Advanced Optimization: Accelerated Gradient Methods, Quasi-Newton Methods

Richard Liu

School of Mathematics, XMU

August 9, 2020

Content

Source

- Michael W. Mahoney, et al. *The Mathematics of Data*
- Amir Beck, *First-Order Methods in Optimization*
- Wen Huang, *Numerical Optimization Course in XMU*

Section 1

Introduction

Introduction

- This chapter will bring with more advanced optimization tools, which requires higher-level mathematical knowledge.
  - Gradient Descent $\rightarrow$ Accelerating Gradient Method
  - Newton Method $\rightarrow$ Quasi-Newton Method
- Better performance but more restrictions on the function f.

Section 2

Smoothness and Convexity

Smoothness

Definition 1: L-smoothness

$L \geq 0, f : \mathbb{R}^n \to (-\infty, \infty]$ is said to be L-smooth over a set $D \subset \mathbb{R}^n$ if it is differentiable over D and $\|\nabla f(x) - \nabla f(y)\|_\star \leq L\|x - y\|, \forall x, y \in D$

- $\|\cdot\|_\star$: Dual Norm, but for 2-norm, they have the same meaning.
- A significant requirement for the following methods to work.
- Tightly connected with convexity. (Not mentioned in this chapter)
- Example: $f(x) = \langle b, x \rangle + c$.

Smoothness Inequality

Lemma 1: Descent Lemma

Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be an L-smooth function with $L \geq 0$ over a given convex set D, then for any $x, y \in D$,
$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2$

Proof

Note that $f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x\rangle\, dt$, simple transformation yields $|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\,| =$
$|\int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle\, dt| \leq$
$\int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_\star \|y - x\| dt \leq \frac{L}{2}\|y - x\|^2$, $\square$

Convexity

Definition 2: σ-Strong Convexity

A function $f : \mathbb{R}^n \to (-\infty, \infty]$ is called σ-strongly convex for a given $\sigma > 0$ if $\mathrm{dom}(f)$ is convex and the following inequality holds for any $x, y \in \mathrm{dom}(f)$, $\lambda \in [0, 1]$,
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\sigma}{2}\lambda(1-\lambda)\|x-y\|^2$$

Proposition 1

$f : \mathbb{R}^n \to (-\infty, \infty]$ is σ-strongly convex function with $\sigma > 0$ iff $f(\cdot) - \frac{\sigma}{2}\|\cdot\|^2$ is convex.

Convexity

Proof

Let $g(x) = f(x) - \frac{\sigma}{2}\|x\|^2$, then we could find that $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$ is equivalent to $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\sigma}{2}\lambda(1-\lambda)\|x-y\|^2$, $\square$

- Meaning?

Convexity: First-Order Characterization

Lemma 2

Let f be a proper closed and convex function with the same domain and range as before, then for a given $\sigma > 0$, the following claims are equivalent.

1. f is σ-strongly convex.
2. $f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\sigma}{2} \|y - x\|^2, \forall x \in \mathrm{dom}(\partial f), y \in \mathrm{dom}(f), g \in \partial f(x)$
3. $\langle g_x - g_y, x - y \rangle \geq \sigma \|x - y\|^2, \forall x, y \in \mathrm{dom}(\partial f), g_x \in \partial f(x), g_y \in \partial f(y)$

Convexity: Uniqueness of min/max-value

Proposition 2

Let f be a proper closed and σ-strongly convex function with the same domain and range as before and $\sigma > 0$, then

1. The minimizer of f exists and is unique.
2. $f(x) - f(x^*) \geq \frac{\sigma}{2} \|x - x^*\|, \forall x \in \mathrm{dom}(f)$, where x^* is the unique minimizer of f.

Proof (Part 1)

We do not prove the existence of the minimizer.

Convexity: Uniqueness of min/max-value

Proof (Part 2)

Suppose $\tilde{x}, \hat{x}$ are two minimizers of f, then $f(\tilde{x}) = f(\hat{x}) = f_{opt}$, then $f_{opt} \leq f(\frac{1}{2}\tilde{x} + \frac{1}{2}\hat{x}) \leq \frac{1}{2}f(\tilde{x}) + \frac{1}{2}f(\hat{x}) - \frac{\sigma}{8}\|\tilde{x} - \hat{x}\|^2 < f_{opt}$, contradiction! $\square$

For x^* is the unique minimizer of f, we have $0 \in \partial f(x^*)$, so the second claim in lemma 2 could be applied to prove the result.$\square$

Section 3

Accelerating Gradient Methods

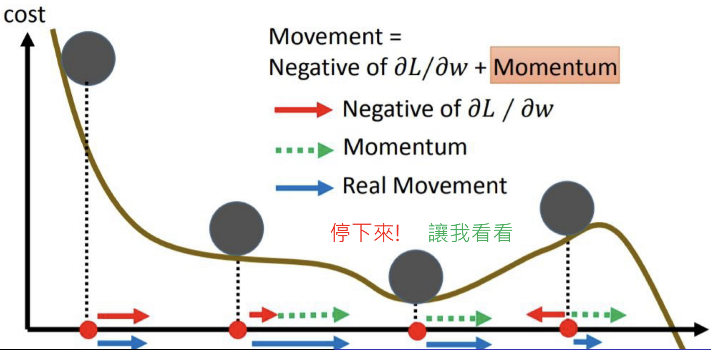
Heavy-Ball Method

- Proposed by Polyak.
- Each iteration has the form $x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$, where $\beta_k (x^k - x^{k-1})$ is called momentum.
- What is momentum?

Momentum: Graph Illustration

Why momentum?

防止因 $\nabla L(\theta)=0$
而卡在鞍点，(动量还在)



Conjugate Gradient

See Chapter 3: Line Search, Linear Conjugate Gradient and Chapter 4: Nonlinear Conjugate Gradient and Trust Region Method.

Nesterov's Accelerated Gradient: Weakly Convex Case

- Each iteration has the form
  $x^{k+1} = x^k - \alpha_k \nabla f(x^k + \beta_k(x^k - x^{k-1})) + \beta_k(x^k - x^{k-1})$
- Suppose that $\alpha_k = \frac{1}{L}$ is a constant, then we can introduce an auxiliary sequence $\{y_k\}$, then the form could be changed into a equation system

$$\begin{cases} x^{k+1} & = y^k - \frac{1}{L}\nabla f(y^k) \\ y^{k+1} & = x^{k+1} + \beta_{k+1}(x^{k+1} - x^k) \end{cases}$$

- The optimal method in all methods using only first-order information.

NAG: Convergence Analysis

Theorem 1

Assume $f(x)$ is convex, $\nabla f(x)$ is smooth with a constant L, the minimum of f is attained at x^*,
$\lambda_0 = 0, \lambda_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_k^2})(\lambda_{k+1}^2 - \lambda_{k+1} = \lambda_k^2)$,
$\beta_k = \frac{\lambda_k - 1}{\lambda_{k+1}}$, then the NAG method with $x^0 = y^0$ yields an iteration sequence $\{x^k\}$ with the following property

$$f(x^T) - f(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{(T+1)^2}, T = 1, 2, \cdots$$

Nesterov's Accelerated Gradient: Strongly Convex Case

- We require the convexity modulus be γ and $\gamma > 0$.
- Change β_{k+1} as $\beta_{k+1} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, where $\kappa = \frac{L}{\gamma}$, where L is the smoothness modulus.

Theorem 2

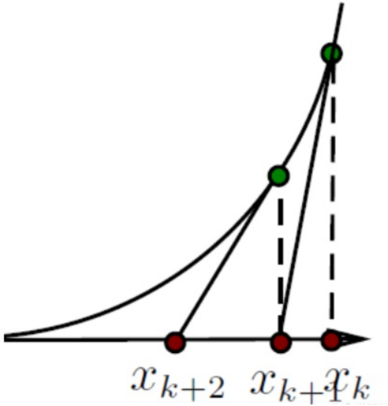
Assume $f(x)$ is convex, $\nabla f(x)$ is smooth with a constant L, the minimum of f is attained at x^*, then the NAG method with $x^0 = y^0$ yields an iteration sequence $\{x^k\}$ with the following property

$$f(x^T) - f(x^*) \leq \frac{L+\gamma}{2}\|x^0 - x^*\|^2 (1 - \frac{1}{\sqrt{\kappa}})^T, T = 1, 2, \cdots$$

Section 4

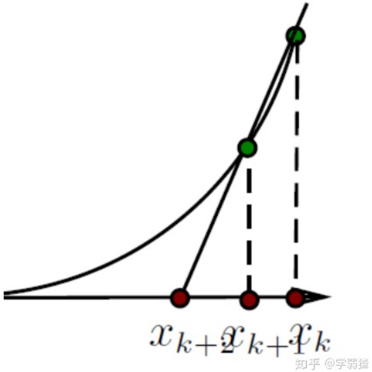
Quasi-Newton Methods

Comparison: Newton Method



- $f''(x_k) = -\frac{f'(x_k)}{x_{k+1} - x_k} \rightarrow$
  $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$

Comparison: Quasi-Newton Method



- $B_k(x_k - x_{k-1}) = f'(x_k) - f'(x_{k-1})$, where B_k is the slope.
- Let $s_k = x_{k+1} - x_k$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, we have $B_k s_{k-1} = y_{k-1}$

BFGS Method: Introduction

- One of the most popular optimization algorithms.
- Core requirements for BFGS Methods:
  - B_k is a SPD matrix.
  - B_k does not change too much.
- This leads to the update formula

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

- LBFGS has almost the same update formula, but the practical algorithm is much more complicated.

Framework for BFGS Method

1. Set an initial iterate x_0, initial SPD matrix B_0.
2. Compute p_k such that $B_k p_k = -\nabla f(x_k)$
3. $x_{k+1} = x_k + \alpha_k p_k$, where α_k is the step-size satisfying Wolfe Conditions.
4. Let $s_k = \alpha_k p_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
5. Update B_k to B_{k+1}
6. Loop until convergence.

An Extension: Wolfe Conditions

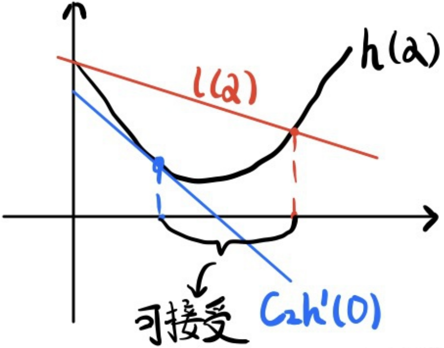
Definition 1: Weak Wolfe Condition

If the step-size α satisfies $h(\alpha) \leq l(\alpha) = h(0) + c_1 \alpha h'(0)$ and $h'(\alpha) \geq c_2 h'(0)$ with $0 < c_1 < c_2 < 1$, then the step-size satisfies weak Weak Wolfe Condition.

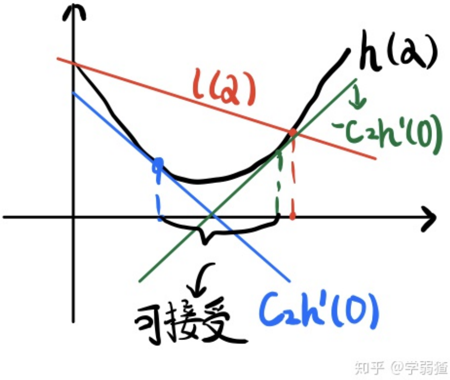
Definition 2: Strong Wolfe Condition

If the step-size α satisfies $h(\alpha) \leq l(\alpha) = h(0) + c_1 \alpha h'(0)$ and $|h'(\alpha)| \leq c_2 |h'(0)|$ with $0 < c_1 < c_2 < 1$, then the step-size satisfies weak Weak Wolfe Condition.

Graph Illustration of Weak Wolfe Conditions



Graph Illustration of Strong Wolfe Conditions



Why Wolfe Condition?

Proposition 3

If the step-size satisfies Wolfe Condition, then $s_k^T y_k > 0$. More importantly, the update formula could guarantee the SPD property of B_{k+1}.

Proof

Wolfe condition yields $\nabla f(x_{k+1})^T p_k \geq c_2 \nabla f(x_k)^T p_k$, which means $\nabla (f(x_{k+1}) - f(x_k))^T \alpha_k p_k \geq (c_2 - 1) \nabla f(x_k)^T \alpha_k p_k$. $c_2 < 1$ and p_k is a descent direction prove the result. $\square$

Convergence Result

Theorem 3

Let x_0 be the initial iterate, $f \in C^2$, $\mathcal{N}_{x_0} = \{x : f(x) \leq f(x_0)\}$ is a convex function. There exist $m, M > 0$ such that $m||z||^2 \leq z^T \nabla^2 f(x) z \leq M||z||^2, \forall z \in \mathbb{R}^n, x \in \mathcal{N}_{x_0}$. Let B_0 be an arbitrary SPD matrix, then BFGS Method will make $\{x_k\}$ converge to the minimizer x^* of f.

Convexity Correction for BFGS

- Problem: No guarantee for non-convex functions.
- Correction (just one optional way):

$$\begin{cases} \tilde{y}_k & = y_k + r_k s_k \\ B_{k+1} & = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{\tilde{y}_k \tilde{y}_k^T}{\tilde{y}_k^T s_k} \end{cases}$$

  One way for r_k is $r_k = (1 + \max(-y_k^T s_k / s_k^T s_k, 0)\|\nabla f(x_k)\|)$

- Core idea: Let $g(x) = f(x) + \frac{1}{2} r_k \|x_k\|^2$, then $\tilde{y}_k = y_k + r_k s_k$. (Make it more convex)

Section 5

Supplementary

Subsection 1

NAG (Weak Convex): Convergence Analysis

Theorem 1

Assume $f(x)$ is convex, $\nabla f(x)$ is smooth with a constant L, the minimum of f is attained at x^*,
$\lambda_0 = 0, \lambda_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_k^2})(\lambda_{k+1}^2 - \lambda_{k+1} = \lambda_k^2)$,
$\beta_k = \frac{\lambda_k - 1}{\lambda_{k+1}}$, then the NAG method with $x^0 = y^0$ yields an iteration sequence $\{x^k\}$ with the following property

$$f(x^T) - f(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{(T+1)^2}, T = 1, 2, \cdots$$

Proof (Part 1)

For any x, y, we have $f(y - \frac{\nabla f(y)}{L}) - f(x) \leq$
$\nabla f(y)^T(y - \frac{\nabla f(y)}{L} - y) + \frac{L}{2}\|y - \nabla f(y)/L - y\|^2 + \nabla f(y)^T(y - x)$

Proof (Part 2)

For this reason, we have

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L}\|\nabla f(y^k)\|^2 + \nabla f(y^k)^T(y^k - x^k) = -\frac{L}{2}\|x^{k+1} - y^k\|^2 - L(x^{k+1} - y^k)^T(y^k - x^k)$$

Similarly we have

$$f(x^{k+1}) - f(x^*) \leq -\frac{L}{2}\|x^{k+1} - y^k\|^2 - L(x^{k+1} - y^k)^T(y^k - x^*)$$

Let $\delta_k = f(x^k) - f(x^*)$, then we have

$$(\lambda_{k+1} - 1)(\delta_{k+1} - \delta_k) + \delta_{k+1} \leq -\frac{L}{2}\lambda_{k+1}\|x^{k+1} - y^k\|^2 - L(x^{k+1} - y^k)^T(\lambda_{k+1}y^k - (\lambda_{k+1} - 1)x^k - x^*)$$

Multiplying by λ_{k+1} yields $\lambda_{k+1}^2\delta_{k+1} - \lambda_k^2\delta_k \leq -\frac{L}{2}[\|\lambda_{k+1}x^{k+1} - (\lambda_{k+1} - 1)x^k - x^*\| - \|\lambda_{k+1}y^k - (\lambda_{k+1} - 1)x^k - x^*\|^2]$

Proof (Part 3)

Note that $\lambda_{k+2}y^{k+1} = \lambda_{k+2}x^{k+1} + (\lambda_{k+1} - 1)(x^{k+1} - x^k)$, rearranging this equality yields
$\lambda_{k+1}x^{k+1} - (\lambda_{k+1} - 1)x^k = \lambda_{k+2}y^{k+1} - (\lambda_{k+2} - 1)x^{k+1}$
This means $\lambda_{k+1}^2\delta_{k+1} - \lambda_k^2\delta_k \leq -\frac{L}{2}(\|u^{k+1}\|^2 - \|u^k\|^2)$, where $u^k = \lambda_{k+1}y^k - (\lambda_{k+1} - 1)x^k - x^*$
So we have $\lambda_T^2\delta_T \leq \frac{L}{2}(\|u^0\|^2 - \|u^T\|^2) \leq \frac{L}{2}\|x^0 - x^*\|^2$, $\lambda_T \geq \frac{T+1}{2}$ yields the conclusion. □

Thank you!