A. Intuition - Power-Law Distribution

Our goal is to compress the auxiliary variables without incurring significant accuracy loss. Unfortunately, selecting the appropriate compression scheme is not clear without any additional information on the parameter distribution. The challenge is that the parameter distribution can change over time, so any static assumption on the approximation is likely to hurt accuracy. Fortunately, in this section we show that there is a potential solution.

In Figure 1, we plot the auxiliary variables sorted according to their normalized absolute values at training epochs 5, 20 and 40. The plots clearly indicate a power-law behavior where only a few parameters have large magnitudes. In Figure 2, we confirm this behavior for every iteration by plotting the midpoint dividing the head and tail. The auxiliary variables have long tails throughout the training process. To the best of our knowledge, this is the first work that empirically shows the existence of a power-law distribution behavior in the gradients and auxiliary variables while training. To dig deeper, we also show the identities of top-100 parameters (the head of power law distribution). The top-k identities change over time, so it is difficult to cluster parameters into predefined, static clusters.

In summary, we need to compress a power law distribution where the top-k identities are constantly changing. Fortunately, the auxiliary variables are updated in a linear fashion. The linear sequence of updates allows us to guarantee that the count-sketch provides an accurate estimate with high probability for each iteration during training. The power law distribution and linear updates make the count-sketch an ideal data structure for this problem.

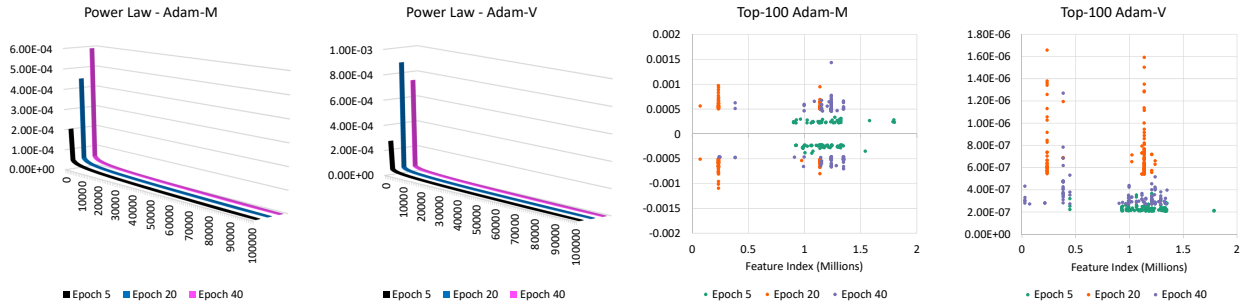

Figure 1. The optimizer's auxiliary variables follow a power-law distribution, but the features associated with top-k values change during training. The X-Axis is the feature ID, while the Y-Axis is the normalized absolute magnitude. The first two charts show the sorted absolute values for the auxiliary variables at different training epochs. The last two charts plot the top 100 features and their magnitudes. We plot the 1st and 2nd moments of the Adam Optimizer for an LSTM weight matrix trained on the Wikitext-2 dataset.

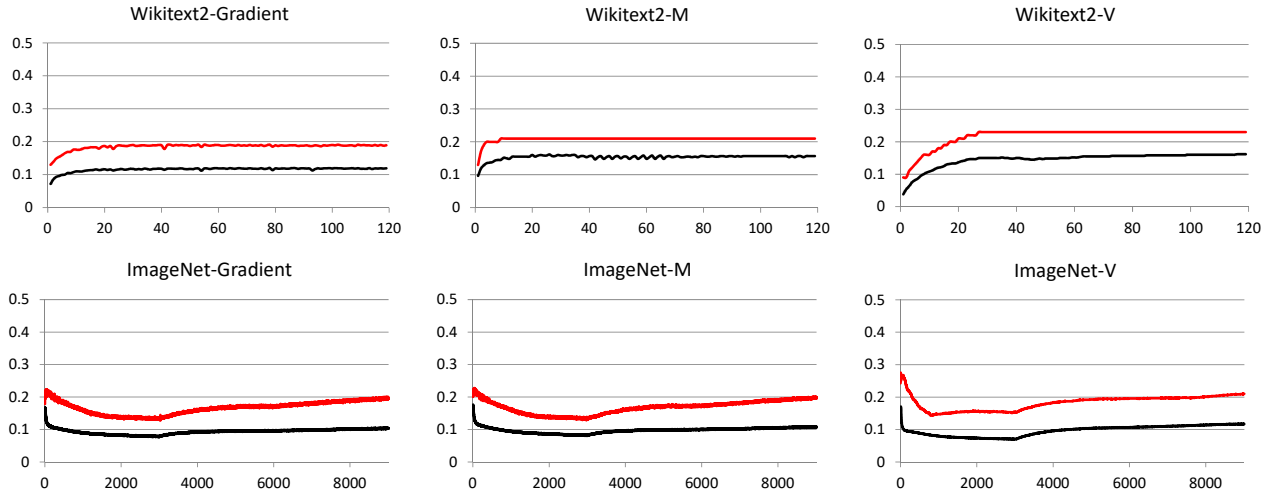

Figure 2. An experiment showing that the model's gradients and the optimizer's auxiliary variables follow a power-law distribution. The X-axis is the number of iterations during training time. The Y-axis is the 50% threshold that marks the midpoint dividing the head and the tail of the distribution. For a uniform distribution, the midpoint is at 0.5. However, the 50% threshold for the gradients and auxiliary variables is less than 0.2 on average, indicating that they follow a power law distribution. The red line marks the maximum threshold for all layers, while the black line represents the average threshold.

B. Proof - Convergence

Count-Sketch Error Bound: (Charikar et al., 2002) Let $\hat{x}_i$ be the Count-Sketch estimate of component i from vector x. For any component x_i, with probability $1-\delta$, a Count-Min Sketch matrix with width $\Theta(\frac{1}{\epsilon_1^2})$ and depth $\Theta(\log(\frac{d}{\delta}))$ satisfies

$$x_i - \epsilon_1 \|x\|_2 \leq \hat{x}_i \leq x_i + \epsilon_1 \|x\|_2 \tag{1}$$

Count-Min Sketch Error Bound: (Cormode & Muthukrishnan, 2005) Let $\hat{x}_i$ be the Count-Min Sketch estimate of component i from vector x. For any component x_i, with probability $1-\delta$, a Count-Min Sketch matrix with width $\Theta(\frac{1}{\epsilon_1})$ and depth $\Theta(\log(\frac{d}{\delta}))$ satisfies

$$x_i \leq \hat{x}_i \leq x_i + \epsilon_1 \|x\|_1 \tag{2}$$

For stochastic non-convex optimization, we measure how the algorithm converges to a stationary point - $\|\nabla f(x_t)\|^2 \leq c$ for some constant c. Notation: batch size b, learning rate η_t, 2nd moment decay rate β_2, count-min sketch error rate ϵ_1, count-min sketch failure probability δ.

Assumptions: Here are the assumptions used in our analysis:

1. Function f is L-Smooth - There exists a constant L such that $\|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\|, \ \forall x, y \in \mathbb{R}^d$
2. Function f has bounded gradients - $[f(x_t)]_i \leq G_i, \ \forall x \in \mathbb{R}^d, i \in [d], \quad G = \|G\|_\infty$
3. The stochastic gradient oracle provides us with an unbiased estimate with fixed variance. Let ξ_t represents the randomness (due to mini-batch sampling) at iteration t.

$$g_{t,i} = [\nabla f(x_t, \xi_t)]_i, \quad \mathbb{E}[g_{t,i}] = [\nabla f(x_t)]_i, \quad \mathbb{E}[(g_{t,i} - [\nabla f(x_t)]_i)^2] \leq \sigma_i$$

For simplicity and to save additional memory by not tracking the 1st moment, let $\beta_1 = 0$. In this form, the optimizer is commonly called RMSPROP. Therefore, the update rule for all $i \in [d]$ is

$$x_{t+1,i} = x_{t,i} - \eta_t \frac{g_{t,i}}{\sqrt{\hat{v}_{t,i}} + \epsilon}, \tag{3}$$

where $\hat{v}_{t,i}$ represents the Count-Min Sketch estimate of component i from vector $v_t = v_{t-1} + (1-\beta_2)(v_{t-1} - g_t^2)$.

Theorem B.1. *Let learning rate $\eta_t = \eta, \forall t \in [T]$ and batch size $b = 1$. Assume β_2, η, and ϵ are selected such that $\eta \leq \frac{\epsilon}{2L}$ and $\sqrt{1-\beta_2} \leq \frac{\epsilon}{4G}$. Given a Count-Min Sketch matrix width $\Theta(\frac{1}{\epsilon_1})$ and depth $\Theta(\log(\frac{dT}{\delta}))$, we have the following bound that holds for Count-Min Sketch Adam with probability $(1-\delta)$*

$$\min_t \mathbb{E}\left[\|\nabla f(x_t)\|^2 \right] \leq O\left(\frac{f(x_0) - f(x_*)}{\eta T} + \sigma^2 + \epsilon_1 d \|G\|_2^2 \right)$$

Proof. Given that the function is L-smooth and by the optimizer update rule, we derive the following:

$$\begin{aligned} f(x_{t+1}) &= f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta_t \sum_{i=0}^{d} \left([\nabla f(x_t)]_i \cdot \frac{g_{t,i}}{\sqrt{\hat{v}_{t,i}} + \epsilon} \right) + \frac{L\eta_t^2}{2} \sum_{i=0}^{d} \frac{g_{t,i}^2}{(\sqrt{\hat{v}_{t,i}} + \epsilon)^2} \end{aligned} \tag{4}$$

Next, we take the expectation of $f(x_{t+1})$, given we that know x_t (assumed fixed):

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) \mid x_t\right] &\leq f(x_t) - \eta_t \sum_{i=0}^{d} \left([\nabla f(x_t)]_i \cdot \mathbb{E}_t \left[\frac{g_{t,i}}{\sqrt{\hat{v}_{t,i}} + \epsilon} \;\middle|\; x_t \right] \right) + \frac{L\eta_t^2}{2} \sum_{i=0}^{d} \mathbb{E}_t \left[\frac{g_{t,i}^2}{(\sqrt{\hat{v}_{t,i}} + \epsilon)^2} \;\middle|\; x_t \right] \\
&= f(x_t) - \eta_t \sum_{i=0}^{d} \left([\nabla f(x_t)]_i \cdot \mathbb{E}_t \left[\frac{g_{t,i}}{\sqrt{\hat{v}_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} + \frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \;\middle|\; x_t \right] \right) \\
&\quad + \frac{L\eta_t^2}{2} \sum_{i=0}^{d} \mathbb{E}_t \left[\frac{g_{t,i}^2}{(\sqrt{\hat{v}_{t,i}} + \epsilon)^2} \;\middle|\; x_t \right] \\
&= f(x_t) - \eta_t \sum_{i=0}^{d} \left([\nabla f(x_t)]_i \cdot \left[\frac{[\nabla f(x_t)]_i}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} + \mathbb{E}_t \left[\frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \;\middle|\; x_t \right] \right] \right) \\
&\quad + \frac{L\eta_t^2}{2} \sum_{i=0}^{d} \mathbb{E}_t \left[\frac{g_{t,i}^2}{(\sqrt{\hat{v}_{t,i}} + \epsilon)^2} \;\middle|\; x_t \right] \\
&\leq f(x_t) - \eta_t \sum_{i=0}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} + \eta_t \sum_{i=0}^{d} \left| [\nabla f(x_t)]_i \right| \cdot \left| \mathbb{E}_t \left[\underbrace{\frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon}}_{T_1} \;\middle|\; x_t \right] \right| \\
&\quad + \frac{L\eta_t^2}{2} \sum_{i=0}^{d} \mathbb{E}_t \left[\frac{g_{t,i}^2}{(\sqrt{\hat{v}_{t,i}} + \epsilon)^2} \;\middle|\; x_t \right]
\end{aligned}$$

The second equality occurs because $g_{t,i}$ is an unbiased estimate of $[\nabla f(x_t)]_i$.

Now, we upper-bound the term T_1:

$$\begin{aligned}
T_1 &= \frac{g_{t,i}}{\sqrt{\hat{v}_{t,i}} + \epsilon} - \frac{g_{t,i}}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \\
&\leq |g_{t,i}| \cdot \left| \frac{1}{\sqrt{\hat{v}_{t,i}} + \epsilon} - \frac{1}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \right| \\
&= \frac{|g_{t,i}|}{(\sqrt{\hat{v}_{t,i}} + \epsilon)(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \cdot \left| \frac{\hat{v}_{t,i} - \beta_2 \hat{v}_{t-1,i}}{\sqrt{\hat{v}_{t,i}} + \sqrt{\beta_2 \hat{v}_{t-1,i}}} \right| \\
&= \frac{|g_{t,i}|}{(\sqrt{\hat{v}_{t,i}} + \epsilon)(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \cdot \left| \frac{(1-\beta_2)\hat{g}_{t,i}^2}{\sqrt{\hat{v}_{t,i}} + \sqrt{\beta_2 \hat{v}_{t-1,i}}} \right| \\
&\leq \frac{|g_{t,i}|}{(\sqrt{\hat{v}_{t,i}} + \epsilon)(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \cdot \left| \frac{(1-\beta_2)(g_{t,i}^2 + \epsilon_1 \left\| g_t^2 \right\|_1)}{\sqrt{\hat{v}_{t,i}} + \sqrt{\hat{v}_{t-1,i}}} \right| \\
&\leq \frac{\sqrt{1-\beta_2}(g_{t,i}^2 + \epsilon_1 \left\| g_t^2 \right\|_1)}{(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)\epsilon}
\end{aligned}$$

From Lemma B.3, we have the second equality. The second inequality occurs because of Lemma B.2, which is derived using the Count-Min Sketch error bound. The third inequality occurs because $\frac{|g_{t,i}|}{\sqrt{\hat{v}_{t,i}} + \sqrt{\hat{v}_{t-1,i}}} \leq \frac{1}{\sqrt{1-\beta_2}}$ and when we drop $\sqrt{\hat{v}_{t,i}}$ from $(\sqrt{\hat{v}_{t,i}} + \epsilon)$.

By substituting the upper-bound for T_1, we arrive at the following:

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) \mid x_t\right] &\leq f(x_t) - \eta_t \sum_{i=0}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} \sum_{i=0}^{d} \left(\mathbb{E}_t\left[\frac{g_{t,i}^2 + \epsilon_1 \left\| g_t^2 \right\|_1}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \;\middle|\; x_t \right] \right) \\
&\quad + \frac{L\eta_t^2}{2\epsilon} \sum_{i=0}^{d} \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{\hat{v}_{t,i}} + \epsilon} \;\middle|\; x_t \right] \\
&\leq f(x_t) - \eta_t \sum_{i=0}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \\
&\quad + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} \sum_{i=0}^{d} \left(\mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \;\middle|\; x_t \right] + \mathbb{E}_t\left[\frac{\epsilon_1 \left\| g_t^2 \right\|_1}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \;\middle|\; x_t \right] \right) \\
&\quad + \frac{L\eta_t^2}{2\epsilon} \sum_{i=0}^{d} \mathbb{E}_t\left[\frac{g_{t,i}^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \;\middle|\; x_t \right] \\
&\leq 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=0}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \\
&\quad + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon} \right) \sum_{i=0}^{d} \frac{\sigma_i^2}{b(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \\
&\quad + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} \sum_{i=0}^{d} \frac{\epsilon_1 \left(\sum_{j=0}^{d} \frac{\sigma_j^2}{b} + [\nabla f(x_t)]_j^2 \right)}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \\
&\leq 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=0}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \\
&\quad + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon} \right) \sum_{i=0}^{d} \frac{\sigma_i^2}{b(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} \sum_{i=0}^{d} \frac{\epsilon_1 (\frac{\sigma^2}{b} + \|G\|_2^2)}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon}
\end{aligned}$$

The first inequality follows because the function has bounded gradients - $[\nabla f(x_t)]_i \leq G$. Now, the second inequality holds because $\hat{v}_{t,i} \geq \beta_2 \hat{v}_{t-1,i}$. In addition, we split the $g_{t,i}^2$ and $\epsilon_1 \left\| g_t^2 \right\|_1$ terms using the linearity of expectation. For the third inequality, we use the result and definitions in Lemma B.1. From the specified parameters for η_t, β_2, and ϵ, we assume the following conditions hold: $\frac{G\sqrt{1-\beta_2}}{\epsilon} \leq \frac{1}{4}$ and $\frac{L\eta_t}{2\epsilon} \leq \frac{1}{4}$.

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) \mid x_t\right] &\leq f(x_t) - \frac{\eta_t}{2} \sum_{i=0}^{d} \frac{[\nabla f(x_t)]_i^2}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon} \right) \sum_{i=0}^{d} \frac{\sigma_i^2}{b(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \\
&\quad + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} \sum_{i=0}^{d} \frac{\epsilon_1 (\frac{\sigma^2}{b} + \|G\|_2^2)}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \\
&\leq f(x_t) - \frac{\eta_t}{2(\sqrt{\beta_2 \epsilon_1 d} G + \epsilon)} \|\nabla f(x_t)\|^2 + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon^2} + \frac{L\eta_t^2}{2\epsilon^2} \right) \frac{\sigma^2}{b} \\
&\quad + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon^2} \left(\frac{\epsilon_1 d \sigma^2}{b} + \epsilon_1 d \|G\|_2^2 \right) \\
&= f(x_t) - \frac{\eta_t}{2(\sqrt{\beta_2 \epsilon_1 d} G + \epsilon)} \|\nabla f(x_t)\|^2 + \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon^2} (1 + \epsilon_1 d) + \frac{L\eta_t^2}{2\epsilon^2} \right) \frac{\sigma^2}{b} \\
&\quad + \frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon^2} (\epsilon_1 d \|G\|_2^2)
\end{aligned}$$

For the standard optimizer, $0 \leq v_{t-1,i} \leq G^2$. For the Count-Min Sketch approximation, $\|v_{t-1}\|_1 = \sum_{i=0}^{d} |v_{t-1,i}| \leq \sum_{i=0}^{d} G^2 = dG^2$. Therefore, this inequality holds $0 \leq v_{t-1,i} \leq \hat{v}_{t-1,i} \leq v_{t-1,i} + \epsilon_1 \|v_{t-1}\|_1 \leq \epsilon_1 dG^2$. In addition, this corollary follows $\frac{1}{\beta_2\sqrt{\epsilon_1 d}G+\epsilon} \leq \frac{1}{\epsilon}$. The second inequality follows given the two inequalities for the Count-Min Sketch approximation.

Now, we take a telescoping sum over all the iterations, and taking the full expectation:

$$\frac{\eta}{2\left(\sqrt{\beta_2\epsilon_1 d}G+\epsilon\right)} \sum_{t=0}^{T} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq f(x_0) - \mathbb{E}[f(x_{T+1})] + \left(\frac{\eta G\sqrt{1-\beta_2}}{\epsilon^2}(1+\epsilon_1 d) + \frac{L\eta^2}{2\epsilon^2}\right)\frac{T\sigma^2}{b} + \frac{T\eta G\sqrt{1-\beta_2}}{\epsilon^2}(\epsilon_1 d \|G\|_2^2)$$

Finally, given that $f(x_*) \leq f(x_{T+1})$ and by multiplying the equation with $\frac{2(\sqrt{\beta_2\epsilon_1 d}G+\epsilon)}{\eta T}$, we arrive at our final result.

$$\frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq 2\left(\sqrt{\beta_2\epsilon_1 d}G+\epsilon\right) \cdot \left[\frac{f(x_0)-f(x_*)}{\eta T} + \left(\frac{G\sqrt{1-\beta_2}}{\epsilon^2}(1+\epsilon_1 d) + \frac{L\eta}{2\epsilon^2}\right)\frac{\sigma^2}{b} + \frac{G\sqrt{1-\beta_2}}{\epsilon^2}(\epsilon_1 d \|G\|_2^2)\right]$$

□

Lemma B.1. *(Zaheer et al., 2018) For all $i \in [d]$ and for the iterates x_t where $t \in [T]$ for Count-Min Sketch Adam, the following inequality holds:*

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

Lemma B.2. *For all $i \in [d]$ and for the iterates x_t where $t \in [T]$ for Count-Min Sketch Adam, the following inequality holds:*

$$\hat{v}_{t,i} - \beta_2 \hat{v}_{t,i} \leq (1-\beta_2)(g_{t,i}^2 + \|g_t\|_1)$$

Proof. Given the error bound for the count-min sketch and the Adam update rule, we have the following:

$$\begin{aligned}
\hat{v}_{t,i} &\leq v_{t,i} + \epsilon_1 \|v_t\|_1 \\
&= v_{t,i} + \epsilon_1 \sum_{i=0}^{d} |v_t, i| \\
&= \beta_2 v_{t,i} + (1-\beta_2) g_{t,i}^2 + \epsilon_1 \sum_{i=0}^{d} \left|\beta_2 v_{t,i} + (1-\beta_2) g_{t,i}^2\right| \\
&= \beta_2 v_{t,i} + (1-\beta_2) g_{t,i}^2 + \epsilon_1 (\sum_{i=0}^{d} |\beta_2 v_{t,i}| + \sum_{i=0}^{d} \left|(1-\beta_2) g_{t,i}^2\right|) \\
&= \beta_2 v_{t,i} + (1-\beta_2) g_{t,i}^2 + \epsilon_1 \beta_2 \|v_{t-1}\|_1 + \epsilon_1 (1-\beta_2) \|g_t\|_1
\end{aligned}$$

By subtracting $\hat{v}_{t,i} - \beta_2 \hat{v}_{t,i}$ and simplifying, we derive the desired inequality.

$$\begin{aligned}
\hat{v}_{t,i} - \beta_2 \hat{v}_{t,i} &\leq \beta_2 v_{t,i} + (1-\beta_2) g_{t,i}^2 + \epsilon_1 \beta_2 \|v_{t-1}\|_1 + \epsilon_1 (1-\beta_2) \|g_t\|_1 - \beta_2 v_{t,i} - \beta_2 \epsilon_1 \|v_{t-1}\|_1 \\
&= (1-\beta_2) g_{t,i}^2 + \epsilon_1 (1-\beta_2) \|g_t\|_1 \\
&= (1-\beta_2)(g_{t,i}^2 + \|g_t\|_1)
\end{aligned}$$

□

Lemma B.3. *For all $i \in [d]$ and for the iterates x_t where $t \in [T]$ for Count-Min Sketch Adam, the following equality holds:*

$$|g_{t,i}| \cdot \left| \frac{1}{\sqrt{\hat{v}_{t,i}} + \epsilon} - \frac{1}{\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon} \right| = \frac{|g_{t,i}|}{(\sqrt{\hat{v}_{t,i}} + \epsilon)(\sqrt{\beta_2 \hat{v}_{t-1,i}} + \epsilon)} \cdot \left| \frac{\hat{v}_{t,i} - \beta_2 \hat{v}_{t-1,i}}{\sqrt{\hat{v}_{t,i}} + \sqrt{\beta_2 \hat{v}_{t-1,i}}} \right|$$

Proof. Let $x = g_{t,i}$, $A = \sqrt{\hat{v}_{t,i}}$, and $B = \sqrt{\beta_2 \hat{v}_{t-1,i}}$

$$\begin{aligned}
|x| \cdot \left| \frac{1}{A+\epsilon} - \frac{1}{B+\epsilon} \right| &= |x| \cdot \left| \frac{B-A}{(A+\epsilon)(B+\epsilon)} \right| \\
&= |x| \cdot \left| \frac{A-B}{(A+\epsilon)(B+\epsilon)} \right| \\
&= \frac{|x|\,(A+B)}{(A+B)} \cdot \left| \frac{A-B}{(A+\epsilon)(B+\epsilon)} \right| \\
&= \frac{|x|\,(A+B)(A-B)}{(A+B)(A+\epsilon)(B+\epsilon)} \\
&= \frac{|x|\,(A^2-B^2)}{(A+B)(A+\epsilon)(B+\epsilon)} \\
&= \frac{|x|}{(A+\epsilon)(B+\epsilon)} \cdot \left| \frac{A^2-B^2}{A+B} \right|
\end{aligned}$$

□

References

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *Intl. Colloquium on Automata, Languages, and Programming*, pp. 693–703. Springer, 2002.

Cormode, G. and Muthukrishnan, S. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. Adaptive methods for nonconvex optimization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9815–9825. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8186-adaptive-methods-for-nonconvex-optimization.pdf.