Optimization Basics: Gradient Descent and Related Extensions

Richard Liu

School of Mathematics, XMU

June 25, 2020

Content

Source

- Shai Shalev-Shwartz, Shai Ben-David, *Understanding Machine Learning: From Theory to Algorithms*
- Wen Huang, *Numerical Optimization Course in XMU*

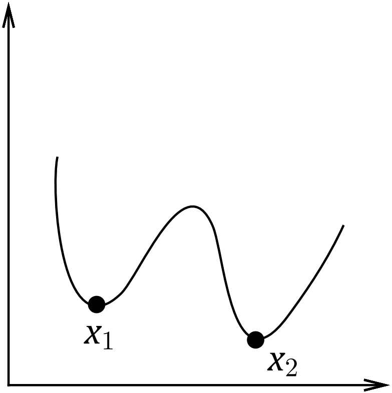
Section 1

Introduction

Introduction

- The formal definition of an optimization problem is to find $\min_{x \in \mathcal{F}} f(x)$ where $\mathcal{F}$ is called feasible domain.
- Problem: Could we find the minimal value?
  - No!
  - Local minima, stationary point and global minima.
- Two important elements: step-size and search direction.
- Illustration?

Graph Illustration



- In this graph, x_1 and x_2 are both stationary points and local minimizers, but only x_2 is a global minimizer.
- If a stationary point is not a local minimizer, then it is called a saddle point.
- Examples?

Introduction

- In this chapter we focus mainly on algorithms for numerical optimization, this is because many real optimization problems do not possess an analytical solution.
- General process: choose a search direction, "walk" a step-size distance, choose a search direction, "walk" a step-size distance, ... (iterative methods)

Theoretical Basics for Optimization

Proposition 1: First Order Necessary Condition

$f \in C^1$, then a necessary condition of x^* to be a local minima is $\nabla f(x^*) = 0$.

Proof

If $\nabla f(x^*) \neq 0$, let $p = -\nabla f(x^*)$, then we have $p^T \nabla f(x^*) < 0$. By continuity, there exists a sufficiently small T such that $p^T \nabla f(x^* + \tau p) < 0, \forall \tau < T$. By Taylor Expansion we have $f(x^* + \mu p) = f(x^*) + \nabla f(x^* + \tau p)^T \mu p < f(x^*), \mu < T$. $\square$

Theoretical Basics for Optimization

Proposition 2: Second Order Necessary Condition

$f \in C^2$, then a necessary condition of x^* to be a local minima is $\nabla^2 f(x^*) \geq 0$.

Proof

Note that if $\nabla^2 f(x^*) \geq 0$ does not hold, then let p be the direction such that $p^T \nabla f(x^*) p < 0$. By continuity, there exists a sufficiently small T such that
$p^T \nabla f(x^* + \tau p) p < 0, \forall \tau < T$. By Taylor Expansion we have
$f(x^* + \mu p) = f(x^*) + \nabla f(x^*)^T \mu p + \frac{1}{2} p^T \nabla^2 f(x + \tau p) p < f(x^*)$,
$\mu < T$. $\square$

Section 2

Gradient Descent with Fixed Step-size

Algorithm Formulation

- The general step of optimization could be written as

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)})$$

- $-\nabla f(\mathbf{w}^{(t)})$ is called a descent direction. This is a first-order method.

Definition 1: Descent Direction

If there exists a sufficient small τ such that $f(x + \tau p) < f(x)$, then p is called a descent direction at point x.

Proposition 3

If $p^T f(x) < 0$, then p is a descent direction.

Convergence Analysis for Convex Functions

Theorem 1

Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be an arbitrary sequence of vectors, let $\mathbf{w}^{(1)} = 0$ and the update rule be $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$, then we have

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \right\rangle \leq \frac{\|\mathbf{w}^*\|}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2$$

Moreover, for every $B, \rho > 0$, if $\|\mathbf{v}_t\| \leq \rho$ for any t and $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, $\|\mathbf{w}^*\| \leq B$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \right\rangle \leq \frac{B\rho}{\sqrt{T}}$$

Convergence Analysis for Convex Functions

Proof

By completing the square, we have

$$
\begin{aligned}
\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \right\rangle &= \frac{1}{\eta} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \eta \mathbf{v}_t \right\rangle \\
&= \frac{1}{2\eta}(-\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + \eta^2 \|\mathbf{v}_t\|^2) \\
&= \frac{1}{2\eta}(-\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2) + \frac{\eta}{2}\|\mathbf{v}_t\|^2
\end{aligned}
$$

So we construct a telescopic sum, which means

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \right\rangle \leq \frac{1}{2\eta}\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \square$$

Convergence Analysis for Convex Functions

We will use the convexity to prove that $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)$ could be arbitrarily small, where $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$. This is because

$$
\begin{aligned}
f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) &\leq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}^{(t)})) - f(\mathbf{w}^*) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)}) \right\rangle \leq \frac{B\rho}{\sqrt{T}}
\end{aligned}
$$

This means, if $T \geq \frac{B^2\rho^2}{\epsilon^2}$, we have $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \epsilon$, this is what we desire.

Subgradient

- Before we introduce stochastic gradient descent (SGD), we introduce another important concept for analysis.

Definition 2: Subgradient

A vector $\mathbf{v}$ is called a subgradient of f at $\mathbf{w}$ if

$$\forall \mathbf{y} \in \mathcal{D}, f(\mathbf{y}) \geq f(\mathbf{w}) + \langle \mathbf{y} - \mathbf{w}, \mathbf{v} \rangle$$

where $\mathcal{D}$ is the domain of function f.

- Exercise: Find the subgradient of function $f = |x|$, where x is a one-dimensional variable.

Stochastic Gradient Descent

- General Step: Choose $\mathbf{v}_t$ at random from a distribution such that $\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$ and then update by $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\mathbf{v}_t$
- We use an example to show the difference of two algorithms and explain why.

An Example

Batch gradient descent

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Repeat {

$$\theta_j := \theta_j - \alpha \underbrace{\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}}_{\frac{\partial}{\partial \theta_j} J_{train}(\theta)}$$

(for every $j = 0, \ldots, n$)

}

Stochastic gradient descent

$$cost(\theta, (x^{(i)}, y^{(i)})) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{train}(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost(\theta, (x^{(i)}, y^{(i)}))$$

1. Randomly shuffle dataset.
2. Repeat {
   for $i=1,\ldots,m$ {
   $$\theta_j := \theta_j - \alpha \underbrace{(h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}}_{\frac{\partial}{\partial \theta_j} cost(\theta, (x^{(i)}, y^{(i)}))}$$
   (for $j=0,\ldots,n$)
   }
   }

$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots$

Andrew Ng

Section 3

Gradient Descent with Unfixed Step-size

Introduction

- Although the previous algorithm could converge to a global minima in convex case, several problems arise.
    - In general, the problem could not preserve convexity in the whole domain.
    - The constants B, ρ are unknown, making us hard to take a proper step size η.
    - All the results should be preserved because we need the average.
- For this reason, a more practical method is to adopt line search framework: look for a proper step-size, update, look for a proper step-size, update, ...

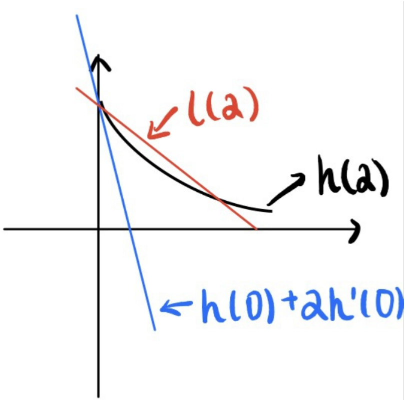
Step-size Selection: Armijo-Goldstein Condition

Definition 3: Armijo-Goldstein Condition

The step size α satisfies $h(\alpha) \leq h(0) + c_1 \alpha h'(0)$ (Armijo Condition), where $h(\alpha) = f(x + \alpha p)$ and α is the largest value in the set $\{t^{(i)} : t^{(i)} \in [\tau_1 t^{(i-1)}, \tau_2 t^{(i-1)}], t^{(0)} = 1$ for any $c_1 \in (0, 1)$ and $0 < \tau_1 \leq \tau_2 < 1$.

- If $\tau_1 = \tau_2$, then the appropriate step-size could be found by a backtracking algorithm (How?).
- Sufficient Descent + not too small.

Graph Illustration of Armijo Condition



Local Convergence Rate Analysis of the Gradient Descent Method

Theorem 2: Linear Local Convergence Rate Analysis

Let $\mathcal{N}_{x_0} = \{x : f(x) \le f(x_0)\}$. $f \in C^2, \mathcal{N}_{x_0}$ is convex, there exist positive constants $0 < m \le M$ such that $m \le \lambda_{\min}(\nabla^2 f(x)) \le \lambda_{\max}(\nabla^2 f(x)) \le M$ for all $x \in \mathcal{N}_{x_0}$, where $\lambda_{\min}(A), \lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A respectively. Let x^* denote the unique minimizer of f in $\mathcal{N}_{x_0}$ and $\{x_k\}$ denote the iterates generated by the Gradient Descent Method with the Byrd Nocedal Conditions. Then we have

$$f(x_k) - f(x^*) \le (1 - \beta \frac{m}{M})^k (f(x_0) - f(x^*))$$

Section 4

Newton Method

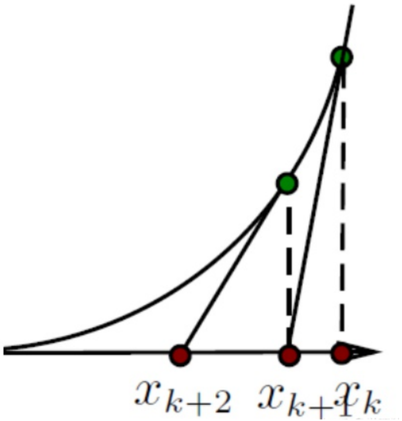
Introduction

- Idea: Find the root of $\nabla f(x) = 0$.
- Because $\nabla f(x+p) \simeq \nabla f(x) + \nabla^2 f(x) p$, we want $f(x+p) \simeq 0$, so we get $p = -(\nabla^2 f(x))^{-1} \nabla f(x)$, so we have the update rule

$$x^{(t+1)} = x^{(t)} - (\nabla^2 f(x^{(t)}))^{-1} \nabla f(x^{(t)})$$

- Second-order method.
- Fast local convergence rate, no global convergence.

Graph Illustration of Newton Method



Subsection 1

Case Study: Logistic Regression

Case Study: Logistic Regression

- Note that the log-likelihood for N observations is

$$l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)$$

where $p_k(x_i; \theta) = P(G = k | X = x_i; \theta)$

- Consider the two-class case, where $y_i = 1$ when $g_i = 1$ and $y_i = 0$ when $g_i = 2$. Then we could write the likelihood as

$$l(\beta) = \sum_{i=1}^{N} [y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})]$$

- By Numerical Optimization, we only need to compute $\frac{\partial l(\beta)}{\partial \beta}$ and $\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}$.

Case Study: Logistic Regression

- By simple calculation, we have

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i(y_i - p(x_i;\beta)), \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1-p(x$$

So for gradient descent method, we have

$$\beta^{new} = \beta^{old} - \eta \frac{\partial l(\beta)}{\partial \beta}$$

for Newton method, we have

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial l(\beta)}{\partial \beta}$$

Section 5

Supplementary

Subsection 1

SGD: Convergence Analysis

SGD: Convergence Analysis

Theorem 3

Let $B, \rho > 0$, f be a convex function and let $\mathbf{w}^* \in \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} f(\mathbf{w})$, $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, T be the number of iterations, $\|\mathbf{v}_t\| \leq \rho$ satisfies with probability 1, then we have

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^*) \leq \frac{B\rho}{\sqrt{T}}$$

Proof (Part 1)

By convexity, we have

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)] \leq \mathbb{E}_{\mathbf{v}_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*))]$$

SGD: Convergence Analysis

Proof (Part 2)

By theorem 1, we have $\mathbb{E}_{\mathbf{v}_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, v_t \right\rangle] \leq \frac{B\rho}{\sqrt{T}}$, which means we only need to prove

$$\mathbb{E}_{\mathbf{v}_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*))] \leq \mathbb{E}_{\mathbf{v}_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, v_t \right\rangle]$$

$$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{v}_{1:T}}[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, v_t \right\rangle]$$

SGD: Convergence Analysis

Proof (Part 3)

By the law of total expectation, we have

$$\mathbb{E}_{\mathbf{v}_{1:T}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle] = \mathbb{E}_{\mathbf{v}_{1:t-1}} \mathbb{E}_{\mathbf{v}_{1:t}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t | \mathbf{v}_{1:t-1} \rangle]$$

$$= \mathbb{E}_{\mathbf{v}_{1:t-1}} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t | \mathbf{v}_{1:t-1}] \rangle \geq \mathbb{E}_{\mathbf{v}_{1:t-1}}[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)]$$

$$= \mathbb{E}_{\mathbf{v}_{1:T}}[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)]$$

The last inequality holds for the reason that $\mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$. $\square$

Subsection 2

Newton Method: Convergence Analysis

Newton Method: Convergence Analysis

Theorem 4

Let x^* be a minimizer of f. Suppose $f \in C^2$, $\nabla f(x^*) = 0$ and $\nabla^2 f(x)$ is positive definite, $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhoold Ω_{x^*} of a solution x^*, which means $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ for $x, y \in \Omega_{x^*}$, then

- If x_0 is sufficiently close to x^*, then $\{x_k\}$ converges to x^*.
- The rate of convergence of $\{x_k\}$ is quadratic.

Proof (Part 1)

Let the current iteration point be x, then we want to firstly find $\|x + p - x^*\|$.

Newton Method: Convergence Analysis

Proof (Part 2)

By the update rule, we have

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

By these two taylor expansions

$\nabla f(x) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x + t(x^* - x))(x - x^*)dt$ and

$\nabla^2 f(x)(x - x^*) = \int_0^1 \nabla^2 f(x)(x - x^*)dt$, we could get

$$\|\nabla^2 f(x)\|\|x+p-x^*\| = \|\int_0^1 [\nabla^2 f(x) - \nabla^2 f(x+t(x^*-x))](x-x^*)dt\|$$

$$\leq \int_0^1 \|\nabla^2 f(x) - \nabla^2 f(x + t(x^* - x))\|dt\|x - x^*\| \leq \frac{1}{2}L\|x - x^*\|^2$$

Newton Method: Convergence Analysis

Proof (Part 3)

Part 2 has shown that, if this algorithm has local convergence, then the convergence rate is quadratic. Let $\|x - x^*\| \leq \gamma$, where γ satisfies $\|\nabla^2 f(x)\|^{-1} \leq 2\|\nabla^2 f(x^*)\|^{-1}$, this means

$$\|x + p - x^*\| \leq L\|\nabla^2 f(x^*)\|^{-1}\|x - x^*\|\|x - x^*\|$$

Let $\|x_0 - x^*\| \leq \frac{1}{2L\|\nabla^2 f(x^*)\|^{-1}}$, by induction we could get the result.

Thank you!