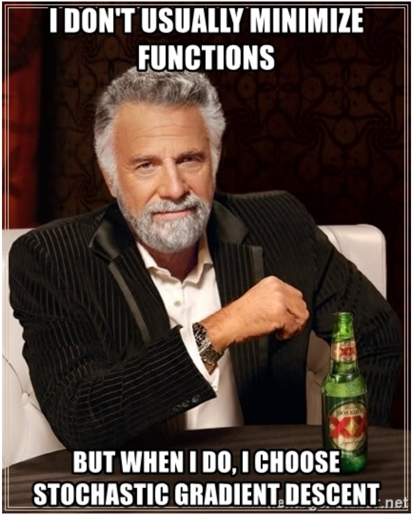
Stochastic Gradient Descent

Axel Böhm

January 24, 2022



I DON'T USUALLY MINIMIZE FUNCTIONS
BUT WHEN I DO, I CHOOSE STOCHASTIC GRADIENT DESCENT

Finite sum structure

Many optimization problems in Data science are sum structured:

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

- known as empirical risk (minimization)
- f_i corresponds to the loss of the i-th observation
- for example: linear regression

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{n} (a_i^T x - b_i)^2$$

- evaluating ∇f can be **expensive** if n is large

Risk minimization

In theory we would even like to minimize the population risk

$$f(x) = \mathbb{E}_{\xi}[f(x, \xi)]$$

- Typically **no access** to f
- most of what follows works in this more general setting

(vanilla) Stochastic gradient descent

sample $i \in 1, \ldots, n$ uniformly at random

$$x_{k+1} = x_k - \alpha \nabla f_i(x_k).$$

- requires only **one** gradient instead of n per iteration.
- we call $g_k := \nabla f_i(x_k)$ a stochastic gradient (estimator)

Unbiased

- Can't really use convexity as before since

$$f(x_k) - f(x^*) \le \langle \nabla f_i(x_k), x^* - x_k \rangle$$

  might not hold in general.
- But holds **in expectation**!
- For this we need that $\nabla f_i(x)$ is unbiased estimator of $\nabla f(x)$

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

Gradient inequality holds in expectation

- We would like to conclude that

$$\mathbb{E}\left[\langle g_k, x^* - x_k\rangle\right] = \langle \mathbb{E}[g_k], \mathbb{E}[x^* - x_k]\rangle$$

but this is not so clear since x_k is also stochastic and in general $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$.

- We use the **conditional Expectation** $\mathbb{E}[\cdot|x_k]$ (read as expectation of $\cdot$ given x_k). Then

$$\mathbb{E}\left[\langle g_k, x^* - x_k\rangle | x_k\right] = \langle \mathbb{E}[g_k|x_k], x^* - x_k\rangle = \langle \nabla f(x_k), x^* - x_k\rangle.$$

- Together with the tower property $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$:

$$\begin{aligned}\mathbb{E}\left[\langle g_k, x^* - x_k\rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle g_k, x^* - x_k\rangle | x_k\right]\right] \\ &= \mathbb{E}\left[\langle \nabla f(x_k), x^* - x_k\rangle\right] \leq \mathbb{E}[f(x^*) - f(x_k)].\end{aligned}$$

Convergence statement: $\mathcal{O}(\epsilon^{-2})$ steps

Assumptions

- f is convex and differentiable
- $\|x_0 - x^*\| \leq D$
- stochastic gradient are bounded in expectation $\mathbb{E}[\|g_k\|^2] \leq B^2$.

Theorem

With the assumptions above and stepsize

$$\alpha = \frac{D}{B\sqrt{k}}$$

yields

$$\mathbb{E}\left[f(\bar{x}_k) - f^*\right] \leq \frac{DB}{\sqrt{k}}.$$

Proof

Proof.

We start as usual (g_k is a stochastic gradient)

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - \alpha g_k - x^*\|^2 \\ &= \|x_k - x^*\|^2 + 2\alpha\langle g_k, x^* - x_k\rangle + \alpha^2\|g_k\|^2.\end{aligned}$$

Now take expectation

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] \leq \mathbb{E}\left[\|x_k - x^*\|^2\right] + 2\alpha\mathbb{E}[f^* - f(x_k)] + \alpha^2\mathbb{E}[\|g_k\|^2].$$

Bound gradients and telescope to finish the proof. □

Comparing constants: SGD vs. GD

- GD: In the bounded (sub-)gradient analysis we assumed $\|\nabla f(x)\|^2 \leq B_{BG}^2$. For finite-sum this gives

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) \right\|^2 \leq B_{BG}^2$$

- SGD: We assumed that the expected squared norm are bounded, i.e.

$$\mathbb{E}[\|\nabla f_i(x)\|^2] = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x)\|^2 \leq B_{SGD}^2$$

By convexity we have that

- $B_{GD}^2 \approx \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) \right\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x)\|^2 \approx B_{SGD}^2$
- but usually comparable

Minibatch SGD

Instead of just using a single element f_i we can use several $S \subset \{1, \ldots, n\}$

$$g_k := \frac{1}{|S|} \sum_{j \in S} \nabla f_j(x_k)$$

Interpolates between

- $|S| = 1 \Leftrightarrow$ (vanilla) SGD, as defined earlier
- $|S| = n \Leftrightarrow$ (batch) GD

Benefit: Gradient computation can parallelized.

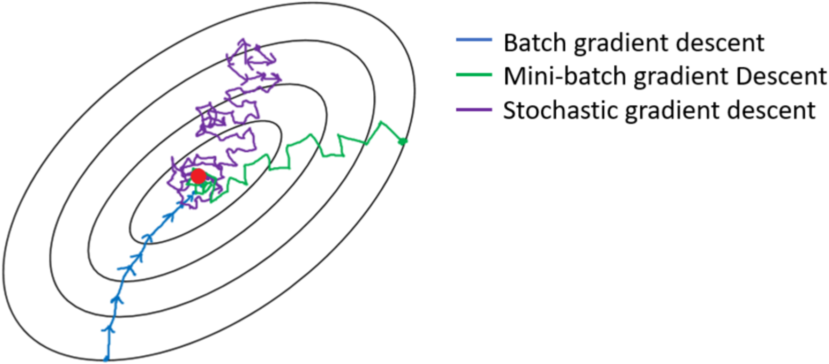
Increasing batch size reduces variance

Taking an average of independent random variables will reduce variance.

$$\mathbb{V}[g_k] = \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{|S|}\sum_{j\in S} f_j(x_k) - \nabla f(x_k)\right\|^2\right]$$
$$= \frac{1}{|S|}\mathbb{E}\left[\|\nabla f_i(x_k) - \nabla f(x_k)\|^2\right]$$
$$= \frac{1}{|S|}\mathbb{V}[\nabla f_i(x_k)]$$

However: We have to use a different analysis to make use of this.

Minibatch illustration



Stochastic Subgradient Method

If we go back to the proof: We did not use smoothness. If we choose unbiased estimate of subgradient $\mathbb{E}[g_k|x_k] \in \partial f(x_k)$ and iterate

$$
\begin{aligned}
&\text{sample } i \in 1, \ldots, n \text{ uniformly at random} \\
&\text{let } g_k \in \partial f_i(x_k) \\
&x_{k+1} = x_k - \alpha g_k.
\end{aligned}
$$

We can get the same $\mathcal{O}(\epsilon^{-2})$ complexity. Smoothness did not provide any benefit (in terms of rate). With a more refined analysis one can get different dependence on constants.

Projected SGD

- Previous proof can be extended (trivially) to the constrained setting
- with same complexity $\mathcal{O}(\epsilon^{-2})$
- but (of course) with an additional projection

$$x_{k+1} = \Pi_C(x_k - \alpha \nabla f_i(x_k))$$

High probability bounds

Theorem

Hoeffding's inequality Let X_i be independent random variables that satisfy

- $\mathbb{E}X_i = 0$
- $\|X_i\| \leq M$.

Then,

$$\mathbb{P}\left[X_1 + \cdots + X_k \geq t\right] \leq e^{-\frac{t^2}{2kM^2}}$$

Azuma-Hoeffding's generalization does not require independence, only $\mathbb{E}[X_k | X_{k-1}, \ldots, X_1] = 0$.

Statement with high probability

Theorem

Let $\delta > 0$ and assumptions as before + iterates remain in bounded set with diameter D (for example constraint set). Then,

$$f(\bar{x}_k) - f^* \leq \frac{DB\delta}{\sqrt{k}}.$$

with probability more than $1 - e^{-\delta^2/8}$.

$\Rightarrow$ choose δ large for bound in higher probability.

Proof.

$$
\begin{aligned}
&\|x_{k+1} - x^*\|^2 \\
&\quad \leq \|x_k - x^*\|^2 + 2\alpha\langle g_k, x^* - x_k\rangle + \alpha^2\|g_k\|^2 \\
&\quad \leq \|x_k - x^*\|^2 + 2\alpha\langle \nabla f(x_k), x^* - x_k\rangle + \alpha^2\|g_k\|^2 + \langle v_k, x^* - x_k\rangle
\end{aligned}
$$

with $v_k = g_k - \nabla f(x_k)$. Continue as usual

$$
f(\bar{x}_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\alpha k} + \alpha B^2 + \frac{1}{k}\sum_{i=1}^{k} X_i
$$

with

$$
X_k := \langle v_k, x^* - x_k\rangle \leq \|v_k\|\|x_k - x^*\| \leq 2BD
$$

and $\mathbb{E}[X_k] = 0$ fulfilling Hoeffding's assumptions. Use it with $t = DB\sqrt{k}\delta$. □