APPENDIX OF THE PAPER "CONVERGENCE ANALYSIS OF ADAM WITH CONSTANT STEP SIZE IN NON-CONVEX SETTING: A SIMPLE PROOF"

Alokendu Mazumder, Bhartendu Kumar, Manan Tayal*, Punit Rathore*

Indian Institute of Science Bangalore, Bengaluru, India

Appendix

To Prove. *The term* $\left((1-\beta_1)\lambda_{min}(\boldsymbol{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\boldsymbol{A}_t)}{\sigma}\right)$ is non-negetive.

Proof. We can construct a lower bound on $\lambda_{min}(\mathbf{A}_t)$ and an upper bound on $\lambda_{min}(\mathbf{A}_t)$ as follows:

$$\lambda_{min}(\mathbf{A}_t) \geq \frac{1}{\epsilon + \sqrt{\underset{1\leq j\leq|\mathbf{v}_t|}{max}(\mathbf{v}_t)_j}} \tag{1}$$

$$\lambda_{max}(\mathbf{A}_t) \leq \frac{1}{\epsilon + \sqrt{\underset{1\leq j\leq|\mathbf{v}_t|}{min}(\mathbf{v}_t)_j}} \tag{2}$$

We remember that $\mathbf{v}_t$ can be rewritten as $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)(\nabla\mathcal{L}(\mathbf{w}_t))^2$, solving this recursion and defining $\rho_t = \underset{1\leq j\leq t, 1\leq k\leq|\mathbf{v}_t|}{min}(\nabla\mathcal{L}(\mathbf{W}_j)^2)_k$ and taking $\gamma_{t-1} = \gamma_t = \gamma$ we have:

$$\lambda_{min}(\mathbf{A}_t) \geq \frac{1}{\epsilon + \sqrt{(1-\beta_2^t)\gamma^2}}$$

$$\lambda_{max}(\mathbf{A}_t) \leq \frac{1}{\epsilon + \sqrt{(1-\beta_2^t)\rho_t}}$$

Where, $\gamma_{t-1} = \underset{1\leq j\leq t-1}{max}\|\nabla\mathcal{L}(\mathbf{w}_j)\|_2$, and $\forall j \in \{1, 2, \ldots t-1\}$. Setting $\rho_t = 0$, we can rewrite the term $\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right)$ as:

$$\begin{aligned}\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right) &\geq \left(\frac{(1-\beta_1)}{\epsilon+\gamma\sqrt{(1-\beta_2^t)}} - \frac{(\beta_1-\beta_1^t)\gamma}{\epsilon\sigma}\right) \\ &\geq \frac{\epsilon\sigma(1-\beta_1) - \gamma(\beta_1-\beta_1^t)(\epsilon+\gamma\sqrt{(1-\beta_2^t)})}{\epsilon\sigma(\epsilon+\gamma\sqrt{(1-\beta_2^t)})} \\ &\geq \gamma(\beta_1-\beta_1^t)\frac{\epsilon\left(\frac{\sigma(1-\beta_1)}{\gamma(\beta_1-\beta_1^t)}-1\right) - \gamma\sqrt{(1-\beta_2^t)}}{\epsilon\sigma(\epsilon+\gamma\sqrt{(1-\beta_2^t)})} \\ &\geq \gamma(\beta_1-\beta_1^t)\left(\frac{\sigma(1-\beta_1)}{\gamma(\beta_1-\beta_1^t)}-1\right)\frac{\epsilon - \left(\frac{\gamma\sqrt{(1-\beta_2^t)}}{\frac{(1-\beta_1\sigma)}{(\beta_1-\beta_1^t)\gamma}-1}\right)}{\epsilon\sigma(\epsilon+\gamma\sqrt{(1-\beta_2^t)})}\end{aligned} \tag{3}$$

*Equal second place/author contribution (alphabetical ordering)

By definition $\beta_1 \in (0,1)$ and hence $(\beta_1 - \beta_1^t) \in (0, \beta_1)$. This implies that $\frac{(1-\beta_1)\sigma}{(\beta_1-\beta_1^t)\gamma} > \frac{(1-\beta_1)\sigma}{\beta_1\gamma} > 1$ where the last inequality follows due to the choice of σ as stated in the beginning of this theorem. This allows us to define a constant $\frac{(1-\beta_1)\sigma}{\beta_1\gamma} - 1 := \psi_1 > 0$ such that $\frac{(1-\beta_1)\sigma}{(\beta_1-\beta_1^t)\gamma} - 1 > \psi_1$. Similarly, our definition of delta allows us to define another constant $\psi_2 > 0$ to get:

$$\left(\frac{\gamma\sqrt{(1-\beta_2^t)}}{\frac{(1-\beta_1\sigma)}{(\beta_1-\beta_1^t)\gamma} - 1} \right) \quad < \quad \frac{\gamma}{\psi_1} = \epsilon - \psi_2 \tag{4}$$

Putting Eq.(4) in Eq.(3), we get:

$$\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1 - \beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma} \right) \geq \qquad \left(\frac{\gamma(\beta_1 - \beta_1^2)\psi_1\psi_2}{\epsilon\sigma(\epsilon+\sigma)} \right) = c > 0$$

$\square$