Optimization for Data Science
ETH Zürich, FS 2023 261-5110-00L

Lecture 10: Mirror Descent, Smoothing, Proximal Algorithms

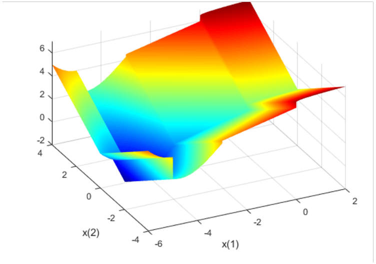
Bernd Gärtner
Niao He

https://www.ti.inf.ethz.ch/ew/courses/ODS23/index.html

April 24, 2023

Recap: Convex Nonsmooth Optimization

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$



- For convex functions, subgradients always exist in the interior.
- Subgradients share lots of similar properties as gradients.
- Subgradient methods can be slow.

NB: For nonconvex nonsmooth functions, finding an approximately stationary point with first-order methods is intractable in general [Zha20].

Recap: Subgradient Descent

Subgradient Descent

$$\mathbf{x}_{t+1} = \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t) = \underset{\mathbf{x}\in X}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2 + \langle \gamma_t \mathbf{g}_t, \mathbf{x}\rangle \right\}, \quad \mathbf{g}_t \in \partial f(\mathbf{x}_t).$$

- **Convergence rate**: $O\left(\frac{B\cdot R}{\sqrt{t}}\right)$ for convex objectives
- **Subgradient complexity**: $O\left(\frac{B\cdot R}{\epsilon^2}\right)$ for convex objectives
- From information-theoretic viewpoint, the rate of subgradient descent cannot really be improved, despite being slow.

$$B := \sup_{\mathbf{x}\in X} \frac{|f(\mathbf{x})-f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2}, R := \max_{\mathbf{x},\mathbf{y}\in X} \|\mathbf{x}-\mathbf{y}\|_2, BR = \|\cdot\|_2\text{-variation of } f \text{ on } X$$

Clicker Question (EduApp)

Consider the example:

$$f(\mathbf{x}) = \sum_{i=1}^{d} |x_i - a_i|,$$

$$X = \Delta_d := \{\mathbf{x} \in \mathbb{R}_+^d : \sum_{i=1}^{d} x_i = 1\}.$$

What's the order of the convergence rate when applying subgradient descent?

- $O\left(\frac{1}{\sqrt{t}}\right)$
- $O\left(\frac{\sqrt{d}}{\sqrt{t}}\right)$
- $O\left(\frac{d}{\sqrt{t}}\right)$
- None of the above

Motivation

In practice, we often have extra information about set X and nonsmooth function f.

- Can we exploit non-Euclidean geometry of convex set X? (instead of Euclidean geometry)
  $\Rightarrow$ **Mirror Descent!**

- Can we exploit additional structure of nonsmooth objective f? (instead of treating it as black box)
  $\Rightarrow$ **Smoothing and Proximal Algorithms!**

General Norms and Dual Norms

- **Norm:** A function $\|\cdot\|: \mathbb{R}^d \to \mathbb{R}_+$ is a <u>norm</u> if
  (a) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$;
  (b) $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$;
  (c) $\|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$.

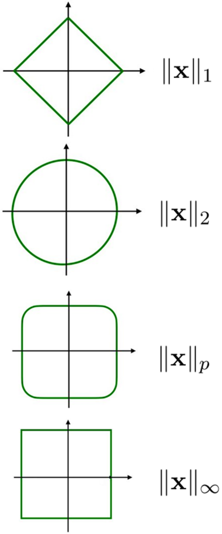
- **Dual norm:**

$$\|\mathbf{y}\|_* := \max_{\|\mathbf{x}\|\le 1} \langle \mathbf{x}, \mathbf{y}\rangle.$$

- Example: for $p \ge 1$ and $1/p + 1/q = 1$,

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}, \|\cdot\|_{p,*} = \|\cdot\|_q$$

- Inequality: $\frac{1}{\sqrt{d}}\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{d}\|\mathbf{x}\|_2$



General Smoothness and Strong Convexity

Smoothness: $f(\mathbf{x})$ is L-smooth on X if $f(\mathbf{x})$ is differentiable and

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in X.$$

Lipschitz continuity: $f(\mathbf{x})$ is B-Lipschitz continuous on X if

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in X.$$

Strong convexity: $f(\mathbf{x})$ is μ-strongly convex on X if for any $\mathbf{g} \in \partial f(\mathbf{x})$,

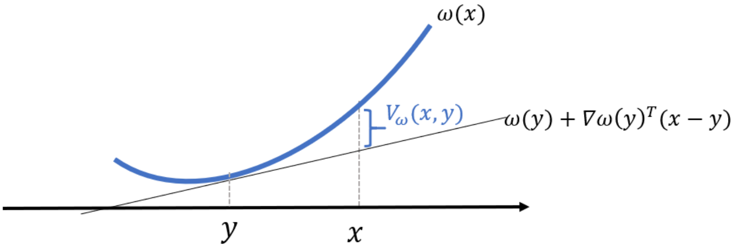
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \mathbf{g}^T(\mathbf{x}-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in X.$$

Bregman Divergence

Let $\omega(\cdot) : \Omega \to \mathbb{R}$ be continuously differentiable on Ω and 1-strongly convex w.r.t. some norm $\|\cdot\|$: $\omega(\mathbf{x}) \geq \omega(\mathbf{y}) + \nabla\omega(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall x, y \in \Omega$.

The Bregman divergence is defined as

$$V_\omega(\mathbf{x},\mathbf{y}) = \omega(\mathbf{x}) - \omega(\mathbf{y}) - \nabla\omega(\mathbf{y})^T(\mathbf{x}-\mathbf{y}), \forall \mathbf{x}, \mathbf{y} \in \Omega.$$



Examples

- Euclidean distance: $\Omega = \mathbb{R}^d$, $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, $\|\cdot\| = \|\cdot\|_2$

$$V_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

- Mahalanobis distance: $\Omega = \mathbb{R}^d$, $\omega(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q \mathbf{x}$ (where $Q \succeq I$), $\|\cdot\| = \|\cdot\|_2$,

$$V_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^T Q(\mathbf{x} - \mathbf{y}).$$

- Kullback-Leibler divergence: $\Omega = \Delta_d$, $\omega(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$, $\|\cdot\| = \|\cdot\|_1$,

$$V_\omega(\mathbf{x}, \mathbf{y}) = \mathrm{KL}(\mathbf{x}|\mathbf{y}) := \sum_{i=1}^d x_i \log \frac{x_i}{y_i}.$$

Clicker Question (EduApp)

Recall the definition of Bregman divergence:

$$V_\omega(\mathbf{x}, \mathbf{y}) = \omega(\mathbf{x}) - \omega(\mathbf{y}) - \nabla\omega(\mathbf{y})^T(\mathbf{x} - \mathbf{y}), \forall \mathbf{x}, \mathbf{y} \in \Omega.$$

Which one of the following statements does not always hold?

A. Nonnegativity: $V_\omega(\mathbf{x}, \mathbf{y}) \geq 0$.

B. Symmetry: $V_\omega(\mathbf{x}, \mathbf{y}) = V_\omega(\mathbf{y}, \mathbf{x})$.

C. Convexity: $V_\omega(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$.

D. $V_\omega(\mathbf{x}, \mathbf{y}) \geq \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$.

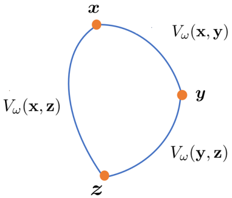
Key Property of Bregman Divergence

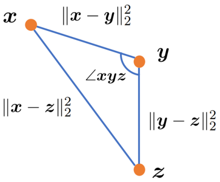
Lemma 10.1 (Three Point Identity)

$$V_\omega(\mathbf{x},\mathbf{z}) = V_\omega(\mathbf{x},\mathbf{y}) + V_\omega(\mathbf{y},\mathbf{z}) - \langle \nabla\omega(\mathbf{z}) - \nabla\omega(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle, \forall \mathbf{x},\mathbf{y},\mathbf{z} \in \Omega$$

- Special case: $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, this is the law of cosine:

$$\|\mathbf{x}-\mathbf{z}\|_2^2 = \|\mathbf{x}-\mathbf{y}\|_2^2 + \|\mathbf{y}-\mathbf{z}\|_2^2 - 2\langle \mathbf{z}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle.$$





- Proof follows by the definition of Bregman divergence (see supplementary).

Mirror Descent

Mirror Descent Algorithm: (Nemirovski & Yudin, 1983)

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in X}\{V_\omega(\mathbf{x},\mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}\rangle\}, \text{ where } \mathbf{g}_t \in \partial f(\mathbf{x}_t).$$

Example:

- Subgradient descent: $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, $V_\omega(\mathbf{x},\mathbf{x}_t) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$.

$$\mathbf{x}_{t+1} = \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t).$$

- Entropic descent: $X = \Delta_d$, $\omega(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$, $V_\omega(\mathbf{x},\mathbf{x}_t) = \mathrm{KL}(\mathbf{x}|\mathbf{x}_t)$.

$$\mathbf{x}_{t+1} \propto \mathbf{x}_t \odot \exp(-\gamma_t \mathbf{g}_t).$$

Here $\odot$ is element-wise multiplication.

Remarks

Mirror Descent is closely related to many classical algorithms in other fields:

- AdaBoost algorithm in machine learning (Freund & Schapire, 1995)
- Winnow algorithm in learning theory (Littlestone, 1988)
- Exponentiated gradient in online learning (Kivinen & Warmuth, 1997)
- Multiplicative update algorithm in game theory in 1950s
- Richardson-Lucy algorithm in imaging processing in 1970s
- Follow-the-regularized-leader (FTRL) in online learning
- Relative Entropy Policy Search in reinforcement learning
- Natural policy gradient (NPG) in reinforcement Learning
- ...

Convergence of Mirror Descent

Let f be convex and $\omega(\cdot)$ be 1-strongly convex on X w.r.t. norm $\|\cdot\|$.

Lemma 10.2

$$\gamma_t(f(\mathbf{x}_t) - f^*) \leq V_\omega(\mathbf{x}^*, \mathbf{x}_t) - V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) + \frac{\gamma_t^2}{2}\|\mathbf{g}_t\|_*^2.$$

Theorem 10.3

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \frac{V_\omega(\mathbf{x}^*, \mathbf{x}_1) + \frac{1}{2}\sum_{t=1}^T \gamma_t^2\|\mathbf{g}_t\|_*^2}{\sum_{t=1}^T \gamma_t}.$$

- Generalizes the previous results for subgradient descent.

Convergence Rate of Mirror Descent

- Suppose f is B-Lipschitz continuous such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$, namely, $\|\mathbf{g}\|_* \leq B$ for any $\mathbf{g} \in \partial f(\mathbf{x})$.

- Define $R^2 := \sup_{\mathbf{x} \in X} V_\omega(\mathbf{x}, \mathbf{x}_1)$, where $R \geq 0$ and set $\gamma_t = \frac{\sqrt{2}R}{B\sqrt{T}}$.

$$\min_{1 \leq t \leq T} f(\mathbf{x}_t) - f^* \leq O\left(\frac{BR}{\sqrt{T}}\right).$$

- Similar results can be obtained when $\gamma_t = \frac{\sqrt{2}R}{B\sqrt{t}}$ or using weighted average.

Convergence of Mirror Descent for Convex Problems

- Generalizes the previous results for subgradient descent.

$$\min_{1\le t\le T} f(\mathbf{x}_t) - f^* = O\left(\frac{BR}{\sqrt{T}}\right),$$

where $R = \sqrt{\max_{\mathbf{x}\in X} V_\omega(\mathbf{x}, \mathbf{x}_1)}$ and $B := \sup_{\mathbf{x}\in X} \frac{|f(\mathbf{x})-f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|}$.

- Subgradient descent: special case with $\|\cdot\| = \|\cdot\|_2$ and $\omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$.

Proof of Lemma 10.2

- Since $\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}\in X}\{V_\omega(\mathbf{x},\mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x}\rangle\}$, by the optimality condition,

$$\langle \nabla\omega(\mathbf{x}_{t+1}) + \gamma_t \mathbf{g}_t - \nabla\omega(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_{t+1}\rangle \geq 0, \forall \mathbf{x} \in X.$$

- From three point identity, we have for $\forall \mathbf{x} \in X$:

$$\langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}\rangle \leq \langle \nabla\omega(\mathbf{x}_{t+1}) - \nabla\omega(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_{t+1}\rangle = V_\omega(\mathbf{x},\mathbf{x}_t) - V_\omega(\mathbf{x},\mathbf{x}_{t+1}) - \underbrace{V_\omega(\mathbf{x}_{t+1},\mathbf{x}_t)}_{\geq \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2}$$

- As a result,

$$\begin{aligned}
\langle \gamma_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^*\rangle &\leq \langle \gamma_t \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}^*\rangle + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle \\
&\leq V_\omega(\mathbf{x}^*,\mathbf{x}_t) - V_\omega(\mathbf{x}^*,\mathbf{x}_{t+1}) - \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \langle \gamma_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle \\
&\leq V_\omega(\mathbf{x}^*,\mathbf{x}_t) - V_\omega(\mathbf{x}^*,\mathbf{x}_{t+1}) + \frac{\gamma_t^2}{2}\|\mathbf{g}_t\|_*^2
\end{aligned}$$

- By convexity of f, we further have the key lemma. $\square$

Subgradient Descent vs. Mirror Descent

$$\min_{1\le t\le T} f(\mathbf{x}_t) - f^* = O\left(\frac{BR}{\sqrt{T}}\right),$$

where $R = \sqrt{\max_{\mathbf{x}\in X} V_\omega(\mathbf{x}, \mathbf{x}_1)}$ and $B := \sup_{\mathbf{x}\in X} \frac{|f(\mathbf{x})-f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|}$.

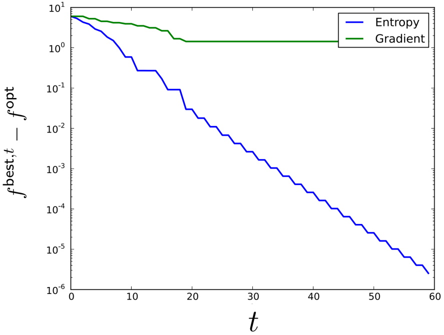
Optimization over simplex:

Assume $\|\mathbf{g}\|_\infty \le 1, \forall \mathbf{g} \in \partial f(\mathbf{x})$ and $X = \Delta_d$. Set $\mathbf{x}_1 = [1/d; \dots; 1/d]$.

- Subgradient Descent: $O\left(\frac{\sqrt{d}}{\sqrt{T}}\right)$,where $B = O(\sqrt{d}), R = O(1)$.
- Mirror Descent: $O\left(\frac{\sqrt{\log d}}{\sqrt{T}}\right)$, where $B = O(1), R = O(\sqrt{\log d})$.

Numerical Illustration: Robust Regression

$$\min_{\mathbf{x}\in\Delta} \; f(\mathbf{x}) = \|Ax - b\|_1 \qquad (A \in \mathbb{R}^{20\times 3000})$$



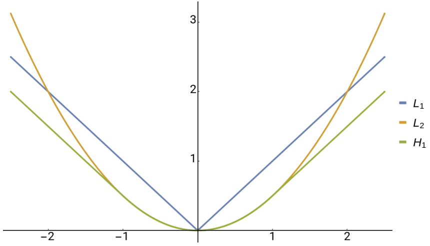
From Boyd's ECE364B lecture

Motivation: absolute value function

Consider the simplest non-smooth and convex function: $f(x) = |x|$.

- Huber function is a smooth approximation of the absolute value function.

$$f_\mu(x) = \begin{cases} \frac{x^2}{2\mu}, |x| \leq \mu \\ |x| - \frac{\mu}{2}, |x| > \mu \end{cases}.$$



- $f_\mu(x) \to f(x)$ as $\mu \to 0$.

$$f(x) - \frac{\mu}{2} \leq f_\mu(x) \leq f(x).$$

- $\nabla f_\mu(x)$ is $\frac{1}{\mu}$-Lipschitz continuous.

Smoothing Idea

Nonsmooth Optimization

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$

$$\Longrightarrow$$

Smooth Optimization

$$\begin{array}{ll} \text{minimize} & f_\mu(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$

- Solving smooth approximation allows for richer and faster algorithms
- Can deal with nonsmooth nonconvex problems
- Desiderata: approximation accuracy, smoothness, computational efficiency

Smoothing Techniques

- Nesterov smoothing (only for convex objectives)
  [Nesterov 2005]
- Moreau-Yosida smoothing/regularization (only for convex objectives)
  [Bauschke et al., 2011]
- Lasry-Lions regularization
  [Lasry and Lions, 1986, Attouch and Aze, 1993]
- Randomized smoothing
  [Duchi et al., 2012]
- ...

A Quick Tour of Convex Conjugate Theory

Definition 10.4

The conjugate function of f is

$$f^{\star}(\mathbf{y}) = \sup_{\mathbf{x}\in\mathbf{dom}(f)} \left\{\mathbf{y}^T\mathbf{x} - f(\mathbf{x})\right\},$$

also called Legendre-Fenchel transformation.

Fenchel's inequality:

$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{x}^T\mathbf{y}, \forall \mathbf{x}, \mathbf{y}$$

A. Legendre

(1752-1833)

Werner Fenchel

(1905-1988)

A Quick Tour of Convex Conjugate Theory

Lemma 10.5 (Chapter C.6, [Nem01])

1. *(Duality) If f is lower semi-continuous (l.s.c.)*[1] *and convex, then $f^{\star\star} = f$.*
2. *(Fenchel's inequality): $\mathbf{x}^T\mathbf{y} = f(\mathbf{x}) + f^\star(\mathbf{y}) \Leftrightarrow \mathbf{y} \in \partial f(\mathbf{x}) \Leftrightarrow \mathbf{x} \in \partial f^\star(\mathbf{y})$.*
3. *If f and g are l.s.c. and convex, then $(f+g)^\star(\mathbf{x}) = \inf_{\mathbf{y}} \{f^\star(\mathbf{y}) + g^\star(\mathbf{x}-\mathbf{y})\}$.*
4. *If f is μ-strongly convex, then $f^\star$ is differentiable and $\frac{1}{\mu}$-smooth.*

[1] Function f is l.s.c. if $f(\mathbf{x}) \leq \liminf_{t\to\infty} f(\mathbf{x}_t)$ for $\mathbf{x}_t \to \mathbf{x}$.

Examples

1. Quadratic: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q \mathbf{x}$ where $Q \succ 0$, $f^\star(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T Q^{-1}\mathbf{y}$.
2. Negative entropy: $f(\mathbf{x}) = \sum_{i=1}^n x_i \log(x_i)$, $f^\star(\mathbf{y}) = \sum_{i=1}^n e^{y_i - 1}$.
3. Negative logarithm: $f(\mathbf{x}) = -\sum_{i=1}^n \log(x_i)$, $f^\star(\mathbf{y}) = -\sum_{i=1}^n \log(-y_i) - n$.
4. Norm: $f(\mathbf{x}) = \|\mathbf{x}\|$, $f^\star(\mathbf{y}) = \begin{cases} 0, & \|\mathbf{y}\|_* \le 1 \\ +\infty, & \|\mathbf{y}\|_* > 1 \end{cases}$.

Smoothing Techniques I: Nesterov's smoothing

$$f_\mu(\mathbf{x}) = \max_{\mathbf{y}\in\mathbf{dom}(f^\star)} \left\{\mathbf{x}^T\mathbf{y} - f^\star(\mathbf{y}) - \mu \cdot d(\mathbf{y})\right\}$$

- Here $f^\star(\mathbf{y})$ is the convex conjugate of f.
- Proximity function: $d(\mathbf{y})$ is 1-strongly convex and nonnegative everywhere.
  - $d(\mathbf{y}) = \frac{1}{2}||\mathbf{y} - \mathbf{y}_0||_2^2$;
  - $d(\mathbf{y}) = \frac{1}{2}\sum w_i(y_i - y_{0,i})^2$ with $w_i \geq 1$;
  - $d(\mathbf{y}) = \omega(\mathbf{y}) - \omega(\mathbf{y}_0) - \nabla\omega(\mathbf{y}_0)^T(\mathbf{y} - \mathbf{y}_0)$ with $\omega(\mathbf{x})$ being 1-strongly convex.

Smoothing Techniques I: Nesterov's smoothing

$$f_\mu(\mathbf{x}) = \max_{\mathbf{y}\in\mathbf{dom}(f^\star)} \left\{\mathbf{x}^T\mathbf{y} - f^\star(\mathbf{y}) - \mu \cdot d(\mathbf{y})\right\}$$

- Smoothness: Function $f_\mu(\mathbf{x})$ is $\frac{1}{\mu}$-smooth.
- Approximation: For convex f with bounded $\mathbf{dom}(f^\star)$, we have

$$f(\mathbf{x}) - \mu D^2 \leq f_\mu(\mathbf{x}) \leq f(\mathbf{x}), \text{ where } D^2 = \max_{\mathbf{y}\in\mathbf{dom}(f^\star)} d(\mathbf{y}).$$

- Tradeoff between approximation error and optimization efficiency:

$$f(\mathbf{x}) - f^* \leq \underbrace{f(\mathbf{x}) - f_\mu(\mathbf{x})}_{\text{approximation error}} + \underbrace{f_\mu(\mathbf{x}) - \min_{\mathbf{x}} f_\mu(\mathbf{x})}_{\text{optimization error}}$$

Smoothing Techniques I: Nesterov's smoothing

- If we apply Accelerated Gradient Descent to solve the smoothed problem:

$$f(\mathbf{x}_t) - f^* \leq O\left(\mu D^2 + \frac{R^2}{\mu t^2}\right).$$

- To achieve accuracy $\epsilon > 0$, need $\mu = O(\frac{\epsilon}{D^2})$.
- The number of AGD iterations is at most $T_\epsilon = O(\frac{R}{\sqrt{\epsilon\mu}}) = O(\frac{RD}{\epsilon})$.
- This is faster than directly applying subgradient methods.

Smoothing Techniques II: Moreau-Yosida Regularization

$$f_\mu(\mathbf{x}) = \min_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2\mu} ||\mathbf{x} - \mathbf{y}||_2^2 \right\}$$

- Here $\mu > 0$ and $f_\mu(\mathbf{x})$ is called the Moreau envelope of $f(\mathbf{x})$.
- Example: Huber function is the Moreau envelope of $f(x) = |x|$:

$$f_\mu(x) = \begin{cases} \frac{x^2}{2\mu}, |x| \leq \mu \\ |x| - \frac{\mu}{2}, |x| > \mu \end{cases} .$$

Smoothing Techniques II: Moreau-Yosida Regularization

$$f_\mu(\mathbf{x}) = \min_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2\mu} ||\mathbf{x} - \mathbf{y}||_2^2 \right\}$$

- Special case of Nesterov's smoothing with $d(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|^2$.

$$\begin{aligned} f_\mu(\mathbf{x}) &= \max_{\mathbf{y}} \left\{ \mathbf{x}^T\mathbf{y} - f^\star(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|_2^2 \right\} \\ &= (f^\star + \frac{\mu}{2}\| \cdot \|_2^2)^\star(\mathbf{x}) \\ &= \inf_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{y}\|_2^2 \right\} \end{aligned}$$

- Smoothness: Function $f_\mu(\mathbf{x})$ is $\frac{1}{\mu}$-smooth.
- Exact Minimization: We have $\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} f_\mu(\mathbf{x})$.

Smoothing Techniques II: Moreau-Yosida Regularization

$$
\begin{aligned}
f_\mu(\mathbf{x}) &= \min_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2\mu}||\mathbf{x}-\mathbf{y}||_2^2 \right\} \\
\mathbf{prox}_{\mu f}(\mathbf{x}) &:= \operatorname*{argmin}_{\mathbf{y}} \left\{ f(\mathbf{y}) + \frac{1}{2\mu}||\mathbf{x}-\mathbf{y}||_2^2 \right\}
\end{aligned}
$$

- Gradient of smooth function: (based on Danskin's theorem or Fenchel duality)

$$
\nabla f_\mu(\mathbf{x}) = \frac{1}{\mu}(\mathbf{x} - \mathbf{prox}_{\mu f}(\mathbf{x}))
$$

- GD on smooth $f_\mu(\mathbf{x})$ reduces to proximal minimization on $f(\mathbf{x})$:

$$
\mathbf{x}_{t+1} = \mathbf{x}_t - \mu \nabla f_\mu(\mathbf{x}_t) \Longleftrightarrow \mathbf{x}_{t+1} = \mathbf{prox}_{\mu f}(\mathbf{x}_t).
$$

Proximal Operators

Definition 10.6

The **proximal operator** of convex function g at $\mathbf{x}$ is defined as

$$\mathbf{prox}_f(\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmin}} \left\{ f(\mathbf{y}) + \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \right\}$$

- For continuous convex function f, $\mathbf{prox}_f(\mathbf{x})$ exists and is unique.
- For many nonsmooth functions, proximal operators can be computed efficiently (*closed form solution, low-cost computation, polynomial time*).

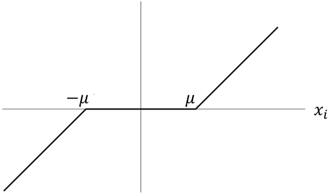
Proximal Operators

Examples:

- If $f(\mathbf{x}) = \delta_X(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in X \\ +\infty, & \mathbf{x} \notin X \end{cases}$, then $\mathbf{prox}_f(\mathbf{x}) = \Pi_X(\mathbf{x})$ is the projection.

- If $f(\mathbf{x}) = \mu\|\mathbf{x}\|_1$, then $\mathbf{prox}_f(\mathbf{x})$ is the soft thresholding operator.

$$\mathbf{prox}_{\mu|\cdot|}(x_i) = \begin{cases} x_i - \mu & \text{if } x_i > \mu \\ 0 & \text{if } |x_i| \leq \mu \\ x_i + \mu & \text{if } x_i < -\mu \end{cases}.$$



Equivalently, $\mathbf{prox}_{\mu\|\cdot\|_1}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \odot \max\{|\mathbf{x}| - \mu, 0\}$.

A non-exhaustive list of proximal operators

Name	Function	Proximal operator	Complexity
ℓ_1-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_1$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \otimes [\lvert\mathbf{x}\rvert - \lambda]_+$	$\mathcal{O}(p)$
ℓ_2-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = [1 - \lambda/\|\mathbf{x}\|_2]_+ \mathbf{x}$	$\mathcal{O}(p)$
Support function	$f(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{C}} \mathbf{x}^T\mathbf{y}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathbf{x} - \lambda \pi_{\mathcal{C}}(\mathbf{x})$	
Box indicator	$f(\mathbf{x}) := \delta_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathcal{O}(p)$
Positive semidefinite cone indicator	$f(\mathbf{X}) := \delta_{\mathbb{S}^p_+}(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X}) = \mathbf{U}[\Sigma]_+\mathbf{U}^T$, where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^T$	$\mathcal{O}(p^3)$
Hyperplane indicator	$f(\mathbf{x}) := \delta_{\mathcal{X}}(\mathbf{x})$, $\mathcal{X} := \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = b\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{\mathcal{X}}(\mathbf{x}) = \mathbf{x} + \left(\frac{b - \mathbf{a}^T\mathbf{x}}{\|\mathbf{a}\|_2}\right)\mathbf{a}$	$\mathcal{O}(p)$
Simplex indicator	$f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x})$, $\mathcal{X} := \{\mathbf{x} : \mathbf{x} \geq 0,\ \mathbf{1}^T\mathbf{x} = 1\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\mathbf{x} - \nu\mathbf{1})$ for some $\nu \in \mathbb{R}$, which can be efficiently calculated	$\tilde{\mathcal{O}}(p)$
Convex quadratic	$f(\mathbf{x}) := (1/2)\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\lambda\mathbb{I} + \mathbf{Q})^{-1}\mathbf{x}$	$\mathcal{O}(p\log p) \to \mathcal{O}(p^3)$
Square ℓ_2-norm	$f(\mathbf{x}) := (1/2)\|\mathbf{x}\|_2^2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (1/(1+\lambda))\mathbf{x}$	$\mathcal{O}(p)$
log-function	$f(\mathbf{x}) := -\log(x)$	$\mathrm{prox}_{\lambda f}(x) = ((x^2 + 4\lambda)^{1/2} + x)/2$	$\mathcal{O}(1)$
$\log\det$-function	$f(\mathbf{x}) := -\log\det(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X})$ is the log-function prox applied to the individual eigenvalues of $\mathbf{X}$	$\mathcal{O}(p^3)$

Source from Volkan Cevher's EE-556 lecture notes. More examples can be found in Parikh & Boyd (2013).

Proximal Point Algorithm

$$\textbf{PPA}: \qquad \mathbf{x}_{t+1} = \mathbf{prox}_{\lambda_t f}(\mathbf{x}_t)$$

Theorem 10.7 (Convergence of PPA)

If f is convex, then for any $T \geq 1$,

$$f(\mathbf{x}_{T+1}) - f^* \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2}{2\sum_{t=1}^{T} \lambda_t}.$$

- Setting $\lambda_t = \lambda$, this implies a $O(1/t)$ convergence rate.

Convergence Proof of Proximal Point Algorithm

Proof.

- First we can prove the following recursion based on optimality of $\mathbf{x}_{t+1}$ (following similar argument as the analysis of Mirror Descent).

$$\lambda_t[f(\mathbf{x}_{t+1}) - f(\mathbf{x})] \leq \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 - \frac{1}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|_2^2 - \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2, \forall \mathbf{x}.$$

- Note that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$.
- Combining these two results leads to the desired result.

$\square$

Smoothing Techniques III: Randomized Smoothing

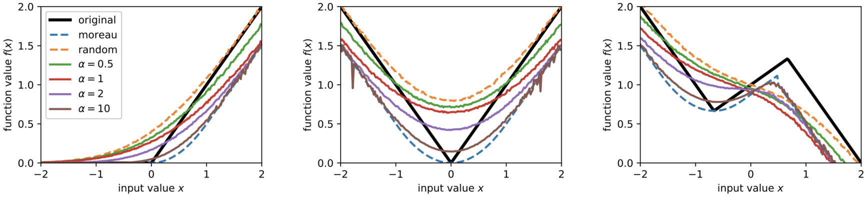
$$f_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{Z}}[f(\mathbf{x} + \mu\mathbf{Z})]$$

where $\mathbf{Z}$ is an isotopic Gaussian or uniform random variable.

- Choosing $\mu = O(\epsilon)$ guarantees ϵ approximation error [Duc12].
- $f_\mu(\mathbf{x})$ is $O(\frac{\sqrt{d}}{\epsilon})$-smooth (dimension dependent) [Duc12].
- Can compute stochastic gradient very efficiently through sampling.

Other Techniques

BMR: Combination of randomized smoothing and Moreau-Yosida smoothing [Sca20]



(a) $f(x) = \max\{0, x\}$. (b) $f(x) = |x|$. (c) f non-convex.

Figure 1: Effect of the parameter α on BMR smoothing (with $\gamma = \min\{1, \alpha^{-1/2}\}$). When $\alpha \to 0$ (resp. $\alpha \to +\infty$), BMR tends to randomized smoothing (resp. Moreau envelope).

Convex Composite Optimization

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad f(\mathbf{x}) + g(\mathbf{x})$$

Assume both f and g are convex.

- $f(\mathbf{x})$ is smooth, $g(\mathbf{x}) = 0$
- $f(\mathbf{x})$ is nonsmooth, $g(\mathbf{x}) = \delta_X(\cdot)$ is indicator function
- $f(\mathbf{x})$ is smooth, $g(\mathbf{x})$ is a "simple" nonsmooth regularizer
- $f(\mathbf{x})$ and $g(\mathbf{x})$ are both "simple" nonsmooth functions
- ....

Application I: Supervised Learning

Most supervised learning problems can be cast into the form:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(\mathbf{x}_i), y_i) + g(\theta)$$

- $\ell(\cdot, \cdot)$ is the loss function, e.g., square loss, hinge loss, logistic loss, etc.
- $h_\theta(\cdot)$ is the predictor, e.g., linear predictor, neural networks, etc.
- $g(\theta)$ is some regularizer, e.g., ℓ_2-norm, ℓ_1-norm, elastic net, etc.
- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are the input data.

Application II: Image Processing

The goal is to recover a clean image $\mathbf{x} \in \mathbb{R}^{n\times m}$ given observation $\mathbf{b} = \mathcal{A}(\mathbf{x}) + \boldsymbol{\epsilon}$.

$$\min_{\mathbf{x}} \ \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_{TV} \qquad \text{(Gaussian noise)}$$

$$\min_{\mathbf{x}} \ \sum_i [\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \log(\langle \mathbf{a}_i, \mathbf{x}\rangle)] + \lambda\|\mathbf{x}\|_{TV} \qquad \text{(Poisson noise)}$$

- $\mathcal{A}(\mathbf{x}) = A\mathbf{x}$ is some linear operator that captures image blur or subsampling.
- Here $\|\mathbf{x}\|_{TV} := \sum_{i,j} |\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}| + |\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}|$ is the total variation norm.

Proximal Gradient Method

Convex composite optimization: $\min_{\mathbf{x}} \; F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$

- f is convex and L-smooth;
- g is convex and proximal-friendly.

Proximal Gradient Method: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

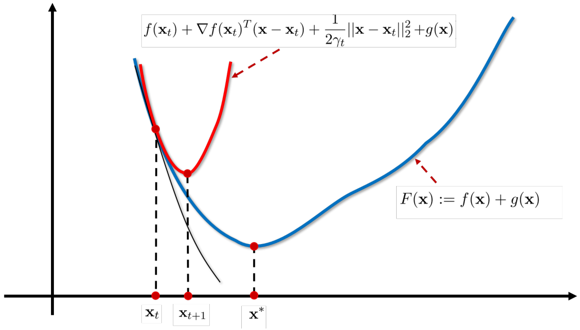
$$\mathbf{x}_{t+1} = \mathbf{prox}_{\gamma_t g}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t)).$$

- Alternates between gradient update and proximal operator.
- Update can be computed efficiently.

Interpretation

Proximal gradient update ≈ majorization-minimization

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} \Big\{ \underbrace{f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^T(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2\gamma_t}\|\mathbf{x}-\mathbf{x}_t\|_2^2}_{\geq f(\mathbf{x}) \quad (\text{if } \gamma_t \leq \frac{1}{L})} + g(\mathbf{x}) \Big\}.$$



Convergence of PGM for Convex Problems

Theorem 10.8

Assume $f(\mathbf{x})$ is convex and L-smooth, $g(\mathbf{x})$ is convex and possibly nonsmooth. Proximal gradient method with fixed step size $\gamma_t = \frac{1}{L}$ satisfies:

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \frac{L||\mathbf{x}_0 - \mathbf{x}^*||_2^2}{2t}.$$

- Behaves as if there is no nonsmooth term $g(\mathbf{x})$.
- Faster than directly applying subgradient method.
- Can be further accelerated with $O(1/t^2)$ rate.

Accelerated Proximal Gradient Method

Accelerated Proximal Gradient: Initialize $\mathbf{x}_0 \in \mathbb{R}^d$ and $\mathbf{y}_0 = \mathbf{x}_0$.

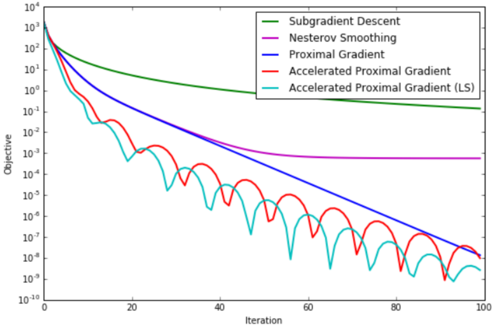
$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{prox}_{\gamma_t g}(\mathbf{y}_t - \gamma_t \nabla f(\mathbf{y}_t)) \\ \mathbf{y}_{t+1} &= \mathbf{x}_{t+1} + \frac{t}{t+3}(\mathbf{x}_{t+1} - \mathbf{x}_t) \end{aligned}$$

- There exist several acceleration schemes, e.g., Nesterov (1983, 2004), Beck and Teboulle (2009), Tseng (2008)
- $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ for convex problems

Example: Lasso

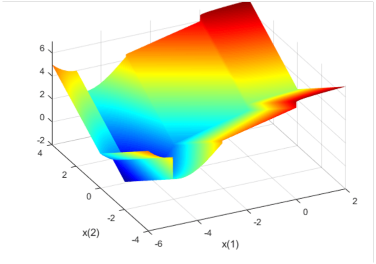
$$\min_{\mathbf{x}} \underbrace{\frac{1}{2}||A\mathbf{x} - b||_2^2}_{f(\mathbf{x})} + \underbrace{\mu||\mathbf{x}||_1}_{g(\mathbf{x})}.$$

Proximal Gradient (a.k.a. ISTA) : $\mathbf{x}_{t+1} = \mathbf{prox}_{\mu\gamma_t\|\cdot\|_1}(\mathbf{x}_t - \gamma_t A^T(A\mathbf{x}_t - b))$.



Summary: Convex Nonsmooth Optimization

- Subgradient Method
- Exploiting non-Euclidean geometry
  - Mirror Descent
- Exploiting nonsmooth structure:
  - Smoothing techniques
  - Proximal point algorithm
  - Proximal gradient methods
  - ....



Additional resources:
Neal Parikh and Stephen Boyd. "Proximal algorithms". Foundations and trends in Optimization 1.3 (2014): 127-239.

Bibliography

Y. Nesterov.
Smooth minimization of non-smooth functions.
Mathematical Programming, 2005.

A. Ben-Tal, A. Nemirovski.
Lectures on modern convex optimization: analysis, algorithms, and engineering applications.
Society for industrial and applied mathematics, 2001.

Jingzhao Zhang, Hongzhou Lin, Suvrit Sra, and Ali Jadbabaie.
Complexity of finding stationary points of nonsmooth nonconvex functions.
International Conference on Machine Learning, 2020.

J. C. Duchi, P. L. Bartlett, M. J. Wainwright.
Randomized smoothing for stochastic optimization.
SIAM Journal on Optimization, 22(2), 674-701, 2012.

K. Scaman, L. Dos Santos, M. Barlier, I. Colin
A Simple and Efficient Smoothing Method for Faster Optimization and Local Exploration.
Advances in Neural Information Processing Systems, 33, pp.6503-6513, 2020

Supplementary Material

Proof of Lemma 10.1

Proof.

This can be easily derived from the definition. We have

$$
\begin{aligned}
V_\omega(\mathbf{x},\mathbf{y}) + V_\omega(\mathbf{y},\mathbf{z}) &= \omega(\mathbf{x}) - \omega(\mathbf{y}) + \omega(\mathbf{y}) - \omega(\mathbf{z}) - \langle \nabla\omega(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle - \langle \nabla\omega(\mathbf{z}), \mathbf{y}-\mathbf{z}\rangle \\
&= V_\omega(\mathbf{x},\mathbf{z}) + \langle \nabla\omega(\mathbf{z}), \mathbf{x}-\mathbf{z}\rangle - \langle \nabla\omega(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle - \langle \nabla\omega(\mathbf{z}), \mathbf{y}-\mathbf{z}\rangle \\
&= V_\omega(\mathbf{x},\mathbf{z}) + \langle \nabla\omega(\mathbf{z}) - \nabla\omega(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle.
\end{aligned}
$$

□

Basic Properties of Proximal Operators

Lemma 10.9

Let g be a convex function, we have

(a) *(Subgradient characterization)*

$$\mathbf{y} = \mathbf{prox}_g(\mathbf{x}) \Longleftrightarrow \mathbf{x} - \mathbf{y} \in \partial g(\mathbf{y}).$$

(b) *(Fixed Point) A point $\mathbf{x}^*$ minimizes $g(\mathbf{x}) \Longleftrightarrow \mathbf{x}^* = \mathbf{prox}_g(\mathbf{x}^*)$.*

(c) *(Non-expansiveness) $\left\|\mathbf{prox}_g(\mathbf{x}) - \mathbf{prox}_g(\mathbf{y})\right\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$.*

Proof follows definition and monotonicity of subgradient.

Interpretation II of Proximal Gradient Methods

Proximal gradient update $\approx$ fixed point iteration

Lemma 10.10

$\mathbf{x}^*$ *is optimal if and only if* $\forall \gamma > 0$: $\mathbf{x}^* = \mathbf{prox}_{\gamma g}(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))$.

Proof.

$$\begin{aligned} & \mathbf{x}^* = \mathbf{prox}_{\gamma g}(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*)) \\ \Leftrightarrow \quad & 0 \in \frac{1}{\gamma}(\mathbf{x}^* - (\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))) + \partial g(\mathbf{x}^*) \\ \Leftrightarrow \quad & 0 \in \nabla f(\mathbf{x}^*) + \partial g(\mathbf{x}^*). \end{aligned}$$

$\square$

Interpretation III of Proximal Gradient Methods

Proximal gradient update $\approx$ forward-backward operator

$$\mathbf{x}_{t+1} = (I + \gamma_t \partial g)^{-1}(I - \gamma_t \nabla f)(\mathbf{x}_t)$$

- $(I - \gamma_t \nabla f)$ is the 'forward' operator;
- $(I + \gamma_t \partial g)^{-1}$, called the resolvent of operator ∂g, is the 'backward' operator.

$$\mathbf{y} = \mathbf{prox}_g(\mathbf{x}) \iff \mathbf{x} \in (I + \partial g)(\mathbf{y}) \iff \mathbf{y} = (I + \partial g)^{-1}(\mathbf{x}).$$

- Also called forward-backward algorithm.

Interpretation IV of Proximal Gradient Methods

Proximal gradient update $\approx$ generalized gradient update

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t G_{\gamma_t}(\mathbf{x}_t)$$

where

$$G_\gamma(\mathbf{x}) := \frac{1}{\gamma}(\mathbf{x} - \mathbf{prox}_{\gamma g}(\mathbf{x} - \gamma \nabla f(\mathbf{x})))$$

- $G_\gamma(\mathbf{x})$ is called the generalized gradient.
- $G_\gamma(\mathbf{x}) = 0$ if and only if $\mathbf{x}$ is optimal.
- $G_\gamma(\mathbf{x}) \in \nabla f(\mathbf{x}) + \partial g(\mathbf{x} - \gamma G_\gamma(\mathbf{x}))$.
  (Easy to check based on the subgradient characterization of proximal operators)

Proof of Theorem 10.8

Lemma 10.11

$$F(\mathbf{x}-\gamma_t G_\gamma(\mathbf{x})) \le F(\mathbf{y})+G_\gamma(\mathbf{x})^T(\mathbf{x}-\mathbf{y})-\frac{\gamma}{2}||G_\gamma(\mathbf{x})||_2^2, \text{ for } \gamma \le 1/L.$$

Applying the inequality at $\mathbf{x} = \mathbf{x}_t$ an $\mathbf{y} = \mathbf{x}^*$, we have:

$$\begin{aligned}
F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) &\le G_\gamma(\mathbf{x}_t)^T(\mathbf{x}_t - \mathbf{x}^*) - \frac{\gamma_t}{2}||G_{\gamma_t}(\mathbf{x}_t)||_2^2 \\
&= \frac{1}{2\gamma_t}\left[||\mathbf{x}_t - \mathbf{x}^*||_2^2 - ||\mathbf{x}_t - \mathbf{x}^* - \gamma_t G_{\gamma_t}(\mathbf{x}_t)||_2^2\right] \\
&= \frac{1}{2\gamma_t}\left[||\mathbf{x}_t - \mathbf{x}^*||_2^2 - ||\mathbf{x}_{t+1} - \mathbf{x}^*||_2^2\right].
\end{aligned}$$

- $F(\mathbf{x}_t)$ is non-increasing (applying $\mathbf{y} = \mathbf{x}_t$).
- $\|\mathbf{x}_t - \mathbf{x}^*\|_2$ is non-increasing ($F(\mathbf{x}_t) \ge F(\mathbf{x}^*)$).
- Taking sums of both sides over all t and setting $\gamma_t = \frac{1}{L}$ leads to desired result.

Proof of Lemma 10.11

- By smoothness of f, we have

$$f(\mathbf{x} - \gamma G_\gamma(\mathbf{x})) \leq f(\mathbf{x}) - \gamma \nabla f(\mathbf{x})^T G_\gamma(\mathbf{x}) + \frac{L\gamma^2}{2} ||G_\gamma(\mathbf{x})||_2^2.$$

- By convexity of f, we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{x})^T (\mathbf{x} - \mathbf{y}).$$

- By convexity of g and the fact that $G_\gamma(\mathbf{x}) - \nabla f(\mathbf{x}) \in \partial g(\mathbf{x} - \gamma G_\gamma(\mathbf{x}))$ we have

$$g(\mathbf{x} - \gamma G_\gamma(\mathbf{x})) \leq g(\mathbf{y}) + (G_\gamma(\mathbf{x}) - \nabla f(\mathbf{x}))^T (\mathbf{x} - \mathbf{y} - \gamma G_\gamma(\mathbf{x})).$$

Combing the above three inequalities lead to the desired result in $(\star)$. $\square$

Proximal Gradient with Backtracking Line-search

In practice, we often do not know L a priori. How to choose stepsize?

We can use backtracking line-search to find the local Lipschitz constant.

- Initialize $L_0 = 1$ and some $\alpha > 1$.
- At each iteration t, we find the smallest integer i such that $L = \alpha^i L_{t-1}$ satisfies the Lipschitz condition:

$$f(\mathbf{x}^+) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)(\mathbf{x}^+ - \mathbf{x}_t) + \frac{L}{2}||\mathbf{x}^+ - \mathbf{x}_t||_2^2$$

  where $\mathbf{x}^+ = \mathbf{prox}_{\frac{g}{L}}(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t))$.
- Then update $L_t = L$ and $\mathbf{x}_{t+1} = \mathbf{x}^+$.