Optimization for Machine Learning CS-439

Lecture 5: Subgradient and Stochastic Gradient Descent

Martin Jaggi

EPFL – `github.com/epfml/OptML_course`

March 23, 2018

Chapter 4

Subgradient Descent

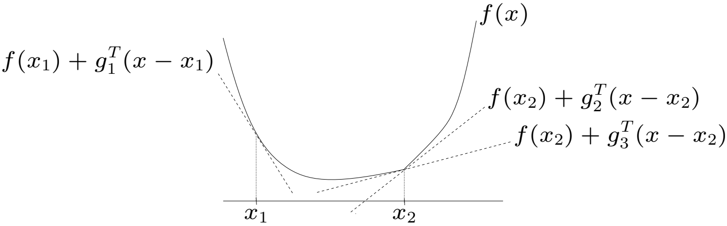
Subgradients

What if f is not differentiable?

Definition

$\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbf{dom}(f)$$



And: $\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$ is the set of subgradients of f at $\mathbf{x}$.

What are subgradients good for?

Convexity

Lemma (Exercise 23)

A function $f : \mathbf{dom}(f) \to \mathbb{R}$ *is convex if and only if* $\mathbf{dom}(f)$ *is convex and* $\partial f(\mathbf{x}) \neq \emptyset$ *for all* $\mathbf{x} \in \mathbf{dom}(f)$.

Lipschitz Continuity

Lemma (Exercise 24)

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be convex,* $B \in \mathbb{R}_+$. *Then the following two statements are equivalent.*

(i) $\|\mathbf{g}\| \leq B$ *for all* $\mathbf{x} \in \mathbb{R}^d$ *and all* $\mathbf{g} \in \partial f(\mathbf{x})$.

(ii) $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

What are subgradients good for?

Subgradient Optimality Condition. Subgradients also allow us to describe cases of optimality for functions which are not necessarily differentiable (and not necessarily convex)

Lemma

Suppose that f is any function over $\mathbf{dom}(f)$, and $\mathbf{x} \in \mathbf{dom}(f)$. If $\mathbf{0} \in \partial f(\mathbf{x})$, then $\mathbf{x}$ is a global minimum.

Proof.

□

The subgradient descent algorithm

An iteration of subgradient descent is defined as

$$
\begin{aligned}
&\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t) \\
&\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \mathbf{g}_t .
\end{aligned}
$$

Bounded subgradients: $\mathcal{O}(1/\varepsilon^2)$ steps

The following result gives the convergence for Subgradient Descent. It is identical to Theorem 2.1, up to relaxing the requirement of differentiability.

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and B-Lipschitz continuous on $\mathbb{R}^d$ with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$. Choosing the constant stepsize

$$\gamma := \frac{R}{B\sqrt{T}},$$

subgradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}.$$

Bounded subgradients: $\mathcal{O}(1/\varepsilon^2)$ steps

Proof.

$\square$

Optimality of first-order methods

With all the convergence rates we have seen so far, a very natural question to ask is if these rates are best possible or not. Surprisingly, the rate can indeed not be improved in general.

Theorem (Nesterov)

For any $T \leq d-1$ and starting point $\mathbf{x}_0$, there is a function f in the problem class of B-Lipschitz functions over $\mathbb{R}^d$, such that any (sub)gradient method has an objective error at least

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \geq \frac{RB}{2(1+\sqrt{T+1})} \ .$$

Chapter 5

Stochastic Gradient Descent

Sum structured objective functions

Consider sum structured objective functions:

$$f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}).$$

Here f_i is typically the cost function of the i-th datapoint, taken from a training set of n elements in total.

The SGD algorithm

An iteration of stochastic gradient descent (SGD) is defined as

sample $i \in [n]$ uniformly at random

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t).$$

The vector $\mathbf{g}_t := \nabla f_i(\mathbf{x}_t)$ is called a stochastic gradient.

Unbiasedness of a stochastic gradient

Why uniform sampling?
In expectation over the random choice of i, $\mathbf{g}_t$ does coincide with the full gradient of f:

$$\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big] = \nabla f(\mathbf{x}_t).$$

- $\mathbf{g}_t$ is an unbiased stochastic gradient.

Why SGD?
n times **cheaper**!

Stochastic vanilla analysis

Idea: follow the vanilla analysis with $\nabla f(\mathbf{x}_t)$ replaced by $\mathbf{g}_t$...

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \overset{\text{NO!!!}}{\leq} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star).$$

but

$$\begin{aligned}
\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) &= \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star). \\
&= \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \\
&= \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right),
\end{aligned}$$

using the definition SGD again. Finally, the telescoping sum:

$$\sum_{t=0}^{T-1} \left(\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)\right) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

Classic GD: For vanilla analysis, we assumed that $\|\nabla f(\mathbf{x})\|^2 \leq B_{\mathsf{GD}}^2$ for all $\mathbf{x} \in \mathbb{R}^d$, where B_{GD} was a constant. So for sum-objective:

$$\Big\|\frac{1}{n}\sum_i \nabla f_i(\mathbf{x})\Big\|^2 \leq B_{\mathsf{GD}}^2 \qquad \forall \mathbf{x}$$

SGD: Assuming same for the expected squared norms of our stochastic gradients, now called B_{SGD}^2.

$$\frac{1}{n}\sum_i \big\|\nabla f_i(\mathbf{x})\big\|^2 \leq B_{\mathsf{SGD}}^2 \qquad \forall \mathbf{x}$$

- get same convergence result, now for expected objective f...

Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mathbf{x}^\star$ a global minimum; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, and that $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$ for all t. Choosing the constant stepsize

$$\gamma := \frac{R}{B\sqrt{T}}$$

stochastic gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[f(\mathbf{x}_t)\big] - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}.$$

Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

Proof. Using convexity and unbiasedness of $\mathbf{g}_t$, we compute

$$\begin{aligned}
\mathbb{E}\big[f(\mathbf{x}_t)\big] - f(\mathbf{x}^\star) &= \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] \\
&\leq \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbf{g}_t\big|\mathbf{x}_t\big]^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big|\mathbf{x}_t\big]\big] \\
&= \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big],
\end{aligned}$$

where the second-to-last step uses linearity of (conditional) expectations, while the last step is known as the tower rule; see Exercise 25.

Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

Now we can again use linearity of expectation and then ( ). We get

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[f(\mathbf{x}_t)\big] - f(\mathbf{x}^\star) \quad &\leq \quad \frac{1}{T}\mathbb{E}\big[\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \\
&= \quad \frac{1}{T}\mathbb{E}\big[\frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\big] \\
&= \quad \frac{1}{T}\left(\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\right) \\
&\leq \quad \frac{RB}{\sqrt{T}},
\end{aligned}
$$

after plugging in our value of γ and the assumption on $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big]$ and $\|\mathbf{x}_0 - \mathbf{x}^\star\|$.

Stochastic Subgradient Descent

For problems which are not necessarily differentiable, we modify SGD to use a subgradient of f_i in each iteration. The update of **stochastic subgradient descent** is given by

sample $i \in [n]$ uniformly at random

let $\mathbf{g}_t \in \partial f_i(\mathbf{x}_t)$

$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t.$

In other words, we are using an unbiased estimate of a subgradient at each step, $\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big] \in \partial f(\mathbf{x}_t)$.

Convergence in $\mathcal{O}(1/\varepsilon^2)$, by using the subgradient property at the beginning of the proof, where convexity was applied.

Constrained optimization

For constrained optimization, Theorem 7 for the convergence in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to constrained problems as well. After every step of SGD, projection back to X is applied as usual. The resulting algorithm is called projected SGD.

Strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

Strengthen the above SGD analysis? Additional assumption of **strong convexity** of the objective f. No constant stepsize γ, but instead use **time-varying stepsize** γ_t decreasing over the time t.

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$; let $\mathbf{x}^\star$ be the unique global minimum of f, and $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \le B^2$ for all $\mathbf{x}$. Choosing the decreasing stepsize

$$\gamma_t := \frac{2}{\mu(t+1)}$$

SGD yields

$$\mathbb{E}\Big[f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\Big) - f(\mathbf{x}^\star)\Big] \le \frac{2B^2}{\mu(T+1)}.$$

Strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

Proof.

$\square$