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

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 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. 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

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

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 $0,1,\dots,t$, as well as using $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big]\leq B^2$, we have

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

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

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,

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

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

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

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

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

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

therefore

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

$\square$

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:

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

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

Chapter 6

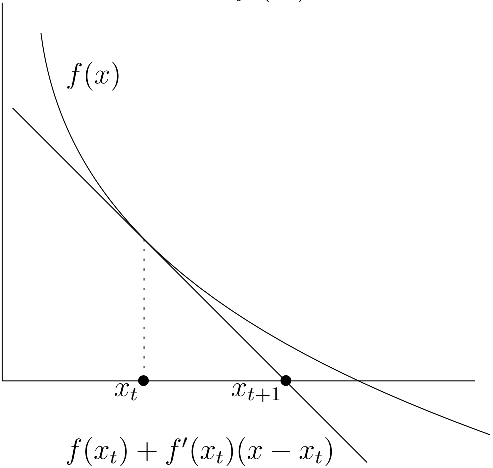
Newton's method

1-dimensional case: Newton-Raphson method

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

Method:

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



Newton's method for convex optimization

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

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 $f : \mathbb{R}^d \to \mathbb{R}$:

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

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 $\sqrt{R} \geq 1/2$),

$$x_T - \sqrt{R} \leq \left(x_0 - \sqrt{R}\right)^{2^T} < \left(\frac{1}{2}\right)^{2^T}, \quad T \geq 0.$$

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

Newton's method for convex optimization

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,

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

□

Affine Invariance

Newton's method is **affine invariant**
(invariant under any invertible affine transformation):

Lemma (Exercise 27)

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be twice differentiable,* $A \in \mathbb{R}^{d \times d}$ *an invertible matrix,* $\mathbf{b} \in \mathbb{R}^d$*. Let* $g : \mathbb{R}^d \to \mathbb{R}$ *be the (bijective) affine function* $g(\mathbf{y}) = A\mathbf{y} + \mathbf{b}, \mathbf{y} \in \mathbb{R}^d$*. Finally, let* $N_h : \mathbb{R}^d \to \mathbb{R}^d$ *denote the Newton step for function* h*, i.e.*

$$N_h(\mathbf{x}) := \mathbf{x} - \nabla^2 h(\mathbf{x})^{-1} \nabla h(\mathbf{x}),$$

whenever this is defined. Then we have $N_{f \circ g} = g^{-1} \circ N_f \circ g$*.*

Affine Invariance

Newton step for $f \circ g$ on $\mathbf{y}_t$: can transform $\mathbf{y}_t$ to $\mathbf{x}_t = g(\mathbf{y}_t)$, perform the Newton step for f on $\mathbf{x}$ and transform the result $\mathbf{x}_{t+1}$ back to $\mathbf{y}_{t+1} = g^{-1}(\mathbf{x}_{t+1})$. I.e., the following diagram commutes:

$$\begin{tikzcd}
\mathbf{x}_t \arrow[r, "N_f"'] & \mathbf{x}_{t+1} \arrow[d, "g^{-1}"'] \\
\mathbf{y}_t \arrow[u, "g"'] \arrow[r, "N_{f\circ g}"] & \mathbf{y}_{t+1}
\end{tikzcd}$$

Hence, while gradient descent suffers if the coordinates are at very different scales, Newton's method doesn't.

Affine Invariance

Invariance to scaling of the input problem

Minimizing the second-order Taylor approximation

Alternative interpretation of Newton's method:
Each step minimizes the local second-order Taylor approximation.

Lemma (Exercise 30)

Let f be convex and twice differentiable at $\mathbf{x}_t \in \mathbf{dom}(f)$, with $\nabla^2 f(\mathbf{x}_t) \succ 0$ being invertible. The vector $\mathbf{x}_{t+1}$ resulting from the Netwon step satisfies

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\operatorname{argmin}}\ f(\mathbf{x}_t)+\nabla f(\mathbf{x}_t)^\top(\mathbf{x}-\mathbf{x}_t)+\frac{1}{2}(\mathbf{x}-\mathbf{x}_t)^\top\nabla^2 f(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t).$$

Once you're close, you're there...

Theorem

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex with a unique global minimum $\mathbf{x}^\star$. Suppose there is an open ball $X \subseteq \mathbf{dom}(f)$ with center $\mathbf{x}^\star$, s.t.

(i) *Bounded inverse Hessians: There exists a real number $\mu > 0$ such that*
$$\|\nabla^2 f(\mathbf{x})^{-1}\| \leq \frac{1}{\mu}, \quad \forall \mathbf{x} \in X.$$

(ii) *Lipschitz continuous Hessians: There exists a real number $L > 0$ such that*
$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Matrix norm is spectral norm. Note: (i) $\Rightarrow$ Hessian invertible at all $\mathbf{x} \in X$.

Then, for $\mathbf{x}_t \in X$ and $\mathbf{x}_{t+1}$ resulting from the Newton step, we have
$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{L}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Super-exponentially fast?

Starting close to the global minimum, we will reach distance at most ε to the minimum within $\mathcal{O}\big(\log\log(1/\varepsilon)\big)$ steps.

Corollary (Exercise 28)

With the assumptions and terminology of the above theorem, and if

$$\|\mathbf{x}_0 - \mathbf{x}^\star\| < \frac{\mu}{L},$$

then Newton's method yields

$$\|\mathbf{x}_T - \mathbf{x}^\star\| < \frac{2\mu}{L}\left(\frac{1}{2}\right)^{2^T}, \quad T \geq 0.$$