Notes on stochastic gradient descent

Amitabha Basu

Tuesday 10th November, 2020

1 Basic properties

We recall that if $f : \mathbb{R}^n \to \mathbb{R}$ has a Lipschitz continuous gradient with Lipschitz constant $L$ — a property that we will term $L$-smooth — then one has the following quadratic upper and lower bounds on the function:

$$\forall x, y \in \mathbb{R}^n : f(x) + \langle \nabla f(x), y-x \rangle - \frac{1}{2}L\|y-x\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}L\|y-x\|^2.$$  \hspace{1cm} (1.1)

So far in our lectures, we have been assuming that we have access to the gradient $\nabla f(x)$ at any point $x$. Often, computing the full gradient can be quite expensive. In practice, one often has an oracle that returns a proxy for the gradient and the only guarantee is that the output of the oracle is random, but its expected value is the true gradient. More precisely, we have the following definition.

Definition 1.1. A stochastic gradient oracle for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ takes as input a point $x \in \mathbb{R}^n$ and outputs a random vector $a \in \mathbb{R}^n$ such that $E[a] = \nabla f(x)$, where the expectation is taken with respect to the randomization of the oracle. Sometimes, this expectation condition is made explicit by saying that the oracle is an unbiased estimator for the gradient.

Example 1.2. The following are two widely used examples of stochastic gradient oracles.

1. The oracle samples a coordinate $i \in \{1, \ldots, n\}$ uniformly at random and returns $n \partial f / \partial x^i e^i$, where $e^i$ is the $i$-th standard unit vector, i.e., $e^i_j = 0$ if $j \neq i$ and $e^i_i = 1$.

2. The function $f(x)$ is given as the sum of many terms, i.e., $f(x) = \frac{1}{D} \sum_{j=1}^D g_j(x)$. This kind of optimization appears in many machine learning and statistical applications. For example, the linear least squares problem that we saw was of the form $f(x) = \frac{1}{D} \sum_{j=1}^D (\langle a^j, x \rangle - b_j)^2$, for some data points $(a^1, b_1), \ldots, (a^D, b_D) \in \mathbb{R}^n \times \mathbb{R}$. In this scenario, instead of computing the gradients of each $g_i$ and adding them all up, the stochastic oracle samples a term $j \in \{1, \ldots, D\}$ uniformly at random and returns $\nabla g_j(x)$.

Stochastic Gradient Descent is a name that is applied to any algorithm of the following type. One starts with an initial iterate $x_0$ and for every iteration $k \geq 0, 1, 2, \ldots$, one queries the stochastic gradient oracle at the current iterate $x_k$ to obtain a (random) vector $a_k$. The next iterate $x_{k+1} := x_k - \alpha_k a_k$. Thus, the sequence of iterates $x_0, x_1, x_2, \ldots$ is a stochastic process.

Note that the stochastic oracle is not guaranteed to return a vector $a$ such that $-a$ is a descent direction at $x$, even though in some cases (like Example 1. above), this is indeed true. It is interesting that one can still prove convergence results even without this requirement with just the assumption that the expected value of the output is the true gradient. However, one has to be careful with the step sizes.
1.1 Convergence analysis

In any analysis with stochastic gradients, one has to assume that the variance or the second moment of the output $a$ of the stochastic oracle is bounded.

**Assumption 1.** There is a constant $V$ such that the variance $\mathbb{V}[a] \leq V$ when the oracle is queried at any point $x \in \mathbb{R}^n$.

**Theorem 1.3.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be any $L$-smooth function bounded below, i.e., $f(x) \geq f^*$ for all $x \in \mathbb{R}^n$. Suppose the stochastic oracle satisfies Assumption 1 and we use step lengths $0 < \alpha_k < \frac{2}{L}$, for all $k \geq 0$. Then for all $T \geq 1$,

$$
\mathbb{E} \left[ \min_{k=0, \ldots, T} \| \nabla f(x_k) \|^2 \right] \leq \frac{f(x_0) - f^* + \frac{VL}{2} \sum_{k=0}^{T} \alpha_k^2}{\sum_{k=0}^{T} \alpha_k (1 - \frac{L\alpha_k}{2})}.
$$

(1.2)

**Proof.** From (1.1), we know that for all $k \geq 0$,

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

We take a conditional expectation of both sides of the inequality, conditioned on $x_k$ being some arbitrary, but fixed, point in $\mathbb{R}^n$ in the stochastic process of the iterates. Thus,

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

$$
= f(x_k) - \alpha_k \mathbb{E}[\langle \nabla f(x_k), a_k \rangle|x_k] + \frac{L}{2} \alpha_k^2 \mathbb{E}[\|a_k\|^2|x_k] + V
$$

$$
= f(x_k) - \alpha_k \mathbb{E}[\nabla f(x_k), \nabla f(x_k)] + \frac{L}{2} \alpha_k^2 \|\nabla f(x_k)\|^2 + \frac{VL}{2} \alpha_k^2
$$

(1.3)

We now take an expectation over $x_k$. By the law of total expectations,

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

Subtracting $f^*$ from both sides and rearranging we obtain

$$
\alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \mathbb{E}[\Delta_k] - \mathbb{E}[\Delta_{k+1}] + \frac{VL}{2} \alpha_k^2
$$

where $\Delta_k = f(x_k) - f^*$ (which is a random variable since $x_k$ is a random variable). Summing the above inequality for $k = 0, \ldots, T$, we obtain

$$
\sum_{k=0}^{T} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \mathbb{E}[\Delta_0] - \mathbb{E}[\Delta_{T+1}] + \frac{VL}{2} \sum_{k=0}^{T} \alpha_k^2 \leq \Delta_0 + \frac{VL}{2} \sum_{k=0}^{T} \alpha_k^2
$$

since $x_0$ is not random and $\Delta_{T+1} \geq 0$. We now use the property that for any set of (possibly dependent) random variables $Y_1, \ldots, Y_m$, $\mathbb{E}[\min_k Y_k] \leq \min_k \mathbb{E}[Y_k]$. Thus, we get

$$
\left(\sum_{k=0}^{T} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right)\right) \mathbb{E}[\min_{k=0, \ldots, T} \| \nabla f(x_k) \|^2] \leq \left(\sum_{k=0}^{T} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right)\right) \min_{k=0, \ldots, T} \mathbb{E}[\| \nabla f(x_k) \|^2]
$$

$$
\leq \sum_{k=0}^{T} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \mathbb{E}[\| \nabla f(x_k) \|^2] \leq \Delta_0 + \frac{VL}{2} \sum_{k=0}^{T} \alpha_k^2.
$$

By hypothesis, $\alpha_k \left(1 - \frac{L\alpha_k}{2}\right) > 0$ for all $k \geq 0$. So dividing through by $\sum_{k=0}^{T} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right)$, we obtain the desired inequality.
**Constant step size** $\alpha_k = \frac{1}{T}$. If we plug into (1.2) the constant step size of $\frac{1}{T}$ used in our deterministic steepest/gradient descent analysis, we obtain
\[
E\left[\min_{k=0,\ldots,T} \|\nabla f(x_k)\|^2 \right] \leq \frac{L(f(x_0) - f^*)}{T + 1} + \frac{V}{2}.
\]
So no matter how many iterations $T$ we go for the expected value of the smallest (squared) norm of the gradients can only be guaranteed to at most the variance bound of the stochastic oracle. We are unable to show it can be made arbitrarily small unless $V = 0$ which would mean we are in the deterministic case.

**Constant step size** $\alpha_k = \frac{1}{L\sqrt{T+1}}$. If one selects the step size to depend on the number of iterations $T$, then one can get something stronger. In particular, plugging in $\alpha_k = \frac{1}{L\sqrt{T+1}}$ in (1.2), we obtain that
\[
E\left[\min_{k=0,\ldots,T} \|\nabla f(x_k)\|^2 \right] \leq \frac{2L(f(x_0) - f^*) + V}{2\sqrt{T + 1} - 1}.
\]
For any real valued random variable $Z$, we have $(E[Z])^2 + \sqrt{\text{Var} Z} = E[Z^2]$ and so $(E[Z])^2 \leq E[Z^2]$. Therefore,
\[
\left(E\left[\min_{k=0,\ldots,T} \|\nabla f(x_k)\| \right] \right)^2 \leq \left(E\left[\min_{k=0,\ldots,T} \|\nabla f(x_k)\|^2 \right] \right) = E\left[\min_{k=0,\ldots,T} \|\nabla f(x_k)\|^2 \right] \leq \frac{2L(f(x_0) - f^*) + V}{2\sqrt{T + 1} - 1}.
\]

**Corollary 1.4.** Let $f: \mathbb{R}^n \to \mathbb{R}$ be any $L$-smooth function bounded below, i.e., $f(x) \geq f^*$ for all $x \in \mathbb{R}^n$. Suppose the stochastic oracle satisfies Assumption 1. For any $T \geq 2$, if we use step lengths $\alpha_k = \frac{1}{L\sqrt{T+1}} < \frac{2}{L}$ for all $k \geq 0$, the stochastic gradient descent algorithm generates iterates with the property that
\[
E\left[\min_{k=0,\ldots,T} \|\nabla f(x_k)\| \right] \leq \sqrt{\frac{2L(f(x_0) - f^*) + V}{2\sqrt{T + 1} - 1}}. \tag{1.4}
\]
In other words, if we wish the expected value of the smallest gradient norm to be less than $\epsilon$, then we require $T = O\left((\frac{1}{\epsilon})^4\right)$. Compare this with the $O\left((\frac{1}{\epsilon})^2\right)$ convergence rate for the deterministic case.

With a slightly more careful analysis, one can use step lengths $\alpha_k = \frac{1}{L\sqrt{\epsilon}}$ and get the same $O\left((\frac{1}{\epsilon})^4\right)$ convergence rate (asymptotically). The advantage in this case is that one does not have to select the number of iterations in advance.

**1.1.1 Using convexity**

As in our previous lectures, we would like to improve the convergence rate if we add the assumption of convexity/strong convexity on top of $L$-smoothness. Like we saw in our deterministic analysis of smooth convex optimization, the following inequality for any differentiable convex function $f: \mathbb{R}^n \to \mathbb{R}$ is critical which gives us better lower bounds than (1.1),
\[
\forall x, y \in \mathbb{R}^n: \quad f(x) + \langle \nabla f(x), y - x \rangle \leq f(y). \tag{1.5}
\]
We also make a stronger assumption compared to Assumption 1; we assume that the stochastic oracle’s output has bounded second moment (this means we are also assuming that all true gradients – at least at the points possibly visited by the algorithm – have norm bounded by some constant).

**Assumption 2.** There is a constant $M$ such that the second moment $E[\|a\|^2] \leq M$ when the oracle is queried at any point $x \in \mathbb{R}^n$.

**Theorem 1.5.** Let $f: \mathbb{R}^n \to \mathbb{R}$ be any convex, $L$-smooth function. Assume that there is a minimizer $x^*$ for $f$ and $f(x^*) = f^*$. Suppose the stochastic oracle satisfies Assumption 2 and we use step lengths $0 < \alpha_k,$
for all \( k \geq 0 \). Let the iterates of the stochastic gradient oracle be \( x_0, x_1, x_2, \ldots \). For every \( k \geq 1 \), define \( \bar{x}_k = \frac{\sum_{j=0}^k \alpha_j x_j}{\sum_{j=0}^k \alpha_j} \), i.e., the average of all iterates up to \( k \). Then, for all \( T \geq 1 \),

\[
\mathbb{E}[f(\bar{x}_T) - f^*] \leq \frac{\|x_0 - x^*\|^2 + M \sum_{k=0}^T \alpha_k^2}{\sum_{k=0}^T \alpha_k}.
\] (1.6)

**Proof.** For all \( k \geq 0 \), we have

\[
\|x_{k+1} - x^*\|^2 = \|x_k - \alpha_k a_k - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k \langle a_k, x_k - x^* \rangle + \alpha_k^2 \|a_k\|^2
\]

We take a conditional expectation of both sides of the inequality, conditioned on \( x_k \) being some arbitrary, but fixed, point in \( \mathbb{R}^n \) in the stochastic process of the iterates.

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

where we have used Assumption 2. Rearranging, we get

\[
2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle \leq \|x_k - x^*\|^2 - \mathbb{E}[\|x_{k+1} - x^*\|^2 | x_k] + \alpha_k^2 M
\]

Taking expectation with respect to \( x_k \) and using the law of total expectation, we obtain

\[
2\alpha_k \mathbb{E}[\langle \nabla f(x_k), x_k - x^* \rangle] \leq \mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2] + \alpha_k^2 M
\] (1.7)

We now appeal to inequality (1.5) with \( x = x_k \) and \( y = x^* \) to get \( f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) = \|x^*\|^2 \). Rearranging and taking expectation on both sides with respect to \( x_k \), we obtain that \( \mathbb{E}[f(x_k) - f^*] \leq \mathbb{E}[\langle \nabla f(x_k), x_k - x^* \rangle] \). Using this in (1.7), we obtain that

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

Summing these inequalities for \( k = 0, \ldots, T \) and dividing by \( \sum_{k=0}^T \alpha_k \), we obtain

\[
\frac{\sum_{k=0}^T \alpha_k \mathbb{E}[f(x_k) - f^*]}{\sum_{k=0}^T \alpha_k} \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] - \mathbb{E}[\|x_{T+1} - x^*\|^2] + \sum_{k=0}^T \alpha_k^2 M}{\sum_{k=0}^T \alpha_k}
\]

Using the convexity of \( f \), linearity of expectation, the facts that \( x_0 \) is not random and \( \|x_{T+1} - x^*\|^2 \geq 0 \), we obtain that

\[
\mathbb{E}[f(\bar{x}) - f^*] \leq \mathbb{E} \left[ \frac{\sum_{k=0}^T \alpha_k f(x_k)}{\sum_{k=0}^T \alpha_k} - f^* \right]
\]

\[
= \frac{\sum_{k=0}^T \alpha_k \mathbb{E}[f(x_k) - f^*]}{\sum_{k=0}^T \alpha_k} \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] - \mathbb{E}[\|x_{T+1} - x^*\|^2] + \sum_{k=0}^T \alpha_k^2 M}{\sum_{k=0}^T \alpha_k}
\]

\[
= \frac{\|x_0 - x^*\|^2 + M \sum_{k=0}^T \alpha_k^2}{\sum_{k=0}^T \alpha_k}
\]

which gives us the desired inequality. \( \square \)

**Step sizes.** The above bound is very similar to the bound we had for the deterministic subgradient algorithm for non smooth convex functions. As in that case, choosing step sizes \( \alpha_k = O(\frac{1}{\sqrt{T+1}}) \) gives us a convergence rate of \( O \left( \left( \frac{1}{T} \right)^2 \right) \) convergence rate to get to an \( \epsilon \)-optimal point.
**Strong convexity.** We can improve the convergence rate by assuming strong convexity. The idea again is that we have an even better lower bound than (1.5) for a $\mu$-strongly convex function $f$:

$$\forall x, y \in \mathbb{R}^n : f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\mu \|y - x\|^2 \leq f(y) \quad (1.8)$$

**Theorem 1.6.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be any $\mu$-strongly convex, $L$-smooth function. Let $x^*$ be a minimizer for $f$ (which is guaranteed to exist if $\mu > 0$) and $f(x^*) = f^*$. Suppose the stochastic oracle satisfies Assumption 2 and we use step lengths $\alpha_k = \frac{\theta}{k + \frac{1}{2}}$, for all $k \geq 0$, with some constant $\theta > \frac{1}{2\mu}$. Let the iterates of the stochastic gradient oracle be $x_0, x_1, x_2, \ldots$. Then there exists a constant $Q$ (depending on $\theta, \mu, M, x_0$) such that for all $T \geq 0$,

$$\mathbb{E}[\|x_T - x^*\|^2] \leq \frac{Q}{T + 1},$$

and consequently,

$$\mathbb{E}[f(x_T) - f^*] \leq \frac{LQ}{2(T + 1)}.$$

**Proof.** Switching the roles of $x$ and $y$ in (1.8) and adding the inequalities together, we obtain that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|y - x\|^2$ for all $x, y \in \mathbb{R}^n$. Setting $x = x_k$ and $y = x^*$, we obtain that $\langle \nabla f(x_k), x_k - x^* \rangle \geq \mu \|x_k - x^*\|^2$ for all $k \geq 0$ (since $\nabla f(x^*) = 0$). We use this in (1.7) that was derived in the previous proof to obtain

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

Rearranging, we obtain

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

Plugging in $\alpha_k = \frac{\theta}{k + \frac{1}{2}}$, we obtain that

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

Setting $Q > \max \left\{ \frac{\theta^2 M}{2\mu \theta - 1}, \|x_0 - x^*\|^2 \right\}$, we verify by induction that

$$\mathbb{E}[\|x_T - x^*\|^2] \leq \frac{Q}{T + 1} \quad \forall T \geq 0.$$

The base case $T = 0$ follows from the fact that $Q > \|x_0 - x^*\|^2$. It suffices to verify that $Q > \frac{\theta^2 M}{2\mu \theta - 1}$ implies

$$\left(1 - \frac{2\theta \mu}{k + 1}\right) \frac{Q}{k + 1} + \frac{\theta^2 M}{(k + 1)^2} \leq \frac{Q}{k + 2} \quad \forall k \geq 0.$$

We then use the upper bound in (1.1) with $y = x_k$ and $x = x^*$ and the fact that $\nabla f(x^*) = 0$ to get the second desired inequality. \qed

**1.1.2 Special case: random coordinate minimization**

We now show that where we get a better guarantee for nonconvex $L$-smooth functions when the stochastic gradient oracle is implemented as the first point under Example 1.2. Thus, for specific stochastic oracles, one may be able to give better guarantees than Theorems 1.3, 1.5 and 1.6.

**Theorem 1.7.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be any $L$-smooth function bounded below, i.e., $f(x) \geq f^*$ for all $x \in \mathbb{R}^n$. Suppose the stochastic gradient oracle is a coordinate partial derivative chosen uniformly at random from
Proof. For each $k \geq 0$, let $i(k) \in \{1, \ldots, n\}$ be the random coordinate chosen at iteration $k$ by the oracle. From (1.1), we know that for all $k \geq 0$,

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

$$= f(x_k) - \frac{1}{L} \langle \nabla f(x_k), \frac{\partial f}{\partial x_{i(k)}} \rangle + \frac{1}{2L} \left( \frac{\partial f}{\partial x_{i(k)}} \right)^2$$

$$= f(x_k) - \frac{1}{L} \left( \frac{\partial f}{\partial x_{i(k)}} \right)^2 + \frac{1}{2L} \left( \frac{\partial f}{\partial x_{i(k)}} \right)^2$$

$$= f(x_k) - \frac{1}{2L} \left( \frac{\partial f}{\partial x_{i(k)}} \right)^2$$

We take a conditional expectation of both sides of the inequality, conditioned on $x_k$ being some arbitrary, but fixed, point in $\mathbb{R}^n$ in the stochastic process of the iterates. Thus,

$$\mathbb{E}[f(x_{k+1}) | x_k] \leq f(x_k) - \frac{1}{2L} \mathbb{E} \left[ \left( \frac{\partial f}{\partial x_{i(k)}} \right)^2 \right]$$

$$= f(x_k) - \frac{1}{2L} \mathbb{E} \| \nabla f(x_k) \|^2$$

We now take an expectation over $x_k$ and subtract $f^*$ from both sides. By the law of total expectations,

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

Rearranging and summing the above inequality for $k = 0, \ldots, T$, we obtain

$$\frac{T + 1}{2Ln} \min_{k=0, \ldots, T} \mathbb{E}[\| \nabla f(x_k) \|^2] \leq \frac{1}{2Ln} \sum_{k=0}^{T} \mathbb{E}[\| \nabla f(x_k) \|^2] \leq \mathbb{E}[f(x_0) - f^*] - \mathbb{E}[f(x_{T+1}) - f^*] \leq f(x_0) - f^*.$$

since $x_0$ is not a random variable. As before, we use standard probability theory to get (see the proof of Theorem 1.3 and the analysis after that for the step sizes)

$$(\mathbb{E}[\min_{k=0, \ldots, T} \| \nabla f(x_k) \|])^2 \leq \mathbb{E}[\min_{k=0, \ldots, T} \| \nabla f(x_k) \|^2] = \mathbb{E}[\min_{k=0, \ldots, T} \| \nabla f(x_k) \|^2] \leq \min_{k=0, \ldots, T} \mathbb{E}[\| \nabla f(x_k) \|^2].$$

Putting everything together, we obtain the desired inequality. 

\[
\]

2 Further Reading

Most of the presentation here was adapted from [4]. See also the series of papers by Ghadimi and Lan [1–3].

References

