

Optimization for Machine Learning

CS-439

Lecture 6: SGD, Newton's method

Martin Jaggi

EPFL – github.com/epfml/OptML_course

April 13, 2018

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, $\mathbb{E}[\mathbf{g}_t | \mathbf{x}_t] \in \partial f(\mathbf{x}_t).$

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

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

SGD yields

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

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

Proof. Step def., and $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$ gives

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_t - \gamma_t \mathbf{g}_t - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \gamma_t^2 \|\mathbf{g}_t\|^2 - 2\gamma_t \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)\end{aligned}$$

Taking conditional expectation on both sides, and using unbiasedness of the stochastic gradient \mathbf{g}_t , we get

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \mid \mathbf{x}_t] \\ = \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \gamma_t^2 \mathbb{E}[\|\mathbf{g}_t\|^2 \mid \mathbf{x}_t] - 2\gamma_t \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*)\end{aligned}$$

Strong convexity with $\mathbf{y} = \mathbf{x}^*$, $\mathbf{x} = \mathbf{x}_t$ yields

$$\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \geq f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2,$$

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

combining the above two, we have

$$\begin{aligned} & \mathbb{E} \left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \mid \mathbf{x}_t \right] \\ & \leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \gamma_t^2 \mathbb{E} \left[\|\mathbf{g}_t\|^2 \mid \mathbf{x}_t \right] - 2\gamma_t \left(f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 \right) \end{aligned}$$

Rearranging and again taking expectation over the randomness of now the entire sequence of steps $0, 1, \dots, t$, as well as using $\mathbb{E}[\|\mathbf{g}_t\|^2] \leq B^2$, we have

$$\begin{aligned} & 2\gamma_t \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \\ & \leq \gamma_t^2 B^2 + (1 - \mu\gamma_t) \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] - \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2] \end{aligned}$$

$$\begin{aligned} & \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \\ & \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] - \frac{\gamma_t^{-1}}{2} \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2] \end{aligned}$$

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

Now using the stepsize $\gamma_t := \frac{2}{\mu(t+1)}$, and multiplying the above inequality by t on both the sides,

$$\begin{aligned} & t\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \\ & \leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4} \left(t(t-1)\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] - t(t+1)\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2] \right) \\ & \leq \frac{B^2}{\mu} + \frac{\mu}{4} \left(t(t-1)\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] - t(t+1)\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2] \right) \end{aligned}$$

Summing from $t = 1, \dots, T$ and telescoping,

$$\begin{aligned} \sum_{t=1}^T t \cdot \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] & \leq \frac{TB^2}{\mu} + \frac{\mu}{4} \left(0 - T(T+1)\mathbb{E}[\|\mathbf{x}_T - \mathbf{x}^*\|^2] \right) \\ & \leq \frac{TB^2}{\mu}. \end{aligned}$$

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

Finally, using Jensen's inequality (since $\frac{2}{T(T+1)} \sum_{t=1}^T t = 1$):

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

therefore

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



Mini-batch SGD

Instead of using a single element f_i , use an average of several of them:

$$\tilde{\mathbf{g}}_t := \frac{1}{m} \sum_{j=1}^m \mathbf{g}_t^j.$$

Extreme cases:

$m = 1 \Leftrightarrow$ SGD as originally defined

$m = n \Leftrightarrow$ full gradient descent

Benefit: Gradient computation can be naively parallelized

Mini-batch SGD

Variance Intuition: Taking an average of many independent random variables reduces the variance. So for larger size of the mini-batch m , $\tilde{\mathbf{g}}_t$ will be closer to the true gradient, in expectation:

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m} \sum_{j=1}^m \mathbf{g}_t^j - \nabla f(\mathbf{x}_t)\right\|^2\right] \\ &= \frac{1}{m} \mathbb{E}\left[\left\|\mathbf{g}_t^1 - \nabla f(\mathbf{x}_t)\right\|^2\right] \\ &= \frac{1}{m} \mathbb{E}\left[\left\|\mathbf{g}_t^1\right\|^2\right] - \frac{1}{m} \left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{B^2}{m} .\end{aligned}$$

Using a modification of the SGD analysis, can use this quantity to relate convergence rate to the rate of full gradient descent.

Chapter 6

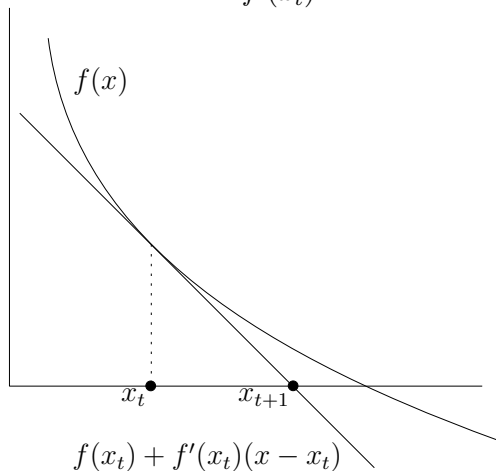
Newton's method

1-dimensional case: Newton-Raphson method

Goal: finding a zero of differentiable $f : \mathbb{R} \rightarrow \mathbb{R}$.

Method:

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \geq 0.$$



Example: Finding the square root

Set $f(x) := x^2 - R$, run Newton-Raphson:

$$x_{t+1} := x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2} \left(x_t + \frac{R}{x_t} \right).$$

Assume we're already close: $x_t - \sqrt{R} < 1/2$ (See Exercise 26).

Then the error goes to 0 **quadratically** (technical: assume $\sqrt{R} \geq 1/2$),

$$x_T - \sqrt{R} \leq \left(x_0 - \sqrt{R} \right)^{2^T} < \left(\frac{1}{2} \right)^{2^T}$$

- ▶ Only $\mathcal{O}(\log \log(1/\varepsilon))$ steps needed!

Newton's method for optimization

1-dimensional case: Find a global minimum x^* of a differentiable convex function $f : \mathbb{R} \rightarrow \mathbb{R}$.

Can equivalently search for a zero of the derivative f' : Apply the Newton-Raphson method to f' . Update step:

$$x_{t+1} := x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - f''(x_t)^{-1} f'(x_t)$$

(needs f twice differentiable)

d -dimensional case: Newton's method for minimizing a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$$

Convergence in one step on quadratics

Lemma

On (nondegenerate) quadratics, with any starting point $\mathbf{x}_0 \in \mathbb{R}^d$, Newton's method yields $\mathbf{x}_1 = \mathbf{x}^*$.

A **nondegenerate** quadratic function is a function of the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top M\mathbf{x} - \mathbf{q}^\top \mathbf{x} + c,$$

where $M \in \mathbb{R}^{d \times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d$, $c \in \mathbb{R}$. Here let $\mathbf{x}^* = M^{-1}\mathbf{q}$ be the unique solution of $\nabla f(\mathbf{x}) = \mathbf{0}$.

Proof.

We have $\nabla f(\mathbf{x}) = M\mathbf{x} - \mathbf{q}$ (this implies $\mathbf{x}^* = M^{-1}\mathbf{q}$) and $\nabla^2 f(\mathbf{x}) = M$. Hence,

$$\mathbf{x}_0 - \nabla^2 f(\mathbf{x}_0)^{-1} \nabla f(\mathbf{x}_0) = \mathbf{x}_0 - M^{-1}(M\mathbf{x}_0 - \mathbf{q}) = M^{-1}\mathbf{q} = \mathbf{x}^*.$$

