Lemma Proofs for ”Adaptive Methods for Nonconvex Optimization”

We analyze the lemmas used in the theorems derived in the paper ”Adaptive Methods for Nonconvex Optimization”, published in NeurIPS 2018. We provide a broader annotated outline of the theorem proofs in another PDF, ”Annotated Proof Outline” on the same github this file is in, and **the interpretation of the results in our slides and video presentation**. We present the ”Annotated Proof Outline” of Theorem 1 and Lemma 1 only as a rudimentary learning aid to follow the algebra structure behind the proof. However, we hope this more developed PDF can also be a learning aid, as we summarize the assumptions made and go in-depth for the proofs of the lemmas.

Preliminaries:

The paper studies the following type of nonconvex stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) \coloneqq \mathbb{E}_{s \sim \mathbb{P}}[\ell(x, s)]$$

Breaking it down, we have that $f(x)$ is the true objective function across $\mathbb{R}^k$, which is equivalent to the expectation of the loss function ℓ that we compute from our minibatch S. (Expectation is taken w.r.t. our entire dataset, i.e. all possible samples). Then, we look to find the model parameters x to minimize this function f.

Stationarity - In the nonconvex setting, we define our convergence criterion to be $\|\nabla f(x)\|^2 \leq \delta$. This same convergence criterion can be made for the convex case. (Since we don’t know the global optimum for nonconvex problems, the suboptimality measure $\|f(x^t) - f(x^*)\| \leq \delta$ can’t really be used here.) However, we do have to note that stationarity doesn’t necessarily imply we’re at a stationary point.

Assumptions:

The following assumptions are made for the proofs:

(1.) ℓ is assumed to be $L - smooth$ in $\mathbb{R}^d$, i.e., there exists a constant L such that:

$$\|\nabla \ell(x, s) - \nabla \ell(y, s)\| \leq L\|x - y\|, \qquad \text{for all } x, y \in \mathbb{R}^d \text{ and } s \in S$$

(This implies the expected loss function f is also $L - smooth$. We can think of the "average" of $L - smooth$ functions to be $L - smooth$.)

(2.) The gradient of ℓ is bounded by a constant G:

$$\|\nabla[\ell(x,s)]_i\| \leq G \qquad \text{for all } x \in \mathbb{R}^d, s \in S, i \in [d]$$

(3.) Finite variance condition, the variance of the stochastic gradients are bounded by a term σ^2:

$$\mathbb{E}\|\nabla\ell(x,s) - \nabla f(x)\|^2 \leq \sigma^2 \qquad \text{for all } x \in \mathbb{R}^d$$

(4.) The sample/stochastic gradient $\nabla\ell$ is an unbiased estimate of the objective function's gradient:

$$\mathbb{E}[\nabla\ell(x,s)] = \nabla f(x) \text{ for all } x \in \mathbb{R}^d, s \in S$$

Notation:

The authors use the following notation:

For any vectors $a, b \in \mathbb{R}^d$, $\sqrt{a}$ is defined to be the element-wise square root, a^2 the element-wise square, $\frac{a}{b}$ element-wise division, and for some $c_i \in \mathbb{R}^d$, the $j-$th coordinate of c_i is denoted to be $c_{i,j}$ or $[c_i]_j$. (note vectors are not bold fonted.)

Also, the authors denote that $g_t = \nabla\ell(x_t, s_t)$, where x_t are the model parameters at iteration t, and s_t the sample at iteration t drawn from $\mathbb{P}$.

Here, we dive more directly into the Lemmas used in the proofs.

Lemma 1:

For the iterates x_t where $t \in [T]$ in the algorithms 1 & 2 in Adaptive Methods for Nonconvex Optimization, it is shown that for all $i \in [d]$,

$$\mathbb{E}[\|g_{t,i}\|^2] \leq \frac{\sigma_i^2}{b} + [\nabla f(x_t)]_i^2$$

That is, the expected stationarity of the gradient's $i-$th term at iteration t is bounded by the variance of the stochastic gradients of the $i-$th term over the mini batch size, plus the squared expected gradient of the $i-$th term of the full objective function f .

Pf.

For ease of use, we define:

$$\zeta_t = \frac{1}{|S_t|}\sum_{s\in S_t}([\nabla \ell(x_t, s)]_i - [\nabla f(x_t)]_i)$$
$$g_{t,i} = \frac{1}{|S_t|}\sum_{s\in S_t}[\nabla \ell(x_t, s)]_i$$

ζ_t can be considered to be the average of the mini-batch gradients' standard deviation (where S_t is our mini-batch). Notice when S_t is the entire training set, $\zeta_t = 0$. The i denotes that we are just taking it with respect to all elements $i \in [d]$.

$g_{t,i}$ denotes the mini-batch gradient. Once again, the i denotes that we are just taking it with respect to all elements $i \in [d]$.

We can also see ζ_t as taking the expectation of our mini-batch gradient, which we assume $\nabla\ell$ is an unbiased estimate of ∇f:

$$\begin{aligned}\mathbb{E}[\zeta_t] &= \frac{1}{|S_t|}\sum_{s\in S_t}(\mathbb{E}[\nabla \ell(x_t, s)]_i - [\nabla f(x_t)]_i)\\ &= \frac{1}{|S_t|}\sum_{s\in S_t}([\nabla f(x_t)]_i - [\nabla f(x_t)]_i) = 0\end{aligned}$$

So, $\mathbb{E}[\zeta_t] = 0$.

Then we have:

$$\begin{aligned}\mathbb{E}_t[g_{t,i}^2] &= \mathbb{E}_t\left[\left\|\frac{1}{|S_t|}\sum_{s\in S_t}[\nabla \ell(x_t, s)]_i\right\|^2\right] &(1)\\ &= \mathbb{E}_t\left[\left\|\frac{1}{|S_t|}\sum_{s\in S_t}([\nabla \ell(x_t, s)]_i - [\nabla f(x_t)]_i + [\nabla f(x_t)]_i\right\|^2\right] &(2)\\ &= \mathbb{E}_t\left[\|\zeta_t + [\nabla f(x_t)]_i\|^2\right] &(3)\\ &= \mathbb{E}_t[\zeta_t^2] + \mathbb{E}[\zeta_t]\cdot[\nabla f(x_t)]_i + [\nabla f(x_t)]_i^2 &(4)\\ &= \mathbb{E}_t[\zeta_t^2] + [\nabla f(x_t)]_i^2 &(5)\end{aligned}$$

(1). Recall our notation - we had that $g_{t,i}^2 = (g_{t,i})^T(g_{t,i}) = \|g_{t,i}\|^2$. We also substitute for $g_{t,i}$.
(2). We add in $0 = -[\nabla f(x_t)]_i + [\nabla f(x_t)]_i$. Notice that the summation doesn't change the value, so we can put it into the summation.

(3). Expand the terms, and use linearity of expectation. Taking the expectation of the true gradient is itself since it's a defined constant (similarly the true gradient squared)
(4). We showed $\mathbb{E}[\zeta_t] = 0$ earlier.

One thing to further note is that we'll denote $|S_t|$ as b, the mini-batch size. We can expand what ζ_t was, and square out the $\frac{1}{|S_t|^2}$ term since it's a constant:

$$\begin{aligned} \mathbb{E}_t[g_{t,i}^2] &= \mathbb{E}_t[\zeta_t^2] + [\nabla f(x_t)]_i^2 \\ &= \frac{1}{b^2}\mathbb{E}_t\left[\left(\sum_{s\in S_t}([\nabla \ell(x_t,s)]_i - [\nabla f(x_t)]_i)\right)^2\right] + [\nabla f(x_t)]_i^2 \\ &\text{Then by the result of Lemma 2:} \\ &= \frac{1}{b^2}\mathbb{E}_t\left[\sum_{s\in S_t}([\nabla \ell(x_t,s)]_i - [\nabla f(x_t)]_i)^2\right] + [\nabla f(x_t)]_i^2 \end{aligned}$$

Notice that $\sum_{s\in S_t}([\nabla \ell(x_t,s)]_i - [\nabla f(x_t)]_i)^2$ is the definition of variance for the stochastic gradients, summed up for all samples in our mini-batch. This can be written as:

$$\mathbb{E}_t\left[\sum_{s\in S_t}([\nabla \ell(x_t,s)]_i - [\nabla f(x_t)]_i)^2\right] = b \cdot ([\nabla \ell(x_t,s)]_i - [\nabla f(x_t)]_i)^2$$

We can substitute this to get the following result:

$$\mathbb{E}_t[g_{t,i}^2] = \frac{1}{b}\mathbb{E}_t\left[\sum_{s\in S_t}([\nabla \ell(x_t,s)]_i - [\nabla f(x_t)]_i)^2\right] + [\nabla f(x_t)]_i^2$$

Finally, we substitute our defined upper bound on the variance of the stochastic gradients, σ^2:

$$\mathbb{E}_t[g_{t,i}^2] \leq \frac{\sigma_i^2}{b} + [\nabla f(x_t)]_i^2$$

□

The core use of Lemma 1 for Theorems 1 and 2 is to provide an upper bound on the stochastic mini-batch gradient evaluations that linearly decreases with batch-size. Lemma 2 is used to derive Lemma 1.

Lemma 2:

For random variables $z_1, \ldots, z_r$ that are *i.i.d.* and have 0 mean,

$$\mathbb{E}[\|z_1 + \cdots + z_r\|]^2 = \mathbb{E}\left[\|z_1\|^2 + \ldots \|z_r\|^2\right]$$

Pf.

$$\mathbb{E}[\|z_1 + \cdots + z_r\|]^2 = \sum_{i,j=1}^{r} \mathbb{E}[z_i z_j] = \mathbb{E}\left[\|z_1\|^2 + \ldots \|z_r\|^2\right]$$

□

The 2nd equality can be thought of as expanding the first term, and then using the property of linearity for expectation. Since the z's are *i.i.d.* and 0 mean, all we're left with is the squared terms.

The above 2 lemmas are very similar to what we proved in HW3.

Finally, we provide a quick interpretation of Theorem 1, the convergence result/bound on ADAM. Theorem 1 and Theorem 2 (YOGI's convergence result) are very similar in both proof structure and interpretable result.

Theorem 1:

Let $\eta_t = \eta$ for all $t \in [T]$. Assume we pick ϵ, β_2, η such that $\eta < \frac{\epsilon}{2L}$ and $(1 - \beta_2) \leq \frac{\epsilon^2}{16G^2}$. Then for x_t generated using the ADAM algorithm listed in Adaptive Methods in Nonconvex Optimization,

$$\mathbb{E}\left\|\nabla f(x_a)\right\|^2 \leq O\left(\frac{f(x_1) - f(x^*)}{\eta T} + \frac{\sigma^2}{b}\right)$$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq 2(\sqrt{2}G + \epsilon) \times \left[\frac{f(x_1) - f(x^*)}{\eta T} + \left(\frac{G\sqrt{1-\beta_2}}{\epsilon^2} + \frac{L\eta}{2\epsilon^2\sqrt{\beta_2}}\right)\frac{\sigma^2}{b}\right]$$

Where x^* is the optimal solution to the stochastic nonconvex problem proposed, i.e. minimizes the loss f, for any x_a chosen where $a \in [T]$.

That is, our expected stationarity of the loss f is shown to be bounded by proportionally by the difference in initial loss to the optimal and inversely by the max iterations T and constant step-size η, plus proportionally to the upper bound of variance of the stochastic gradients and inversely by the mini-batch size b (They technically show for $b = 1$, but the result of $b = 1$ provides a corollary for $b > 1$).