Supplementary materials for “Nostalgic Adam: Weighting more of the past gradients when designing the adaptive learning rate”

Haiwen Huang
School of Mathematical Sciences
Peking University, Beijing, 100871
smshhw@pku.edu.cn

Chang Wang
School of Mathematical Sciences
Peking University, Beijing, 100871
1500010660@pku.edu.cn

Bin Dong
Beijing International Center for Mathematical Research, Peking University
Center for Data Science, Peking University
Beijing Institute of Big Data Research
Beijing, China
dongbin@math.pku.edu.cn

Introudction

In this supplementary material, we use the same notations as in the paper “Nostalgic Adam: Weighting more of the past gradients when designing the adaptive learning rate”. We are going to prove a more general convergence theorem. In our original paper, we propose NosAdam, as shown in Algorithm 1. But in fact, NosAdam can be considered as a particular case of a more general algorithm, in which we replaces g_t^2 in the calculation of v_t by g_p, and $v_t^{1/2}$ in the update equation by $v_t^{1/p}$. We call this algorithm p-NosAdam, as shown in Algorithm 2. NosAdam is p-NosAdam when $p = 2$.

In the remaining part of this material, we are going to prove the convergence theorem of p-NosAdam when $p > 1$. From Theorem 1, we can see that the regret bound is $O(T^{\max(\frac{1}{p},\frac{p-1}{p})})$.

Convergence of p-NosAdam

Theorem 1 (Convergence of p-NosAdam)**.** *Let B_t and b_k be the sequences defined in p-NosAdam, $\alpha_t = \alpha/t^{1/p}, p > 1$, $\beta_{1,1} = \beta_1, \beta_{1,t} \leq \beta_1$ for all t. Assume that $\mathscr{F}$ has bounded diameter D_∞ and $||\nabla f_t(x)||_\infty \leq G_\infty$ for all t and $x \in \mathscr{F}$. Furthermore, let $\beta_{2,t}$ be such that the following conditions are satisfied:*

$$1. \frac{B_t}{t} \leq \frac{B_{t-1}}{t-1}$$

$$2. \frac{B_t}{tb_t^p} \geq \frac{B_{t-1}}{(t-1)b_{t-1}^p}$$

Then for $\{x_t\}$ generated using p-NosAdam, we have the following bound on the regret

$$R_T \leq \frac{D_\infty^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} T^{\frac{1}{p}} v_{T,i}^{\frac{1}{p}} + \frac{D_\infty^2}{2(1-\beta_1)} \sum_{t=1}^{T}\sum_{i=1}^{d} \frac{\beta_{1,t} v_{t,i}^{\frac{1}{p}}}{\alpha_t} + \frac{\alpha(\beta_1+1)}{(1-\beta_1)^3} \sum_{i=1}^{d} (\sum_{t=1}^{T} b_t g_{t,i}^p)^{\frac{p-1}{p}} (\frac{B_T}{Tb_T^p})^{\frac{1}{p}}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Algorithm 1 Nostalgic Adam Algorithm

```
Input: $x \in F$, $m_0 = 0$, $V_0 = 0$
for $t = 1$ to $T$ do
    $g_t = \nabla f_t(x_t)$
    $\beta_{2,t} = B_{t-1}/B_t$, where $B_t = \sum_{k=1}^{t} b_k$ for $t \geq 1$, $b_k \geq 0$, and $B_0 = 0$
    $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
    $V_t = \beta_{2,t} V_{t-1} + (1-\beta_{2,t}) g_t^2$
    $\hat{x}_{t+1} = x_t - \alpha_t m_t / \sqrt{V_t}$
    $x_{t+1} = \mathcal{P}_{\mathcal{F}, \sqrt{V_t}}(\hat{x}_{t+1})$
end for
```

Algorithm 2 p-NosAdam Algorithm

```
Input: $x \in F$, $m_0 = 0$, $V_0 = 0$
for $t = 1$ to $T$ do
    $g_t = \nabla f_t(x_t)$
    $\beta_{2,t} = B_{t-1}/B_t$, where $B_t = \sum_{k=1}^{t} b_k$ for $t \geq 1$, $b_k \geq 0$, and $B_0 = 0$
    $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
    $V_t = \beta_{2,t} V_{t-1} + (1-\beta_{2,t}) g_t^p$ for $p > 1$
    $\hat{x}_{t+1} = x_t - \alpha_t m_t / V_t^{1/p}$
    $x_{t+1} = \mathcal{P}_{\mathcal{F}, V_t^{1/p}}(\hat{x}_{t+1})$
end for
```

Proof of Theorem 1:

Recall that

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{F}} \sum_{t=1}^{T} f_t(x). \tag{1}$$

Let $x^* = \text{argmin}_{x \in \mathcal{F}} \sum_{t=1}^{T} f_t(x)$. Therefore $R_T = \sum_{t=1}^{T} f_t(x_t) - f_t(x^*)$.

To prove this theorem, we will use the following lemmas.

Lemma 2.

$$\begin{aligned}
\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \sum_{t=1}^{T} [& \frac{1}{2\alpha_t(1-\beta_{1t})}(||V_t^{1/2p}(x_t - x^*)||^2 \\
& - ||V_t^{1/2p}(x_{t+1} - x^*)||^2) + \frac{\alpha_t}{2(1-\beta_{1t})}||V_t^{-1/2p} m_t||^2 \\
& + \frac{\beta_{1t}}{2(1-\beta_{1t})}\alpha||V_t^{-1/2p} m_{t+1}||^2 + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}||V_t^{1/2p}(x_t - x^*)||^2]
\end{aligned}$$

Proof of Lemma 2:

We begin with the following observation:

$$x_{t+1} = \Pi_{\mathscr{F}, V_t^{1/p}}(x_t - \alpha_t V_t^{-1/p} m_t) = \min_{x \in \mathscr{F}} ||V_t^{1/2p}(x - (x_t - \alpha_t V_t^{-1/p} m_t))||$$

Using Lemma 4 in [1] with $u_1 = x_{t+1}$ and $u_2 = x^*$, we have the following:

$$\begin{aligned}
||V_t^{1/2p}(x_{t+1} - x^*)||^2 &\leq ||V_t^{1/2p}(x_t - \alpha_t V_t^{-1/p} m_t - x^*)||^2 \\
&= ||V_t^{1/2p}(x_t - x^*)||^2 + \alpha_t^2 ||V_t^{-1/2p} m_t||^2 - 2\alpha_t \langle m_t, x_t - x^* \rangle \\
&= ||V_t^{1/2p}(x_t - x^*)||^2 + \alpha_t^2 ||V_t^{-1/2p} m_t||^2 \\
&\qquad - 2\alpha_t \langle \beta_{1t} m_{t-1} + (1-\beta_{1t}) g_t, x_t - x^* \rangle
\end{aligned}$$

Rearranging the above inequality, we have

$$\begin{aligned}\langle g_t, x_t - x^*\rangle \leq & \frac{1}{2\alpha_t(1-\beta_{1t})}[||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2] \\ & + \frac{\alpha_t}{2(1-\beta_{1t})}||V_t^{-1/2p}m_t||^2 - \frac{\beta_{1t}}{1-\beta_{1t}}\langle m_{t-1}, x_t - x^*\rangle \\ \leq & \frac{1}{2\alpha_t(1-\beta_{1t})}[||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2] \\ & + \frac{\alpha_t}{2(1-\beta_{1t})}||V_t^{-1/2p}m_t||^2 + \frac{\alpha_t\beta_{1t}}{2(1-\beta_{1t})}||V_t^{-1/2p}m_{t-1}||^2 \\ & + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}||V_t^{1/2p}(x_t - x^*)||^2\end{aligned}$$

The second inequality follows from simple application of Cauchy-Schwarz and Youngs inequality. We now use the standard approach of bounding the regret at each step using convexity of the function f_t in the following manner:

$$\begin{aligned}&\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \sum_{t=1}^{T}\langle g_t, x_t - x^*\rangle \\ &\leq \sum_{t=1}^{T}[\frac{1}{2\alpha_t(1-\beta_{1t})}(||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2) + \frac{\alpha_t}{2(1-\beta_{1t})}||V_t^{-1/2p}m_t||^2 \\ &+ \frac{\beta_{1t}}{2(1-\beta_{1t})}\alpha||V_t^{-1/2p}m_{t+1}||^2 + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}||V_t^{1/2p}(x_t - x^*)||^2]\end{aligned}$$

This completes the proof of Lemma 2.

Base on this Lemma, we are going to find the corresponding upper bound for each term in the above regret bound inequality.

For the first term $\sum_{t=1}^{T}[\frac{1}{2\alpha_t(1-\beta_{1t})}(||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2)$, we have Lemma 3.

Lemma 3. *When B_t/t is non-increasing, then $V_t/\alpha_t^2 - V_{t-1}/\alpha_{t-1}^2$ is semi-positive, and*

$$\sum_{t=1}^{T}\frac{1}{2\alpha_t(1-\beta_{1t})}(||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2) \leq \frac{T^{1/p}}{2(1-\beta_1)}\frac{V_t^{1/p}}{\alpha}D_\infty^2$$

Proof of Lemma 3:

$$\begin{aligned}\frac{V_t}{\alpha_t^p} =&\frac{t}{\alpha^p}\sum_{j=1}^{t}\Pi_{k=1}^{t-j}\beta_{2,t-k+1}(1-\beta_{2,j})g_j^p \\ =&\frac{t}{\alpha^p}\sum_{j=1}^{t}\frac{B_{t-1}}{B_t}\cdots\frac{B_j}{B_{j+1}}\frac{B_j - B_{j-1}}{B_j}g_j^p \\ =&\frac{t}{B_t\alpha^p}\sum_{j=1}^{t}b_j g_j^p \geq \frac{t-1}{B_{t-1}\alpha^2}\sum_{j=1}^{t-1}b_j g_j^p \\ =&\frac{V_{t-1}}{\alpha_{t-1}^p}\end{aligned}$$

which means $V_t/\alpha_t^2 - V_{t-1}/\alpha_{t-1}^2$ is semi-positive.

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{1}{2\alpha_t(1-\beta_{1t})}(||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2) \\
\leq& \frac{1}{2(1-\beta_1)} \sum_{t=1}^{T} \frac{1}{\alpha_t}(||V_t^{1/2p}(x_t - x^*)||^2 - ||V_t^{1/2p}(x_{t+1} - x^*)||^2) \\
\leq& \frac{1}{2(1-\beta_1)}(||V_1^{1/2p}(x_1 - x^*)||^2 - ||V_T^{1/2p}(x_{T+1} - x^*)||^2) \\
+& \frac{1}{2(1-\beta_1)} \sum_{t=2}^{T}(\frac{V_t^{1/p}}{\alpha_t} - \frac{V_{t-1}^{1/p}}{\alpha_{t-1}})(x_t - x^*)^2 \\
\leq& \frac{1}{2(1-\beta_1)}||V_1^{1/2p}||^2 D_\infty^2 + \frac{1}{2(1-\beta_1)} \sum_{t=2}^{T}(\frac{V_t^{1/p}}{\alpha_t} - \frac{V_{t-1}^{1/p}}{\alpha_{t-1}}) D_\infty^2 \\
=& \frac{T^{1/p}}{2(1-\beta_1)} \frac{V_t^{1/p}}{\alpha} D_\infty^2
\end{aligned}
$$

The third inequation use the knowledge that $V_t/\alpha_t^2 - V_{t-1}/\alpha_{t-1}^2 \geq 0$.

This completes the proof of Lemma 3.

For the second and the third terms in Lemma 2, we have Lemma 4.

Lemma 4.

$$
\frac{\alpha_t}{2(1-\beta_{1t})}||V_t^{-1/2p} m_t||^2 + \frac{\alpha_t \beta_{1t}}{2(1-\beta_{1t})}||V_t^{-1/2p} m_{t-1}||^2 \leq \frac{p}{2(p-1)} \frac{\alpha(1+\beta_1)}{(1-\beta_1)^3} S_T^{\frac{p-1}{p}} (\frac{B_T}{T b_T^p})^{\frac{1}{p}}
$$

Proof of Lemma 4

For the second term in Lemma 2 :

$$
\begin{aligned}
&\sum_{t=1}^{T} \alpha_t ||V_t^{-1/2p} m_t||^2 \\
=& \sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_t||^2 + \alpha_T \sum_{i=1}^{d} \frac{m_{T,i}^2}{v_{T,i}^{1/p}} \\
\leq& \sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_t||^2 + \alpha_T \sum_{i=1}^{d} \frac{(\sum_{j=1}^{T}(1-\beta_{1j})\Pi_{k=1}^{T-j}\beta_{1(T-k+1)} g_{j,i})^2)}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j}) g_{j,i}^2)^{1/p}} \\
\leq& \sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_t||^2 \\
&\quad + \alpha_T \sum_{i=1}^{d} \frac{(\sum_{j=1}^{T}(1-\beta_{1j})\Pi_{k=1}^{T-j}\beta_{1(T-k+1)})(\sum_{j=1}^{T}(1-\beta_{1j})\Pi_{k=1}^{T-j}\beta_{1(T-k+1)} g_{j,i}^2)}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j}) g_{j,i}^2)^{1/p}} \\
\leq& \sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_t||^2 + \alpha_T \sum_{i=1}^{d} \frac{(\sum_{j=1}^{T}\beta_1^{T-j})(\sum_{j=1}^{T}\beta_1^{T-j} g_{j,i}^2)}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j}) g_{j,i}^2)^{1/p}} \\
\leq& \sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_t||^2 + \frac{\alpha_T}{1-\beta_1} \sum_{i=1}^{d} \frac{\sum_{j=1}^{T}\beta_1^{T-j} g_{j,i}^2}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j}) g_{j,i}^2)^{1/p}} \\
\leq& \sum_{i=1}^{d}\sum_{t=1}^{T} \frac{\alpha_t}{1-\beta_1} \frac{\sum_{j=1}^{t}\beta_1^{t-j} g_{j,i}^2}{(\sum_{j=1}^{t}\Pi_{k=1}^{t-j}\beta_{2(T-k+1)}(1-\beta_{2j}) g_{j,i}^2)^{1/p}}
\end{aligned}
$$

For the third term of in Lemma 2 :

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t ||V_t^{-1/2p} m_{t-1}||^2 \\
=&\sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_{t-1}||^2 + \alpha_T \sum_{i=1}^{d} \frac{m_{T-1,i}^2}{v_{T,i}^{1/p}} \\
\leq&\sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_{t-1}||^2 + \alpha_T \sum_{i=1}^{d} \frac{(\sum_{j=1}^{T-1}(1-\beta_{1j})\Pi_{k=1}^{T-1-j}\beta_{1(T-k)}g_{j,i})^2)}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j})g_{j,i}^2)^{1/p}} \\
\leq&\sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_{t-1}||^2 + \alpha_T \sum_{i=1}^{d} \frac{(\sum_{j=1}^{T}\beta_1^{T-1-j})(\sum_{j=1}^{T-1}\beta_1^{T-1-j}g_{j,i}^2)}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j})g_{j,i}^2)^{1/p}} \\
\leq&\sum_{t=1}^{T-1} \alpha_t ||V_t^{-1/2p} m_{t-1}||^2 + \frac{\alpha_T}{1-\beta_1} \sum_{i=1}^{d} \frac{\sum_{j=1}^{T-1}\beta_1^{T-1-j}g_{j,i}^2}{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{2(T-k+1)}(1-\beta_{2j})g_{j,i}^2)^{1/p}} \\
\leq&\sum_{i=1}^{d}\sum_{t=1}^{T} \frac{\alpha_t}{1-\beta_1} \frac{\sum_{j=1}^{t-1}\beta_1^{t-1-j}g_{j,i}^2}{(\sum_{j=1}^{t}\Pi_{k=1}^{t-j}\beta_{2(T-k+1)}(1-\beta_{2j})g_{j,i}^2)^{1/p}}
\end{aligned}$$

What's more:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha_t}{1-\beta_1} \frac{\sum_{j=1}^{t}\beta_1^{t-j}g_j^p}{[\sum_{j=1}^{t}\Pi_{k=1}^{t-j}\beta_{2,t-k+1}(1-\beta_{2,j})g_j^p]^{\frac{1}{p}}} = \sum_{t=1}^{T} \frac{\alpha}{1-\beta_1} \frac{\sum_{j=1}^{t}\beta_1^{t-j}g_j^p}{[\sum_{j=1}^{t}\frac{t}{B(t)}b_j g_j^p]^{\frac{1}{p}}} \\
\leq&\sum_{t=1}^{T} \frac{\alpha}{1-\beta_1}(\frac{B_t}{t})^{\frac{1}{p}} \sum_{j=1}^{t} \frac{\beta_1^{t-j}g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}} = \sum_{j=1}^{T}\sum_{t=j}^{T} \frac{\beta_1^{t-j}g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}}(\frac{B_t}{t})^{\frac{1}{p}}\frac{\alpha}{1-\beta_1} \\
\leq&\sum_{j=1}^{T}\sum_{t=j}^{T} \frac{\beta_1^{t-j}g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}}(\frac{B_j}{j})^{\frac{1}{p}}\frac{\alpha}{1-\beta_1} \\
\leq&\frac{\alpha}{(1-\beta_1)^2}\sum_{j=1}^{T} \frac{g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}}(\frac{B_j}{j})^{\frac{1}{p}}
\end{aligned}$$

The first inequality comes from $\sum_{k=1}^{j} b_k g_k^p \leq \sum_{k=1}^{t} b_k g_k^p$. The second inequality comes from that B_t/t is non-increasing with respect to t. The last inequality follows from then inequality $\sum_{t=j}^{T}\beta_1^{t-j} \leq 1/(1-\beta_1)$.

Similar to the proof above:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha_t}{1-\beta_1} \frac{\sum_{j=1}^{t-1}\beta_1^{t-1-j}g_j^p}{[\sum_{j=1}^{t}\Pi_{k=1}^{t-j}\beta_{2,t-k+1}(1-\beta_{2,j})g_j^p]^{\frac{1}{p}}} = \sum_{t=1}^{T} \frac{\alpha}{1-\beta_1} \frac{\sum_{j=1}^{t-1}\beta_1^{t-1-j}g_j^p}{[\sum_{j=1}^{t}\frac{t}{B(t)}b_j g_j^p]^{\frac{1}{p}}} \\
\leq&\sum_{t=1}^{T} \frac{\alpha}{1-\beta_1}(\frac{B_t}{t})^{\frac{1}{p}} \sum_{j=1}^{t-1} \frac{\beta_1^{t-1-j}g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}} = \sum_{j=1}^{T-1}\sum_{t=j+1}^{T} \frac{\beta_1^{t-1-j}g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}}(\frac{B_t}{t})^{\frac{1}{p}}\frac{\alpha}{1-\beta_1} \\
\leq&\sum_{j=1}^{T-1}\sum_{t=j+1}^{T} \frac{\beta_1^{t-1-j}g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}}(\frac{B_j}{j})^{\frac{1}{p}}\frac{\alpha}{1-\beta_1} \\
\leq&\frac{\alpha}{(1-\beta_1)^2}\sum_{j=1}^{T-1} \frac{g_j^p}{(\sum_{k=1}^{j}b_k g_k^p)^{\frac{1}{p}}}(\frac{B_j}{j})^{\frac{1}{p}}
\end{aligned}$$

Let $S_j = \sum_{k=1}^{j} b_k g_k^p$,

$$
\begin{aligned}
&\sum_{j=1}^{T} \frac{g_j^p}{(\sum_{k=1}^{j} b_k g_k^p)^{\frac{1}{p}}} (\frac{B_j}{j})^{\frac{1}{p}} = \sum_{j=1}^{T} \frac{g_j^p}{S_j^{\frac{1}{p}}} (\frac{B_j}{j})^{\frac{1}{p}} \\
\leq &\sum_{j=1}^{T} g_j^p (\frac{B_j}{j})^{\frac{1}{p}} \frac{p}{p-1} \frac{(S_j^{\frac{p-1}{p}} - S_{j-1}^{\frac{p-1}{p}})}{S_j - S_{j-1}} = \frac{p}{p-1} \sum_{j=1}^{T} (S_j^{\frac{p-1}{p}} - S_{j-1}^{\frac{p-1}{p}}) (\frac{B_j}{j b_j^p})^{\frac{1}{p}} \\
= &\frac{p}{p-1} S_T^{\frac{p-1}{p}} (\frac{B_T}{T b_T^p})^{\frac{1}{p}} + \frac{p}{p-1} \sum_{j=1}^{T-1} [-(\frac{B_{j+1}}{(j+1) b_{j+1}^p})^{\frac{1}{p}} + (\frac{B_j}{j b_j^p})^{\frac{1}{p}}] S_j^{\frac{p-1}{p}} \\
\leq &\frac{p}{p-1} S_T^{\frac{p-1}{p}} (\frac{B_T}{T b_T^p})^{\frac{1}{p}}
\end{aligned}
$$

The first inequality comes from Lemma 5 when $p > 1$. The last inequality comes from the second constraint, which tells us that $B_j/(j b_j^p)$ is non-decreasing with respect to j. This completes the proof of the lemma.

This finally completes the proof of Lemma 4.

To complete Lemma 4, we need to finally prove the next Lemma.

Lemma 5.

$$
\frac{1}{S_j^{\frac{1}{p}}} \leq \frac{p}{p-1} \frac{(S_j^{\frac{p-1}{p}} - S_{j-1}^{\frac{p-1}{p}})}{S_j - S_{j-1}}
$$

Proof of Lemma 5

When $p > 1$,$s > 0$,$x \geq 0$

$$
\begin{aligned}
&1 \leq \frac{p}{p-1} [1 - \frac{1}{p} (\frac{s}{s+x})^{\frac{p-1}{p}}] \\
\Rightarrow &x \leq \frac{p}{p-1} [(s+x) - s^{\frac{p-1}{p}} (s+x)^{\frac{1}{p}}] \\
\Rightarrow &\frac{1}{(s+x)^{\frac{1}{p}}} \leq \frac{p}{p-1} \frac{((s+x)^{\frac{p-1}{p}} - s^{\frac{p-1}{p}})}{x} \\
\Rightarrow &\frac{1}{S_j^{\frac{1}{p}}} \leq \frac{p}{p-1} \frac{(S_j^{\frac{p-1}{p}} - S_{j-1}^{\frac{p-1}{p}})}{S_j - S_{j-1}}
\end{aligned}
$$

This completes the proof of Lemma 5.

Finally, back to our theorem, using the inequalities in Lemma 4 and Lemma 3 to substitute the first three terms in Lemma 2 completes the proof of our theorem.

Convergence of NosAdam-HH

Corollary 5.1 (Convergence of NosAdam-HH)**.** *Suppose*$\beta_{1,t} = \beta_1 \lambda^{t-1}$, $b_k = k^{-\gamma}, \gamma \geq 0$ *, thus* $B_t = \sum_{k=1}^{t} k^{-\gamma}$*, and* $\beta_{2,t} = B_{t-1}/B_t < 1$ *in Algorithm 1. Then* B_t *and* b_t *satisfy the constraints in Therorem 1, and we have*

$$
\begin{aligned}
R_T \leq \frac{D_\infty^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T} v_{T,i}^{\frac{1}{2}} + \frac{D_\infty^2 G_\infty \beta_1}{2(1-\beta_1)} \frac{1}{(1-\lambda)^2} \cdot d \\
+ \frac{2\alpha\beta_1}{(1-\beta_1)^3} G_\infty \sqrt{T}
\end{aligned}
$$

Proof. We only need to verify the two constraints on B_t and b_t in Theorem 1 are satisfied.

Since $B_t = \sum_{k=1}^{t} k^{-\gamma}$, $\frac{B_t}{t} = \frac{\sum_{k=1}^{t} k^{-\gamma}}{t}$. Therefore

$$\frac{B_t}{t} \leq \frac{B_{t-1}}{t-1}$$

is equivalent to

$$\sum_{k=1}^{t} k^{-\gamma}(t-1) \leq \sum_{k=1}^{t-1} k^{-\gamma} t,$$

and is again equivalent to

$$(t-1)t^{-\gamma} \leq \sum_{k=1}^{t-1} k^{-\gamma}.$$

And this is true since $\gamma \geq 0$.

For the second constraint, since $1/b_T \geq 1/b_{T-1}$, we just need to prove that

$$\frac{B_T}{Tb_T} \geq \frac{B_{T-1}}{(T-1)b_{T-1}}$$

When we transform it into the form of Riemann sum, we will find that

$$\sum_{t=1}^{T} (\frac{t}{T})^{-\gamma} \frac{1}{T} = \sum_{k=1}^{T(T-1)} (\frac{a_k}{T})^{-\gamma} \frac{1}{T(T-1)},$$

$$\sum_{t=1}^{T-1} (\frac{t}{T-1})^{-\gamma} \frac{1}{T-1} = \sum_{k=1}^{T(T-1)} (\frac{b_k}{T-1})^{-\gamma} \frac{1}{T(T-1)},$$

and

$$\frac{a_k}{T} = \frac{\lceil \frac{k}{T-1} \rceil}{T} \geq \frac{\lceil \frac{k}{T} \rceil}{T-1} = \frac{b_k}{T-1}.$$

Thus

$$\frac{B_T}{Tb_T} = \sum_{t=1}^{T} (\frac{t}{T})^{-\gamma} \frac{1}{T} \geq \sum_{t=1}^{T-1} (\frac{t}{T-1})^{-\gamma} \frac{1}{T-1} = \frac{B_{T-1}}{(T-1)b_{T-1}}.$$

□

References

[1] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.