Lecture Notes CAO

Juan Vera

October 2024

2 Motivation - the gradient method

The gradient method

To minimize a convex differentiable function f : choose an initial point x_0 and repeat

$$x_{k+1} = x_k - \eta_k \nabla f(x_k),$$

where step size η_k is chosen by line search or is some (small) fixed constant.

Advantages:

- every iteration is inexpensive,
- does not require second derivatives.

Disadvantages:

- often slow;
- does not handle non-differentiable problems.

We will study methods addressing the shortcomings of the gradient method:

1. subgradient method
2. proximal gradient method
3. DouglasRachford method
4. ADMM

2.1 Analysis of gradient method for fixed step $\eta_k = \eta > 0$

Assumption 1. *In this section we assume*

1. *f is convex and differentiable with* $\operatorname{dom} f = \mathbb{R}^n$
2. *∇f is L-Lipschitz continuous (L-C), with $L > 0$:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \text{ for all } x, y \in \operatorname{dom} f$$

3. *optimal value* $f^* = \inf f(x)$ *is finite and attained at* x^*

4. $0 < \eta \leq 1/L$.

For any x let $x^+ = x - \eta \nabla f(x)$.

Lemma 1. *For all* x, $f(x^+) \leq f(x) - \frac{\eta}{2}\|\nabla f(x)\|^2$.

Proof. Let $d = \nabla f(x)$. Let $g(t) = f(x - td)$. Then $g'(t) = d^T f(x - td)$. From L-C we have

$$g'(t) - g'(0) = d^T(\nabla f(x) - \nabla f(x - td)) \leq \|d\|\|\nabla f(x) - \nabla f(x - td)\| \leq Lt\|d\|^2 = Lt\|\nabla f(x)\|^2.$$

Then,

$$\begin{aligned} f(x^+) = g(\eta) = g(0) + \int_0^\eta g'(t)dt &\leq f(x) + L\tfrac{\eta^2}{2}\|\nabla f(x)\|^2 + \eta g'(0) \\ &= f(x) + \eta(L\tfrac{\eta}{2} - 1)\|\nabla f(x)\|^2 \leq f(x) - \frac{\eta}{2}\|\nabla f(x)\|^2. \end{aligned}$$

□

Lemma 2. *For all* x,

$$f(x^+) - f^* \leq \tfrac{1}{2\eta}(\|x - x^*\|^2 - \|x^+ - x^*\|^2).$$

Proof. By Lemma 1

$$\begin{aligned} f(x^+) - f^* &\leq f(x) - f^* - \tfrac{\eta}{2}\|\nabla f(x)\|^2 \\ &\leq \nabla f(x)^T(x - x^*) - \tfrac{\eta}{2}\|\nabla f(x)\|^2 \\ &= \tfrac{1}{2\eta}\left(\|x - x^*\|^2 - \|x - x^* - \eta\nabla f(x)\|^2\right) \\ &= \tfrac{1}{2\eta}\left(\|x - x^*\|^2 - \|x^+ - x^*\|^2\right). \end{aligned}$$

□

Descend properties of the gradient method

If $\nabla f(x) \neq 0$ and $0 < \eta \leq 1/L$.

- Lemma 1 shows
$$f(x^+) < f(x)$$
- Lemma 2 shows
$$\|x^+ - x^*\| < \|x - x^*\|$$

Function value and distance to optimum are decreasing! - how fast?

Theorem 1. *Number of iterations to reach* $f(x_k) - f^* \leq \varepsilon$ *is* $O(1/\varepsilon)$.

Proof. From Lemma 2

$$\begin{aligned} \sum_{i=1}^k (f(x_i) - f^*) &\leq \tfrac{1}{2\eta}\sum_{i=1}^k (\|x_{i-1} - x^*\|^2 - \|x_i - x^*\|^2) \\ &= \tfrac{1}{2\eta}(\|x_0 - x^*\|^2 - \|x_k - x^*\|^2) \\ &\leq \tfrac{1}{2\eta}\|x_0 - x^*\|^2. \end{aligned}$$

From Lemma 1, $f(x_k)$ is non-increasing and thus

$$f(x_k) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f^*) \leq \frac{1}{2k\eta}\|x_0 - x^*\|^2.$$

□

2.2 Related results

Theorem 2. *If f is strongly convex (on top of assumption 1) the number of iterations to reach $f(x_k) - f^* \leq \varepsilon$ is $O(\log(1/\varepsilon))$.*

Limits on convergence rate of first-order methods

First order methods: Any iterative algorithm that selects x_{k+1} in the set

$$x_0 + span\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)\}.$$

Problem class: Any function satisfying assumption 1

Theorem 3. *[Nesterov, Theorem 2.1.7 Lectures on Convex Optimization (2018)] For every integer $k \leq (n-1)/2$ and every x_0, there exist functions in the problem class such that any first-order method*

$$f(x_k) - f^* \geq \frac{3}{32}\frac{L\|x_0 - x^*\|^2}{(k+1)^2}.$$

Bibliography

- A. Beck, First-Order Methods in Optimization (2017), chapter 5.
- Yu. Nesterov, Lectures on Convex Optimization (2018), section 2.1.
- B. T. Polyak, Introduction to Optimization (1987), section 1.4.
- L. Vandenberghe, lecture slides ECE236C, Optimization Methods for Large-Scale Systems.

3 Subgradient method

The subgradient method

To minimize a nondifferentiable convex function f : choose an initial point x_0 and repeat

$$x_{k+1} = x_k - \eta_k g_k, \ k = 0, 1, \ldots$$

where $g_k \in \partial f(x_k)$ is a subgradient of f at x_k.

Step size rules:

- fixed step: η_k constant

- fixed length: $\eta_k\|g_k\| = \|x_{k+1} - x_k\|$ constant
- diminishing: $t_k \to 0$ and $\sum_{k=0}^{\infty} t_k = \infty$.

Assumption 2 (Subgradient method)**.** *In this section we assume*

1. *f is convex with* $\operatorname{dom} f = \mathbb{R}^n$
2. *f is G-Lipschitz continuous:*

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

3. *optimal value $f^* = \inf f(x)$ is finite and attained at x^**

Lemma 3. *Assumption 22 is equivalent to $\|g\| \leq G$ for all x and all $g \in \partial f(x)$.*

Proof. Assume $\|g\| \leq G$ for all x and all $g \in \partial f(x)$. WLOG assume $f(x) \geq f(y)$. Let $g_x \in \partial f(x)$. We have

$$|f(x) - f(y)| = f(x) - f(y) \leq g_x^T(x - y) \leq G\|x - y\|,$$

where the last inequality follows by the Cauchy-Schwarz inequality.

Now, assume f is G-Lipschitz continuous. Take x and $g \in \partial f(x)$. Take $y = x + \frac{1}{\|g\|}g$, we have $f(y) \geq f(x) + g^T(y - x) = f(x) + \|g\|$ and thus $\|g\| \leq f(y) - f(x) \leq G\|y - x\| = G$. □

3.1 Analysis of the subgradient method

The subgradient method is not a descent method. Thus we are interested in the quantity $f_k^* = \min_{i=0,\dots,k} f(x_i)$ the best value obtained at iteration k.

Lemma 4. *For all k,*

$$f_k^* - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^{k} \eta_i^2\|g_i\|^2}{2\sum_{i=0}^{k} \eta_i}.$$

Proof. For all $i \leq k$

$$\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \|x_i - \eta_i g_i - x^*\|^2 \\
&= \|x_i - x^*\|^2 - 2\eta_i g_i^T(x_i - x^*) + \eta_i^2\|g_i\|^2 \\
&\leq \|x_i - x^*\|^2 - 2\eta_i(f(x_i) - f^*) + \eta_i^2\|g_i\|^2 \\
&\leq \|x_i - x^*\|^2 - 2(f_k^* - f^*)\eta_i + \eta_i^2\|g_i\|^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
2(f_k^* - f^*)\sum_{i=0}^{k} \eta_i &\leq \|x_0 - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \sum_{i=0}^{k} \eta_i^2\|g_i\|^2 \\
&\leq \|x_0 - x^*\|^2 + \sum_{i=0}^{k} \eta_i^2\|g_i\|^2
\end{aligned}$$

□

3.1.1 Fixed step size: $\eta_k = \eta$ with η constant

From Lemma 4

$$f_k^* - f^* \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\eta} + \frac{\eta G^2}{2}.$$

We can not guarantee convergence of f_k^*. When k is large f_k^* is ηG^2-suboptimal.

3.1.2 Fixed step length: $\eta_k = \eta/\|g_k\|$ with η constant

From Lemma 4

$$f_k^* - f^* \leq \frac{G\|x_0 - x^*\|^2}{2(k+1)\eta} + \frac{\eta G}{2}.$$

We can not guarantee convergence of f_k^*. When k is large f_k^* is ηG-suboptimal.

3.1.3 Disminishing step size: $\eta_k \to 0$ with $\sum_{i=0}^{\infty} \eta_i = \infty$

From Lemma 4

$$f_k^* - f^* \leq \frac{\|x_0 - x^*\|^2}{2\sum_{i=0}^k \eta_i} + \frac{G^2}{2}\frac{\sum_{i=0}^k \eta_i^2}{\sum_{i=0}^k \eta_i}.$$

It can be shown that $\frac{\sum_{i=0}^k \eta_i^2}{\sum_{i=0}^k \eta_i} \to 0$ (exercise) thus $f_k^* \to f^*$.

In practice some typical diminishing step sizes are $\eta_i = \frac{\eta}{i+1}$ and $\eta_i = \frac{\eta}{\sqrt{i+1}}$.

3.2 Optimal step sizes

So far we only considered convergence. What about rates of convergence?

3.2.1 Optimal step size when f^* is known

Assume f^* known. Look at proof of Lemma 4. We have

$$\|x_{i+1} - x^*\|^2 \leq \|x_i - x^*\|^2 - 2\eta_i(f(x_i) - f^*) + \eta_i^2 G^2.$$

Taking $\eta_i = (f(x_i) - f^*)/G^2$ we obtain for all $i \leq k$,

$$(f_k^* - f^*)^2 \leq (f(x_i) - f^*)^2 \leq G^2(\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2)$$

And therefore,

$$f_k^* - f^* \leq \frac{G\|x_0 - x^*\|}{\sqrt{k+1}}.$$

3.2.2 Optimal step size for fixed number of iterations

Assume $\|x_0 - x^*\| \leq R$ known, and k given. Let $s_i = \eta_i\|g_i\|$. From Lemma 4

$$f_k^* - f^* \leq \frac{G}{2}\frac{R^2 + \sum_{i=0}^k s_i^2}{\sum_{i=0}^k s_i}.$$

The RHS is minimized by taking $s_i = \frac{R}{\sqrt{k+1}}$, obtaining the bound

$$f_k^* - f^* \leq \frac{GR}{\sqrt{k+1}}.$$

3.2.3 Optimality of rate of convergence

In both cases we have obtained $f_k^* - f^* \leq \frac{GR}{\sqrt{k+1}}$. Such bound guarantees accuracy $f_k^* - f^* \leq \varepsilon$ in $k = O(1/\varepsilon^2)$ iterations.

Question

Is the bound

$$f_k^* - f^* \leq \frac{GR}{\sqrt{k+1}}.$$

the best we can do?

We will see that actually subgradient method is optimal in terms of convergence rate as this is the best possible bound. For this we need to define the problem/algorithm class that we are looking at:

Problem class: Minimize a function satisfying assumption 2.

- We are given x_0 such that $\|x_0 - x^*\| \leq R$.
- We know G the Lipschitz constant of f on $\{x : \|x - x^*\| \leq R\}$.
- f is given by an oracle: for any x the oracle returns $f(x)$ and $g \in \partial f(x)$.

Algorithm Class: Any iterative algorithm that selects x_{i+1} in the set

$$x_0 + span\{g_0, g_1, \ldots, g_i\},$$

and stops after k iterations.

Theorem 4. *For every integer $k < n$, there exist a function in the problem class such that any algorithm in the algorithm class*

$$f_k^* - f^* \geq \frac{GR}{2(2+\sqrt{k+1})}.$$

Proof. Exercise. □

3.3 Subgradient method to find point in intersection of convex sets

Problem: find point in the intersection of m closed convex sets $C_1, \ldots, C_m$

Model as optimization problem: Let $f_j(x) = \inf_{u \in C_j} \|x - u\|$ be the distance from x to C_j. Let $f(x) = \max_{j=1,\ldots,m} f_j(x)$. To find point on intersection solve,

$$\min f(x).$$

We have that each f_j is convex (exercise) and thus f is also convex. Also $f^* = 0$ if and only if intersection nonempty.

We need to compute subgradients. $g \in \partial f(x)$ if $g \in \partial f_j(x)$ where C_j is the farthest set from x. Now, we can find a subgradient $g \in \partial f_j(x)$. If $x \in C_j$ take $g = 0$. If $x \notin C_j$ take

$g = \frac{1}{\|x - P_j(x)\|}(x - P_j(x)) = \frac{1}{f_j(x_k)}(x - P_j(x))$ where $P_j(x)$ is the projection on C_j.

Exercise: Find $\partial f(x)$ for all x.

As we are interested in the case of the non-empty intersection we consider the optimal step size when $f^* = 0$. For the considered subgradients in $\partial f(x)$ we have $\|g\| = 1$ unless x is in $C = \cap_{j=1}^m C_j$. From Section 3.2.1 we have that the optimal step size is $\eta_i = f(x_i)$. Thus at iteration i, find farthest set C_j from x_i and take

$$x_{i+1} = x_i - \eta_i g_i = x_i - \frac{f(x_i)}{f_j(x_i)}(x_i - P_j(x_i)) = P_j(x_i).$$

I.e. at each step project current point into the farthest set.

3.4 Projected subgradient method

We can extend the subgradient method to solve constrained problems. Let f be a (nondifferentiable) convex function and C be a closed convex set.

The Projected subgradient method

To minimize a nondifferentiable convex function f on closed convex set C: choose an initial point x_0 and repeat

$$x_{k+1} = P_C(x_k - \eta_k g_k),\ k = 0, 1, \ldots$$

where $g_k \in \partial f(x_k)$ is a subgradient of f at x_k, P_C is the Euclidean projection on C and η_k chosen by same rules for the unconstrained version.

To apply the projected subgradient it is necessary to be able to compute $P_C(y)$ for any given y. In general this is a difficult task. But there are available analytical forms for *simple* sets.

3.4.1 Examples of projections

1. Halfspace: $C = \{x : a^T x \leq b\}$

$$P_C(x) = \begin{cases} x & \text{if } x \in C \\ x + \frac{b - a^T x}{\|a\|^2} a & \text{otherwise.} \end{cases}$$

2. Ball: $C = \{x : \|x\| \leq R\}$

$$P_C(x) = \begin{cases} x & \text{if } x \in C \\ \frac{R}{\|x\|} x & \text{otherwise.} \end{cases}$$

Works for many common norms (e.g. 1-norm, 2-norm, ∞-norm).

3.4.2 Analysis of projected subgradient

Assume C is closed convex set and assumption 2.

Lemma 5. *For any x and any $y \in C$ we have*

$$\|P_C(x) - y\| \leq \|x - y\|.$$

Proof. Exercise □

Next we see that Lemma 4 still holds. We can replace the first two equalities by

$$\begin{aligned}\|x_{i+1} - x^*\|^2 &= \|P_C(x_i - \eta_i g_i) - x^*\|^2 \\ &\leq \|x_i - \eta_i g_i - x^*\|^2 \\ &= \|x_i - x^*\|^2 - 2\eta_i g_i^T (x_i - x^*) + \eta_i^2 \|g_i\|^2.\end{aligned}$$

And thus the same analysis as for the unrestricted case applies.

Bibliography

- S. Boyd, Lecture slides and notes for EE364b, Convex Optimization II.
- Yu. Nesterov, Lectures on Convex Optimization (2018), section 3.2.3.
- B. T. Polyak, Introduction to Optimization (1987), section 5.3.
- L. Vandenberghe, lecture slides ECE236C, Optimization Methods for Large-Scale Systems.