ORIE 7391: Faster: Algorithmic Ideas for Speeding Up Optimization

Convex Optimization

Professor Udell

Operations Research and Information Engineering
Cornell

February 9, 2022

Quiz

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

Outline

Convexity: definitions

Q: Define convex set? convex function?

Convexity: definitions

Q: Define convex set? convex function?

- A set $S \subseteq \mathbf{R}^n$ is convex if it contains every chord: for all $\theta \in [0, 1]$, w, $v \in S$,

$$\theta w + (1 - \theta)v \in S$$

Convexity: definitions

Q: Define convex set? convex function?

- A set $S \subseteq \mathbf{R}^n$ is convex if it contains every chord: for all $\theta \in [0,1]$, w, $v \in S$,

$$\theta w + (1-\theta)v \in S$$

- A function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff its superlevel set $S = \{(x,t) : \ t \geq f(x)\}$ is convex.

Convexity: definitions

Q: Define convex set? convex function?

- A set $S \subseteq \mathbf{R}^n$ is convex if it contains every chord: for all $\theta \in [0, 1]$, w, $v \in S$,

$$\theta w + (1 - \theta) v \in S$$

- A function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff its superlevel set $S = \{(x, t) : \; t \geq f(x)\}$ is convex.
- A differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff it satisfies the first order condition

$$f(v) - f(w) \geq \nabla f(w)^\top (v - w) \qquad \forall w, v \in \mathbf{R}^n$$

Convexity: definitions

Q: Define convex set? convex function?

- A set $S \subseteq \mathbf{R}^n$ is convex if it contains every chord: for all $\theta \in [0, 1]$, w, $v \in S$,

$$\theta w + (1 - \theta) v \in S$$

- A function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff its superlevel set $S = \{(x, t) : \ t \geq f(x)\}$ is convex.
- A differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff it satisfies the first order condition

$$f(v) - f(w) \geq \nabla f(w)^\top (v - w) \qquad \forall w, v \in \mathbf{R}^n$$

- A twice differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff its Hessian is always **positive semidefinite**: $\lambda_{\min}(\nabla^2 f) \geq 0$

Convexity examples

Q: Which of these functions are convex?

Quadratic approximation

Suppose $f : \mathbf{R} \to \mathbf{R}$ is differentiable. For any $x \in \mathbf{R}$, write its second order expansion about x:

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x).$$

If f is a quadratic function, $\nabla^2 f(x) = H$ is constant.

Quadratic approximation

Suppose $f : \mathbf{R} \to \mathbf{R}$ is differentiable. For any $x \in \mathbf{R}$, write its second order expansion about x:

$$f(y) \approx f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x).$$

If f is a quadratic function, $\nabla^2 f(x) = H$ is constant.

Quadratic approximations are useful because quadratics are easy to minimize:

$$y^\star = \operatorname*{argmin}_y f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x$$

$$\nabla f(x) + H(y-x) = 0$$

$$y^\star = x + H^{-1}(-\nabla f(x)).$$

Quadratic approximation

Suppose $f : \mathbf{R} \to \mathbf{R}$ is differentiable. For any $x \in \mathbf{R}$, write its second order expansion about x:

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x).$$

If f is a quadratic function, $\nabla^2 f(x) = H$ is constant.

Quadratic approximations are useful because quadratics are easy to minimize:

$$y^\star \;=\; \operatorname*{argmin}_y f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x$$

$$\nabla f(x) + H(y - x) = 0$$

$$y^\star = x + H^{-1}(-\nabla f(x)).$$

If we approximate the Hessian of f by $H = \frac{1}{t} I$ for some $t > 0$ and choose x^+ to minimize the quadratic approximation, we obtain the **gradient descent** update with step size t:

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

Quadratic bounds

A function $f : \mathbf{R} \to \mathbf{R}$ is L-**smooth** if

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

Equivalently, the operator $\frac{1}{L}\nabla f$ is L-**Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y-x\|$$

and $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbf{dom}\, f$.

Quadratic bounds

A function $f : \mathbf{R} \to \mathbf{R}$ is L-**smooth** if

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

Equivalently, the operator $\frac{1}{L}\nabla f$ is L-**Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

and $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbf{dom}\, f$.

A function $f : \mathbf{R} \to \mathbf{R}$ is μ-**strongly convex** if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Equivalently, the operator $\frac{1}{\mu}\nabla f$ is μ-**coercive**:

$$(\nabla f(y) - \nabla f(x))^T (y - x)\| \geq \mu\|y - x\|^2$$

and $\nabla^2 f(x) \succeq \mu I$ for all $x \in \mathbf{dom}\, f$.

Quadratic bounds

A function $f : \mathbf{R} \to \mathbf{R}$ is L-**smooth** if

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

Equivalently, the operator $\frac{1}{L}\nabla f$ is L-**Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

and $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbf{dom}\, f$.

A function $f : \mathbf{R} \to \mathbf{R}$ is μ-**strongly convex** if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Equivalently, the operator $\frac{1}{\mu}\nabla f$ is μ-**coercive**:

$$(\nabla f(y) - \nabla f(x))^T (y - x)\| \geq \mu\|y - x\|^2$$

and $\nabla^2 f(x) \succeq \mu I$ for all $x \in \mathbf{dom}\, f$.

Quadratic bounds

A function $f : \mathbf{R} \to \mathbf{R}$ is L-**smooth** if

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

Equivalently, the operator $\frac{1}{L}\nabla f$ is L-**Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

and $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbf{dom}\, f$.

A function $f : \mathbf{R} \to \mathbf{R}$ is μ-**strongly convex** if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Equivalently, the operator $\frac{1}{\mu}\nabla f$ is μ-**coercive**:

$$(\nabla f(y) - \nabla f(x))^T (y - x)\| \geq \mu\|y - x\|^2$$

and $\nabla^2 f(x) \succeq \mu I$ for all $x \in \mathbf{dom}\, f$.

Q: For $A \succeq 0$, the quadratic function $f(x) = \frac{1}{2}x^T A x$ is ?-smooooth and ?-strongly convex

Quadratic bounds

A function $f : \mathbf{R} \to \mathbf{R}$ is L-**smooth** if

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

Equivalently, the operator $\frac{1}{L}\nabla f$ is L-**Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

and $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbf{dom}\, f$.

A function $f : \mathbf{R} \to \mathbf{R}$ is μ-**strongly convex** if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Equivalently, the operator $\frac{1}{\mu}\nabla f$ is μ-**coercive**:

$$(\nabla f(y) - \nabla f(x))^T (y - x)\| \geq \mu\|y - x\|^2$$

and $\nabla^2 f(x) \succeq \mu I$ for all $x \in \mathbf{dom}\, f$.

Q: For $A \succeq 0$, the quadratic function $f(x) = \frac{1}{2}x^T A x$ is ?-smoooth and ?-strongly convex

A: $\lambda_{max}(A)$-smooth and $\lambda_{min}(A)$-strongly convex

Outline

Convexity

Gradient descent

Classical analysis of GD

Polyak-Lojasiewiczcondition

Unconstrained minimization

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

- $f : \mathbf{R}^n \to \mathbf{R}$ convex, ctsly differentiable (hence $\mathbf{dom}\, f$ open)
- assume optimal value $f^\star = \inf_x f(x)$ is attained (and finite)
- assume a starting point $x^{(0)}$ such that $x^{(0)} \in \mathbf{dom}\, f$ is known

unconstrained minimization methods

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

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

- can be seen as iterative methods for solving optimality condition

$$\nabla f(x^\star) = 0$$

Rates of convergence

- linear convergence:

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

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

Gradient descent

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

idea: go downhill to get to a (the?) minimum!

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: evaluating $f(x)$ takes $\mathcal{O}(nd)$ flops each time ...

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

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

  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

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

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^+:

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

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

Q: can we can always satisfy the Armijo rule for some t?
A: yes! see gradient descent demo

Outline

Quadratic upper and lower bounds

what problem does the gradient descent step $x := x + t\nabla f(x)$ solve?

Quadratic upper and lower bounds

what problem does the gradient descent step $x := x + t\nabla f(x)$ solve? answer:

$$\text{minimize}_y \quad f(x) + \nabla f(x)^T(y - x) + \frac{1}{2t}\|x - y\|^2$$

So gradient descent will work best if Hessian is "almost" scaled multiple of the identity.

Formally, find $\alpha \in \mathbf{R}$ and $\beta \in \mathbf{R}$ so that for all x, $y \in \mathbf{dom}\, f$,

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \leq f(y) \leq f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)
- clearly $\alpha \leq \beta$

Example I: quadratic

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \leq f(y) \leq f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

suppose $A \in \mathbf{S}_+^n$, and consider $f(x) = \frac{1}{2}x^T Ax$. is f smooth and strongly convex?

Example I: quadratic

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \leq f(y) \leq f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

suppose $A \in \mathbf{S}^n_+$, and consider $f(x) = \frac{1}{2}x^T Ax$. is f smooth and strongly convex?

- strongly convex, with parameter $\alpha = \lambda_{\min}(A)$ (if $\lambda_{\min}(A) > 0$)
- smooth, with parameter $\beta = \lambda_{\max}(A)$

Example I: quadratic

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \le f(y) \le f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

suppose $A \in \mathbf{S}_+^n$, and consider $f(x) = \frac{1}{2}x^T Ax$. is f smooth and strongly convex?

- strongly convex, with parameter $\alpha = \lambda_{\min}(A)$ (if $\lambda_{\min}(A) > 0$)
- smooth, with parameter $\beta = \lambda_{\max}(A)$

proof:

$$\begin{aligned}\frac{1}{2}y^T Ay &= \frac{1}{2}(x+(y-x))^T A(x+(y-x)) \\ &= \frac{1}{2}x^T x + (Ax)^T(y-x) + \frac{1}{2}(y-x)^T A(y-x)\end{aligned}$$

Example II: smoothed absolute value

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \le f(y) \le f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

consider

$$\textbf{huber}(x) = \begin{cases} \frac{1}{2}x^2 & x^2 \le 1 \\ |x| - \frac{1}{2} & \text{otherwise} \end{cases}$$

is **huber** smooth and strongly convex?

Example II: smoothed absolute value

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \leq f(y) \leq f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

consider

$$\textbf{huber}(x) = \begin{cases} \frac{1}{2}x^2 & x^2 \leq 1 \\ |x| - \frac{1}{2} & \text{otherwise} \end{cases}$$

is **huber** smooth and strongly convex?

- not strongly convex
- smooth, with parameter $\beta = 1$

Example III: regularized absolute value

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \leq f(y) \leq f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

consider $f(x) = |x| + x^2$. is f smooth and strongly convex?

Example III: regularized absolute value

$$f(x)+\nabla f(x)^T(y-x)+\frac{\alpha}{2}\|x-y\|^2 \leq f(y) \leq f(x)+\nabla f(x)^T(y-x)+\frac{\beta}{2}\|x-$$

- lower bound is called **strong convexity** (with parameter α)
- upper bound is called **smoothness** (with parameter β)

consider $f(x) = |x| + x^2$. is f smooth and strongly convex?

- strongly convex, with parameter $\alpha = 1$
- not smooth

Roadmap

we'll analyze gradient descent for a few cases:

- is f α-strongly convex, or not?
- is f β-smooth, or not?
- is f Lipschitz differentiable, or not?
- do we use a fixed step size, or line-search?

a question: are these rates "the best possible"? how could we tell? compared to what?

we'll do the β-smooth case first, then work up to the others

Monotone gradient

first, another convexity condition:

- a differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ is convex iff for all x, $y \in \mathbf{dom}\, f$,

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0$$

- ∇f is called a **monotone mapping**
- strict inequality $\implies$ **strictly monotone mapping**

Monotone gradient: proof

- if f is differentiable and convex, then

  $f(y) \geq f(x)+\nabla f(x)^T(y-x) \qquad f(x) \geq f(y)+\nabla f(y)^T(x-y)$

  add these to get

  $$0 \geq (\nabla f(x) - \nabla f(y))^T(y-x)$$

- if ∇f is monotone, then for any x, $y \in \mathbf{dom}\, f$, let $g(t) = f(x + t(y-x))$. for $t \geq 0$,

  $$g'(t) = \nabla f(x + t(y-x))^T(y-x) \geq g'(0).$$

  so

  $$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)dt &\geq g(0) + g'(0) \\ &= f(x) + \nabla f(x)(y-x) \end{aligned}$$

Smoothness: equivalent definitions

for convex $f : \mathbf{R}^n \to \mathbf{R}$, the following properties are all equivalent:

1. $\frac{\beta}{2}x^T x - f(x)$ is convex
2. f is β-**smooth**: for all x, $y \in \mathbf{dom}\, f$,
$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2}\|x - y\|^2$$
3. (if f is twice differentiable) $\nabla^2 f(x) \preceq \beta I$
4. ∇f is **Lipschitz continuous** with parameter β: $\forall x, y \in \mathbf{dom}\, f$,
$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$
5. ∇f is **co-coercive** with parameter $\frac{1}{\beta}$: for all x, $y \in \mathbf{dom}\, f$,
$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2$$

Smoothness: equivalent definitions

for convex $f : \mathbf{R}^n \to \mathbf{R}$, the following properties are all equivalent:

1. $\frac{\beta}{2} x^T x - f(x)$ is convex
2. f is β-**smooth**: for all x, $y \in \mathbf{dom}\, f$,
$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|x - y\|^2$$
3. (if f is twice differentiable) $\nabla^2 f(x) \preceq \beta I$
4. ∇f is **Lipschitz continuous** with parameter β: $\forall x, y \in \mathbf{dom}\, f$,
$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$
5. ∇f is **co-coercive** with parameter $\frac{1}{\beta}$: for all x, $y \in \mathbf{dom}\, f$,
$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

proof: 2) is first order condition for convexity of 1); 3) is second order condition for convexity of 1); 4) $\implies$ 2) by integration; 5)

Smoothness: proofs of equivalence

1. f is β-**smooth**: for all x, $y \in \mathbf{dom}\, f$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2}\|x - y\|^2$$

2. ∇f is **Lipschitz continuous** with parameter β: for all x, $y \in \mathbf{dom}\, f$,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$$

proof. integrate and use Lipschitz ∇f (last line) to prove smoothness:

$$\begin{aligned}
& f(y) - f(x) - \nabla f(x)^T (y - x) \\
= & \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt - \nabla f(x)^T (y - x) \\
= & \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) dt \\
\leq & \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_2 \|(y - x)\|_2 dt
\end{aligned}$$

Smoothness: proofs of equivalence

1. ∇f is **Lipschitz continuous** with parameter β: for all x, $y \in \mathbf{dom}\, f$,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

2. ∇f is **co-coercive** with parameter $\frac{1}{\beta}$: for all x, $y \in \mathbf{dom}\, f$,

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

Smoothness: proofs of equivalence

1. ∇f is **Lipschitz continuous** with parameter β: for all x, $y \in \mathbf{dom}\, f$,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

2. ∇f is **co-coercive** with parameter $\frac{1}{\beta}$: for all x, $y \in \mathbf{dom}\, f$,

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

proof. use Cauchy-Shwarz (first ineq) and co-coercivity (second)

$$\begin{aligned} \|\nabla f(x) - \nabla f(y))\| \|(x - y)\| &\geq (\nabla f(x) - \nabla f(y))^T (x - y) \\ &\geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2 \\ \|(x - y)\| &\geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\| \end{aligned}$$

to get Lipschitz continuity

Analysis of gradient descent for smooth functions

three assumptions for analysis:

- $f : \mathbf{R}^n \to \mathbf{R}$ convex and differentiable with $\mathbf{dom}\, f = \mathbf{R}^n$
- f is smooth with parameter $\beta > 0$
- optimal value $f^\star = \inf_x f(x)$ finite and attained at $x^\star$

Algorithm: GD with constant step size.
pick constant step size $0 < t \leq \frac{1}{\beta}$, and repeat

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

(nb: not implementable without guess for β)

Analysis of gradient descent for smooth functions

use quadratic upper bound with $y = x^+ = x - t\nabla f(x)$:

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)(x^+ - x) + \frac{\beta}{2}\|x^+ - x\|^2 \\ &= f(x) - t\|\nabla f(x)\|^2 + t^2\frac{\beta}{2}\|\nabla f(x)\|^2 \end{aligned}$$

if constant step size $0 < t \leq \frac{1}{\beta}$,

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|^2 \\ &\leq f(x^\star) - \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|^2 \\ &= f^\star + \frac{1}{2t}(\|x - x^\star\|^2 - \|x - x^\star - t\nabla f(x)\|^2) \\ &= f^\star + \frac{1}{2t}(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2) \end{aligned}$$

(second line uses first order convexity condition)

Analysis of gradient descent for smooth functions

take average over iteration counter $i = 1, \ldots, k$:

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f^\star &\leq \frac{1}{k}\sum_{i=1}^{k} \frac{1}{2t}\left(\|x^{(i)} - x^\star\|^2 - \|x^{(i+1)} - x^\star\|^2\right) \\ &\leq \frac{1}{2tk}\left(\|x^{(0)} - x^\star\|^2 - \|x^{(k+1)} - x^\star\|^2\right) \\ &\leq \frac{1}{2tk}\|x^{(0)} - x^\star\|^2 \end{aligned}$$

since $f(x^{(k)})$ is non-increasing,

$$f(x^{(k)}) - f^\star \leq \frac{1}{2tk}\|x^{(0)} - x^\star\|^2$$

so number of iterations k to reach $f(x^{(k)}) - f^\star \leq \epsilon$ is $\mathcal{O}(1/\epsilon)$

Analysis of gradient descent for smooth functions

now, with line search!

- t chosen by line search w/params $(a, b) = (\frac{1}{2}, \frac{1}{2})$ (to simplify proofs), so $x^+ = x - t\nabla f(x)$ satisfies
$$f(x^+) < f(x) - \frac{t}{2}\|\nabla f(x)\|^2.$$
- from smoothness of f, we know $t = \frac{1}{\beta}$ would work
- so linesearch returns $t \geq \frac{b}{\beta} = \frac{1}{2\beta}$

Algorithm: GD with line search.
pick line search parameters $(a, b) = (\frac{1}{2}, \frac{1}{2})$ and $x^{(0)} \in \mathbf{R}^n$, and repeat

1. compute $\nabla f(x^{(k)})$
2. find $t^{(k)}$ by line search
3. update
$$x^{(k+1)} = x^{(k)} - t\nabla f(x^{(k)})$$

Analysis of gradient descent for smooth functions

using line search condition,

$$
\begin{aligned}
f(x^+) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|^2 \\
&\leq f(x^\star) - \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|^2 \\
&= f^\star + \frac{1}{2t}(\|x - x^\star\|^2 - \|x - x^\star - t\nabla f(x)\|^2) \\
&= f^\star + \frac{1}{2t}(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2) \\
&\leq f^\star + \beta(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2)
\end{aligned}
$$

second line uses first order convexity condition,
last line uses $\frac{1}{t} \leq \frac{\beta}{b} = 2\beta$

Analysis of gradient descent for smooth functions

take average over iteration counter $i = 1, \ldots, k$:

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f^\star &\leq \frac{1}{k}\sum_{i=1}^{k} \beta(\|x^{(i)} - x^\star\|^2 - \|x^{(i+1)} - x^\star\|^2) \\
&\leq \frac{\beta}{k}(\|x^{(0)} - x^\star\|^2 - \|x^{(k+1)} - x^\star\|^2) \\
&\leq \frac{\beta}{k}\|x^{(0)} - x^\star\|^2
\end{aligned}$$

since $f(x^{(k)})$ is non-increasing,

$$f(x^{(k)}) - f^\star \leq \frac{\beta}{k}\|x^{(0)} - x^\star\|^2$$

so number of iterations k to reach $f(x^{(k)}) - f^\star \leq \epsilon$ is $\mathcal{O}(1/\epsilon)$

Strong convexity: equivalent definitions

for convex $f : \mathbf{R}^n \to \mathbf{R}$, the following properties are all equivalent:

1. $f(x) - \frac{\alpha}{2}x^T x$ is convex
2. f is α-**strongly convex**: for all $x, y \in \mathbf{dom}\, f$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2}\|x - y\|^2$$

3. (if f is twice differentiable) $\nabla^2 f(x) \succeq \alpha I$
4. ∇f is **coercive** with parameter α: for all $x, y \in \mathbf{dom}\, f$,

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \alpha \|x - y\|^2$$

proof: 2) is first order condition for convexity of 1); 3) is second order condition for convexity of 1); 4) is monotone gradient condition for 1)

Strong convexity + smoothness

if f is α-strongly convex and β-smooth,
then $h(x) = f(x) - \alpha/2\|x\|^2$ is convex and $(\beta - \alpha)$-smooth:

$$(\nabla h(x) - \nabla h(y))^T(x - y) \geq \frac{1}{\beta - \alpha}\|\nabla h(x) - \nabla h(y)\|^2$$

expand h to show

$$(\nabla f(x) - \nabla f(y))^T(x - y) \quad \geq \quad \frac{\alpha\beta}{\alpha + \beta}\|x - y\|^2 + \frac{1}{\alpha + \beta}\|\nabla f(x) - \nabla f($$

Analysis of gradient descent for SSC functions

assumptions for analysis:

- $f : \mathbf{R}^n \to \mathbf{R}$ convex and differentiable with $\mathbf{dom}\, f = \mathbf{R}^n$
- f is smooth with parameter $\beta > 0$
- f is strongly convex with parameter $\alpha > 0$
- optimal value $f^\star = \inf_x f(x)$ finite and attained at $x^\star$

Algorithm: GD with constant step size.
pick constant step size $0 < t \leq \frac{2}{\alpha+\beta}$, and repeat

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

note $\alpha < \beta \implies \frac{1}{\beta} < \frac{2}{\alpha+\beta}$,
so SSC allows larger step sizes than just smoothness

Analysis of gradient descent for SSC functions

$$
\begin{aligned}
\|x^+ - x^\star\|^2 &= \|x - t\nabla f(x) - x^\star\|^2 \\
&= \|x - x^\star\|^2 + t^2\|\nabla f(x)\|^2 - 2t\nabla f(x)^T(x - x^\star) \\
&\le \|x - x^\star\|^2 + t^2\|\nabla f(x)\|^2 \\
&\quad -2t\left(\frac{\alpha\beta}{\alpha+\beta}\|x - x^\star\|^2 + \frac{1}{\alpha+\beta}\|\nabla f(x)\|^2\right) \\
&= \left(1 - t\left(\frac{2\alpha\beta}{\alpha+\beta}\right)\right)\|x - x^\star\|^2 + t\left(t - \frac{2}{\alpha+\beta}\right)\|\nabla f(
\\
&\le \left(1 - t\left(\frac{2\alpha\beta}{\alpha+\beta}\right)\right)\|x - x^\star\|^2
\end{aligned}
$$

(first inequality uses coercivity + co-coercivity,
last uses $t \le \frac{2}{\alpha+\beta}$)

Analysis of gradient descent for SSC functions

- distance to optimum decreases by $c = 1 - t(\frac{2\alpha\beta}{\alpha+\beta})$ every iteration
$$\|x^{(k)} - x^\star\|^2 \le c^k \|x^{(0)} - x^\star\|^2$$
i.e., "linear convergence"
- if $t = \frac{2}{\alpha+\beta}$, $c = (\frac{\kappa-1}{\kappa+1})^2$, where $\kappa = \frac{\beta}{\alpha} \ge 1$ is condition number
- using quadratic upper bound, get bound on function value
$$f(x^{(k)}) - f^\star \le \frac{\beta}{2}\|x^{(k)} - x^\star\|^2 \le \frac{\beta c^k}{2}\|x^{(0)} - x^\star\|^2$$
- so number of iterations k to reach $f(x^{(k)}) - f^\star \le \epsilon$ is $\mathcal{O}(\log(1/\epsilon))$

Conclusion

we showed that the gradient method with appropriate step sizes converges, and guarantees

- for f convex and β-smooth,

$$f(x^{(k)}) - f^\star \le \frac{\beta}{2k}\|x^{(0)} - x^\star\|^2$$

- for f convex, β-smooth, and α-strongly convex,

$$f(x^{(k)}) - f^\star \le \frac{\beta c^k}{2}\|x^{(0)} - x^\star\|^2,$$

  where $c = (\frac{\kappa-1}{\kappa+1})^2$, $\kappa = \frac{\beta}{\alpha} \ge 1$ is condition number

References

- Lieven Vandenberghe, UCLA EE236C: Gradient methods
- Sebastian Bubeck, Convex Optimization: Algorithms and Complexity

Outline

Convexity

Gradient descent

Classical analysis of GD

Polyak-Lojasiewiczcondition

The Polyak-Lojasiewiczcondition

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

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

The Polyak-Lojasiewiczcondition

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

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

proof: plug the points $(x, x^\star)$ into the strong convexity condition:

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

since $f(x) - f^\star \geq 0$, we can establish μ-coercivity of ∇f:

$$\begin{aligned}
\nabla f(x)^T(x - x^\star) &\geq \frac{\mu}{2}\|x - x^\star\|^2 \\
\|\nabla f(x)\|\|x - x^\star\| &\geq \frac{\mu}{2}\|x - x^\star\|^2 \\
\|\nabla f(x)\| &\geq \frac{\mu}{2}\|x - x^\star\|
\end{aligned}$$