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

Lecture 9: Subgradient Methods

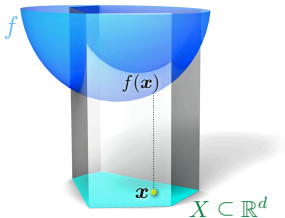
Bernd Gärtner
Niao He

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

April 18, 2023

Overview

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



- $X = \mathbb{R}^d$, $f(\mathbf{x})$ is convex and smooth:
  Gradient Descent, Coordinate Descent, Newton Method, etc.

- $X \subsetneq \mathbb{R}^d$, X is convex, $f(\mathbf{x})$ is convex and smooth:
  Projected Gradient Descent, Frank-Wolfe
  X is a closed convex set.

this one is very sure, need to pay attention to it.

Q: What if $f(\mathbf{x})$ is non-smooth and possibly non-differentiable?

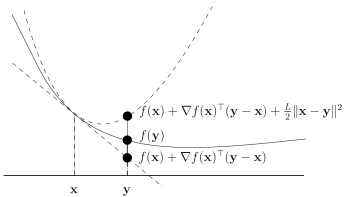
Smooth vs. Nonsmooth

Informally, in the context of optimization:

- Smooth: f is differentiable and has Lipschitz continuous gradients on $\mathbf{dom}(f)$:

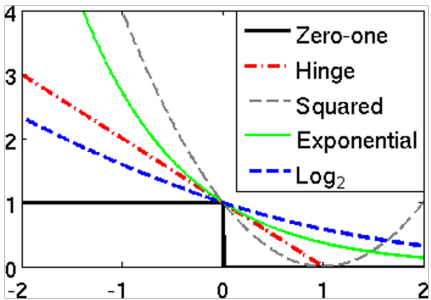
$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})||_2 \leq L||\mathbf{x} - \mathbf{y}||_2$$

$$\Rightarrow \quad f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$



- Nonsmooth: the negative description
  - f may not always be differentiable
  - Gradients of f, even if exist, may not always be Lipschitz continuous

Which loss function(s) used in classification are nonsmooth?



$s = y \cdot y'$

- $y \in \{+1, -1\}$ is true label
- y' is predicted label or score

check this part for the further result.

- 0-1 Loss: $f(s) = \begin{cases} 1, & s < 0 \\ 0, & s \geq 0 \end{cases}$
- Hinge losss: $f(s) = \max(0, 1 - s)$
- Squared loss: $f(s) = (s-1)^2$ this one is smooth.
- Exponential loss: $f(s) = e^{-s}$ this one is smooth.
- Logistic loss: $f(s) = \log(1 + e^{-s})$ this one is also smooth.

Numerous Applications in Machine Learning

we need to infd way to handle these non-smooth functions.

Non-smoothness arises everywhere in machine learning:

- Loss function: e.g., hinge loss, perceptron loss, ℓ_1-loss, etc.
- Regularization: e.g., ℓ_1-norm, total variation, elastic net, etc.
- Activation function in neural network: e.g., ReLU, Leaky ReLU etc.
- Likelihood: Laplacian noise, etc.

In contrast, non-smooth functions have points or regions where the derivatives do not exist, are discontinuous, or exhibit other irregular behaviors.

Examples: Regression and Classification

Least Absolute Shrinkage and Selection Operator (LASSO)

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\mathbf{x}_i)^2 + \lambda\|\boldsymbol{w}\|_1$$

where $\|\boldsymbol{w}\|_1 := \sum_{j=1}^{d}|w_j|$ is the ℓ_1-regularization.

Soft Margin Support Vector Machines (SVM)

$$\min_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\max\{1 - y_i(\boldsymbol{w}^\top\mathbf{x}_i + b), 0\}$$

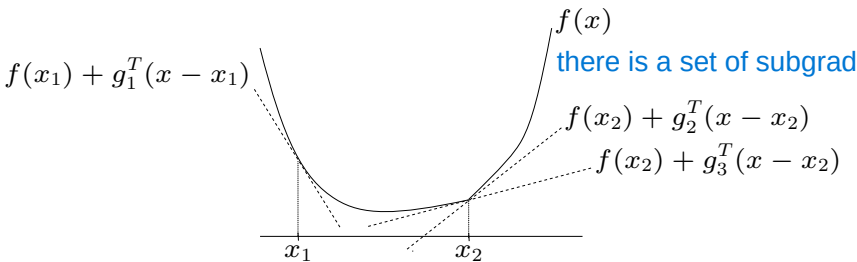
Lecture Outline

Subgradients

Definition 9.1

$\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbf{dom}(f)$$



$\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$ is the subdifferential, the set of subgradients of f at $\mathbf{x}$.

Examples

it is a set for it.

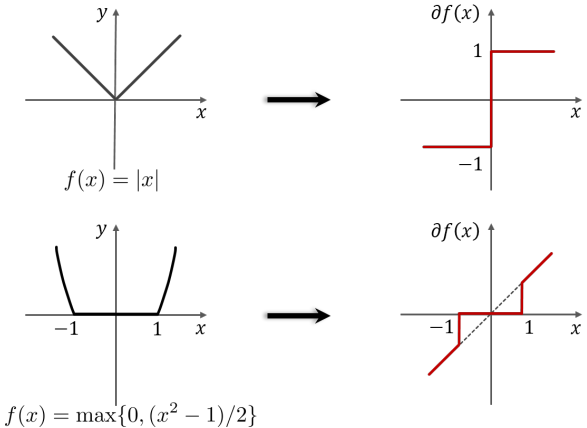


Figure: Examples of subdifferential sets

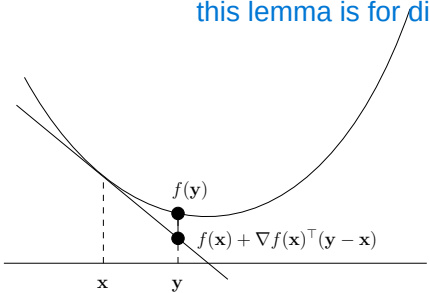
Subgradients of differentiable functions

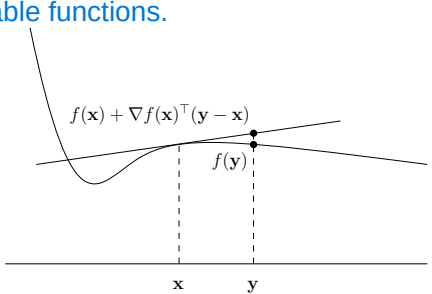
Lemma 9.2 (Exercise)

If f is differentiable at $\mathbf{x} \in \mathbf{dom}(f)$, then $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.

Either exactly one subgradient $\nabla f(\mathbf{x})\ldots$ $\quad$ $\ldots$ or no subgradient at all.

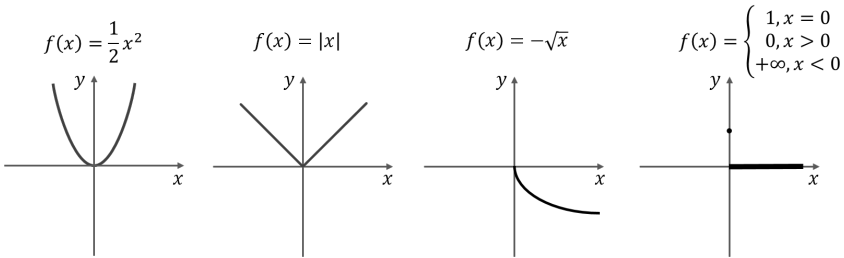
this lemma is for differentiable functions.





Clicker Question

Which of the (convex) function(s) below does NOT have a subgradient at $x = 0$?



c might be the answer based on the definition above.

Subgradient characterization of convexity

"convex = subgradients everywhere"

Lemma 9.3

(i) *If f is convex, then $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x}$ in the (relative) interior*[1] *of $\mathbf{dom}(f)$.*

(ii) *If $\mathbf{dom}(f)$ is convex and $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in \mathbf{dom}(f)$, then f is convex.*

Proof hint: (i) Hyperplane separation theorem (ii) following definition of subgradient.

[1] See Supplementary slide for formal definition and examples.

Hyperplane Separation Theorem

Theorem 9.4 ([Roc97, Theorem 11.3])

Two nonempty convex sets can be separated by a hyperplane if their (relative) interiors do not intersect.

this set is convex

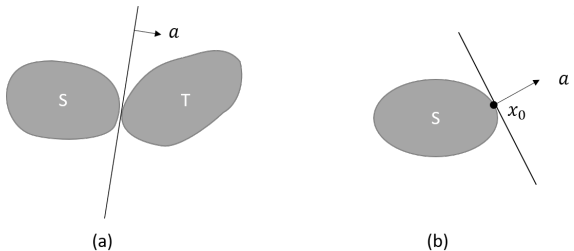


Figure: (a) Separation of two sets, (b) Separation of a convex set and a boundary point

Differentiability of convex functions

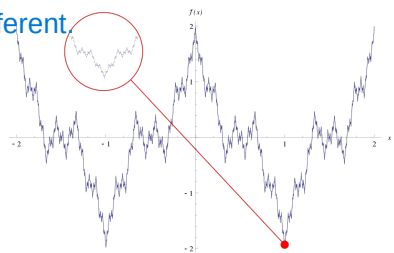
Q: How "wild" can a non-differentiable convex function be?

Theorem 9.5 ([Roc97, Theorem 25.5])

A convex function f is differentiable almost everywhere on $\mathbf{dom}(f)$.

this function is very different.

- Not true for nonconvex functions.
- Recall Weierstrass function is continuous everywhere but differentiable nowhere.



Subgradient optimality condition

Lemma 9.6

Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ *and* $\mathbf{x} \in \mathbf{dom}(f)$. *If* $\mathbf{0} \in \partial f(\mathbf{x})$, *then* $\mathbf{x}$ *is a global minimum*.

this lemma seems important for finding the global minimum.

Proof.

By definition of subgradients, $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$ gives

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all $\mathbf{y} \in \mathbf{dom}(f)$, so $\mathbf{x}$ is a global minimum. □

Calculus of Subgradient and Subdifferential

three rules for the canculation.

- Conic combination: Let $h(\mathbf{x}) = \beta_1 f_1(\mathbf{x}) + \beta_2 f_2(\mathbf{x})$ with $\beta_1, \beta_2 \geq 0$,

$$\partial h(\mathbf{x}) = \beta_1 \partial f_1(\mathbf{x}) + \beta_2 \partial f_2(\mathbf{x}).$$

- Affine transformation: Let $h(\mathbf{x}) = f(A\mathbf{x} + b)$,

$$\partial h(\mathbf{x}) = A^\top \partial f(A\mathbf{x} + b).$$

- Pointwise maximum: Let $h(\mathbf{x}) = \max_{i=1,\ldots,m} f_i(\mathbf{x})$,

$$\partial h(x) = \operatorname{conv}\{\partial f_i(\mathbf{x}) : i \text{ such that } f_i(\mathbf{x}) = h(\mathbf{x})\} \text{ (convex hull)}$$

Many other rules...In practice, it is easy to obtain a subgradient.

Lecture Outline

Recap

$\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbf{dom}(f)$$

So far, we have shown that

- For convex functions, subgradients always exist in the interior.
- Subgradients share lots of similar properties as gradients.
- Subgradients enjoy nice calculus and can be computed easily.
- $dist(0, \partial f(\mathbf{x}))$ characterizes the suboptimality.

Clicker Question: Make a Guess

For a convex function, is negative subgradient a descent direction at $\mathbf{x} \neq \mathbf{x}^*$?

A Always.

B Never.

C It depends. C

Note: $\mathbf{d}$ is a descent direction at $\mathbf{x}$ if you can decrease the function value along that direction with small stepsize, namely, $f'(\mathbf{x}; \mathbf{d}) := \lim_{\delta \to 0^+} \frac{f(\mathbf{x}+\delta\mathbf{d}) - f(\mathbf{x})}{\delta} < 0$.

Remark

- Negative subgradient may not be a descent direction.

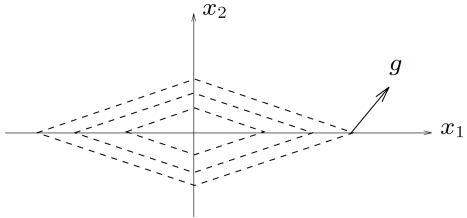


Figure: Contours of function $f(x_1, x_2) = |x_1| + 2|x_2|$

- At $\mathbf{x} = (1, 0)$, $\partial f(\mathbf{x}) = \{(1, a) : a \in [-2, 2]\}$.
- Consider $\mathbf{g} = (1, 0)$, $\mathbf{d} = -\mathbf{g}$ is a descent direction.
- Consider $\mathbf{g} = (1, 2)$, $\mathbf{d} = -\mathbf{g}$ is not a descent direction.

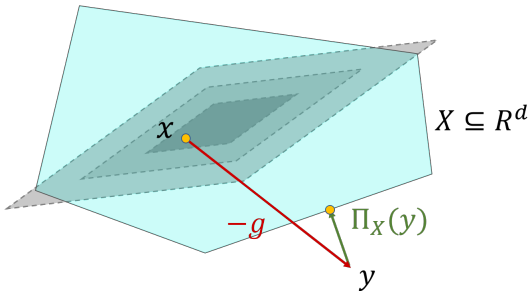
Q: If we replace GD by subgradient, will it converge?

Solving Convex Nonsmooth Problems

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

Assume the optimal value $f^* < \infty$ and the optimal set $X^* \neq \emptyset$.

Idea: move along negative subgradient and project onto X after every step



Subgradient Descent

Subgradient descent: choose $\mathbf{x}_1 \in \mathbb{R}^d$.

$$\mathbf{x}_{t+1} \quad := \quad \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t), \text{ where } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$

- $\mathbf{g}_t$ is a subgradient of f at $\mathbf{x}_t$.
- $\gamma_t > 0$ is a proper (time-varying) stepsize.
- $\Pi_X(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|_2^2$ is the projection.

Subgradient Descent

Subgradient descent: choose $\mathbf{x}_1 \in \mathbb{R}^d$.

$$\mathbf{x}_{t+1} := \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t), \text{ where } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$

- When f is differentiable and $X = \mathbb{R}^d$, this reduces to Gradient Descent.
- When f is differentiable and $X \subset \mathbb{R}^d$, this reduces to Projected Gradient Descent.
- When f is non-differentiable, we see that it is not always a descent method.

Q: Does it converge? If so, how fast?

Clicker Question

Q3: Let f be convex. Consider the update $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{g}_t$, where $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ and $\|\mathbf{g}_t\|_2 \leq 1$ for any t. Which one of the following stepsizes might not lead to convergence: $\lim_{t\to\infty} f(\mathbf{x}_t) = f^*$?

A

A $\gamma_t = 0.001$

B $\gamma_t = \frac{10}{\sqrt{t}}$

C $\gamma_t = \frac{100}{t}$

Illustration

$$\min_{\mathbf{x}} \quad \|A\mathbf{x} - b\|_1$$

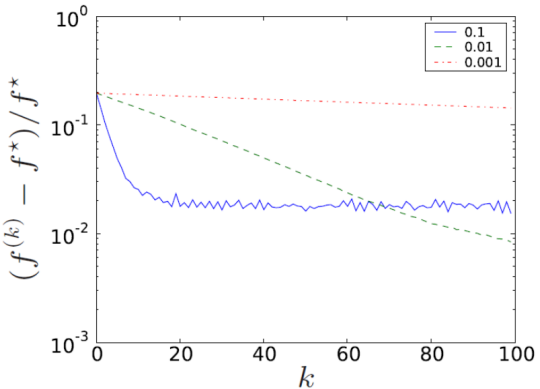


Figure: Convergence of Subgradient Descent under different stepsize

Basic "Descent" Lemma

Lemma 9.7

If f is convex, then for any optimal solution $\mathbf{x}^* \in X^*$,

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t(f(\mathbf{x}_t) - f^*) + \gamma_t^2\|\mathbf{g}_t\|_2^2. \quad (\star)$$

Proof.

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\Pi_X(\mathbf{x}_t - \gamma_t\mathbf{g}_t) - \mathbf{x}^*\|_2^2 \\ &\leq \|\mathbf{x}_t - \gamma_t\mathbf{g}_t - \mathbf{x}^*\|_2^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t\mathbf{g}_t^T(\mathbf{x}_t - \mathbf{x}^*) + \gamma_t^2\|\mathbf{g}_t\|^2 \end{aligned}$$

Due to convexity of f, we have $f^* \geq f(\mathbf{x}_t) + \mathbf{g}_t^T(\mathbf{x}^* - \mathbf{x}_t)$, i.e.

$$\mathbf{g}_t^T(\mathbf{x}_t - \mathbf{x}^*) \geq f(\mathbf{x}_t) - f^*.$$

Main Theorem on Convergence

Theorem 9.8

If f is convex, then the subgradient method satisfies:

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

this one seems not appear in the exams

- Denote the best solution among $x_1, \ldots, x_t$ as $\mathbf{x}_T^{\text{best}}$ such that $f(\mathbf{x}_t^{\text{best}}) = \min_{1\leq\tau\leq t} f(\mathbf{x}_\tau)$.
- Same bound holds for the weighted average solution $\bar{\mathbf{x}}_T = \frac{\sum_{t=1}^T \gamma_t \mathbf{x}_t}{\sum_{t=1}^T \gamma_t}$.
- Proof follows by telescoping sum of $(\star)$ and invoking convexity.

Asymptotic Convergence Under Different Stepsizes

Assume $\|\mathbf{g}_t\|_2 \leq B, \forall t$.

- Constant stepsize: $\gamma_t \equiv \gamma$

$$\lim_{t\to\infty} f(\mathbf{x}_t^{\text{best}}) \leq f^* + B^2\gamma/2.$$

- Scaled stepsize: $\gamma_t = \frac{\gamma}{\|\mathbf{g}_t\|_2}$

$$\lim_{t\to\infty} f(\mathbf{x}_t^{\text{best}}) \leq f^* + B\gamma/2.$$

- Square-summable stepsize: $\sum_{t=1}^{\infty}\gamma_t^2 < +\infty$ and $\sum_{t=1}^{\infty}\gamma_t = +\infty$

$$\lim_{t\to\infty} f(\mathbf{x}_t^{\text{best}}) = f^*.$$

- Diminishing stepsize: $\gamma_t \to 0$ and $\sum_{t=1}^{\infty}\gamma_t = +\infty$

$$\lim_{t\to\infty} f(\mathbf{x}_t^{\text{best}}) = f^*. \quad \text{(why?)}$$

Asymptotic Convergence Under Polyak's Stepsize

choosing step size is an art, since it is very different with the one.

- Minimizing the surrogate function in $(\star)$ yields the optimal stepsize (Polyak, 1987):

$$\gamma_t = \frac{f(\mathbf{x}_t) - f^*}{\|\mathbf{g}_t\|_2^2}$$

- This guarantees strict error reduction:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{(f(\mathbf{x}_t) - f^*)^2}{\|g(\mathbf{x}_t)\|_2^2}$$

- It follows that $f(\mathbf{x}_t) \to f^*$ since $\sum_{t=1}^{\infty}(f(\mathbf{x}_t) - f^*)^2 \leq B^2\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 < \infty$.
- We can further prove that $\{\mathbf{x}_t\} \to \mathbf{x}^*$ for some $\mathbf{x}^*$ (See supplementary material).

Polyak's Stepsize

- Useful when the optimal value f^* is known
  - Overparametrized neural networks, $h(\mathbf{a}_i; \mathbf{x}^*) = b_i, \forall i = 1, \dots, n$

$$\min_{\mathbf{x}} \frac{1}{n} \sum_{i=1}^{n} (h(\mathbf{a}_i; \mathbf{x}) - b_i)^2.$$

- In practice, the optimal value is often not available. One can replace f^* by an online estimate, e.g., $\hat{f}_t := \min_{0 \le \tau \le t} f(\mathbf{x}_\tau) - \delta$.

Convergence Rate for Convex Lipschitz Problems

Corollary 9.9

Assume f is convex and B-Lipschitz continuous, namely, $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in X$, and X is convex compact with $R^2 := \max_{\mathbf{x},\mathbf{y} \in X} \|\mathbf{x} - \mathbf{y}\|_2 < +\infty$. Let $\gamma_t \equiv \frac{R}{B\sqrt{T}}$, then the subgradient descent satisfies

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

there is a upper bound for it.

- Lipschitz continuity implies $\|\mathbf{g}\|_2 \leq B$ for all $\mathbf{g} \in \partial f(\mathbf{x})$. (Excecise)
- Similar results can be obtained by setting $\gamma_t = \frac{R}{B\sqrt{t}}$. (Excecise)
- Subgradient method converges sublinearly for convex nonsmooth problems.
- For an accuracy $\epsilon > 0$, need $O(\frac{B^2R^2}{\epsilon^2})$ number of iterations or subgradients.

Convergence Rate for Strongly Convex Lipschitz Problems

strong convex is a bit of different.

Theorem 9.10

Let f be μ-strongly convex and B-Lipschitz continuous on X, then with $\gamma_t = \frac{2}{\mu(t+1)}$, subgradient descent satisfies

$$\min_{1 \le t \le T} f(\mathbf{x}_t) - f^* \le \frac{2B^2}{\mu \cdot (T+1)}.$$

- For an accuracy $\epsilon > 0$, need $O(\frac{B^2}{\mu\epsilon})$ number of iterations or subgradients.
- Proof follows by extending the descent lemma and invoking the strong convexity. Full proof is provided in supplementary material.

Summary

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

Subgradient Descent

remember this table, put this one on the note.

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

	Convex	Strongly Convex
Convergence rate	$O\left(\frac{B \cdot R}{\sqrt{t}}\right)$	$O\left(\frac{B^2}{\mu t}\right)$
Subgradient complexity	$O\left(\frac{B \cdot R}{\epsilon^2}\right)$	$O\left(\frac{B^2}{\mu \epsilon}\right)$

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

Lecture Outline

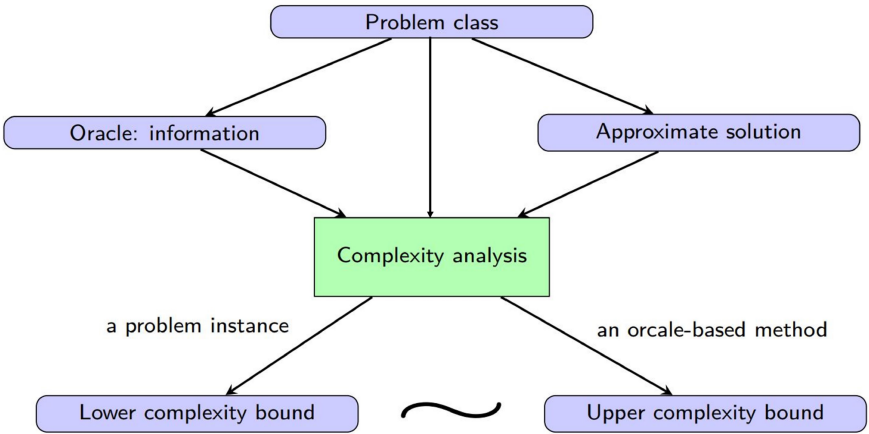
Subgradient Descent vs. Gradient Descent

Setting	Algorithm	**Convex**	**Strongly Convex**
Nonsmooth	**Subgradient method**	$O\left(\frac{B \cdot R}{\sqrt{t}}\right)$	$O\left(\frac{B^2}{\mu t}\right)$
Smooth	**Gradient descent**	$O\left(\frac{L \cdot R^2}{t}\right)$	$O\left(\left(1-\frac{\mu}{L}\right)^t\right)$
	Accelerated gradient descent	$O\left(\frac{L \cdot R^2}{t^2}\right)$	$O\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^t\right)$

this table is a summary for it.

Q: The rates of subgradient method seem much worse than gradient descent. Can we further improve on it?

Problem Class, Oracles, Complexity



Lower Complexity Bound for Nonsmooth Convex Optimization

- In the worst case, the sublinear rates $O(1/\sqrt{t})$ and $O(1/t)$ for convex and strongly convex Lipschitz problems cannot be improved, for algorithms using only subgradient oracles.
- Subgradient descent is "optimal" for such problem classes.

Theorem 9.11 (Nemirovski & Yudin, 1983)

For any $1 \leq t \leq d$, $\mathbf{x}_1 \in \mathbb{R}^d$,there exists a B-Lipschitz continuous and convex function f, a convex set X with diameter R, such that for any first-order method that generates:

$$\mathbf{x}_t \in \mathbf{x}_1 + span(\mathbf{g}_1, ..., \mathbf{g}_{t-1}), \text{ where } \mathbf{g}_i \in \partial f(\mathbf{x}_i), i = 1, \ldots, t-1,$$

we have $\min_{1 \leq s \leq t} f(\mathbf{x}_s) - f^* \geq \frac{B \cdot R}{4(1+\sqrt{t})}$.

Bibliography

R. Tyrrell Rockafellar.
Convex Analysis.
Princeton Landmarks in Mathematics. Princeton University Press, 1997.

A. Nemirovski and D. Yudin.
Problem complexity and method efficiency in optimization.
Wiley-Interscience, 1983.

A. Juditsky, and A. Nemirovski.
First order methods for nonsmooth convex large-scale optimization.
Optimization for Machine Learning, 2011.

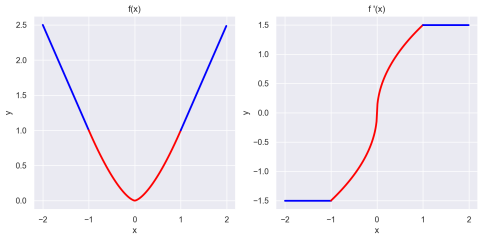
Supplementary Material

Miscellaneous

Let f be convex and differentiable.

- **Lipschitz continuity $\not\Rightarrow$ smoothness**

$$f(x) = \begin{cases} |x|^{3/2}, & |x| \leq 1 \\ \frac{3}{2}|x| - \frac{1}{2}, & |x| \geq 1 \end{cases}$$



- **Smoothness $\not\Rightarrow$ Lipschitz continuity**

$$f(x) = x^2, x \in (-\infty, \infty); \quad f(x) = x \log x, x \in [1, \infty)$$

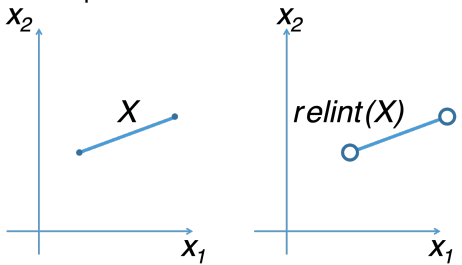
Clarification on Relative Interior

Let X be convex with nonempty interior.

- The interior of X is dense in the closure of X. Note in general for nonconvex set, its interior and closure can differ dramatically: for example, if X irrationals on $[0, 1]$, its interior is $\emptyset$ and its closure is $[0, 1]$.
- The relative interior of X is defined as

$$\text{reint}(X) = \{x : \exists r > 0, \text{ such that } B(x, r) \cap \text{Aff}(X) \subseteq X\},$$

which is the set of interior points relative to the affine subspace that contains X.



Clarification on iterate convergence under Polyak stepsize

We can show that under Polyak's stepsize, there exists an optimal solution $\mathbf{x}^* \in X^*$ such that $\limsup_{t\to\infty} \|\mathbf{x}_t - \mathbf{x}^*\|_2 \to 0$.

Recall that the strict error reduction holds for any $\mathbf{x}^* \in X^*$

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{\left(f(\mathbf{x}_t) - f^*\right)^2}{\|g(\mathbf{x}_t)\|_2^2}$$

- First of all, note that $\|\mathbf{x}_t - \mathbf{x}^*\|_2, t \geq 1$ are bounded and non-increasing. There exists a subsequence $\{\mathbf{x}_{t_k}\}$ with accumulation point $\hat{\mathbf{x}}$.
- Since we already show that $f(\mathbf{x}_t) \to f^*$, it further implies that $\hat{\mathbf{x}} \in X^*$.
- Therefore, we also have

$$\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \leq \|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\|_2^2 - \sum_{j=t_k}^{t} \frac{\left(f(\mathbf{x}_j) - f^*\right)^2}{\|g(\mathbf{x}_j)\|_2^2}.$$

Let $t \to \infty$, we have $\limsup_{t\to\infty} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \leq \|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\|_2^2 - \sum_{j=t_k}^{\infty} \frac{(f(\mathbf{x}_j)-f^*)^2}{\|g(\mathbf{x}_j)\|_2^2}$.

Now let $k \to \infty$, we have $\limsup_{t\to\infty} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \leq \lim_{k\to\infty} \|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\|_2^2 = 0$.

Descent Direction

Definition 9.12

The direction $\mathbf{d} \in \mathbb{R}^d$ is a descent direction if

$$f'(\mathbf{x}; \mathbf{d}) := \lim_{\delta \to 0^+} \frac{f(\mathbf{x} + \delta \mathbf{d}) - f(\mathbf{x})}{\delta} < 0.$$

Note

- If f is differentiable, $f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d}$.
- If f is convex and $\mathbf{x} \in \text{int}(\mathbf{dom}(f))$, $f'(\mathbf{x}, \mathbf{d}) = \max_{\mathbf{g} \in \partial f(\mathbf{x})} \mathbf{g}^T \mathbf{d}$. [Roc97, Theorem 23.4]

Complete Proof of Theorem 9.10

Proof. We first establish the following descent lemma following similar proof as Lemma 9.7 by invoking strong convexity.

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq (1 - \mu\gamma_t)\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t(f(\mathbf{x}_t) - f^*) + \gamma_t^2\|\mathbf{g}_t\|_2^2. \ (\star\star)$$

From $(\star\star)$, we have

$$\begin{aligned}(f(\mathbf{x}_t) - f^*) &\leq \frac{1 - \mu\gamma_t}{2\gamma_t}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{1}{2\gamma_t}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \frac{\gamma_t}{2}\|\mathbf{g}_t\|_2^2 \\ &= \frac{\mu(t-1)}{4}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{\mu(t+1)}{4}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \frac{1}{\mu(t+1)}\|\mathbf{g}_t\|_2^2\end{aligned}$$

Hence, $\sum_{t=1}^T t(f(\mathbf{x}_t) - f^*) \leq -\frac{\mu T(T+1)}{4}\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2 + \frac{T}{\mu}B^2$.

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \frac{TB^2/\mu}{\sum_{t=1}^T t} = \frac{2B^2}{\mu \cdot (T+1)}$$

Lower Bound for Convex Lipschitz Problem

Theorem 9.13 (Nemirovski & Yudin, 1983)

For any $1 \leq t \leq d$, $\mathbf{x}_1 \in \mathbb{R}^d$,there exists a B-Lipschitz continuous and convex function f, a convex set X with diameter R, such that for any first-order method that generates:

$$\mathbf{x}_t \in \mathbf{x}_1 + span(\mathbf{g}_1, ..., \mathbf{g}_{t-1}), \text{ where } \mathbf{g}_i \in \partial f(\mathbf{x}_i), i = 1, \ldots, t-1,$$

we always have

$$\min_{1 \leq s \leq t} f(\mathbf{x}_s) - f^* \geq \frac{B \cdot R}{4(1 + \sqrt{t})}.$$

Lower Bound for Strongly Convex Lipschitz Problem

Theorem 9.14 (Nemirovski & Yudin, 1983)

For any $1 \leq t \leq d$, $\mathbf{x}_1 \in \mathbb{R}^d$,there exists a μ-strongly convex, B-Lipschitz continuous function f and a convex set X, such that for any first-order method that generates:

$$\mathbf{x}_t \in \mathbf{x}_1 + span(\mathbf{g}_1, ..., \mathbf{g}_{t-1}), \text{ where } \mathbf{g}_i \in \partial f(\mathbf{x}_i), i = 1, \ldots, t-1,$$

we always have

$$\min_{1 \leq s \leq t} f(\mathbf{x}_s) - f^* \geq \frac{B^2}{8\mu t}.$$

Proof of Lower Bound I

Consider the hard instance $\min_{\mathbf{x}\in X} f(\mathbf{x})$, where

$$f(\mathbf{x}) = C \cdot \max_{1\le i\le t} x_i + \frac{\mu}{2}\|\mathbf{x}\|_2^2,$$

$$X = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le R/2\}$$

- The subdifferential set of function f:

$$\partial f(\mathbf{x}) = \mu\mathbf{x} + C \cdot \mathrm{conv}\{\boldsymbol{e}_i : i \text{ that } x_i = \max_{1\le j\le t} x_j\}$$

  - Subgradient oracle: Given an input x, it returns $\mathbf{g} = C \cdot \boldsymbol{e}_i + \mu\mathbf{x}$, with $i = \min\{i : x_i = \max_{1\le j\le t} x_j\}$.

- The optimal solution and optimal value:

$$(\mathbf{x}^*)_i = \begin{cases} -\frac{C}{\mu t} & 1 \le i \le t \\ 0 & t < i \le d \end{cases} \quad \text{and } f^* = -\frac{C^2}{2\mu t}.$$

Proof of Lower Bound II

- W.l.o.g., set $\mathbf{x}_1 = 0$.
- By induction, we can show that $\mathbf{x}_t \in span(\boldsymbol{e}_1, ..., \boldsymbol{e}_{t-1})$.
- This implies for $1 \leq s \leq t$, $f(\mathbf{x}_s) \geq 0$.

$$\min_{1 \leq s \leq t} f(\mathbf{x}_s) - f^* \geq \frac{C^2}{2\mu t}.$$

- If $C = \frac{B}{2}$, $\mu = \frac{B}{R}$, then $f(\mathbf{x})$ is B-Lipschitz continuous and μ-strongly convex,

$$\min_{1 \leq s \leq t} f(\mathbf{x}_s) - f^* \geq \frac{C^2}{2\mu t} = \frac{B^2}{8\mu t}.$$

- If $C = \frac{B\sqrt{t}}{1+\sqrt{t}}$, $\mu = \frac{2B}{R(1+\sqrt{t})}$, then $f(\mathbf{x})$ is B-Lipschitz continuous,

$$\min_{1 \leq s \leq t} f(\mathbf{x}_s) - f^* \geq \frac{C^2}{2\mu t} = \frac{BR}{4(1+\sqrt{t})}.$$