ORIE 6326: Convex Optimization

Unconstrained minimization

Professor Udell

Operations Research and Information Engineering
Cornell

April 30, 2017

Slides adapted from Stanford EE364a

Outline

- terminology and assumptions
- gradient descent method
- steepest descent method
- Newton's method
- self-concordant functions
- implementation

Unconstrained minimization

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

- $f : \mathbf{R}^n \to \mathbf{R}$ convex, continuously differentiable (hence $\mathbf{dom}\, f$ open)
- we assume optimal value $p^\star = \inf_x f(x)$ is attained (and finite)
- we 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

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

- can be interpreted as iterative methods for solving optimality condition

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

Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

- other notations: $x^+ = x + t\Delta x$, $x := x + t\Delta x$, $x \leftarrow x + t\Delta x$
- Δx is the **step**, or **search direction**
- t is the **step size**, or **step length**
- from convexity, $f(x^+) \geq f(x) + \nabla f(x)\Delta x$, so

$$f(x^+) < f(x) \text{ implies } \nabla f(x)^T \Delta x < 0$$

(*i.e.*, Δx is a **descent direction**)

Algorithm 1 General descent method.

---

given a starting point $x \in$ **dom** f.
repeat
1. Determine a descent direction Δx.
2. **Line search.** Choose a step size $t > 0$.
3. **Update.** $x := x + t\Delta x$.

until stopping criterion is satisfied.

Step size choices

- constant step size $t^{(k)} = t$
- decreasing step size $t^{(k)} = 1/k$
- line search

Line search types

exact line search: $t = \operatorname{argmin}_{t>0} f(x + t\Delta x)$

- how?

Line search types

exact line search: $t = \operatorname{argmin}_{t>0} f(x + t\Delta x)$

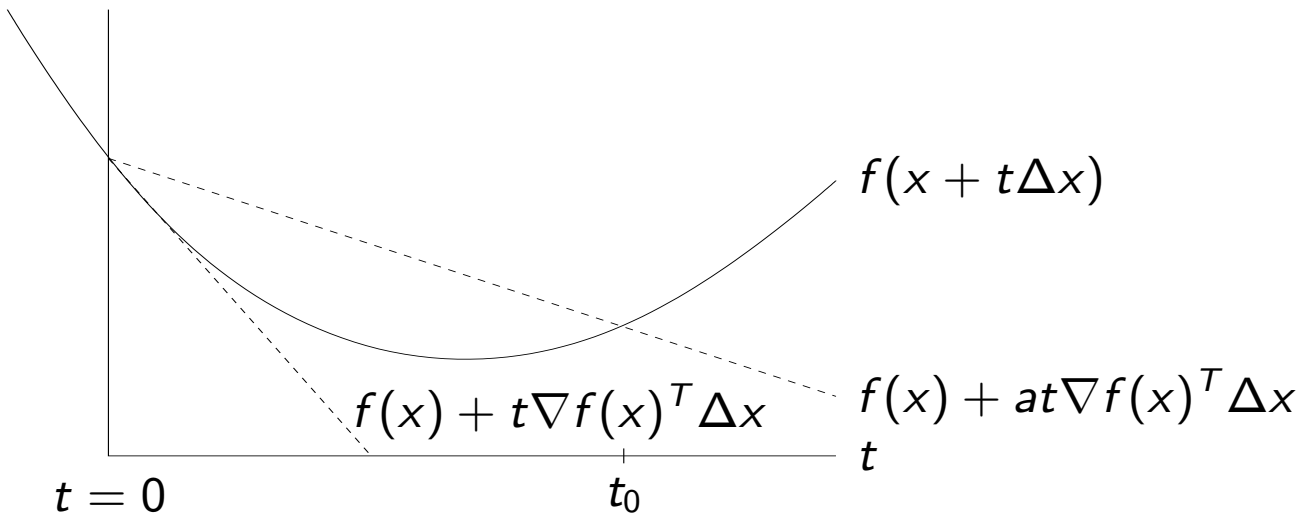
- how? bisection to find a zero of $g'(t)$ where $g(t) = f(x + t\Delta x)$
- takes $O(\log(1/\epsilon))$ steps to find t with $|g'(t)| \leq \epsilon$

backtracking line search (with parameters $a \in (0, 1/2]$, $b \in (0, 1)$)

- starting at $t = 1$, repeat $t := bt$ until

$$f(x + t\Delta x) < f(x) + at\nabla f(x)^T \Delta x$$

- graphical interpretation: backtrack until $t \leq t_0$



Gradient descent method

general descent method with $\Delta x = -\nabla f(x)$

Algorithm 2 Gradient descent method.

---

given a starting point $x \in \mathbf{dom}\, f$.
repeat

1. $\Delta x := -\nabla f(x)$.
2. **Line search.** Choose a step size $t > 0$.
3. **Update.** $x := x + t\Delta x$.

until stopping criterion is satisfied.

- stopping criterion usually of the form $\|\nabla f(x)\|_2 \leq \epsilon$
- very simple, but often very slow

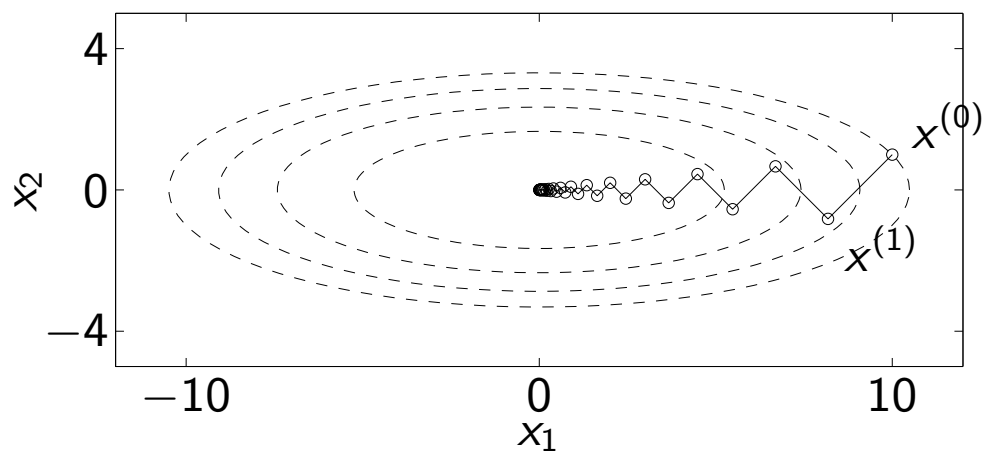
quadratic problem in $\mathbf{R}^2$

$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \qquad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

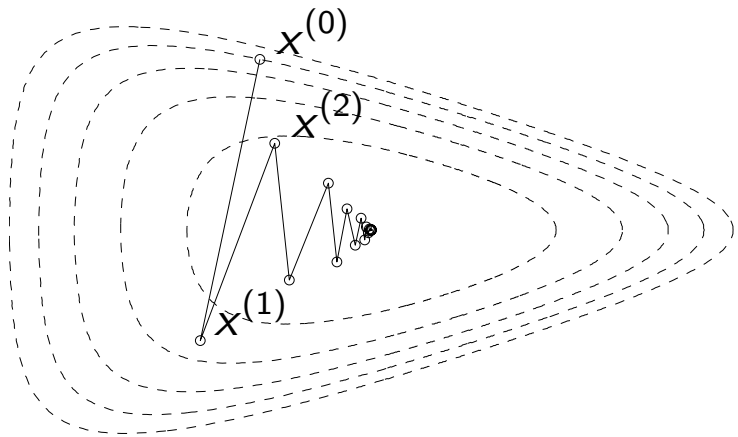
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1} \right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1} \right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$:

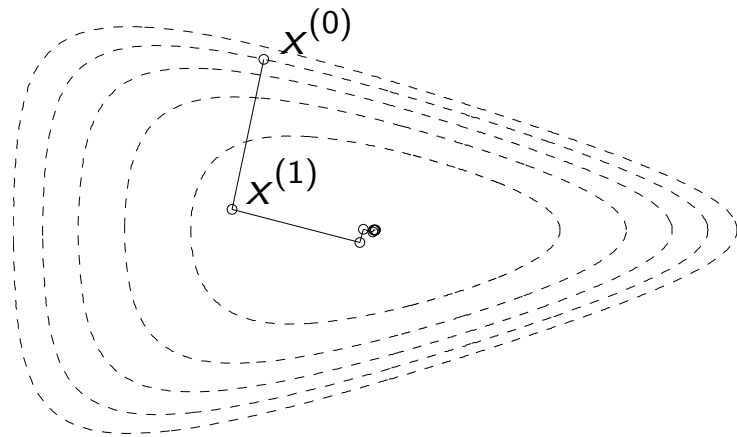


nonquadratic example

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



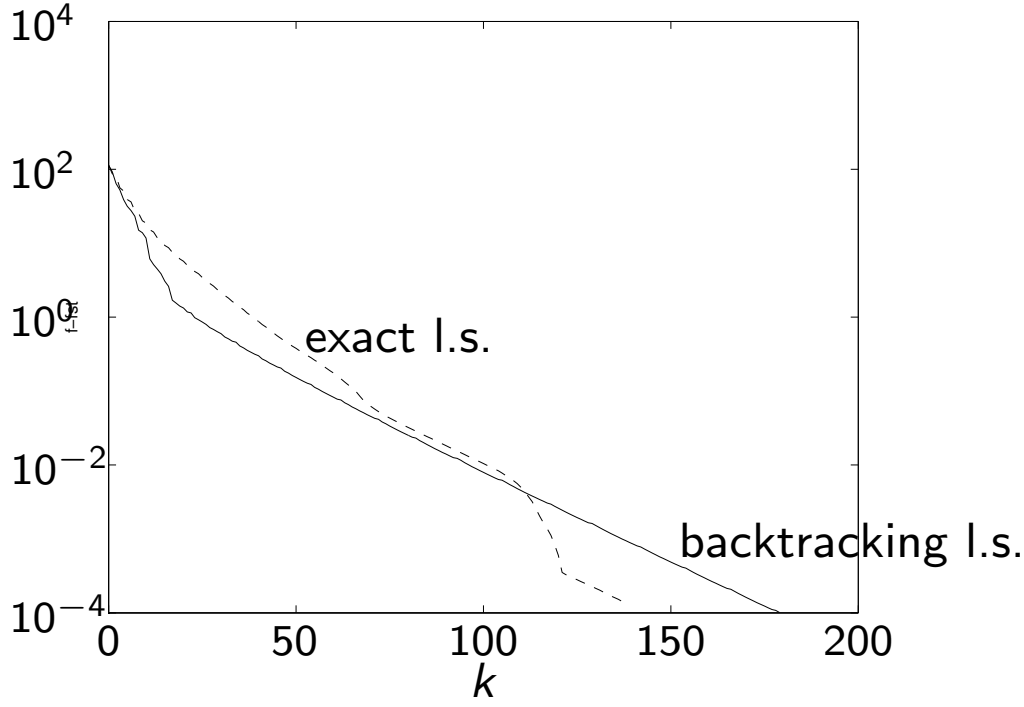
backtracking line search



exact line search

a problem in $\mathbf{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$



'linear' convergence, *i.e.*, a straight line on a semilog plot

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, $y \in \mathbf{dom}\, f$,

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

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

Example I: quadratic

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

- 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 $f(x) = \frac{1}{2}x^TAx$. is f smooth and strongly convex?

Example I: quadratic

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

- 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 $f(x) = \frac{1}{2}x^TAx$. 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

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

- 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 $f(x) = \frac{1}{2}x^TAx$. 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:

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

$$\frac{\lambda_{\min}(A)}{2}\|y-x\|^2 \leq \frac{1}{2}(y-x)^TA(y-x) \leq \frac{\lambda_{\max}(A)}{2}\|y-x\|^2$$

Example II: smoothed absolute value

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

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

consider

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

is huber smooth and strongly convex?

Example II: smoothed absolute value

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

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

consider

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

is huber smooth and strongly convex?

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

Example III: regularized absolute value

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

- 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

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

- 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) $\implies$ 4) using Cauchy-Schwarz; 2) $\implies$ 5) using first order condition (Lemma 3.5 of Bubeck). only 2) $\implies$ 5) requires convexity.

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:

$$\begin{aligned}
& f(y) - f(x) - \nabla f(x)^T(y-x) \\
= & \int_0^1 \nabla f(x+t(y-x))^T(y-x)dt - \nabla f(x)^T(y-x) \\
= & \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x))^T(y-x)dt \\
\leq & \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|_2\|(y-x)\|_2 dt \\
\leq & \int_0^1 \beta t\|(y-x)\|_2^2 dt = \frac{\beta}{2}\|(y-x)\|_2^2
\end{aligned}$$

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 $p^\star = \inf_x f(x)$ 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 $0 < t \leq \frac{1}{\beta}$,

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

(second line uses first order convexity condition)

Analysis of gradient descent for smooth functions

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

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

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

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

so number of iterations k to reach $f(x^{(k)}) - p^\star \leq \epsilon$ 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,

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

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 $i = 1, \ldots, k$:

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

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

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

so number of iterations k to reach $f(x^{(k)}) - p^\star \leq \epsilon$ 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

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

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 $p^\star = \inf_x f(x)$ 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

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

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

Analysis of gradient descent for SSC functions

- distance to optimum decreases by $c = 1 - t(\frac{2\alpha\beta}{\alpha+\beta})$ every iteration

$$\|x^{(k)} - x^\star\|^2 \leq c^k \|x^{(0)} - x^\star\|^2$$

  i.e., "linear convergence"
- if $t = \frac{2}{\alpha+\beta}$, $c = (\frac{\kappa-1}{\kappa+1})^2$, where $\kappa = \frac{\beta}{\alpha} \geq 1$ is condition number
- using quadratic upper bound, get bound on function value

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

- so number of iterations k to reach $f(x^{(k)}) - p^\star \leq \epsilon$ 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,

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

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

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

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

References

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