The original PDF proof is from "Adaptive Methods for Nonconvex Optimization", published in NeurIPS 2018.
Here, we provide a rudimentary annotated proof outline on top of the outline they provide. A more in-depth look and proof of the lemmas are provided in the other PDF of this github. The interpretable results are provided in the slides and video presentation (although we briefly do touch upon the results here as well).

# Appendix

## A Proof of Theorem 1

The first proof is to demonstrate that ADAM converges to delta stationarity with certain assumptions.
The proof assumes that beta1 = 0, which demonstrates the result of RMSProp, and b = 1, which we compute the stochastic gradient of mini-batch =1

We analyze the convergence of ADAM for general minibatch size here. Theorem 1 is obtained by setting $b = 1$. Recall that the update of ADAM is the following

$$x_{t+1,i} = x_{t,i} - \eta_t \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon},$$

for all $i \in [d]$. Since the function $f$ is $L$-smooth, we have the following:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta_t \sum_{i=1}^{d} \left( [\nabla f(x_t)]_i \times \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \frac{g_{t,i}^2}{(\sqrt{v_{t,i}} + \epsilon)^2} \qquad (2) \end{aligned}$$

L-smooth definition

expand inner product

The second step follows simply from ADAM's update. We take the expectation of $f(x_{t+1})$ in the above inequality:

$$\begin{aligned} \mathbb{E}_t[f(x_{t+1})] &\leq f(x_t) - \eta_t \sum_{i=1}^{d} \left( [\nabla f(x_t)]_i \times \mathbb{E}_t \left[ \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} \right] \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}_t \left[ \frac{g_{t,i}^2}{(\sqrt{v_{t,i}} + \epsilon)^2} \right] \\ &= f(x_t) - \eta_t \sum_{i=1}^{d} \left( [\nabla f(x_t)]_i \times \mathbb{E}_t \left[ \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} + \frac{g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} \right] \right) \\ &\quad + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}_t \left[ \frac{g_{t,i}^2}{(\sqrt{v_{t,i}} + \epsilon)^2} \right] \\ &= f(x_t) - \eta_t \sum_{i=1}^{d} \left( [\nabla f(x_t)]_i \times \left[ \frac{[\nabla f(x_t)]_i}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} + \mathbb{E}_t \left[ \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} \right] \right] \right) \\ &\quad + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}_t \left[ \frac{g_{t,i}^2}{(\sqrt{v_{t,i}} + \epsilon)^2} \right] \\ &\leq f(x_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} + \eta_t \sum_{i=1}^{d} |[\nabla f(x_t)]_i| \left| \underbrace{\mathbb{E}_t \left[ \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} \right]}_{T_1} \right| \\ &\quad + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}_t \left[ \frac{g_{t,i}^2}{(\sqrt{v_{t,i}} + \epsilon)^2} \right] \qquad (3) \end{aligned}$$

expectation w.r.t a constant- we assume we currently know x_t and compute expected value of f(x_t+1)

add 0 term

take expectation of this term for the next step. Notice E[g_t] is an unbiased estimate of the true gradient at x_t

don't forget to distribute to the E_t[] term on the right.

The second equality follows from the fact that $g_t$ is an unbiased estimate of $\nabla f(x_t)$ i.e., $\mathbb{E}[g_t] = \nabla f(x_t)$. This is possible because $v_{t-1,i}$ is independent of $S_t$ sampled at time step $t$. The terms $T_1$ in the above inequality needs to be bounded in order to show convergence. We obtain the following bound on the term $T_1$:

$$\begin{aligned} T_1 &= \frac{g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} \\ &\leq |g_{t,i}| \times \left| \frac{1}{\sqrt{v_{t,i}} + \epsilon} - \frac{1}{\sqrt{\beta_2 v_{t-1,i}} + \epsilon} \right| \\ &= \frac{|g_{t,i}|}{(\sqrt{v_{t,i}} + \epsilon)(\sqrt{\beta_2 v_{t-1,i}} + \epsilon)} \times \left| \frac{v_{t,i} - \beta_2 v_{t-1,i}}{\sqrt{v_{t,i}} + \sqrt{\beta_2 v_{t-1,i}}} \right| \\ &= \frac{|g_{t,i}|}{(\sqrt{v_{t,i}} + \epsilon)(\sqrt{\beta_2 v_{t-1,i}} + \epsilon)} \times \frac{(1 - \beta_2) g_{t,i}^2}{\sqrt{v_{t,i}} + \sqrt{\beta_2 v_{t-1,i}}} \end{aligned}$$

take g_t,i out of the parentheses

multiple algebra steps to get from 2 -> 3
1.) multiply by xy/xy = 1
2.) subtract epsilon terms
3.)multiply by (a+b)/(a+b) = 1
(check handwritten notes)

by definition of adam update

The third equality is due to the definition of $v_{t-1,i}$ and $v_{t,i}$ in ADAM i.e., $v_{t,i} = \beta_2 v_{t-1,i} + (1-\beta_2) g_{t,i}^2$. We further bound $T_1$ in the following manner:

$$T_1 \le \frac{|g_{t,i}|}{(\sqrt{v_{t,i}}+\epsilon)(\sqrt{\beta_2 v_{t-1,i}}+\epsilon)} \times \frac{(1-\beta_2) g_{t,i}^2}{\sqrt{\beta_2 v_{t-1,i} + (1-\beta_2) g_{t,i}^2} + \sqrt{\beta_2 v_{t-1,i}}}$$

once again. definition of adam update, but this time in the denominator of term 2

$$\le \frac{1}{(\sqrt{v_{t,i}}+\epsilon)(\sqrt{\beta_2 v_{t-1,i}}+\epsilon)} \times \sqrt{1-\beta_2}\, g_{t,i}^2$$

multiple algebra steps: check (b) in notes

v_t-1 dropped in denom as explained below

$$\le \frac{\sqrt{1-\beta_2}\, g_{t,i}^2}{(\sqrt{\beta_2 v_{t-1,i}}+\epsilon)\epsilon}.$$

Here, the third inequality is obtained by dropping $v_{t,i}$ from the denominator to obtain an upper bound. The second inequality is due to the fact that

If Beta2 = 0, then 1 <= 1
if beta2 --> 1, then we notice that the v_t-1 term goes is bounded by infinity, since 1-beta1 goes to 0.

$$\frac{|g_{t,i}|}{\sqrt{\beta_2 v_{t-1,i} + (1-\beta_2) g_{t,i}^2}} \le \frac{1}{\sqrt{1-\beta_2}}.$$

Note that the bound of coordinates of gradient of $\ell$ automatically provides a bound on $[\nabla f(x_t)]_i$ i.e., $|[\nabla f(x_t)]_i| \le G$ for all $i \in [d]$ . Substituting the above bound on $T_1$ in Equation (3) and using the bound on $[\nabla f(x_t)]_i$, we have the following:

I was first thinking, By finite variance condition. We showed something similar in hw3 - the expected l2 norm of the gradient is bounded by some constant dependent on variance & dimensionality, denoted here as G, but im not sure that's right.

I was thinking this is assumed to be true by using the Lipschitz constant since the loss is L-smooth, but im not too sure, since they dont use L to refer to it.

$$\mathbb{E}_t[f(x_{t+1})] \le f(x_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon} + \frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} \sum_{i=1}^{d} \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon}\right]$$
$$+ \frac{L\eta_t^2}{2\epsilon} \sum_{i=1}^{d} \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{v_{t,i}}+\epsilon}\right]$$

replace grad of f with G as the upper bound, take out sqrt(1-beta2)/epsilon out of the expectation since constants

$$\le f(x_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon} + \frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} \sum_{i=1}^{d} \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon}\right]$$
$$+ \frac{L\eta_t^2}{2\epsilon} \sum_{i=1}^{d} \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon}\right]$$

We replace v_t,i with beta2*v_t-1,i if it occurs in the deinominator, where v_t,i > beta2*v_t-1,i (C - in notes. )
We want to upper bound the expected loss, so a smaller value in the denominator means that our bound still holds.

$$\le f(x_t) - \left(\eta_t - \frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} - \frac{L\eta_t^2}{2\epsilon}\right) \sum_{i=1}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon}$$
$$+ \left(\frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon}\right) \sum_{i=1}^{d} \frac{\sigma_i^2}{b\sqrt{\beta_2 v_{t-1,i}}+\epsilon}.$$

Just algebra/substitution here. These 2 terms come from lemma 3 as well, the negative squared gradient portion. It's originally in the 2nd term, we see that it turns out to have the same terms as the first summation, so they move the 2 coefficients to here.

eta_t -eta_t/4 - eta_t/4 = eta_t/2

The first inequality follows from the fact that $|[\nabla f(x_t)]_i| \le G$. The third inequality follows from Lemma 1. The application of Lemma 1 is possible because $v_{t-1,i}$ is independent of random variables in $|S_t|$. The second inequality is due to the following inequality : $v_{t,i} \ge \beta_2 v_{t-1,i}$. This is obtained from the definition of $v_{t,i}$ in ADAM i.e., $v_{t,i} = \beta_2 v_{t-1,i} + (1-\beta_2) g_{t,i}^2$. From the parameters $\eta_t$, $\epsilon$ and $\beta_2$ stated in our theorem, we see that the following conditions hold: $\frac{L\eta_t}{2\epsilon} \le \frac{1}{4}$ and

$$\frac{G\sqrt{1-\beta_2}}{\epsilon} \le \frac{1}{4}.$$

To show convergence, the authors conveniently picked these parameters to derive constant upper bounds. If you scroll up to their theorem, Their original assumptions were that we pick beta_2, eta, and epsilon such that eta <= (epsilon/(2*L)) and (1-beta_2) <= (epsilon^2)/(16*G^2) In practice, L and G are usually not known. However these assumptions prove the theorem. One thing to note is that in class, we showed that picking 1/L as our stepsize, we guarantee convergence eventually in GD, Thus similarly, an even lower bound to guarantee convergence is saying eta <= epsilon/2*L, where 1/L is clearly greater than epsilon/2*L.
I believe they pick beta2 similarly, in the sense that it's proven to converge, but that lies out of the scope of the paper

Using these inequalities in Equation (3), we obtain

substitute the inequalities

$$\mathbb{E}_t[f(x_{t+1})] \le f(x_t) - \frac{\eta_t}{2} \sum_{i=1}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon} + \left(\frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon}\right) \sum_{i=1}^{d} \frac{\sigma_i^2}{b(\sqrt{\beta_2 v_{t-1,i}}+\epsilon)}$$

After replacing all instances v_t-1,i with upper bound G^2, we average over all our minibatch/samples

$$\le f(x_t) - \frac{\eta_t}{2(\sqrt{\beta_2}G+\epsilon)} \|\nabla f(x_t)\|^2 + \left(\frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon^2} + \frac{L\eta_t^2}{2\epsilon^2}\right) \frac{\sigma^2}{b}$$

variances of individual gradients from our sample

The second inequality follows from the fact that $0 \le v_{t-1,i} \le G^2$. Using telescoping sum and rearranging the inequality, we obtain

by L-smooth?

for the next step,I believe they just get rid of this term since 1/epsilon >= that term. That's where the epsilon^2 comes from.

$$\frac{\eta}{2(\sqrt{\beta_2}G+\epsilon)} \sum_{t=1}^{T} \mathbb{E}\|\nabla f(x_t)\|^2 \le f(x_1) - \mathbb{E}[f(x_{T+1})] + \left(\frac{\eta G \sqrt{1-\beta_2}}{\epsilon^2} + \frac{L\eta^2}{2\epsilon^2}\right) \frac{T\sigma^2}{b}. \quad (4)$$

First, we move the l2 norm squared of the grad to the left hand side, and the expectation of the loss at x_t+1 to the rhs.
Then by telescoping sum, the sum of the expected (I want to say that they omitted the Expectation of it the entire time ,but that doesnt seem likely) l2 norm squared gradients is less than or equal to the sum of the RHS - it's telescoping since E[f(x_t) = f(x_t)], and so we cancel the first 2 terms on the RHS except our first loss evaluation f(x_1), and our last, which is at iteration T+1. The constant terms are added T times of course.

Now, we can get an expression for the averaged expected gradients by multiplying below, and with the definition of stationary point at x*

Multiplying with $\frac{2(\sqrt{\beta_2}G+\epsilon)}{T\eta}$ on both sides and using the fact that $f(x^*) \leq f(x_{t+1})$, we obtain the following:

In green: Our averaged expected l2 squared gradient up to iteration T.

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

which gives us the desired result.

in red: parameters that put the upper bound on the green. Notice that all other terms are essentially constants in terms of bounding our gradient value.

# B Proof of Theorem 2

The proof follows along similar lines as Theorem 1 with some important differences. We, again, analyze the convergence of YOGI for general minibatch size here. Theorem 2 is obtained by setting $b = 1$. We start with the following observation:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1}-x_t\rangle + \frac{L}{2}\|x_{t+1}-x_t\|^2 \\ &= f(x_t) - \eta_t \sum_{i=1}^{d}\left([\nabla f(x_t)]_i \times \frac{g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\frac{g_{t,i}^2}{(\sqrt{v_{t,i}}+\epsilon)^2} \end{aligned} \tag{5}$$

The first step follows from the $L$-smoothness of the function $f$. The second step follows from the definition of YOGI update step i.e.,

$$x_{t+1,i} = x_{t,i} - \eta_t \frac{g_{t,i}}{\sqrt{v_{t,i}}+\epsilon},$$

for all $i \in [d]$. Taking the expectation at time step $t$ in Equation (2), we get the following:

$$\begin{aligned} \mathbb{E}_t[f(x_{t+1})] &\leq f(x_t) - \eta_t \sum_{i=1}^{d}\left([\nabla f(x_t)]_i \times \mathbb{E}_t\left[\frac{g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\right]\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}_t\left[\frac{g_{t,i}^2}{(\sqrt{v_{t,i}}+\epsilon)^2}\right] \\ &= f(x_t) - \eta_t \sum_{i=1}^{d}\left([\nabla f(x_t)]_i \times \mathbb{E}_t\left[\frac{g_{t,i}}{\sqrt{v_{t,i}}+\epsilon} - \frac{g_{t,i}}{\sqrt{v_{t-1,i}}+\epsilon} + \frac{g_{t,i}}{\sqrt{v_{t-1,i}}+\epsilon}\right]\right) \\ &\qquad + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}_t\left[\frac{g_{t,i}^2}{(\sqrt{v_{t,i}}+\epsilon)^2}\right] \\ &\leq f(x_t) - \eta_t \sum_{i=1}^{d}\frac{[\nabla f(x_t)]_i^2}{\sqrt{v_{t-1,i}}+\epsilon} + \eta_t \sum_{i=1}^{d}|[\nabla f(x_t)]_i| \underbrace{\left|\mathbb{E}_t\left[\frac{g_{t,i}}{\sqrt{v_{t,i}}+\epsilon} - \frac{g_{t,i}}{\sqrt{v_{t-1,i}}+\epsilon}\right]\right|}_{T_1} \\ &\qquad + \underbrace{\frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}_t\left[\frac{g_{t,i}^2}{(\sqrt{v_{t,i}}+\epsilon)^2}\right]}_{T_2}. \end{aligned} \tag{6}$$

The second equality follows from the fact that $g_t$ is an unbiased estimate of $\nabla f(x_t)$ i.e., $\mathbb{E}[g_t] = \nabla f(x_t)$. The key difference here in comparison to proof of Theorem 1 is that the deviation to bound in $T_1$ is from $\frac{g_{t,i}}{\sqrt{v_{t-1,i}}+\epsilon}$ as opposed to $\frac{g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}}+\epsilon}$ in proof of ADAM. Our aim is to bound the terms $T_1$ and $T_2$ in the above inequality. We bound the term $T_1$ in the following manner:

$$\begin{aligned} T_1 &\leq |g_{t,i}|\left|\frac{1}{\sqrt{v_{t,i}}+\epsilon} - \frac{1}{\sqrt{v_{t-1,i}}+\epsilon}\right| \\ &= \frac{|g_{t,i}|}{(\sqrt{v_{t,i}}+\epsilon)(\sqrt{v_{t-1,i}}+\epsilon)}\left|\frac{v_{t,i}-v_{t-1,i}}{\sqrt{v_{t,i}}+\sqrt{v_{t-1,i}}}\right| \\ &= \frac{|g_{t,i}|}{(\sqrt{v_{t,i}}+\epsilon)(\sqrt{v_{t-1,i}}+\epsilon)} \times \frac{(1-\beta_2)g_{t,i}^2}{\sqrt{v_{t,i}}+\sqrt{v_{t-1,i}}} \leq \frac{\sqrt{1-\beta_2}\,g_{t,i}^2}{(\sqrt{v_{t-1,i}}+\epsilon)\epsilon}. \end{aligned}$$

The second equality is from the update rule of YOGI which is $v_{t,i} = v_{t-1,i} - (1-\beta_2)\text{sign}(v_{t-1,i} - g_{t,i}^2)g_{t,i}^2$. The last inequality is due to the fact that

$$\frac{|g_{t,i}|}{\sqrt{v_{t,i}} + \sqrt{v_{t-1,i}}} \leq \frac{1}{\sqrt{1-\beta_2}}.$$

The above inequality in turn follows from the fact that either $\frac{|g_{t,i}|}{\sqrt{v_{t-1,i}}} \leq 1$ when $v_{t-1,i} \geq g_{t,i}^2$ or $\frac{|g_{t,i}|}{\sqrt{v_{t,i}}} \leq \frac{1}{\sqrt{1-\beta_2}}$ when $v_{t-1,i} < g_{t,i}^2$. We next bound the term $T_2$ as follows:

$$T_2 = \frac{L\eta_t^2}{2}\sum_{i=1}^d \mathbb{E}_t\left[\frac{g_{t,i}^2}{(\sqrt{v_{t,i}}+\epsilon)^2}\right] \leq \frac{L\eta_t^2}{2\epsilon\sqrt{\beta_2}}\sum_{i=1}^d \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{v_{t-1,i}}+\epsilon}\right].$$

The inequality is due to the following : $v_{t,i} \geq \beta_2 v_{t-1,i}$. To see this, first note that $v_{t,i} = v_{t-1,i} - (1-\beta_2)sign(v_{t-1,i} - g_{t,i}^2)g_{t,i}^2$. If $v_{t-1,i} \leq g_{t,i}^2$, then it is easy to see that $v_{t,i} \geq v_{t-1,i}$. Consider the case where $v_{t-1,i} > g_{t,i}^2$, then we have

$$v_{t,i} = v_{t-1,i} - (1-\beta_2)g_{t,i}^2 \geq \beta_2 v_{t-1,i}.$$

Therefore, $v_{t,i} \geq \beta_2 v_{t-1,i}$. Substituting the above bounds on $T_1$ and $T_2$ in Equation (6), we obtain the following bound:

$$\begin{aligned}
\mathbb{E}_t[f(x_{t+1})] &\leq f(x_t) - \eta_t \sum_{i=1}^d \frac{[\nabla f(x_t)]_i^2}{\sqrt{v_{t-1,i}+\epsilon}} + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon}\sum_{i=1}^d \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{v_{t-1,i}}+\epsilon}\right] \\
&\qquad + \frac{L\eta_t^2}{2\epsilon\sqrt{\beta_2}}\sum_{i=1}^d \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{v_{t-1,i}}+\epsilon}\right] \\
&\leq f(x_t) - \left(\eta_t - \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} - \frac{L\eta_t^2}{2\epsilon\sqrt{\beta_2}}\right)\sum_{i=1}^d \frac{[\nabla f(x_t)]_i^2}{\sqrt{v_{t-1,i}}+\epsilon} \\
&\qquad + \left(\frac{\eta_t G^2(1-\beta_2)}{2\epsilon} + \frac{L\eta_t^2}{2\epsilon\sqrt{\beta_2}}\right)\sum_{i=1}^d \frac{\sigma_i^2}{b\sqrt{v_{t-1,i}}+\epsilon}.
\end{aligned}$$

The first inequality follows from the fact that $|[\nabla f(x_t)]_i| \leq G$. The second inequality follows from Lemma 1. Now, from our theorem result, we observe that,

$$\begin{aligned}
\frac{G\sqrt{1-\beta_2}}{\epsilon} &\leq \frac{1}{4}, \\
\frac{L\eta_t}{2\epsilon\sqrt{\beta_2}} &\leq \frac{1}{4}.
\end{aligned}$$

Using these inequalities in Equation (6), we obtain

$$\begin{aligned}
\mathbb{E}_t[f(x_{t+1})] &\leq f(x_t) - \frac{\eta_t}{2}\sum_{i=1}^d \frac{[\nabla f(x_t)]_i^2}{\sqrt{v_{t-1,i}}+\epsilon} + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon\sqrt{\beta_2}}\right)\sum_{i=1}^d \frac{\sigma_i^2}{b\sqrt{v_{t-1,i}}+\epsilon} \\
&\leq f(x_t) - \frac{\eta_t}{2(\sqrt{2}G+\epsilon)}\|\nabla f(x_t)\|^2 + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon^2} + \frac{L\eta_t^2}{2\epsilon^2\sqrt{\beta_2}}\right)\frac{\sigma^2}{b}
\end{aligned}$$

The second inequality follows from the fact that $0 \leq v_{t-1,i} \leq 2G^2$. Using telescoping sum and rearranging the inequality, we obtain

$$\frac{\eta}{2(\sqrt{2}G+\epsilon)}\sum_{t=1}^T \mathbb{E}\|\nabla f(x_t)\|^2 \leq f(x_1) - \mathbb{E}[f(x_{T+1})] + \left(\frac{\eta G\sqrt{1-\beta_2}}{\epsilon^2} + \frac{L\eta^2}{2\epsilon^2\sqrt{\beta_2}}\right)\frac{T\sigma^2}{b}. \quad (7)$$

Multiplying with $\frac{2(\sqrt{2}G+\epsilon)}{\eta}$ on both sides and using the fact that $f(x^*) \leq f(x_{t+1})$ gives us the desired result.

# C Auxiliary Lemma

The following result is useful for bounding the variance of the updates of the algorithms in this paper.

**Lemma 1.** *For the iterates $x_t$ where $t \in [T]$ in Algorithm 1 and 2, the following inequality holds:*

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

*for all $i \in [d]$.*

*Proof.* Let us define the following notation for the ease of exposition:

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

zeta_t is defined to be the average gradient standard deviation using minibatch S_t. Notice when S_t is our full dataset, we dont have deviation, thus zeta_t = 0. In the next step, squaring this gives our variance.

Using this notation, we obtain the following bound:

$$\mathbb{E}_t[g_{t,i}^2] = \mathbb{E}_t[\|\zeta_t + \nabla f(x_t)\|^2]$$

Just subititute above into here.

$$= \mathbb{E}_t[\zeta_t^2] + [\nabla f(x_t)]_i^2$$

Linearity step- also, right term is a constant/known. The true gradient squared isn't a R.V.

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

substitute, take out the b^2, minibatch size

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

by lemma 2, the expectation of a sum squared with independent r.v.'s and expectation 0
(since E[r.v.] - truegrad = truegrad - truegrad=0)
is equivalent to taking the expectation of the squared sums

$$\leq \frac{\sigma_i^2}{b} + [\nabla f(x_t)]_i^2.$$

The inner sum is exactly the definition of variance. Then the gradient squared at dimension i is bounded by the sum of the sample variances at dimension i (which are all the same, so we get b*sigma_i^2 and then we divide by one of the b's.)

The second equality is due to the fact that $\zeta_t$ is a mean 0 random variable. The third equality follows from Lemma 2. The last inequality is due to the fact that $\mathbb{E}_{s \sim \mathbb{P}}[([\nabla \ell(x_t, s)]_i - [\nabla f(x_t)]_i)^2] \leq \sigma_i^2$ for all $x \in \mathbb{R}^d$. ◻

**Lemma 2.** *For random variables $z_1, \ldots, z_r$ are independent and mean 0, we have*

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

*Proof.* We have the following:

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

The second equality follows from the fact that $z_i$'s are independent and mean 0. ◻