Optimization for Machine Learning CS-439

Lecture 10: Accelerated Gradient Descent

Martin Jaggi

EPFL – `github.com/epfml/OptML_course`

May 18, 2018

Re-visiting gradient descent

Property of f	Learning Rate γ	Number of steps
$\lVert\mathbf{x}_0 - \mathbf{x}^\star\rVert \leq R$, $\lVert\nabla f(\mathbf{x})\rVert \leq L$ for all $\mathbf{x}$	$\frac{R}{L\sqrt{T}}$	$\mathcal{O}(1/\varepsilon^2)$
f is L-smooth	$\frac{1}{L}$	$\mathcal{O}(1/\varepsilon)$
f is L-smooth and μ-strongly convex	$\frac{1}{L}$	$\mathcal{O}(\log(1/\varepsilon))$

Improving gradient descent

Problem: Can we do any better? In particular, can we accelerate gradient descent?

Solution: Nesterov's accelerated gradient methods come to the rescue.

Momentum

Idea:

Use **momentum** from "movement" so far

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t) + \nu \big[\mathbf{x}_t - \mathbf{x}_{t-1} \big]$$

$\nu > 0$ is called the momentum parameter

Accelerated Gradient Method - AGD

$$
\begin{aligned}
\mathbf{x}_0 &:= \mathbf{y}_0 := \mathbf{z}_0 \\
\mathbf{x}_{t+1} &:= \tau \mathbf{z}_t + (1-\tau)\mathbf{y}_t \\
\mathbf{y}_{t+1} &:= \mathbf{x}_{t+1} - \frac{1}{L}\nabla f(\mathbf{x}_{t+1}) \\
\mathbf{z}_{t+1} &:= \mathbf{z}_t - \gamma \nabla f(\mathbf{x}_{t+1})
\end{aligned}
$$

Accelerated Gradient Method - Analysis

Problem: What about the values of γ and τ?

Solution: We start with analysis and set them so as to get the best results.

AGD - Analysis for smooth convex functions, cont.

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L, $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $|f(\mathbf{x}_0) - f(\mathbf{x}^\star)| \leq d$. Then, after $T = 4R\sqrt{\frac{L}{d}}$ steps and setting $\gamma = \frac{R}{\sqrt{dL}}$ and τ such that $\frac{1-\tau}{\tau} = \gamma L$, the average of the first T iterates satisfies

$$f\Big(\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_t\Big) - f(\mathbf{x}^\star) \leq \frac{d}{2}$$

AGD - Analysis for smooth convex functions, cont.

Proof.

(i)

Recall from Lecture 3 that the updates of the type $\mathbf{y}_{t+1} := \mathbf{x}_{t+1} - \frac{1}{L}\nabla f(\mathbf{x}_{t+1})$ are always monotone decreasing:

$$f(\mathbf{y}_t) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

(ii)

Use the fact that $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ to obtain

$$\gamma\nabla f(\mathbf{x}_{t+1})^\top(\mathbf{z}_t - \mathbf{x}^\star) = \frac{\gamma^2}{2}\|\nabla f(\mathbf{x}_{t+1})\|^2 + \frac{1}{2}\|\mathbf{z}_t - \mathbf{x}^\star\|^2 - \frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{x}^\star\|^2$$

Using the first equation we get

$$\gamma\nabla f(\mathbf{x}_{t+1})^\top(\mathbf{z}_t - \mathbf{x}^\star) \leq \gamma^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})) + \frac{1}{2}\|\mathbf{z}_t - \mathbf{x}^\star\|^2 - \frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{x}^\star\|^2 \qquad (1)$$

AGD - Analysis for smooth convex functions, cont.

Use convexity and set $\frac{1-\tau}{\tau} = \gamma L$ to obtain

$$
\begin{aligned}
\gamma \nabla f(\mathbf{x}_t)^\top((\mathbf{x}_{t+1} - \mathbf{x}^\star) - (\mathbf{z}_t - \mathbf{x}^\star)) \quad &= \gamma \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{z}_t) \\
&= \tfrac{1-\tau}{\tau} \gamma \nabla f(\mathbf{x}_t)^\top(\mathbf{y}_t - \mathbf{x}_{t+1}) \\
&\leq \gamma^2 L (f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})) \qquad (2)
\end{aligned}
$$

Add (1) and (2) to obtain

$$
\begin{aligned}
\gamma \nabla f(\mathbf{x}_{t+1})^\top(\mathbf{x}_{t+1} - \mathbf{x}^\star) \leq \quad & \gamma^2 L (f(\mathbf{y}_t) - f(\mathbf{y}_{t+1})) + \\
& \tfrac{1}{2}\|\mathbf{z}_t - \mathbf{x}^\star\|^2 - \tfrac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{x}^\star\|^2
\end{aligned}
$$

AGD - Analysis for smooth convex functions, cont.

We know that

$$f\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t\Big) - f(\mathbf{x}^\star) \leq \frac{1}{T}\nabla f(\mathbf{x}_{t+1})^\top(\mathbf{x}_{t+1} - \mathbf{x}^\star)$$

Using the telescoping sum in the previous slide, the proposed substitutions give the desired result. □

AGD - Analysis for smooth convex functions, cont.

Theorem

By repeatedly restarting the AGD algorithm, we can find an ε-optimal solution in $\mathcal{O}(1/\sqrt{\varepsilon})$ updates.

Proof.

Use the previous theorem (Exercise). □

AGD - Analysis for strongly convex smooth functions

Theorem

Along with the previous assumptions, if we assume that the function f is μ-strongly convex, then we can find a point $\mathbf{x}$ with $\mathcal{O}(\sqrt{\frac{L}{\mu}})$ updates such that

$$\|\mathbf{x} - \mathbf{x}^\star\|^2 \leq \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

Proof.

Use the results of previous theorem with $\varepsilon = \frac{\mu}{4}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$ to find a point $\mathbf{x}$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \frac{\mu}{4}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \tag{3}$$

This will take $\mathcal{O}(\sqrt{\frac{L}{\mu}})$ update steps. □

AGD - Analysis for strongly convex smooth functions, cont.

Proof.

Use strong convexity of f to obtain

$$\frac{\mu}{2}\|\mathbf{x}-\mathbf{x}^\star\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^\star) \tag{4}$$

Combine (3) and (4) to get the desired result. □

AGD - Analysis for strongly convex smooth functions, cont.

Theorem

Convergence in Iterate -
By repeatedly starting the AGD algorithm, for a μ-strongly convex and L-smooth function, we can find an ε-optimal solution in the value of iterate in $\mathcal{O}(\log(1/\varepsilon))$ updates where the constant in the big-$\mathcal{O}$ is $\sqrt{\frac{L}{\mu}}$ compared to vanilla GD where the constant is $\frac{L}{\mu}$.

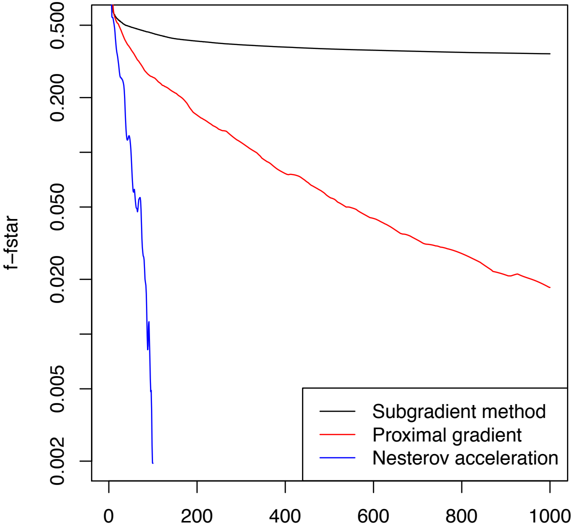
Overview of Accelerated Gradient Method

Properties of f	GD steps	AGD steps
f is L-smooth	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(1/\sqrt{\varepsilon})$
f is L-smooth and μ-strongly convex	$\mathcal{O}(\frac{L}{\mu}\log(1/\varepsilon))$	$\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(1/\varepsilon))$

Table: A comparison of Gradient descent and Accelerated Gradient Method for convex functions - number of updates to obtain an ε-optimal solution

Acceleration in practice

Application to a Lasso problem



Acceleration in practice

Excellent illustration and simulation:

`https://distill.pub/2017/momentum/`

Potential issues

- requires tuning of a new hyperparameter (the momentum param)