DSA3102: Analysis of Stochastic Gradient Descent

Vincent Y. F. Tan

October 30, 2023

In this short document, we provided an analysis of stochastic gradient descent.

In many machine learning problems, we have a set of training samples. The loss function we want to minimize is the sum of losses associated to each of the training samples. In other words, the objective function can be written as the following finite sum

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{1}$$

where $x \in \mathbb{R}^n$ is the decision variable and $m \in \mathbb{N}$ is the number of training samples. Let us assume that each of the f_i's is differentiable.

Stochastic gradient descent (SGD) dates back to the seminal works of Robbins and Monro [RM51] and Kiefer and Wolfowitz [KW52] and is now used extensively in machine learning, especially the training of deep neural networks, so it is worth knowing. SGD updates the decision variable as follows:

$$x^{(k+1)} = x^{(k)} - t_k \nabla f_{i_k}(x^{(k)}) \tag{2}$$

where i_k is an index drawn randomly from $[m]$ and $t_k > 0$ is the step size that is allowed to vary with the iteration number k. Thus, in SGD, we randomly pick an index $i_k \in [m]$, compute the gradient of f_{i_k} and evaluate it at the current iterate $x^{(k)}$, then move in the direction of the negative gradient of f_{i_k} for a step length of t_k.

To analyze SGD, we will make the following assumptions.

1. f is bounded below, i.e., $p^* = \inf_{x \in \mathbf{dom}\ f} f(x) > -\infty$.
2. ∇f is M-Lipschitz, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \qquad \forall\, x, y \in \mathbf{dom}\ f. \tag{3}$$

   If f is twice continuously differentiable, this is equivalent to the Hessian of f satisfying

$$\nabla^2 f(x) \preceq MI. \tag{4}$$

   The condition in (3) and (4) are sometimes called *M-smoothness*.

3. The variance is bounded, i.e.,

$$\mathbb{E}\left[\|\nabla f_i(x)\|^2\right] \leq \sigma^2. \tag{5}$$

   Here, the randomness in the expectation is over the index $i \in [m]$.

Equipped with the above assumptions, we can start analyzing SGD. Recall that the *descent lemma* states that for an M-smooth function,

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \nabla f(x^{(k)})^T(x^{(k+1)} - x^{(k)}) + \frac{M}{2}\|x^{(k+1)} - x^{(k)}\|^2. \tag{6}$$

This comes from Taylor's theorem which states that for any $x, y \in \mathbf{dom}\ f$, there exist z in the line segment connecting x and y such that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x). \tag{7}$$

Now, we can particularize this to $y = x^{(k+1)}$, $x = x^{(k)}$ and upper bound $\nabla^2 f(z)$ uniformly by MI; see (4).

Now, plugging in the stochastic gradient iteration (2) into the descent lemma (6), we obtain

$$f(x^{(k+1)}) \leq f(x^{(k)}) - t_k \nabla f(x^{(k)})^T \nabla f_{i_k}(x^{(k)}) + t_k^2 \frac{M}{2} \|\nabla f_{i_k}(x^{(k)})\|^2. \tag{8}$$

Taking the expectation with respect to the random i_k on both sides (assuming $\Pr(i_k = i) = 1/m$ for each $i \in [m]$), we obtain

$$\mathbb{E}\left[f(x^{(k+1)})\right] \leq f(x^{(k)}) - t_k \nabla f(x^{(k)})^T \mathbb{E}\left[\nabla f_{i_k}(x^{(k)})\right] + t_k^2 \frac{M}{2} \mathbb{E}\left[\|\nabla f_{i_k}(x^{(k)})\|^2\right]. \tag{9}$$

Since the stochastic gradient is unbiased, i.e., $\mathbb{E}\left[\nabla f_{i_k}(x^{(k)})\right] = \nabla f(x^{(k)})$, we have

$$\mathbb{E}\left[f(x^{(k+1)})\right] \leq f(x^{(k)}) - t_k \|\nabla f(x^{(k)})\|^2 + t_k^2 \frac{M}{2} \mathbb{E}\left[\|\nabla f_{i_k}(x^{(k)})\|^2\right]. \tag{10}$$

We may now use the bound on the variance in (5) to obtain the following upper bound

$$\mathbb{E}\left[f(x^{(k+1)})\right] \leq f(x^{(k)}) - t_k \|\nabla f(x^{(k)})\|^2 + t_k^2 \frac{M}{2} \sigma^2. \tag{11}$$

This means, by a slight rearrangement, that

$$t_k \|\nabla f(x^{(k)})\|^2 \leq f(x^{(k)}) - \mathbb{E}\left[f(x^{(k+1)})\right] + t_k^2 \frac{M}{2} \sigma^2. \tag{12}$$

Now, assuming we run SGD for $T \in \mathbb{N}$ iterations, we may sum up both sides and use the law of iterated expectations to get

$$\sum_{k=1}^{T} t_{k-1} \mathbb{E}\left[\|\nabla f(x^{(k-1)})\|^2\right] \leq \sum_{k=1}^{T} \mathbb{E}\left[f(x^{(k)}) - f(x^{(k+1)})\right] + \sum_{k=1}^{T} t_k^2 \frac{M}{2} \sigma^2. \tag{13}$$

From here, we will bound $\mathbb{E}\left[\|\nabla f(x^{(k-1)})\|^2\right]$ on the left by its minimum value across the iterations, use a telescoping sum on the first term on the right, and notice that $M\sigma^2/2$ has no dependence on k to obtain

$$\min_{k=0,1,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right] \left(\sum_{k=0}^{T-1} t_k\right) \leq f(x^{(0)}) - \mathbb{E}\big[f(x^{(T)})\big] + \frac{M\sigma^2}{2} \sum_{k=0}^{T-1} t_k^2. \tag{14}$$

Using the the fact that $\mathbb{E}[f(x^{(k)})] \geq p^*$, we obtain the final bound

$$\min_{k=0,1,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right] \leq \frac{f(x^{(0)}) - p^*}{\sum_{k=0}^{T-1} t_k} + \frac{M\sigma^2}{2} \frac{\sum_{k=0}^{T-1} t_k^2}{\sum_{k=0}^{T-1} t_k}. \tag{15}$$

A few remarks on this bound are in order.

- Note that if $\sigma = 0$, SGD essentially reduces to gradient descent. We may set the step size $t_k = t$ to be a constant, say $t \in [0, 1/M]$ so that the convergence rate is $O(1/T)$. This is similar to vanilla gradient descent. Note that here, we have not made any assumptions on strong convexity. If f is strongly convex, we have shown that the convergence rate is linear, i.e., c^T for some $c \in (0, 1)$ related to the condition number of $\nabla^2 f$.

- However, in SGD, $\sigma^2 > 0$. The convergence rate is thus dominated by the second term. Specifically, the term that is of interest in for the convergence rate is $(\sum_{k=0}^{T-1} t_k^2)/(\sum_{k=0}^{T-1} t_k)$. Setting t_k to $1/k$, and observing that $\sum_{k=0}^{T-1} t_k = \Theta(\log T)$ and $\sum_{k=0}^{T-1} t_k^2 = \Theta(1)$, we obtain that convergence rate is of order $O(1/\log T)$, which is slow.
- We may play with other choices of the step size. For example, if we choose $t_k = \alpha/\sqrt{k}$ for some $\alpha > 0$, $\sum_{k=0}^{T-1} t_k^2 = \Theta(\log T)$ and $\sum_{k=0}^{T-1} t_k = \Theta(\sqrt{T})$ and the convergence rate is $O((\log T)/\sqrt{T})$, which is significantly better than in the previous point.
- Similar to vanilla gradient descent, we will get significant better convergence rates if we assume strong convexity, but, to the best of my knowledge, not as good as vanilla gradient descent.
- We bounded the term on the left of (13) by the minimum $\min_{k=0,1,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right]$. Another option is to run SGD for a *random* number of iterations τ where we run for $\tau = k$ iterations with probability

$$\Pr(\tau = k) \propto \frac{t_k}{\sum_{\ell=0}^{T-1} t_\ell}. \tag{16}$$

  By the law of total probability, the expected squared-norm of the gradient of ∇f is

$$\mathbb{E}\left[\|\nabla f(x^{(\tau)})\|^2\right] = \left(\sum_{\ell=0}^{T-1} t_\ell\right)^{-1} \sum_{k=0}^{T-1} t_k \, \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right]. \tag{17}$$

  Combining (13) and (17) and using the same steps as in (14) and (15), we obtain

$$\mathbb{E}\left[\|\nabla f(x^{(\tau)})\|^2\right] \leq \frac{f(x^{(0)}) - p^*}{\sum_{k=0}^{T-1} t_k} + \frac{M\sigma^2}{2} \frac{\sum_{k=0}^{T-1} t_k^2}{\sum_{k=0}^{T-1} t_k}. \tag{18}$$

  The above conclusions concerning the choice of the step size still apply.
- Finally, SGD approximates the full gradient with 1 sample, i.e.,

$$\nabla f_{i_k}(x^{(k)}) \approx \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x^{(k)}). \tag{19}$$

  While the bias is 0, the variance can be very large. A more common thing to do in practice is to use *batches*, also called "mini-batches", i.e.,

$$\frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla f_i(x^{(k)}) \approx \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x^{(k)}), \tag{20}$$

  where $\mathcal{B}_k \subset [m]$ is the subset of training samples for which we have resources to compute the gradient on at iteration k. This is particularly useful for vectorization or parallelization across different machines. For example, if you have 16 cores, set $|\mathcal{B}_k| = 16$ and you can compute 16 gradients at once. Of course, you can also analyze SGD with mini-batches, but we will leave that as an exercise for the interested reader.

References

[KW52] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400, 1951.