- Info: Tutorial 1 Monday (zoom)
Tutorial 2 Tuesday 3pm

Optimization for Machine Learning

Lecture 3: Stochastic Gradient Descent

Sebastian Stich

CISPA – `https://cms.cispa.saarland/optml24/`

April 30, 2024

Quiz Week 2 (1)

$O(\frac{1}{\varepsilon})$ $o(\frac{1}{T})$

What does $f(n) \in \mathcal{O}(g(n))$ (for $n \to \infty$) mean?

$O(\,)$, $\Omega(\cdot)$, $\Theta(\cdot)$

Examples:

- $10n^2 \in \mathcal{O}(n^2)$? ✓
- $n^2 \in \mathcal{O}(n^3)$? ✓ $\Longleftarrow$
- $n^3 + n^2 + n + 1 \in \mathcal{O}(n^3)$? ✓

$f(n) \in O(g(n))$

" $\lim_{n \to \infty} \frac{f(n)}{g(n)} \to 0$ "

"big-O notation"

Formally:

What about $\epsilon \to 0$?

- Consider $n = \frac{1}{\epsilon}$.
- $\frac{1}{\epsilon^2} \in \mathcal{O}\left(\frac{1}{\epsilon^3}\right)$?

"

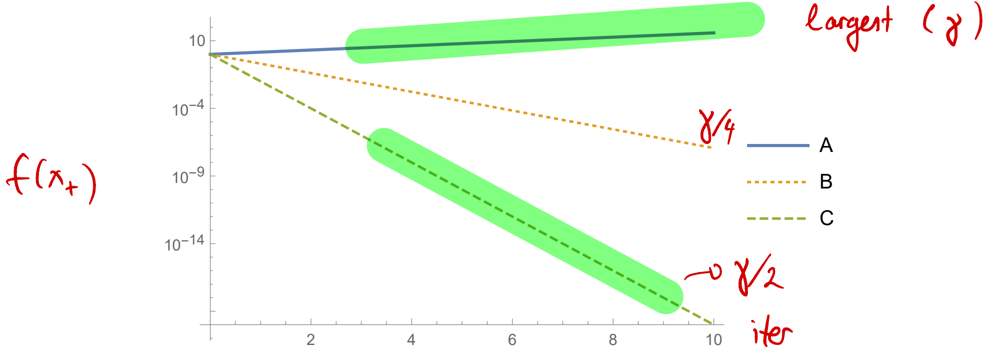
$\frac{1}{\varepsilon^2} = n^2 \in O(n^3) = O\left(\frac{1}{\varepsilon^3}\right)$

Quiz Week 2 (2)

Consider gradient descent on a smooth and convex function $f\colon \mathbb{R}^d \to \mathbb{R}$,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)\,,$$

for a stepsize $\gamma > 0$.



The figure shows three runs of gradient descent, with the stepsizes $\{\gamma, \gamma/2, \gamma/4\}$, for a (fixed) value of γ. Which curve does correspond to which stepsize?

Chapter 6

Stochastic Gradient Descent

Stochastic gradient descent

Example: $\text{loss}(NN(\text{image}_i), \text{label}_i)$

$\| NN(\text{image}_i) - \text{label}_i \|^2$

Many objective functions are sum structured:

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

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).

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

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 **iterations** $t = 0, 1, \ldots,$ 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.

Stochastic Optimization

The finite sum structure is not necessary. All results we discuss in this course do also hold for stochastic optimization problems:

$$f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(\mathbf{x}, \xi)]$$

- $\mathcal{D}$ a distribution "real world data"
- for every ξ, access to stochastic gradients $\nabla F(\mathbf{x}, \xi)$
- finite-sum is a special case:

$$\mathcal{D} = \underbrace{\{1, \ldots, n\}}_{n \text{ events} \to \frac{1}{n} \text{ probability}}, \quad f(x) = E_{\xi \sim \mathcal{D}} F(x, \xi) = \sum_{i=1}^{n} \underbrace{\frac{1}{n}}_{\text{prob.}} \cdot f_i(x)$$

- algorithm:

> sample $\xi_t \sim \mathcal{D}$ uniformly at random
>
> $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla F(\mathbf{x}_t, \xi_t).$

Unbiasedness

Consider a stochastic gradient $\mathbf{g}_t$, for a random index $i_t \in [n]$.

$$\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t)\,,$$

We cannot use our previous inequalities as they might not hold, depending on how the stochastic gradient $\mathbf{g}_t$ turns out.

We will show (and exploit): many inequalities holds in expectation.

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

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

Convexity in expectation

Note, for any fixed vector $\mathbf{y} \in \mathbb{R}^d$:

$$\mathbb{E}[\mathbf{g}_t^\top \mathbf{y}] = \mathbb{E}[\mathbf{g}_t]^\top \mathbf{y} = \nabla f(\mathbf{x}_t)^\top \mathbf{y} \,.$$

Hence, for a convex function $f \colon \mathbb{R}^d \to \mathbb{R}$:

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

Quadratic upper bound with stochastic updates?

Can we also use expectation with the quadratic upper bound?

Recall, a step of SGD: $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathbf{g}_t$.

$$
\begin{aligned}
& \mathbb{E}\left[f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] \\
&= \mathbb{E}\left[f(\mathbf{x}_t) + \underbrace{\nabla f(\mathbf{x}_t)^\top (-\gamma \mathbf{g}_t)}_{\text{linear}} + \frac{L}{2}\|-\gamma \mathbf{g}_t\|^2\right] \\
&= \quad f(\mathbf{x}_t) - \gamma \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) + \underbrace{\frac{\gamma^2 L}{2} \mathbb{E}\left[\|\mathbf{g}_t\|^2\right]}_{??}
\end{aligned}
$$

What is $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right]$? We need one more assumption!

Case 1: **Bounded Gradients**

Bounded Gradient Assumption

Assume that there exists a constant $B \geq 0$, such that:

$$\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \leq B^2$$

for all t.

+ This simplifies the proofs to a certain degree, while still comprehensively addressing most of the additional complexity presented by stochastic gradients..
- Might not hold. (Example: quadratic functions)

$f(x) = \frac{1}{2}x^2 \qquad \nabla f(x) = x$

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

Theorem (Lecture-3).1

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 $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le B^2$ for all t. Choosing the constant stepsize

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

stochastic gradient descent yields

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

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

Proof I

$x_{t+1} = x_t - \gamma g_t$

$\|a-b\|^2 = \|a\|^2 - 2ab + \|b\|^2$

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

$E\, g_t = \nabla f(x_t)$ $\qquad$ $E\|g_t\|^2 \le B^2$

$$
\le \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\gamma \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) + \gamma^2 B^2
$$

convexity

$$
\le \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\gamma(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) + \gamma^2 B^2 \tag{1}
$$

Proof II

We re-arrange and prepare to apply the telescoping sum trick:

$$2\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right]}{\gamma} + \gamma B^2$$

This does not seem to work! However, we can take also take expectation over $\mathbf{x}_t$:

$$2\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right] \leq \frac{\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2}{\gamma} + \gamma B^2$$

Note: this argument can be made more rigorous. See lecture notes or other sources for details.

Proof III

By telescoping (and dividing by T):

best γ!

$$\frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}f(\mathbf{x}_t)-f(\mathbf{x}^\star)\leq\frac{\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\gamma T}+\gamma B^2\leq\frac{R^2}{\gamma T}+\gamma B^2$$

$+1$

$$\frac{R^2 B\sqrt{T}}{R\,T}+\frac{R B^2}{B\sqrt{T}}=\frac{2RB}{\sqrt{T}}\ \checkmark$$

We now observe that the choice $\gamma=\frac{R}{B\sqrt{T}}$ indeed implies the theorem.

$$\min_\gamma \frac{R^2}{\gamma T}+\gamma B^2 \quad \xrightarrow{\text{derivative}} \quad -\frac{R^2}{\gamma^2 T}+B^2\overset{!}{=}0 \Rightarrow \gamma=\frac{R}{B\sqrt{T}}$$

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

Theorem (Lecture-3).2

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 assume that* $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \leq B^2$ *for all* t*. With decreasing step size*

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

stochastic gradient descent yields

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

- weighted averaging puts more importance on recent iterates!

Proof I

The proof is starting in the same way. Except that we can use strong convexity:

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

Equation (1) will change into:

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

And therefore

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

Proof II

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

$$t \cdot \mathbb{E}\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big(t(t-1)\mathbb{E}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big)$$

$$\leq \frac{B^2}{\mu} + \frac{\mu}{4}\Big(t(t-1)\mathbb{E}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big).$$

≤ 1

Now we get telescoping...

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

Finally, use $\frac{2}{T(T+1)}\sum_{t=1}^{T} t = 1$, and Jensen's inequality.

Discussion

$O\left(\frac{1}{T}\right)$

$O\left(\frac{1}{\sqrt{T}}\right) < \varepsilon \to T \geq \frac{1}{\varepsilon^2}$

- strong convexity helps: $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ convergence, vs. $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ ← convex

  $\varepsilon = 0.01$ $\approx \frac{1}{\varepsilon} \approx 100$ steps $\frac{1}{\varepsilon^2} \approx 10'000$ steps

- stochastic gradients make the convergence more difficult: $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ convergence vs. $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ in the deterministic setting for gradient descent! (recall Exercise Sheet 2)

  $\|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^{t+1} \cdot \|x_0 - x^*\|^2$

  $\log \frac{1}{\varepsilon} \approx 3$

- Note: The $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ convergence is optimal!
- Weighted averaging is a common & useful trick to adapt telescoping sum proofs ot the strongly-convex case!

Case 2: **Bounded Variance**

Bounded Variance Assumption

$(\mathbb{E}\|g_t\|^2 \leq B^2)$

Assume that there exists a constant $\sigma \geq 0$, such that:

$$\mathbb{E}\left[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2\right] \leq \sigma^2$$

for all t.

+ Standard and widely-accepted model in complexity theory.
- Might not hold on all (but much fewer) problems of interest.
- (Convergence proof: we will cover some examples next week—you could try yourself as an exercise!)

Mini-batch SGD

Mini-batch SGD

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

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

where $\mathbf{g}_t^j$ denotes a stochastic gradient drawn uniformly and independently at random.
m denotes the **batch size**.

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

mini-batch size m

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:

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

$$\underbrace{E\left\langle \frac{1}{m}\sum_{j=2}^{m} g_t^j - \frac{1}{m}\nabla f(x_t), \frac{1}{m} g_t^1 - \frac{1}{m}\nabla f(x_t)\right\rangle}_{0} + \underbrace{E\left\|\frac{1}{m} g_t^1 - \frac{1}{m}\nabla f(x_t)\right\|^2 + E\left\|\frac{1}{m}\sum_{j=2}^{m} g_t^j - \frac{1}{m}\nabla f(x_t)\right\|^2}_{m \cdot E\left\|\frac{1}{m} g_t^1 - \frac{1}{m}\nabla f(x_t)\right\|^2}$$

$$= \frac{1}{m}\mathbb{E}\left[\left\|\mathbf{g}_t^1 - \nabla f(\mathbf{x}_t)\right\|^2\right] \leq \frac{\sigma^2}{m}.$$

- variance reduction by a factor of at least m

Lecture 3 Recap

- SGD: the most important building block in ML/DL optimization!
  - low per-iteration cost
  - ideal if low-accuracy approximations suffice (say, $\epsilon \geq 0.01$)
- SGD convergence proof under the bounded gradient assumption
  - we will discuss next week a proof with the bounded variance assumption
- variance-reduction effect of mini-batches
- weighted averaging to make telescoping work

Discussion

Discussion

Discussion