Modern Optimization

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

1. Stochastic gradient descent

Outline

Definition 2.1

We define the **sum structured objective functions** *as an objective function which is separable:*

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

Example 2.1

Finding optimal coefficients from a multiple linear regression problem is an example of such setup.

Definition 2.2

Stochastic gradient descent *is an algorithm iterating the following sequence*

$$\text{sample } i \text{ uniformly at random } in\{1, \cdots, n\}$$
$$x^{k+1} = x^k - \gamma_k \nabla f_i(x^k)$$

Remark 2.1

It is trivial to see that the classical gradient descent update would read

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

Consequently the update is roughly n times cheaper in SGD than in classical (batch) gradient descent.

Proposition 2.1

Let $g_k := \nabla f_i(x^k)$ *with* i *sampled uniformly at random in* $\{1, \cdots, n\}$ *be the (stochastic) gradient at iteration* k *of a convex function* $f = \frac{1}{n} \sum f_i$*. Then* g_k *is an unbiased estimator of the gradient of* f*, namely*

$$\mathbb{E}\left[g_k^T(x - x^*) | x^k = x\right] = \nabla f(x)^T(x - x^*).$$

Moreover, it holds

$$\mathbb{E}\left[g_k^T(x^k - x^*)\right] \geq \mathbb{E}\left[f(x^k) - f(x^*)\right].$$

Theorem 2.1

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be a convex differentiable function such that* $f = \frac{1}{n} \sum f_i$*. Assume* f *has a minimizer* x^* *and suppose that there exists a* $B > 0$ *such that* $\mathbb{E}\left[\|g_k\|^2\right] \leq B^2$*. If the stepsize is chosen constant and such that*

$$\gamma_k = \gamma = \frac{\|x^0 - x^*\|}{B\sqrt{K}}$$

then the iterates generated by the stochastic gradient descent satisfy

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[f(x^k)\right] - f(x^*) \leq \frac{\|x^0 - x^*\|_2 B}{\sqrt{K}}.$$

Definition 2.3

Projected stochastic gradient descent *can be used to solve constrained optimization, with* Ω *a closed convex set,*

$$\min_{x\in\Omega} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

by projecting the gradient update step as

$$x^{k+1} = P_\Omega(x^k - \gamma_k \nabla f_i(x^k))$$

for an i *sampled uniformly at random from* $\{1, \cdots, n\}$*.*

Proposition 2.2

The projected gradient descent's convergence can be analysed in the same way as the unconstrained variant and hence enjoys a square root convergence for bounded gradient convex functions.

Remark 2.2

One can also redesign some stochastic subgradient descent with similar convergence results. We will not detail this any further.

Definition 2.4

A **mini-batch stochastic gradient descent algorithm** *is defined by the following sequences, for an integer* $1 \geq m \geq n$

$$\text{sample } S \text{ of dimension } m \text{ uniformly at random } in\{1, \cdots, n\}$$

$$g_k := \frac{1}{m} \sum_{i \in S} \nabla f_i(x^k)$$

$$x^{k+1} = x^k - \gamma_k g_k.$$

Remark 2.3

It is worth mentioning

- $m = 1$ yields the previous definition of stochastic gradient descent.
- $m = n$ is equivalent to the classical gradient descent.
- m is called the mini-batch size
- all chosen gradients can be computed in parallel. The update of x can be done as the gradients come in.

Proposition 2.3

The variance of the mini-batch stochastic gradient descent decreases linearly with the mini-batch size. More precisely:
Let $S \subset \{1, \cdots, n\}$ *be a subset sampled uniformly at random and let* g_k *denotes the stochastic mini batch gradient over* S. *Then*

$$\mathbb{E}\left[\left\|g_k - \nabla f(x^k)\right\|^2\right] \leq \frac{B^2}{|S|}.$$

Proposition 2.4

In general, adding smoothness and nothing else does not improve the convergence rates of stochastic gradient descent methods.
See: Ganghui Lan, An Optimal Method for Stochastic Composite Optimization, Mathematical programming, 2012.

Proposition 2.5

A decay rate of $\mathcal{O}(1/K)$ can be achieved if f is a least square regularization function:
Under some smoothness and R bounded data assumptions and f being the least mean square function, using the constant stepzie

$$\gamma < \frac{1}{R^2}$$

the iterates generated by stochastic gradient descent satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f(x^*)\right] \leq \frac{1}{2K}\left(\frac{\sigma\sqrt{d}}{1-\sqrt{\gamma R^2}}+\frac{R\|x^0-x^*\|_2}{\sqrt{\gamma R^2}}\right)^2.$$

In case $\gamma = \frac{1}{4R^2}$ the bound becomes

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f(x^*)\right] \leq \frac{2}{K}\left(\sigma\sqrt{d}+R\|x^0-x^*\|_2\right)^2.$$

See: Francis Bach, Eric Moulines, Non-strongly convex smooth stochastic approximation with convergence $\mathcal{O}(1/n)$, Neural Information Processing Systems, 2013.