annotated version

# 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

$f_1(x) + f_2(x) \ldots + f_n(x)$

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

$$
\begin{aligned}
&\text{sample } i \in [n] \text{ uniformly at random} \\
&\text{let } \mathbf{g}_t \in \partial f_i(\mathbf{x}_t) \\
&\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t .
\end{aligned}
$$

$\partial f_1(x) \qquad \partial f_n(x)$

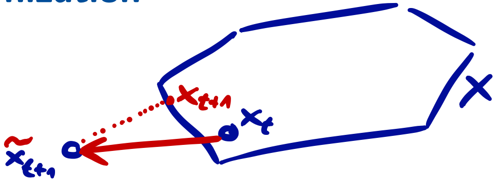
$\in \qquad \in$

$g_1 + g_2 \ldots + g_n$

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, 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 $\gamma$, but instead use **time-varying stepsize** $\gamma_t$ 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 $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le B^2$ 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.** Step def., and $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ gives

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

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

$$\begin{aligned} &\mathbb{E}\big[\, \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \,\big|\, \mathbf{x}_t \big] \\ &= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \gamma_t^2 \underbrace{\mathbb{E}\big[\, \|\mathbf{g}_t\|^2 \,\big|\, \mathbf{x}_t \big]}_{B^2} - 2\gamma_t \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star) \end{aligned}$$

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

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

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

combining the above two, we have

$$\mathbb{E}\Big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \,\Big|\, \mathbf{x}_t\Big]$$

$$\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \gamma_t^2 \mathbb{E}\Big[\,\|\mathbf{g}_t\|^2 \,\Big|\, \mathbf{x}_t\Big] - 2\gamma_t\Big(f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2\Big)$$

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

$$2\gamma_t \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)]$$
$$\leq \gamma_t^2 B^2 + (1 - \mu\gamma_t)\mathbb{E}\big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\big] - \mathbb{E}\big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\big]$$

$$\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)]$$
$$\leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\mathbb{E}\big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\big] - \frac{\gamma_t^{-1}}{2}\mathbb{E}\big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\big]$$

# 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,

$$
\begin{aligned}
&t\mathbb{E}\big[f(\mathbf{x}_t)-f(\mathbf{x}^\star)\big] \\
&\quad\le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big(t(t-1)\mathbb{E}\big[\,\|\mathbf{x}_t-\mathbf{x}^\star\|^2\,\big] - t(t+1)\mathbb{E}\big[\,\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\,\big]\Big) \\
&\quad\le \frac{B^2}{\mu} + \frac{\mu}{4}\Big(t(t-1)\mathbb{E}\big[\,\|\mathbf{x}_t-\mathbf{x}^\star\|^2\,\big] - t(t+1)\mathbb{E}\big[\,\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\,\big]\Big)
\end{aligned}
$$

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

$$
\sum_{t=1}^{T} t\cdot\mathbb{E}\big[f(\mathbf{x}_t)-f(\mathbf{x}^\star)\big] \le \frac{TB^2}{\mu} + \frac{\mu}{4}\Big(0 - \cancel{T(T+1)\mathbb{E}\big[\,\|\mathbf{x}_{T+1}-\mathbf{x}^\star\|^2\,\big]}\Big)
$$

$$
\le \frac{TB^2}{\mu}.
$$

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

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

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

therefore

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

$\square$

# Mini-batch SGD

min $f = \frac{1}{n} \sum f_i$

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

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

stoch. grad at $x_t$

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:

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

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

# Chapter 6

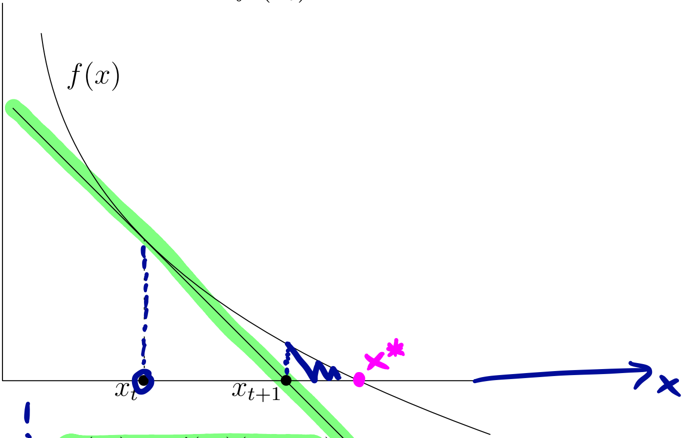
## Newton's method

# 1-dimensional case: Newton-Raphson method

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

Method:

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




$$0 \stackrel{!}{=} f(x_t) + f'(x_t)(x - x_t)$$

## Example: Finding the square root

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

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

f

f'

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 $\mathcal{O}\big(\log\log(1/\varepsilon)\big)$ steps needed!

Proof:

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

# Newton's method for ~~convex~~ optimization

$\min_x f(x)$

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

find $x$ s.t. $\nabla f(x) = 0$

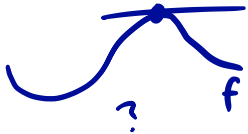
Can equivalently search for a zero of the derivative $f'$: Apply the Newton-Raphson method to $f'$. Update step:

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

(needs $f$ twice differentiable)

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

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



# Newton's method for convex optimization

solves quadratics in one step!

## Lemma

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

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, $\mathbf{q} \in \mathbb{R}^d, c \in R$. Here let $\mathbf{x}^\star = M^{-1}\mathbf{q}$ 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 $\mathbf{x}^\star = M^{-1}\mathbf{q}$) and $\nabla^2 f(\mathbf{x}) = M$. Hence,

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

□