Subgradient method

Geoff Gordon & Ryan Tibshirani
Optimization 10-725 / 36-725

Remember gradient descent

We want to solve

$$\min_{x \in \mathbb{R}^n} f(x),$$

for f convex and differentiable

Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

If ∇f Lipschitz, gradient descent has convergence rate $O(1/k)$

Downsides:

- Can be slow $\leftarrow$ later
- Doesn't work for nondifferentiable functions $\leftarrow$ today

Outline

Today:

- Subgradients
- Examples and properties
- Subgradient method
- Convergence rate

Subgradients

Remember that for convex $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{all } x, y$$

I.e., linear approximation always underestimates f

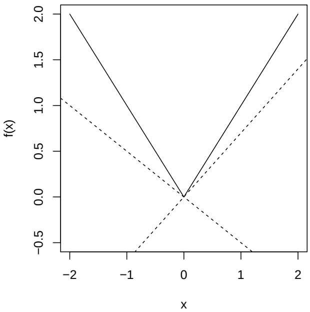
A **subgradient** of convex $f : \mathbb{R}^n \to \mathbb{R}$ at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x), \quad \text{all } y$$

- Always exists
- If f differentiable at x, then $g = \nabla f(x)$ uniquely
- Actually, same definition works for nonconvex f (however, subgradient need not exist)

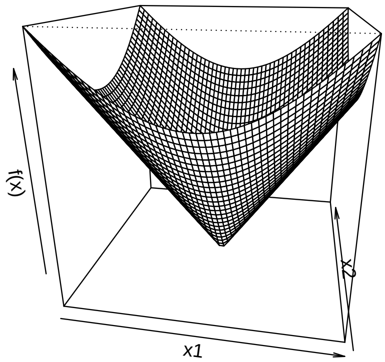
Examples

Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$



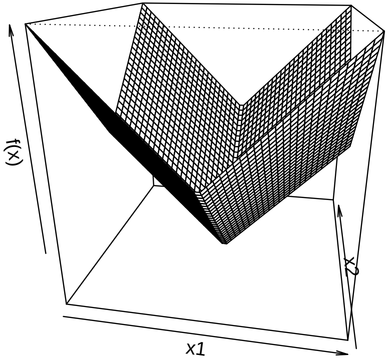
- For $x \neq 0$, unique subgradient $g = \mathrm{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|$ (Euclidean norm)



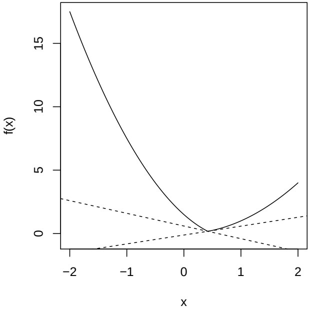
- For $x \neq 0$, unique subgradient $g = x/\|x\|$
- For $x = 0$, subgradient g is any element of $\{z : \|z\| \leq 1\}$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$



- For $x_i \neq 0$, unique ith component $g_i = \text{sign}(x_i)$
- For $x_i = 0$, ith component g_i is an element of $[-1, 1]$

Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable, and consider $f(x) = \max\{f_1(x), f_2(x)\}$



- For $f_1(x) > f_2(x)$, unique subgradient $g = \nabla f_1(x)$
- For $f_2(x) > f_1(x)$, unique subgradient $g = \nabla f_2(x)$
- For $f_1(x) = f_2(x)$, subgradient g is any point on the line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$

Subdifferential

Set of all subgradients of convex f is called the **subdifferential:**

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$$

- $\partial f(x)$ is closed and convex (even for nonconvex f)
- Nonempty (can be empty for nonconvex f)
- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$

Connection to convex geometry

Convex set $C \subseteq \mathbb{R}^n$, consider indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$,

$$I_C(x) = I\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

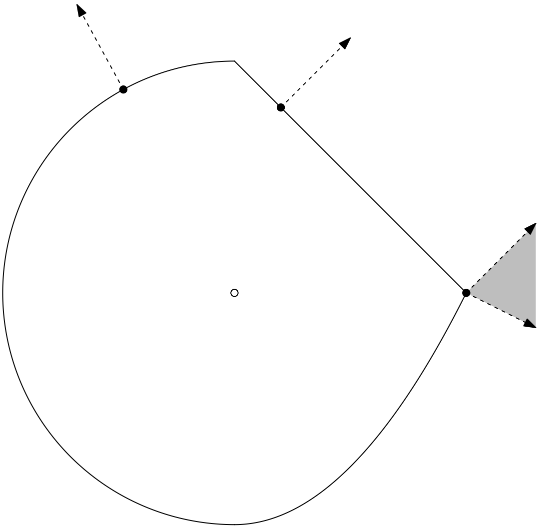
For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x,

$$\mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

Why? Recall definition of subgradient g,

$$I_C(y) \geq I_C(x) + g^T(y - x) \quad \text{for all } y$$

- For $y \notin C$, $I_C(y) = \infty$
- For $y \in C$, this means $0 \geq g^T(y - x)$



Subgradient calculus

Basic rules for convex functions:

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- Affine composition: if $g(x) = f(Ax + b)$, then

$$\partial g(x) = A^T \partial f(Ax + b)$$

- Finite pointwise maximum: if $f(x) = \max_{i=1,\ldots m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv}\Big(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \Big),$$

  the convex hull of union of subdifferentials of all active functions at x

- General pointwise maximum: if $f(x) = \max_{s \in \mathcal{S}} f_s(x)$, then

$$\partial f(x) \supseteq \mathrm{cl}\Big\{\mathrm{conv}\Big(\bigcup_{s: f_s(x) = f(x)} \partial f_s(x)\Big)\Big\}$$

  and under some regularity conditions (on $\mathcal{S}$, f_s), we get $=$

- Norms: important special case, $f(x) = \|x\|_p$. Let q be such that $1/p + 1/q = 1$, then

$$\partial f(x) = \Big\{y : \|y\|_q \leq 1 \text{ and } y^T x = \max_{\|z\|_q \leq 1} z^T x\Big\}$$

  Why is this a special case? Note

$$\|x\|_p = \max_{\|z\|_q \leq 1} z^T x$$

Why subgradients?

Subgradients are important for two reasons:

- Convex analysis: optimality characterization via subgradients, monotonicity, relationship to duality
- Convex optimization: if you can compute subgradients, then you can minimize (almost) any convex function

Optimality condition

For convex f,

$$f(x^\star) = \min_{x \in \mathbb{R}^n} f(x) \quad \Leftrightarrow \quad 0 \in \partial f(x^\star)$$

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$

Why? Easy: $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note analogy to differentiable case, where $\partial f(x) = \{\nabla f(x)\}$

Soft-thresholding

Lasso problem can be parametrized as

$$\min_x \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1$$

where $\lambda \geq 0$. Consider simplified problem with $A = I$:

$$\min_x \frac{1}{2}\|y - x\|^2 + \lambda\|x\|_1$$

Claim: solution of simple problem is $x^\star = S_\lambda(y)$, where S_λ is the **soft-thresholding operator:**

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

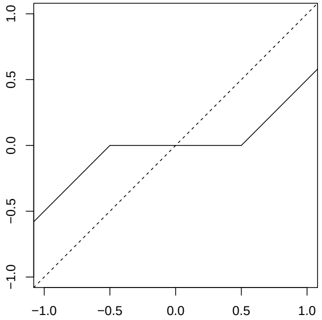
Why? Subgradients of $f(x) = \frac{1}{2}\|y - x\|^2 + \lambda\|x\|_1$ are

$$g = x - y + \lambda s,$$

where $s_i = \mathrm{sign}(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$

Now just plug in $x = S_\lambda(y)$ and check we can get $g = 0$

Soft-thresholding in one variable:



Subgradient method

Given convex $f : \mathbb{R}^n \to \mathbb{R}$, not necessarily differentiable

Subgradient method: just like gradient descent, but replacing gradients with subgradients. I.e., initialize $x^{(0)}$, then repeat

$$x^{(k)} = x^{(k-1)} - t_k \cdot g^{(k-1)}, \quad k = 1, 2, 3, \ldots,$$

where $g^{(k-1)}$ is any subgradient of f at $x^{(k-1)}$

Subgradient method is not necessarily a descent method, so we keep track of best iterate $x^{(k)}_{\text{best}}$ among $x^{(1)}, \ldots x^{(k)}$ so far, i.e.,

$$f(x^{(k)}_{\text{best}}) = \min_{i=1,\ldots k} f(x^{(i)})$$

Step size choices

- Fixed step size: $t_k = t$ all $k = 1, 2, 3, \ldots$
- Diminishing step size: choose t_k to satisfy

$$\sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty,$$

i.e., square summable but not summable

Important that step sizes go to zero, but not too fast

Other options too, but important difference to gradient descent: all step sizes options are pre-specified, not adaptively computed

Convergence analysis

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex, also:

- f is Lipschitz continuous with constant $G > 0$,

$$|f(x) - f(y)| \leq G\|x - y\| \quad \text{for all } x, y$$

  Equivalently: $\|g\| \leq G$ for any subgradient of f at any x

- $\|x^{(1)} - x^*\| \leq R$ (equivalently, $\|x^{(0)} - x^*\|$ is bounded)

Theorem: For a fixed step size t, subgradient method satisfies

$$\lim_{k\to\infty} f(x^{(k)}_{\text{best}}) \leq f(x^\star) + G^2 t/2$$

Theorem: For diminishing step sizes, subgradient method satisfies

$$\lim_{k\to\infty} f(x^{(k)}_{\text{best}}) = f(x^\star)$$

Basic inequality

Can prove both results from same basic inequality. Key steps:

- Using definition of subgradient,

$$\|x^{(k+1)} - x^\star\|^2 \le \|x^{(k)} - x^\star\|^2 - 2t_k(f(x^{(k)}) - f(x^\star)) + t_k^2\|g^{(k)}\|^2$$

- Iterating last inequality,

$$\|x^{(k+1)} - x^\star\|^2 \le \|x^{(1)} - x^\star\|^2 - 2\sum_{i=1}^{k} t_i(f(x^{(i)}) - f(x^\star)) + \sum_{i=1}^{k} t_i^2\|g^{(i)}\|^2$$

- Using $\|x^{(k+1)} - x^\star\| \geq 0$ and $\|x^{(1)} - x^\star\| \leq R$,

$$2\sum_{i=1}^{k} t_i(f(x^{(i)}) - f(x^\star)) \leq R^2 + \sum_{i=1}^{k} t_i^2 \|g^{(i)}\|^2$$

- Introducing $f(x^{(k)}_{\text{best}})$,

$$2\sum_{i=1}^{k} t_i(f(x^{(i)}) - f(x^\star)) \geq 2\Big(\sum_{i=1}^{k} t_i\Big)(f(x^{(k)}_{\text{best}}) - f(x^\star))$$

- Plugging this in and using $\|g^{(i)}\| \leq G$,

$$f(x^{(k)}_{\text{best}}) - f(x^\star) \leq \frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2\sum_{i=1}^{k} t_i}$$

Convergence proofs

For constant step size t, basic bound is

$$\frac{R^2 + G^2 t^2 k}{2tk} \to \frac{G^2 t}{2} \text{ as } k \to \infty$$

For diminishing step sizes t_k,

$$\sum_{i=1}^{\infty} t_i^2 < \infty, \quad \sum_{i=1}^{\infty} t_i = \infty,$$

we get

$$\frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i} \to 0 \text{ as } k \to \infty$$

□

Convergence rate

After k iterations, what is complexity of error $f(x_{\text{best}}^{(k)}) - f(x^\star)$?

Consider taking $t_i = R/(G\sqrt{k})$, all $i = 1, \ldots k$. Then basic bound is

$$\frac{R^2 + G^2 \sum_{i=1}^k t_i^2}{2\sum_{i=1}^k t_i} = \frac{RG}{\sqrt{k}}$$

Can show this choice is the best we can do (i.e., minimizes bound)

I.e., subgradient method has convergence rate $O(1/\sqrt{k})$

I.e., to get $f(x_{\text{best}}^{(k)}) - f(x^\star) \le \epsilon$, need $O(1/\epsilon^2)$ iterations

Intersection of sets

Example from Boyd's lecture notes: suppose we want to find $x^\star \in C_1 \cap \ldots \cap C_m$, i.e., find point in intersection of closed, convex sets $C_1, \ldots C_m$

First define

$$f(x) = \max_{i=1,\ldots m} \operatorname{dist}(x, C_i),$$

and now solve

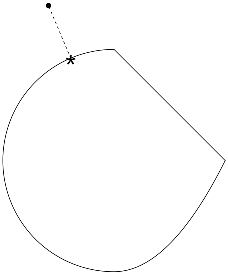
$$\min_{x \in \mathbb{R}^n} f(x)$$

Note that $f(x^\star) = 0 \;\Rightarrow\; x^\star \in C_1 \cap \ldots \cap C_m$

Recall distance to set C,

$$\operatorname{dist}(x, C) = \min\{\|x - u\| : u \in C\}$$

For closed, convex C, there is a unique point minimizing $\|x - u\|$ over $u \in C$. Denoted $u^\star = P_C(x)$, so $\mathrm{dist}(x, C) = \|x - P_C(x)\|$



Let $f_i(x) = \mathrm{dist}(x, C_i)$, each i. Then $f(x) = \max_{i=1,\ldots m} f_i(x)$, and

- For each i, and $x \notin C_i$, $\nabla f_i(x) = \frac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|}$
- If $f(x) = f_i(x) \neq 0$, then $\frac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|} \in \partial f(x)$

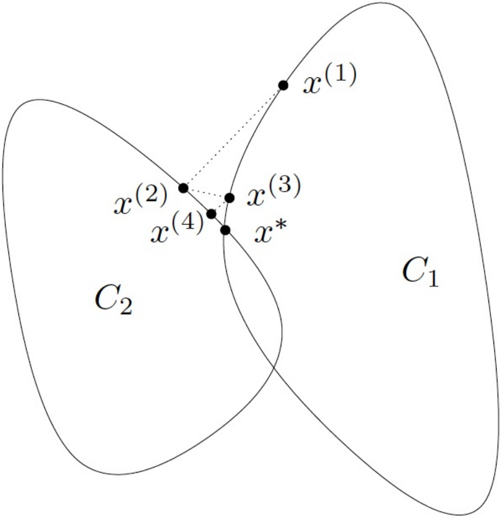
Now apply subgradient method with step size $t_k = f(x^{(k-1)})$ (Polyak step size, can show that we get convergence)

Hence at iteration k, find C_i so that $x^{(k-1)}$ is farthest from C_i. Then update

$$
\begin{aligned}
x^{(k)} &= x^{(k-1)} - f(x^{(k-1)}) \frac{x^{(k-1)} - P_{C_i}(x^{(k-1)})}{\|x^{(k-1)} - P_{C_i}(x^{(k-1)})\|} \\
&= P_{C_i}(x^{(k-1)})
\end{aligned}
$$

Here we used
$f(x^{(k-1)}) = \mathrm{dist}(x^{(k-1)}, C_i) = \|x^{(k-1)} - P_{C_i}(x^{(k-1)})\|$

For two sets, this is exactly the famous **alternating projections** method, i.e., just keep projecting back and forth



(From Boyd's notes)

Can we do better?

Strength of subgradient method: broad applicability

Downside: $O(1/\sqrt{k})$ rate is really slow ... can we do better?

Given starting point $x^{(0)}$. Setup:

- Problem class: convex functions f with solution $x^\star$, with $\|x^{(0)} - x^\star\| \leq R$, f Lipschitz with constant $G > 0$ on $\{x : \|x - x^{(0)}\| \leq R\}$
- Weak oracle: given x, oracle returns a subgradient $g \in \partial f(x)$
- Nonsmooth first-order methods: iterative methods that start with $x^{(0)}$ and update $x^{(k)}$ in

$$x^{(0)} + \mathrm{span}\{g^{(0)}, g^{(1)}, \ldots g^{(k-1)}\}$$

  subgradients $g^{(0)}, g^{(1)}, \ldots g^{(k-1)}$ come from weak oracle

Lower bound

Theorem (Nesterov): For any $k \leq n-1$ and starting point $x^{(0)}$, there is a function in the problem class such that any nonsmooth first-order method satisfies

$$f(x^{(k)}) - f(x^\star) \geq \frac{RG}{2(1+\sqrt{k+1})}$$

Proof: We'll do the proof for $k = n-1$ and $x^{(0)} = 0$; the proof is similar otherwise. Let

$$f(x) = \max_{i=1,\ldots n} x_i + \frac{1}{2}\|x\|^2$$

Solution: $x^\star = (-1/n, \ldots -1/n)$, $f(x^\star) = -1/(2n)$

For $R = 1/\sqrt{n}$, f is Lipschitz with $G = 1 + 1/\sqrt{n}$

Oracle: returns $g = e_j + x$, where j is smallest index such that $x_j = \max_{i=1,\ldots n} x_i$

Claim: for any $i \in 1, \dots n-1$, the ith iterate satisfies

$$x_{i+1}^{(i)} = \dots = x_n^{(i)} = 0$$

Start with $i = 1$: note $g^{(0)} = e_1$. Then:

- $\mathrm{span}\{g^{(0)}, g^{(1)}\} \subseteq \mathrm{span}\{e_1, e_2\}$
- $\mathrm{span}\{g^{(0)}, g^{(1)}, g^{(2)}\} \subseteq \mathrm{span}\{e_1, e_2, e_3\}$
- ...
- $\mathrm{span}\{g^{(0)}, g^{(1)}, \dots g^{(i-1)}\} \subseteq \mathrm{span}\{e_1, \dots e_i\}$ v

Therefore $f(x^{(n-1)}) \geq 0$, recall $f(x^\star) = -1/(2n)$, so

$$f(x^{(n-1)}) - f(x^\star) \geq \frac{1}{2n} = \frac{RG}{2(1+\sqrt{n})}$$

□

Improving on the subgradient method

To improve, we must go beyond nonsmooth first-order methods

There are many ways to improve for general nonconvex problems, e.g., localization methods, filtered subgradients, memory terms

Instead, we'll focus on minimizing functions of the form

$$f(x) = g(x) + h(x)$$

where g is convex and differentiable, h is convex

For a lot of problems (i.e., functions h), we can recover $O(1/k)$ rate of gradient descent with a simple algorithm, having big practical consequences

References

- S. Boyd, Lecture Notes for EE 264B, Stanford University, Spring 2010-2011
- Y. Nesterov (2004), *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Chapter 3
- B. Polyak (1987), *Introduction to Optimization*, Optimization Software Inc., Chapter 5
- R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press, Chapters 23–25
- L. Vandenberghe, Lecture Notes for EE 236C, UCLA, Spring 2011-2012