Optimization for Machine Learning
CS-439

Lecture 5: Subgradient and Stochastic Gradient Descent

Martin Jaggi

EPFL – `github.com/epfml/OptML_course`

March 22, 2019

Chapter 4

Subgradient Descent, continuted

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.

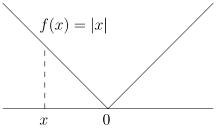
Theorem (Nesterov)

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

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

Smooth (non-differentiable) functions?

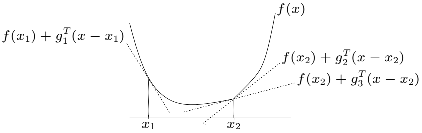
They don't exist (Exercise 26)!



At 0, graph can't be below a tangent paraboloid.

Can we still improve over $O(1/\varepsilon^2)$ steps for Lipschitz functions?

Yes, if we also require strong convexity (graph is above not too flat tangent paraboloids).



Strongly convex functions

"Not too flat"

Straightforward generalization to the non-differentiable case:

Definition

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called strongly convex (with parameter μ) if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f),\ \forall \mathbf{g} \in \partial f(\mathbf{x}).$$

Strongly convex functions: characterization via “normal” convexity

Lemma (Exercise 28)

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ *be convex,* $\mathbf{dom}(f)$ *open,* $\mu \in \mathbb{R}_+, \mu > 0$. f *is strongly convex with parameter* μ *if and only if* $f_\mu : \mathbf{dom}(f) \to \mathbb{R}$ *defined by*

$$f_\mu(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|^2, \quad \mathbf{x} \in \mathbf{dom}(f)$$

is convex.

Tame strong convexity

For fast convergence, we consider additional assumptions.

Smoothness? - Not an option in the non-differentiable case (Exercise 26).

Instead: assume that all subgradients $\mathbf{g}_t$ that we encounter during the algorithm are bounded in norm.

May be realistic if...

- we start close to optimality
- we run projected subgradient descent over a compact set X

May also fail!

- Over $\mathbb{R}^d$, strong convexity and bounded subgradients contradict each other! (Exercise 30).

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

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex with parameter $\mu > 0$ and let $\mathbf{x}^\star$ be the unique global minimum of f. With decreasing step size

$$\gamma_t := \frac{2}{\mu(t+1)}, \quad t > 0,$$

subgradient descent yields

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2B^2}{\mu(T+1)},$$

where $B = \max_{t=1}^{T} \|\mathbf{g}_t\|$.

$\uparrow$

convex combination of iterates

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps II

Proof.

Vanilla analysis ($\mathbf{g}_t \in \partial f(\mathbf{x}_t)$):

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma_t}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma_t}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right).$$

Lower bound from strong convexity:

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Putting it together (with $\|\mathbf{g}_t\|^2 \leq B^2$):

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t^{-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2.$$

Summing over $t = 1, \ldots, T$: we used to have telescoping ($\gamma_t = \gamma, \mu = 0$)...

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps III

Proof.

So far we have:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 .$$

To get telescoping, we would need $\gamma_t^{-1} = \gamma_{t+1}^{-1} - \mu$.

Works with $\gamma_t^{-1} = \mu(1+t)$, but not $\gamma_t^{-1} = \mu(1+t)/2$ (the choice here).

Exercise 31: what happens with $\gamma_t^{-1} = \mu(1+t)$?

Now: what happens with $\gamma_t^{-1} = \mu(1+t)/2$ (the choice here)?

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps IV

Proof.

So far we have:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t^{-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2.$$

Plug in $\gamma_t^{-1} = \mu(1+t)/2$ and multiply with t on both sides:

$$\begin{aligned} t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) &\le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\left(t(t-1)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - (t+1)t\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \\ &\le \frac{B^2}{\mu} + \frac{\mu}{4}\left(t(t-1)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - (t+1)t\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right). \end{aligned}$$

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps V

Proof.

We have

$$
\begin{aligned}
t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) &\leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\left(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\right) \\
&\leq \frac{B^2}{\mu} + \frac{\mu}{4}\left(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\right).
\end{aligned}
$$

Now we get telescoping...

$$
\sum_{t=1}^{T} t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{TB^2}{\mu} + \frac{\mu}{4}\left(0 - T(T+1)\left\|\mathbf{x}_{T+1} - \mathbf{x}^\star\right\|^2\right) \leq \frac{TB^2}{\mu}.
$$

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps VI

Proof.

Almost done:

$$\underline{\sum_{t=1}^{T} t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right)} \leq \frac{TB^2}{\mu} + \frac{\mu}{4}\left(0 - T(T+1)\left\|\mathbf{x}_{T+1} - \mathbf{x}^\star\right\|^2\right) \leq \frac{TB^2}{\mu}.$$

Since

$$\frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1,$$

Jensen's inequality yields

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2}{T(T+1)} \underline{\sum_{t=1}^{T} t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right)}.$$

Tame strong convexity: Discussion

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2B^2}{\mu(T+1)},$$

Weighted average of iterates achieves the bound (later iterates have more weight)

Bound is independent of initial distance $\|\mathbf{x}_0 - \mathbf{x}^\star\|\ldots$

...but not really: B typically depends on $\|\mathbf{x}_0 - \mathbf{x}^\star\|$ (for example, $B = \mathcal{O}(\|\mathbf{x}_0 - \mathbf{x}^\star\|)$ for quadratic functions)

Recall: we can only hope that B is small (can be checked while running the algorithm)

What if we don't know the parameter μ of strong convexity?

$\rightarrow$ **Bad luck!** In practice, try some μ's, pick best solution obtained

Chapter 5

Stochastic Gradient Descent

Stochastic gradient descent

Many objective functions are sum structured:

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

Example: f_i is the cost function of the i-th observation, taken from a training set of n observation.

Evaluating $\nabla f(\mathbf{x})$ of a sum-structured function is expensive (sum of n gradients).

Stochastic gradient descent: the algorithm

choose $\mathbf{x}_0 \in \mathbb{R}^d$.

> sample $i \in [n]$ uniformly at random
> $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t)$.

for **times** $t = 0, 1, \ldots,$ and **stepsizes** $\gamma_t \geq 0$.

Only update with the gradient of f_i instead of the full gradient!

Iteration is n times cheaper than in full gradient descent.

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

$\mathbf{g}_t$ is a vector of d random variables, but we will also simply call this a random variable.

Unbiasedness

Can't use convexity

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$$

on top of the vanilla analysis, as this may hold or not hold, depending on how the stochastic gradient $\mathbf{g}_t$ turns out.

We will show (and exploit): the inequality holds in expectation.

Fot this, we use that by definition, $\mathbf{g}_t$ is an **unbiased estimate** of $\nabla f(\mathbf{x}_t)$:

$$\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t = \mathbf{x}\big] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

The inequality $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ holds in expectation

For any fixed $\mathbf{x}$, linearity of conditional expectations (Exercise 32) yields

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

Event $\{\mathbf{x}_t = \mathbf{x}\}$ can occur only for $\mathbf{x}$ in some finite set X ($\mathbf{x}_t$ is determined by the choices of indices in all iterations so far). Partition Theorem (Exercise 32):

$$\begin{aligned}
\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star)\big] &= \sum_{\mathbf{x}\in X} \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}-\mathbf{x}^\star)\big|\mathbf{x}_t = \mathbf{x}\big]\ \mathrm{prob}(\mathbf{x}_t = \mathbf{x}) \\
&= \sum_{\mathbf{x}\in X} \nabla f(\mathbf{x})^\top(\mathbf{x}-\mathbf{x}^\star)\ \mathrm{prob}(\mathbf{x}_t = \mathbf{x}) = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^\star)\big].
\end{aligned}$$

Hence,

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star)\big] = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^\star)\big] \underset{\text{convexity}}{\geq} \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big].$$

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 $\|\mathbf{x}_0 - \mathbf{x}^\star\| \le R$, and that $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \le B^2$ for all t. Choosing the constant stepsize

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

stochastic gradient descent yields

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

Same procedure as every week... except

- we assume bounded stochastic gradients in expectation;
- error bound holds in expectation.

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

Proof.

Vanilla analysis (this time, $\mathbf{g}_t$ is the stochastic gradient):

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

Taking expectations and using "convexity in expectation":

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

Result follows as every week (optimize γ) . . . $\square$

Convergence rate comparison: SGD vs GD

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:

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

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

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

So by Jensen's inequality for $\|.\|^2$

- $B_{\mathsf{GD}}^2 \approx \left\| \frac{1}{n} \sum_i \nabla f_i(\mathbf{x}) \right\|^2 \leq \frac{1}{n} \sum_i \left\| \nabla f_i(\mathbf{x}) \right\|^2 \approx B_{\mathsf{SGD}}^2$
- B_{GD}^2 can be smaller than B_{SGD}^2, but often comparable.
  Very similar if larger mini-batches are used.