CME 307 / MS&E 311: Optimization

Gradient descent

Professor Udell

Management Science and Engineering
Stanford

February 7, 2024

Outline

Classification

Unconstrained minimization

Analysis via Polyak-Lojasiewicz condition

Background: classification

classification problem: m data points

- feature vector $a_i \in \mathbf{R}^n$, $i = 1, \ldots, m$
- label $b_i \in \{-1, 1\}$, $i = 1, \ldots, m$

choose decision boundary $a^T x = 0$ to separate data points into two classes

- $a^T x > 0 \implies$ predict class 1
- $a^T x < 0 \implies$ predict class -1

classification is correct if $b_i a_i^T x > 0$

Background: classification

classification problem: m data points

- feature vector $a_i \in \mathbf{R}^n$, $i = 1, \ldots, m$
- label $b_i \in \{-1, 1\}$, $i = 1, \ldots, m$

choose decision boundary $a^T x = 0$ to separate data points into two classes

- $a^T x > 0 \implies$ predict class 1
- $a^T x < 0 \implies$ predict class -1

classification is correct if $b_i a_i^T x > 0$

- projective transformation transforms affine boundary to linear boundary
- classification is invariant to scalar multiplication of x

Logistic regression

(regularized) **logistic regression** minimizes the **finite sum**

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \log(1 + \exp\left(-b_i a_i^T x\right)) + r(x) \\ \text{variable} & x \in \mathbf{R}^n \end{array}$$

where

- $b_i \in \{-1, 1\}$, $a_i \in \mathbf{R}^n$
- $r : \mathbf{R}^n \to \mathbf{R}$ is a **regularizer**, *e.g.*, $\|x\|^2$ or $\|x\|_1$

Support vector machine

support vector machine (SVM) minimizes the **finite sum**

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} \max(0, 1 - b_i a_i^T x) + \gamma \|x\|^2 \\ \text{variable} & x \in \mathbf{R}^n \end{array}$$

where $b_i \in \{-1, 1\}$ and $a_i \in \mathbf{R}^n$.

Support vector machine

support vector machine (SVM) minimizes the **finite sum**

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \max(0, 1 - b_i a_i^T x) + \gamma \|x\|^2 \\ \text{variable} & x \in \mathbf{R}^n \end{array}$$

where $b_i \in \{-1, 1\}$ and $a_i \in \mathbf{R}^n$. not differentiable!

Support vector machine

support vector machine (SVM) minimizes the **finite sum**

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \max(0, 1 - b_i a_i^T x) + \gamma \|x\|^2 \\ \text{variable} & x \in \mathbf{R}^n \end{array}$$

where $b_i \in \{-1, 1\}$ and $a_i \in \mathbf{R}^n$. not differentiable!

how to solve?

Support vector machine

support vector machine (SVM) minimizes the **finite sum**

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \max(0, 1 - b_i a_i^T x) + \gamma \|x\|^2 \\ \text{variable} & x \in \mathbf{R}^n \end{array}$$

where $b_i \in \{-1, 1\}$ and $a_i \in \mathbf{R}^n$. not differentiable!

how to solve?

- use **subgradient** method
- transform to **conic form**
- solve **dual** problem instead
- **smooth** the objective

Outline

Unconstrained minimization

$$\text{minimize} \quad f(x)$$

- $f : \mathbf{R}^n \to \mathbf{R}$ differentiable
- assume optimal value $f^\star = \inf_x f(x)$ is attained (and finite)
- assume a starting point $x^{(0)}$ is known

unconstrained minimization methods

- produce sequence of points $x^{(k)}$, $k = 0, 1, \ldots$ with

$$f(x^{(k)}) \to f^\star$$

(we hope)

Gradient descent

$$\text{minimize} \quad f(x)$$

idea: go downhill

Algorithm Gradient descent

Given: $f : \mathbf{R}^d \to \mathbf{R}$, stepsize t, maxiters
Initialize: $x = 0$ (or anything you'd like)
For: $k = 1, \ldots,$ maxiters

- update x:

$$x \leftarrow x - t\nabla f(x)$$

Gradient descent: choosing a step-size

- **constant step-size.** $t^{(k)} = t$ (constant)
- **decreasing step-size.** $t^{(k)} = 1/k$
- **line search.** try different possibilities for $t^{(k)}$ until objective at new iterate

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

  decreases enough.

tradeoff: line search requires evaluating $f(x)$ (can be expensive)

Line search

define $x^+ = x - t\nabla f(x)$

- exact line search: find t to minimize $f(x^+)$
- the **Armijo rule** requires t to satisfy

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2$$

for some $c \in (0, 1)$, *e.g.*, $c = .01$.

Line search

define $x^+ = x - t\nabla f(x)$

- exact line search: find t to minimize $f(x^+)$
- the **Armijo rule** requires t to satisfy

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2$$

for some $c \in (0, 1)$, *e.g.*, $c = .01$.

a simple **backtracking line search** algorithm:

- set $t = 1$
- if step decreases objective value sufficiently, accept x^+:

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2 \quad \Longrightarrow \quad x \leftarrow x^+$$

otherwise, halve the stepsize $t \leftarrow t/2$ and try again

Line search

define $x^+ = x - t\nabla f(x)$

- exact line search: find t to minimize $f(x^+)$
- the **Armijo rule** requires t to satisfy

$$f(x^+) \le f(x) - ct\|\nabla f(x)\|^2$$

  for some $c \in (0,1)$, *e.g.*, $c = .01$.

a simple **backtracking line search** algorithm:

- set $t = 1$
- if step decreases objective value sufficiently, accept x^+:

$$f(x^+) \le f(x) - ct\|\nabla f(x)\|^2 \quad \Longrightarrow \quad x \leftarrow x^+$$

  otherwise, halve the stepsize $t \leftarrow t/2$ and try again

Q: can we can always satisfy the Armijo rule for some t?

Line search

define $x^+ = x - t\nabla f(x)$

- exact line search: find t to minimize $f(x^+)$
- the **Armijo rule** requires t to satisfy

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2$$

for some $c \in (0, 1)$, *e.g.*, $c = .01$.

a simple **backtracking line search** algorithm:

- set $t = 1$
- if step decreases objective value sufficiently, accept x^+:

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2 \quad \implies \quad x \leftarrow x^+$$

otherwise, halve the stepsize $t \leftarrow t/2$ and try again

Q: can we can always satisfy the Armijo rule for some t?

A: yes! see gradient descent demo

Demo: gradient descent

https://github.com/stanford-cme-307/demos/blob/main/gradient-descent.ipynb

How well does GD work?

for $x \in \mathbf{R}^n$,

- $f(x) = x^T x$
- $f(x) = x^T A x$ for $A \succeq 0$
- $f(x) = \|x\|_1$ (nonsmooth but differentiable **almost** everywhere)
- $f(x) = 1/x$ on $x > 0$ (strictly convex but not strongly convex)

https://github.com/stanford-cme-307/demos/blob/main/gradient-descent-contours.ipynb

Outline

Classification

Unconstrained minimization

Analysis via Polyak-Lojasiewicz condition

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \to \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$$

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \to \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$$

Theorem ([Karimi, Nutini, and Schmidt (2016)])

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \to \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \to \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \to \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$$

Theorem ([Karimi, Nutini, and Schmidt (2016)])

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \to \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \to \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

so logistic loss (on a compact set) and quadratic loss are Polyak-Lojasiewicz even when $m < n$

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \to \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$$

Theorem ([Karimi, Nutini, and Schmidt (2016)])

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \to \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \to \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

so logistic loss (on a compact set) and quadratic loss are Polyak-Lojasiewicz even when $m < n$

Q: Are all Polyak-Lojasiewicz functions convex?

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \to \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$$

Theorem ([Karimi, Nutini, and Schmidt (2016)])

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \to \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \to \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

so logistic loss (on a compact set) and quadratic loss are Polyak-Lojasiewicz even when $m < n$

Q: Are all Polyak-Lojasiewicz functions convex?
A: No. A river valley is Polyak-Lojasiewicz but not convex.
why use Polyak-Lojasiewicz? Polyak-Lojasiewicz is weaker than strong convexity and yields simpler proofs

PL and invexity

Theorem

Every Polyak-Lojasiewicz function is invex. (That is, any stationary point of a Polyak-Lojasiewicz function is globally optimal.)

PL and invexity

Theorem

Every Polyak-Lojasiewicz function is invex. (That is, any stationary point of a Polyak-Lojasiewicz function is globally optimal.)

proof: if $\nabla f(\bar{x}) = 0$, then

$$0 = \frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(\bar{x}) - f^\star) \geq 0$$

$\implies f(\bar{x}) = f^\star$ is the global optimum.

strong convexity $\implies$ Polyak-Lojasiewicz

Theorem

If f is μ-strongly convex, then f is μ-Polyak-Lojasiewicz.

strong convexity $\implies$ Polyak-Lojasiewicz

Theorem

If f is μ-strongly convex, then f is μ-Polyak-Lojasiewicz.

proof: minimize the strong convexity condition over y:

$$\begin{aligned}
\min_y f(y) &\geq \min_y \left(f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2 \right) \\
f^\star &\geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2 \\
\frac{1}{2} \|\nabla f(x)\|^2 &\geq \mu(f(x) - f^\star)
\end{aligned}$$

as minimum occurs for $y - x = -\nabla f(x)/\mu$

Types of convergence

- objective converges

$$f(x^{(k)}) \to f^\star$$

- iterates converge

$$x^{(k)} \to x^\star$$

under

- strong convexity: objective converges $\implies$ iterates converge
  proof: use strong convexity with $x = x^\star$ and $y = x^{(k)}$:

$$f(x^{(k)}) - f^\star \geq \frac{\mu}{2}\|x^{(k)} - x^\star\|^2$$

- Polyak-Lojasiewicz: not necessarily true ($x^\star$ may not be unique)

Rates of convergence

- linear convergence with rate c

$$f(x^{(k)}) - f^\star \leq c^k(f(x^{(0)}) - f^\star)$$

  - looks like a line on a semi-log plot
  - example: gradient descent on smooth strongly convex function
- sublinear convergence
  - looks slower than a line (curves up) on a semi-log plot
  - example: $1/k$ convergence

$$f(x^{(k)}) - f^\star \leq \mathcal{O}(1/k)$$

  - example: gradient descent on smooth convex function
  - example: stochastic gradient descent

Gradient descent converges linearly

Theorem

If $f : \mathbf{R}^n \to \mathbf{R}$ is μ-Polyak-Lojasiewicz, L-smooth, and $x^\star = \operatorname{argmin}_x f(x)$ exists, then gradient descent with stepsize L

$$x^{(k+1)} = x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})$$

converges linearly to $f^\star$ with rate $(1 - \frac{\mu}{L})$.

Gradient descent converges linearly: proof

proof: plug in update rule to L-smoothness condition

$$
\begin{aligned}
f(x^{(k+1)}) - f(x^{(k)}) &\leq \nabla f(x^{(k)})^T (x^{(k+1)} - x^{(k)}) + \frac{L}{2}\|x^{(k+1)} - x^{(k)}\|^2 \\
&\leq (-\frac{1}{L} + \frac{1}{2L})\|\nabla f(x^{(k)})\|^2 \\
&\leq -\frac{1}{2L}\|\nabla f(x^{(k)})\|^2 \\
&\leq -\frac{\mu}{L}(f(x^{(k)}) - f^\star) \triangleright \text{(using PL)}
\end{aligned}
$$

decrement proportional to error $\implies$ linear convergence:

$$
\begin{aligned}
f(x^{(k)}) - f^\star &\leq (1 - \frac{\mu}{L})(f(x^{(k-1)}) - f^\star) \\
&\leq (1 - \frac{\mu}{L})^k (f(x^{(0)}) - f^\star)
\end{aligned}
$$

Practical convergence

- Gradient descent with optimal stepsize converges even faster.

$$f(x^{(k+1)}) = \inf_{\alpha} f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)} - \frac{1}{L} \nabla f(x^{(k)}))$$

Practical convergence

- Gradient descent with optimal stepsize converges even faster.

$$f(x^{(k+1)}) = \inf_{\alpha} f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)} - \frac{1}{L} \nabla f(x^{(k)}))$$

- Local vs global convergence

Practical convergence

- Gradient descent with optimal stepsize converges even faster.

$$f(x^{(k+1)}) = \inf_{\alpha} f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)} - \frac{1}{L} \nabla f(x^{(k)}))$$

- Local vs global convergence
- What does this proof technique tell us about the convergence of gradient descent on non-convex functions? On functions that are convex but not strongly convex?

Quiz

- A strongly convex function always satisfies the Polyak-Lojasiewicz condition
  A. true
  B. false
- Suppose $f : \mathbf{R} \to \mathbf{R}$ is L-smooth and satisfies the Polyak-Lojasiewicz condition. Then any stationary point $\nabla f(x) = 0$ of f is a global optimum: $f(x) = \mathrm{argmin}_y f(y) =: f^\star$.
  A. true
  B. false
- Suppose $f : \mathbf{R} \to \mathbf{R}$ is L-smooth and satisfies the Polyak-Lojasiewicz condition. Then gradient descent on f converges linearly from any starting point.
  A. true
  B. false