@bibtex Files

Summary

Suppose, our target function is the sum of functions.

$$\min_{\theta \in \mathbb{R}^p} g(\theta) := \frac{1}{n} \sum_{i=1}^n f_i(\theta)$$

This problem usually arises in Deep Learning, where the gradient of the loss function is calculating over the huge number of data points, which could be very expensive in terms of the iteration cost (calculation of gradient is linear in n).

Thus, we can switch from the full gradient calculation to its unbiased estimator:

$$\theta_{k+1} = \theta_k - \alpha_k \nabla f_{i_k}(\theta),$$

where we randomly choose i_k index of point at each iteration uniformly:

$$\mathbb{E}[\nabla f_{i_k}(\theta)] = \sum_{i=1}^n p(i_k = i) \nabla f_i(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\theta) = \nabla g(\theta)$$

Iterations could be n times cheaper! But convergence requires $\alpha_k \to 0$.

Convergence

General setup

We consider classic finite-sample average minimization:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

Let us consider stochastic gradient descent assuming ∇f is Lipschitz:

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) \tag{SGD}$$

Lipschitz continiity implies:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

using (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2$$

Now let's take expectation with respect to i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2]$$

Using linearity of expectation:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \mathbb{E}[\nabla f_{i_k}(x_k)] \rangle + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Since uniform sampling implies unbiased estimate of gradient: $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$
:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Polyak-Lojasiewicz conditions

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*), \forall x \in \mathbb{R}^p \tag{PL}$$

This inequality simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. Note, that strong convexity implies PL, but not vice versa. Using PL we can write:

$$\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

This bound already indicates, that we have something like linear convergence if far from solution and gradients are similar, but no progress if close to solution or have high variance in gradients at the same time.

Stochastic subgradient descent

$$x_{k+1} = x_k - \alpha_k g_{i_k} \tag{SSD}$$

for some $g_{i_k} \in \partial f_{i_k}(x_k)$.

For convex f we have

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] = \|x_k - x^*\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle + \alpha_k^2 \mathbb{E}[\|g_{i_k}\|^2]$$

Here we can see, that step-size α_k controls how fast we move towards solution. And squared step-size α_k^2 controls how much variance moves us away. Usually, we bound $\mathbb{E}[|g_{i_k}|^2]$ by some constant B^2.

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] = \|x_k - x^*\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle + \alpha_k^2 B^2$$

If we also have strong convexity:

$$\mathbb{E}[\|x_k - x^*\|^2] \leq (1 - 2\alpha_k \mu)\|x_{k-1} - x^*\|^2 + \alpha_k^2 B^2$$

And finally, with $\alpha_k = \alpha < \frac{2}{\mu}$:

$$\mathbb{E}[\|x_k - x^*\|^2] \leq (1 - 2\alpha_k \mu)^k R^2 + \frac{\alpha B^2}{2\mu},$$

where $R = |x_0 - x^*|$

Bounds

Conditions	$\|\mathbb{E}[f(x_k)] - f(x^*)\| \leq$	Type of convergence
Convex, Lipschitz-continuous gradient (L)	$\mathcal{O}\left(\dfrac{1}{\sqrt{k}}\right)$	Sublinear
μ-Strongly convex, Lipschitz-continuous gradient (L)	$\mathcal{O}\left(\dfrac{1}{k}\right)$	Sublinear
Convex, non-smooth	$\mathcal{O}\left(\dfrac{1}{\sqrt{k}}\right)$	Sublinear
μ-Strongly convex, non-smooth	$\mathcal{O}\left(\dfrac{1}{k}\right)$	Sublinear

Code

Open in Colab

References

- Lecture by Mark Schmidt @ University of British Columbia
- Convergence theorems on major cases of GD, SGD (projected version included)

https://colab.research.google.com/github/MerkulovDaniil/sber219/blob/main/notebooks/9_01.ipynb#scrollTo=UUJA6iqI7MLu