial f(Lx) + \partial g(x) \subseteq \partial((f \circ L)(x) + g(x))$$

which by Fermat’s rule is equivalent to $x$ solution to (1)

- Note: (1) can have solution but no $x$ exists that satisfies (2)
A necessary and sufficient optimality condition

Let $f : \mathbb{R}^m \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$, $L \in \mathbb{R}^{m \times n}$ with $f$, $g$ closed convex and assume $\text{relint \ dom}(f \circ L) \cap \text{relint \ dom} g \neq \emptyset$ then:

$$\text{minimize } f(Lx) + g(x)$$  \hfill (1)

is solved by $x \in \mathbb{R}^n$ if and only if $x$ satisfies

$$0 \in L^T \partial f(Lx) + \partial g(x)$$  \hfill (2)

- Subdifferential calculus equality rules say:

  $$0 \in L^T \partial f(Lx) + \partial g(x) = \partial((f \circ L)(x) + g(x))$$

  which by Fermat’s rule is equivalent to $x$ solution to (1)

- Algorithms search for $x$ that satisfy $0 \in L^T \partial f(Lx) + \partial g(x)$
Evaluating subgradients of convex functions

• Obviously need to evaluate subdifferentials to solve

\[ 0 \in L^T \partial f(Lx) + \partial g(x) \]

• Explicit evaluation:
  • If function is differentiable: \( \nabla f \) (unique)
  • If function is nondifferentiable: compute element in \( \partial f \)

• Implicit evaluation:
  • Proximal operator (specific element of subdifferential)
Proximal operator

• Proximal operator of (convex) $g$ defined as:

$$\text{prox}_\gamma g(z) = \arg\min_x (g(x) + \frac{1}{2\gamma} \|x - z\|^2)$$

where $\gamma > 0$ is a parameter

• Evaluating prox requires solving optimization problem

• Objective is strongly convex $\Rightarrow$ solution exists and is unique
Prox evaluates the subdifferential

• Fermat’s rule on prox definition: \( x = \text{prox}_{\gamma g}(z) \) if and only if

\[
0 \in \partial g(x) + \gamma^{-1}(x - z) \iff \gamma^{-1}(z - x) \in \partial g(x)
\]

Hence, \( \gamma^{-1}(z - x) \) is element in \( \partial g(x) \)

• A subgradient \( \partial g(x) \) where \( x = \text{prox}_{\gamma g}(z) \) is computed

• Often used in algorithms when \( g \) nonsmooth (no gradient exists)
Conjugate functions

- The conjugate function of $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as
  
  $$f^*(s) := \sup_{x} (s^T x - f(x))$$

- Implicit definition via optimization problem
Conjugate interpretation

- Conjugate $f^*(s)$ defines affine minorizer to $f$ with slope $s$:

\[
\begin{align*}
\text{Conjugate interpretation} & \\
\text{• Conjugate } f^*(s) \text{ defines affine minorizer to } f \text{ with slope } s: & \\
& \quad \text{where } f^*(s) \text{ decides the constant offset to have support at } x^* \\
& \text{• “Affine minorizor generator: Pick slope } s, \text{ get offset for support”} \\
& \text{• Why? Consider } f^*(s) = \sup_x (s^T x - f(x)) \text{ with maximizer } x^*: \\
& \quad f^*(s) = s^T x^* - f(x^*) \quad \Leftrightarrow \quad f^*(s) \geq s^T x - f(x) \text{ for all } x \\
& \quad \Leftrightarrow \quad f(x) \geq s^T x - f^*(s) \text{ for all } x \\
& \text{• Support at } x^* \text{ since } f(x^*) = s^T x^* - f^*(s)
\end{align*}
\]
Fenchel Young’s equality

- Going back to conjugate interpretation:

\[ f(x) \geq s^T x - f^*(s) \text{ for all } x, s \]

- Fenchel-Young’s equality and equivalence:

\[ f(x^*) = s^T x^* - f^*(s) \text{ holds if and only if } s \in \partial f(x^*) \]
A subdifferential formula

Assume $f$ closed convex, then $\partial f(x) = \text{Argmax}_s (s^T x - f^*(s))$

- Since $f^{**} = f$, we have $f(x) = \sup_s (x^T s - f^*(s))$ and
  $$s^* \in \text{Argmax}_s (x^T s - f^*(s)) \iff f(x) = x^T s^* - f^*(s^*)$$
  $$\iff s^* \in \partial f(x)$$

- The last equivalence is Fenchel-Young
Subdifferential of conjugate – Inversion formula

Suppose $f$ closed convex, then $s \in \partial f(x) \iff x \in \partial f^*(s)$

• Consequence of Fenchel-Young
• Another way to write the result is that for closed convex $f$:

\[ \partial f^* = (\partial f)^{-1} \]

(Definition of inverse of set-valued $A$: $x \in A^{-1}u \iff u \in Ax$)
Strong convexity

- Let $\sigma > 0$
- A function $f$ is $\sigma$-strongly convex if $f - \frac{\sigma}{2} \| \cdot \|_2^2$ is convex
- Alternative equivalent definition of $\sigma$-strong convexity:
  \[
  f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) - \frac{\sigma}{2} \theta (1 - \theta) \| x - y \|_2^2
  \]
  holds for every $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$
- Strongly convex functions are strictly convex and convex
- Example: $f$ 2-strongly convex since $f - \| \cdot \|_2^2$ convex:

\[
f(x) \quad \text{and} \quad f(x) - \| x \|_2^2
\]
First-order condition for strong convexity

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable
- $f$ is $\sigma$-strongly convex with $\sigma > 0$ if and only if
  \[
f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\sigma}{2} \|x - y\|_2^2
  \]
  for all $x, y \in \mathbb{R}^n$

- Function $f$ has for all $x \in \mathbb{R}^n$ a quadratic minorizer that:
  - has curvature defined by $\sigma$
  - coincides with function $f$ at $x$
  - defines normal $(\nabla f(x), -1)$ to epigraph of $f$
Smoothness

• A function is called $\beta$-smooth if its gradient is $\beta$-Lipschitz:
  \[ \|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2 \]
  for all $x, y \in \mathbb{R}^n$ (it is not necessarily convex)

• Alternative equivalent definition of $\beta$-smoothness
  \[
  f(\theta x + (1 - \theta)y) \geq \theta f(x) + (1 - \theta)f(y) - \frac{\beta}{2} \theta(1 - \theta)\|x - y\|^2 \\
  f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) + \frac{\beta}{2} \theta(1 - \theta)\|x - y\|^2
  \]

hold for every $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$

• Smoothness does not imply convexity

• Example:
First-order condition for smoothness

• $f$ is $\beta$-smooth with $\beta \geq 0$ if and only if

\[
\begin{align*}
f(y) &\leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|x - y\|^2_2 \\
f(y) &\geq f(x) + \nabla f(x)^T (y - x) - \frac{\beta}{2} \|x - y\|^2_2
\end{align*}
\]

for all $x, y \in \mathbb{R}^n$

• Quadratic upper/lower bounds with curvatures defined by $\beta$
• Quadratic bounds coincide with function $f$ at $x$
First-order condition for smooth convex

- $f$ is $\beta$-smooth with $\beta \geq 0$ and convex if and only if
  
  $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \| x - y \|^2_2$

  
  $f(y) \geq f(x) + \nabla f(x)^T (y - x)$

  for all $x, y \in \mathbb{R}^n$

- Quadratic upper bound and affine lower bound
- Bounds coincide with function $f$ at $x$
- Quadratic upper bound is called descent lemma
Duality correspondance

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Then the following are equivalent:

(i) $f$ is closed and $\sigma$-strongly convex
(ii) $\partial f$ is maximally monotone and $\sigma$-strongly monotone
(iii) $\nabla f^*$ is $\sigma$-cocoercive
(iv) $\nabla f^*$ is maximally monotone and $\frac{1}{\sigma}$-Lipschitz continuous
(v) $f^*$ is closed convex and satisfies descent lemma (is $\frac{1}{\sigma}$-smooth)

where $\nabla f^* : \mathbb{R}^n \to \mathbb{R}^n$ and $f^* : \mathbb{R}^n \to \mathbb{R}$

Comments:

• Relation (i) $\Leftrightarrow$ (v) most important for us
• Since $f = f^{**}$ the result holds with $f$ and $f^*$ interchanged
• Full proof available on course webpage
Composite Optimization
Composite optimization

We consider composite optimization problems of the form

$$\min_x f(Lx) + g(x)$$
Optimality conditions and dual problem

• Assume $f, g$ closed convex and that CQ holds
• Problem $\min_x (f(Lx) + g(x))$ is solved by $x$ iff
\[
0 \in L^T \partial f(Lx) + \partial g(x)
\]
where dual variable $\mu$ has been defined
• Primal dual necessary and sufficient optimality conditions:
\[
\begin{align*}
\mu &\in \partial f(Lx) \\
-L^T \mu &\in \partial g(x) \\
Lx &\in \partial f^*(\mu) \\
-L^* \mu &\in \partial g(x)
\end{align*}
\]
\[
\begin{align*}
\mu &\in \partial f(Lx) \\
x &\in \partial g^*(-L^T \mu) \\
Lx &\in \partial f^*(\mu) \\
x &\in \partial g^*(-L^T \mu)
\end{align*}
\]
• Dual optimality condition
\[
0 \in \partial f^*(\mu) + \partial (g^* \circ -L^T)(\mu)
\] (1)
\[
\text{solves dual problem } \min_\mu f^*(\mu) + g^*(-L^T \mu)
\]
• If CQ-D holds, all dual problem solutions satisfy (1)
• Dual searches for $\mu$ such that $L^T \mu \in \partial f(x)$ and $-L^T \mu \in \partial g(x)$
Solving the primal via the dual

• Why solve dual? Sometimes easier to solve than primal
• Only interesting if primal solution can be recovered
• Assume $f, g$ closed convex and CQ
• Assume optimal dual $\mu$ known: $0 \in \partial f^*(\mu) + \partial (g^* \circ -L^T)(\mu)$
• Optimal primal $x$ must satisfy any and all primal-dual conditions:

\[
\begin{align*}
\mu &\in \partial f(Lx) & Lx &\in \partial f^*(\mu) \\
-L^T \mu &\in \partial g(x) & -L^* \mu &\in \partial g(x) \\
\mu &\in \partial f(Lx) & Lx &\in \partial f^*(\mu) \\
x &\in \partial g^*(-L^T \mu) & x &\in \partial g^*(-L^T \mu)
\end{align*}
\]

• If one of these uniquely characterizes $x$, then must be solution:
  • $\partial g^*$ is differentiable at $-L^T \mu$ for dual solution $\mu$
  • $\partial f^*$ is differentiable at dual solution $\mu$ and $L$ invertible
  • ...
Algorithms
Proximal gradient method

• Consider minimize $\min_x f(x) + g(x)$ where
  • $f$ is $\beta$-smooth $f : \mathbb{R}^n \to \mathbb{R}$ (not necessarily convex)
  • $g$ is closed convex
• Due to $\beta$-smoothness of $f$, we have
  \[
  f(y) + g(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \| y - x \|^2 + g(y)
  \]
  for all $x, y \in \mathbb{R}^n$, i.e., r.h.s. is majorizing function for fixed $x$
• Majorization minimization with majorizer if $\gamma_k \in [\epsilon, \beta^{-1}]$, $\epsilon > 0$:
  \[
  x_{k+1} = \arg\min_y \left( f(x_k) + \nabla f(x_k)^T (y - x) + \frac{1}{2\gamma_k} \| y - x_k \|^2_2 + g(y) \right)
  \]
  \[
  = \arg\min_y \left( g(y) + \frac{1}{2\gamma_k} \| y - (x_k - \gamma_k \nabla f(x_k)) \|^2_2 \right)
  \]
  \[
  = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))
  \]
gives proximal gradient method
Proximal gradient – Fixed-points

- Denote $T_{PG}^\gamma := \text{prox}_{\gamma g}(I - \gamma \nabla f)$, gives algorithm $x_{k+1} = T_{PG}^\gamma x_k$
- Proximal gradient fixed-point set definition

$$\text{fix}T_{PG}^\gamma = \{x : x = T_{PG}^\gamma x\} = \{x : x = \text{prox}_{\gamma g}(x - \gamma \nabla f(x))\}$$

i.e., set of points for which $x_{k+1} = x_k$

Let $\gamma > 0$. Then $\bar{x} \in \text{fix}T_{PG}^\gamma$ if and only if $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$.

- Consequence: fixed-point set same for all $\gamma > 0$
- We call inclusion $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$ fixed-point characterization
  - For convex problems: global solutions
  - For nonconvex problems: critical points
Applying proximal gradient to primal problems

Problem $\min_x f(x) + g(x)$:

- Assumptions:
  - $f$ β-smooth
  - $g$ closed convex and prox friendly\(^1\)
  - $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$
- Algorithm: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$

Problem $\min_x f(Lx) + g(x)$:

- Assumptions:
  - $f$ β-smooth (implies $f \circ L$ $\beta \|L\|_2^2$-smooth)
  - $g$ closed convex and prox friendly\(^1\)
  - $\gamma_k \in [\epsilon, \frac{2}{\beta \|L\|_2^2} - \epsilon]$
- Gradient $\nabla (f \circ L)(x) = L^T \nabla f(Lx)$
- Algorithm: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k L^T \nabla f(Lx_k))$

\(^1\)Prox friendly: proximal operator cheap to evaluate, e.g., $g$ separable
Applying proximal gradient to dual problem

Dual problem minimize $f^*(\nu) + g^*(-L^T\nu)$:

- Assumptions:
  - $f$ closed convex and prox friendly
  - $g$ $\sigma$-strongly convex (which implies $g^* \circ -L^T \frac{2\|\nu\|}{\sigma}$-smooth)
  - $\gamma_k \in [\epsilon, \frac{2\sigma}{\|L\|^2} - \epsilon]

- Gradient: $\nabla (g^* \circ -L^T)(\nu) = -L \nabla g^*(-L^T\nu)$

- Prox (Moreau): $\text{prox}_{\gamma_k f^*}(\nu) = \nu - \gamma_k \text{prox}_{\gamma_k^{-1} f} (\gamma_k^{-1} \nu)$

- Algorithm:

  $\nu_{k+1} = \text{prox}_{\gamma_k f^*} (\nu_k - \gamma_k \nabla (g^* \circ -L^T)(\nu_k))$

  $= (I - \gamma_k \text{prox}_{\gamma_k^{-1} f} (\gamma_k^{-1} \circ I))(\nu_k + \gamma_k L \nabla g^*(-L^T\nu_k))$

- Problem must be convex to have dual!

- Enough to know prox of $f$
What problems cannot be solved (efficiently)?

Problem \[ \min_x f(x) + g(x) \]

- Assumptions: \( f \) and \( g \) convex and nonsmooth
- No term differentiable, another method must be used:
  - Subgradient method
  - Douglas-Rachford splitting
  - Primal-dual methods

Problem \[ \min_x f(x) + g(Lx) \]

- Assumptions:
  - \( f \) smooth
  - \( g \) nonsmooth convex
  - \( L \) arbitrary structured matrix
- Can apply proximal gradient method, but

\[
\text{prox}_{\gamma_k(g \circ L)}(z) = \arg\min_x g(Lx) + \frac{1}{2\gamma} \|x - z\|^2_2
\]

often not “prox friendly”, i.e., it is expensive to evaluate
Training problems

- Training problem format

\[
\min_\theta \sum_{i=1}^{N} L(m(x_i; \theta), y_i) + \sum_{j=1}^{n} g_j(\theta_j)
\]

where \( f \) is data misfit term and \( g \) is regularizer

- Regularizers \((\theta = (w, b))\)
  - Tikhonov \( g(\theta) = \|w\|_2^2 \) is prox-friendly
  - Sparsity inducing 1-norm \( g(\theta) = \|w\|_1 \) is prox-friendly

- Data misfit terms (with \( m(x; \theta) = \phi(x)^T \theta \) for convex problems)
  - Least squares \( L(u, y) = \|u - y\|_2^2 \) smooth, hence \( f \) smooth
  - Logistic \( L(u, y) = \log(1 + e^u) - yu \) smooth, hence \( f \) smooth
  - SVM \( L(u, y) = \max(0, 1 - yu) \) not smooth, hence \( f \) not smooth

- Proximal gradient method
  - Least squares: can efficiently solve primal
  - Logistic regression: can solve primal
  - SVM: add strongly convex regularization and solve dual
    - Strongly convex regulariztion to have one conjugate smooth
    - If bias term not regularized, only strongly convex in \( w \)
    - SVM with \( \| \cdot \|_1 \)-regularization not solvable with prox-grad
Dual training problem

• Convex training problem

\[
\begin{align*}
\text{minimize} \quad & \sum_{i=1}^{N} L\left(\phi(x_i)^T \theta, y_i\right) + \sum_{j=1}^{n} g_j(\theta_j) \\
\text{subject to} \quad & f(X\theta) \\
& g(\theta)
\end{align*}
\]

has dual

\[
\begin{align*}
\text{minimize} \quad & \sum_{i=1}^{N} L^*(\mu_i) + \sum_{j=1}^{n} g^*_j((-X^T \mu)_j) \\
\text{subject to} \quad & f^*(\mu) \\
& g^*(-X^T \mu)
\end{align*}
\]

where the conjugate of \( L \) is w.r.t. first argument

• Dual has same structure as primal, finite-sum plus separable
Training problem structure

• Primal training problem

\[
\min_{\theta} \sum_{i=1}^{N} L(m(x_i; \theta), y_i) + \sum_{j=1}^{n} g_j(\theta_j)
\]

– Primal: \(f(X\theta)\)

• Dual training problem

\[
\min_{\theta} \sum_{i=1}^{N} L^*(\mu_i) + \sum_{j=1}^{n} g^*_j((-X^T\mu)_j)
\]

– Dual: \(g(\theta)\)

• Common structure, finite sum plus separable:

\[
\min_{\theta} \sum_{i=1}^{N} f_i((X\theta)_i) + \sum_{j=1}^{n} \psi_j(\theta_j)
\]

– Primal: \(f_i = L(m(x_i; \cdot), y_i)\) (one summand per training example)

– Dual: \(f_i = g^*_j((-X^T\cdot)_j), \psi_j = L^*\)
Exploiting structure

• Common structure, finite sum plus separable:

\[
\minimize_{\theta} \sum_{i=1}^{N} f_i((X\theta)_i) + \sum_{j=1}^{n} \psi_j(\theta_j)
\]

• Stochastic gradient descent exploits finite-sum structure:
  • Computes stochastic gradient of smooth part \( f \)
  • Pick summand \( f_i \) at random and perform gradient step
  • Primal formulations: Pick training example and compute gradient
  • Deep learning: evaluated via backpropagation

• Coordinate gradient descent exploits separable structure:
  • Coordinate-wise updates if nonsmooth \( \phi_j \) separable
  • Requires efficient coordinate-wise evaluations of \( \nabla f \)
On exam

• The convex analysis part
• Algorithms:
  • Be able to use descent lemma and draw simple conclusions
  • No stochastic analysis of algorithms
  • Know what formulations that can be solved efficiently using:
    • proximal gradient method
    • coordinate and stochastic gradient methods
  in terms of problem structure and problem assumptions