Optimization for Data Science Lecture 08 and 09: Optimal Gradient Methods

Kimon Fountoulakis

School of Computer Science
University of Waterloo

03/10/2019

Outline

- Accelerated gradient method for convex functions.

Descent methods

- So far we have talked about algorithms that generate a sequence $\{x_k\}_{k\in\mathbb{N}_0}$ such that

$$f(x_{k+1}) < f(x_k)$$

- This type of sequences gave us the following rates

	Convex	$\delta-$**strongly-convex**
Smooth $f(x)$	$\mathcal{O}\left(\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\delta}\log\frac{1}{\epsilon}\right)$

Optimal gradient methods

- For convex and strongly-convex functions the previous rates are not optimal!!

- It can be shown that optimal algorithms that only use gradients to produce a sequence $\{x_k\}_{k\in\mathbb{N}_0}$ have the following rate:

	Convex	$\delta-$**strongly-convex**
Smooth $f(x)$	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\delta}}\log\frac{1}{\epsilon}\right)$

Nesterov's (optimal) accelerated method

- Let γ_k be such that $\prod_{i=0}^{k}(1-\gamma_i) \geq \gamma_k^2 \quad \forall k \geq 0$ and $\gamma_k \in [0,1]$.

- One option for γ_k such that this is true is:

  - $\gamma_0 = \gamma_1 = \gamma_2 = \gamma_3 = 0$ and $\gamma_i := \dfrac{2}{i}$ for $i \geq 4$.

- Let $\lambda_k := \prod_{i=0}^{k}(1-\gamma_i)$.

Nesterov's (optimal) accelerated method

- Pick and arbitrary $x_0 \in \mathbb{R}^n$.
- Set $z_0 := x_0, \gamma_0 = 0$
- Until the termination criterion is not satisfied do:
  - $y_{k-1} := (1 - \gamma_k)x_{k-1} + \gamma_k z_{k-1}$
  - $z_k := z_{k-1} - \dfrac{\gamma_k}{\lambda_k} \nabla f(y_{k-1})$
  - $x_k := y_{k-1} - \dfrac{1}{L} \nabla f(y_{k-1})$

Optimal gradient methods

- This result was initially an existential result, i.e., it was proved that there must exist gradient based methods that have optimal rates of convergence.
- However, an actual algorithm with optimal rate was first discovered by Y. Nesterov in 1983.

Optimal gradient methods

- Nesterov showed that to obtain algorithms with optimal rate we have to relax the "descent" constraint that the output sequence $\{x_k\}_{k\in\mathbb{N}_0}$ must satisfy

$$f(x_{k+1}) < f(x_k)$$

- In other words, we will have to think of methods that do not decrease monotonically the objective function.

- Today we will discuss a new type of sequences that allows us to obtain algorithms with optimal rate.

Estimate Sequence

- **Definition:** a sequence $\{\phi_k, \lambda_k, x_k\}_{k\in\mathbb{N}_0}$, where $\phi_k : \mathbb{R}^n \to \mathbb{R}$, $\lambda_k \in [0,1]$ and $x_k \in \mathbb{R}^n$ is said to be an **estimate sequence** if it satisfies the following:
  - "upper bound": $\forall k \in \mathbb{N}_0$ and $\forall x \in \mathbb{R}^n$ we have $\phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x)$
  - "lower bound": $\forall x \in \mathbb{R}^n$ we have $f(x_k) \leq \phi_k(x)$

Estimate Sequence

- Function ϕ_k serves as an approximation to f, which becomes tighter as $k \to \infty$.

- Example: since both inequalities hold for any x. Let's set $x := x^*$ (the minimizer of f). Then using the "lower bound" inequality in the definition of an estimate sequence we get:

$$f(x_k) - f^* \leq \phi_k(x^*) - f^*$$

Estimate Sequence

- Therefore, if $\phi_k(x^*) - f^* \to 0$ as $k \to \infty$, then we must also have that $f(x_k) - f^* \to 0$.

Estimate Sequence

- Can we prove that an estimate sequence exists?

- **Theorem:** for every convex function f with Lipschitz continuous gradient, then for an arbitrary $x_0 \in \mathbb{R}^n$ there exists an estimate sequence $\{\phi_k, \lambda_k, x_k\}_{k \in \mathbb{N}_0}$ with $\phi_0(x) := f(x_0) + \frac{L}{2}\|x - x_0\|_2^2$ and $\lambda_k \leq \frac{c}{k^2}$ for some constant $c > 0$.

Estimate Sequence

- Why is an estimate sequence useful?

- For such a sequence we have the following:

$$f(x_k) - f^* \leq \mathcal{O}\left(\frac{1}{k^2}\right)$$

- This is 1/k times faster than what we get with gradient descent!!

Estimate Sequence

- This also means that after

$$t = \mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$$

- iterations we guarantee that $f(x_k) - f^* \leq \epsilon$. This is much faster than

$$t = \mathcal{O}\left(\frac{L}{\epsilon}\right)$$

- which is what we get for gradient descent for convex functions.

Estimate Sequence

- Proving the convergence rate theorem for the estimate sequence is **non-trivial!!**

- But we will do this anyway because it's cool!!