A PROOF OF THEOREM 1

Lemma 1 (McMahan & Streeter (2010)). *For any $Q \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \subset \mathbb{R}^d$, suppose $u_1 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_1)\|$ and $u_2 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_2)\|$ then we have $\|Q^{1/2}(u_1 - u_2)\| \leq \|Q^{1/2}(z_1 - z_2)\|$.*

Proof. We provide the proof here for completeness. Since $u_1 = min_{x \in \mathcal{F}} \left\|Q^{1/2}(x - z_1)\right\|$ and $u_2 = min_{x \in \mathcal{F}} \left\|Q^{1/2}(x - z_2)\right\|$ and from the property of projection operator we have the following:

$$\langle z_1 - u_1, Q(z_2 - z_1)\rangle \geq 0 \text{ and } \langle z_2 - u_2, Q(z_1 - z_2)\rangle \geq 0.$$

Combining the above inequalities, we have

$$\langle u_2 - u_1, Q(z_2 - z_1)\rangle \geq \langle z_2 - z_1, Q(z_2 - z_1)\rangle . \tag{3}$$

Also, observe the following:

$$\langle u_2 - u_1, Q(z_2 - z_1)\rangle \leq \frac{1}{2}\left[\langle u_2 - u_1, Q(u_2 - u_1)\rangle + \langle z_2 - z_1, Q(z_2 - z_1)\rangle\right].$$

The above inequality can be obtained from the fact that

$$\langle (u_2 - u_1) - (z_2 - z_1), Q((u_2 - u_1) - (z_2 - z_1))\rangle \geq 0 \text{ as } Q \in \mathcal{S}_+^d$$

and rearranging the terms. Combining the above inequality with Equation 3, we have the required the result. □

Proof. For simplicity, vectors is also denoted in common lowercase in the proof. We begin with the following observation:

$$x_{t+1} = \Pi_{\mathcal{F}, \text{diag}(\eta_t^{-1})}(x_t - \eta_t \odot g_t) = \min_{x \in \mathcal{F}} \|\eta_t^{-1/2} \odot (x - (x_t - \eta_t \odot g_t))\|.$$

Furthermore, as $\mathcal{F}$ is closed and convex, we can get $x^* = \arg\min_{x \in \mathcal{F}} \sum_{t=1}^T f_t(x)$. Using Lemma 1 with $u_1 = x_{t+1}$ and $u_2 = x^*$, we have the following:

$$\begin{aligned}\|\eta_t^{-1/2} \odot (x_{t+1} - x^*)\|^2 &\leq \|\eta_t^{-1/2} \odot (x_t - \eta_t \odot g_t - x^*)\|^2 \\ &= \|\eta_t^{-1/2} \odot (x_t - x^*)\|^2 + \|\eta_t^{1/2} \odot g_t\|^2 - 2\langle g_t, x_t - x^*\rangle.\end{aligned}$$

Rearranging the above inequality, we have

$$\langle g_t, x_t - x^*\rangle \leq \frac{1}{2}\left[\|\eta_t^{-1/2} \odot (x_t - x^*)\|^2 - \|\eta_t^{-1/2} \odot (x_{t+1} - x^*)\|^2\right] + \frac{1}{2}\|\eta_t^{1/2} \odot g_t\|^2. \tag{4}$$

We now use the standard approach of bounding the regret at each step using convexity of the function f_t in the following manner:

$$\begin{aligned}&\sum_{t=1}^T f_t(x_t) - f_t(x^*) \leq \sum_{t=1}^T \langle g_t, x_t - x^*\rangle \\ &\leq \frac{1}{2}\sum_{t=1}^T \left[\|\eta_t^{-1/2} \odot (x_t - x^*)\|^2 - \|\eta_t^{-1/2} \odot (x_{t+1} - x^*)\|^2 + \|\eta_t^{1/2} \odot g_t\|^2\right] \\ &= \frac{1}{2}\Bigg[\sum_{t=2}^T \left[\|\eta_t^{-1/2} \odot (x_t - x^*)\|^2 - \|\eta_{t-1}^{-1/2} \odot (x_t - x^*)\|^2\right] \\ &\quad + \|\eta_1^{-1/2} \odot (x_1 - x^*)\|^2 - \|\eta_t^{-1/2} \odot (x_{t+1} - x^*)\|^2 + \sum_{t=1}^T \|\eta_t^{1/2} \odot g_t\|^2\Bigg] \\ &= \frac{1}{2}\Bigg[\sum_{t=2}^T \sum_{i=1}^d (x_{t,i} - x_i^*)^2(\eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}) + \sum_{i=1}^d \eta_{1,i}^{-1}(x_{1,i} - x_i^*)^2 \\ &\quad - \sum_{i=1}^d \eta_{t,i}^{-1}(x_{t+1,i} - x_i^*)^2 + \sum_{t=1}^T \sum_{i=1}^d g_{t,i}^2 \eta_{t,i}\Bigg].\end{aligned} \tag{5}$$

The first inequality is due to the convexity of functions $\{f_t\}$. The second inequality follows from the bound in Equation 4. For further bounding this inequality, we need the following intermediate result.

Lemma 2. *For the parameter settings and conditions assumed in Algorithm 1, we have*

$$\sum_{t=2}^{T} \left|\eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}\right| \leq \frac{2}{\eta(1-\lambda)^3}\left[(5-4\lambda)\sqrt{T} + 2\lambda - 1\right].$$

Proof. For simplicity, we ignore subscript i in this lemma. Let $\eta_{t+1} = \eta_t + \Delta t$, $c_t = \frac{g_t m_t}{\left|g_t^i\right| \max(|m_t|)+\epsilon}$, $b_t = \lambda^t c_t$ and $a_t = \frac{1}{\sqrt{t}}$, thus we have

$$\begin{aligned}
\eta_t &= \frac{\eta}{\sqrt{t}}(1+b_t), \\
\frac{\Delta t}{\eta} &= \frac{\eta_{t+1}-\eta_t}{\eta} = \frac{1}{\sqrt{t+1}}\left(1+b_{t+1}\right) - \frac{1}{\sqrt{t}}\left(1+b_t\right) \\
&= \left(\frac{1}{\sqrt{t+1}} - \frac{1}{\sqrt{t}}\right) + a_{t+1}b_{t+1} - a_t b_t \\
&= a_{t+1}\left(b_{t+1}-b_t\right) + \left(1+b_t\right)\left(a_{t+1}-a_t\right).
\end{aligned}$$

We observe that,

$$\left|a_{t+1}\left(b_{t+1}-b_t\right)\right| = \frac{\left|\lambda^{t+1}c_{t+1} - \lambda^t c_t\right|}{\sqrt{t+1}} \leq \frac{\lambda^{t+1}\left|c_{t+1}\right| + \lambda^t\left|c_t\right|}{\sqrt{t+1}} \leq \frac{2\lambda^t}{\sqrt{t}}.$$

Also, observe the following:

$$\left|\left(1+b_t\right)\left(a_{t+1}-a_t\right)\right| \leq (1+\lambda)\left(\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+1}}\right) \leq \frac{2}{t\sqrt{t}}.$$

The above inequality can be obtained from the fact that $|1+b_t| \leq 1+|b_t| \leq 1+\lambda\left|c_t\right| \leq 1+\lambda$.

Hence, we have,

$$\frac{|\Delta t|}{\eta} \leq \frac{2}{\sqrt{t}}(\frac{1}{t} + \lambda^t)$$

By definition,

$$\eta\frac{1-\lambda}{\sqrt{t}} \leq \eta_t \leq \eta\frac{1+\lambda}{\sqrt{t}}$$

And then,

$$\left|\eta_{t+1}^{-1} - \eta_t^{-1}\right| = \left|\frac{\eta_{t+1}-\eta_t}{\eta_{t+1}\eta_t}\right| \leq \frac{|\Delta t|\,(t+1)}{\eta^2(1-\lambda)^2} \leq \frac{2(t+1)(\frac{1}{t}+\lambda^t)}{\sqrt{t}\eta(1-\lambda)^2}.$$

Finally, we have,

$$\begin{aligned}
&\sum_{t=2}^{T}\left|\eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}\right| \leq \frac{2}{\eta(1-\lambda)^2}\left[\sum_{t=2}^{T}\frac{t}{t-1}\frac{1}{\sqrt{t-1}} + \sum_{t=2}^{T}\lambda^{t-1}\frac{t}{\sqrt{t-1}}\right] \\
&\leq \frac{2}{\eta(1-\lambda)^2}\left[2\sum_{t=2}^{T}\frac{1}{\sqrt{t-1}} + \sum_{t=2}^{T}\lambda^{t-1} + \sum_{t=2}^{T}\lambda^{t-1}\sqrt{t-1}\right] \\
&\leq \frac{2}{\eta(1-\lambda)^3}\left[(5-4\lambda)\sqrt{T} + 2\lambda - 1\right].
\end{aligned}$$

The last inequality is due to the following upper bound:

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 1 + \int_{t=1}^{T}\frac{\mathrm{d}t}{\sqrt{t}} = 2\sqrt{T} - 1.$$

□

We now return to the proof of Theorem 1. Using the D_∞ bound on the feasible region and making use of the above property in Equation 5 and Lemma 2, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t\left(x_t\right) - f_t\left(x^*\right) \\
&\leq \frac{1}{2}\left[\sum_{t=2}^{T}\sum_{i=1}^{d}(x_{t,i}-x_i^*)^2\left|\eta_{t,i}^{-1}-\eta_{t-1,i}^{-1}\right| + \sum_{i=1}^{d}\eta_{1,i}^{-1}(x_{1,i}-x_i^*)^2 + \sum_{t=1}^{T}\sum_{i=1}^{d} g_{t,i}^2\eta_{t,i}\right] \\
&\leq \frac{D_\infty^2 d}{\eta(1-\lambda)^3}\left[(5-4\lambda)\sqrt{T}+2\lambda-1\right] + \frac{D_\infty^2 d}{2\eta(1-\lambda)} + G_2^2 d\eta(2\sqrt{T}-1).
\end{aligned}$$

It is easy to see that the regret of AdaRem is upper bounded by $O(\sqrt{T})$. □

B Experiment Details

B.1 Hyper-parameters grid search

We run all experiments with cosine learning rate without a warmup stage and train for 100 epochs with a minibatch size of 1024 on 16 GPUs.

B.1.1 Adaptive optimizers' performance on deep convolutional network

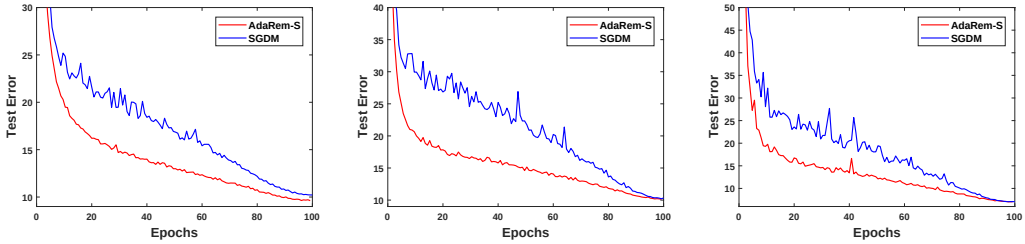
We set the base learning rate of 0.4 for SGDM just as (He et al., 2019), AdaRem and AdaRem-S. For Adam, we set the base learning rate as 0.004, and choose the EMA parameter of the second momentum of gradient β_2 from {0.99,0.999}. For AdamW and AdaBound, we adopt the same hyper-parameters as Adam. For AdaBound, we choose the final_lr from {0.1,0.4}. For RMSProp, we choose the base learning rate from {0.04,0.004,0.0004}, and β_2 from {0.99,0.999}. The momentum parameter β of AdaRem is set as 0.999. Weight decay parameter is choosen from {0.0001,0.0003} for all methods. Additional details can be seen in the Table 4.

Table 4: Hyper-parameters' setting of various optimization methods for ResNet18 on ImageNet. The bold number indicates the best one of the hyper-parameters to be selected and ϵ is a term to improve numerical stability. β_1 is the EMA parameter of the first momentum of gradient and β_2 is the EMA parameter of the second momentum of gradient.

Model	Hyper-parameter					
	lr	β_1	β_2	weight_decay	ϵ	final_lr
SGDM	0.4	**0.9**,0.999	$\sim$	**0.0001**,0.0003	$\sim$	$\sim$
Adam	0.004	**0.9**,0.999	0.99,**0.999**	**0.0001**,0.0003	1e-8	$\sim$
AdamW	0.004	**0.9**,0.999	0.999	**0.0001**,0.0003	1e-8	$\sim$
AdaBound	0.004	**0.9**,0.999	0.999	**0.0001**,0.0003	1e-8	0.1,**0.4**
RMSProp	0.04,0.004,**0.0004**	$\sim$	**0.99**,0.999	**0.0001**,0.0003	1e-8	$\sim$
AdaRem	0.4	0.9,**0.999**	$\sim$	0.0001,**0.0003**	1e-8	$\sim$

B.1.2 AdaRem-S vs. SGDM across various architectures

For AdaRem-S, we choose momentum parameter from {0.995,0.999} of ResNet50, and the sphere radius from {10,100} of MobileNetV2. MobileNetV2-0.5 and MobileNetV2-1.0 employ the same hyper-parameter's setting. Additional details can be seen in the Table 5.



(a) Top-5 Error for MobileNetV2 (b) Top-5 Error for ResNet18 (c) Top-5 Error for ResNet50

Figure 6: Top-5 error of three networks on ImageNet.

Table 5: Hyper-parameters' setting of AdaRem-S and SGD for ResNet-18, ResNet-50 and MobileNetV2 on ImageNet. The bold number indicates the best one of the hyper-parameters to be selected.

Model	optimizer	Hyper-parameter				
		lr	momentum	weight_decay	R	ϵ
ResNet-18	SGDM	0.4	0.9	1e-4	$\sim$	$\sim$
	AdaRem-S	0.4	0.999	1e-4	10	1e-8
ResNet-50	SGDM	0.4	0.9	1e-4	$\sim$	$\sim$
	AdaRem-S	0.4	**0.995**,0.999	1e-4	10	1e-8
MobileNetV2	SGDM	0.4	0.9	1e-4,**4e-5**	$\sim$	$\sim$
	AdaRem-S	0.4	0.999	1e-4,**4e-5**	10,**100**	1e-8
ShuffleNetV2	SGDM	0.4	0.9	1e-4,**4e-5**	$\sim$	$\sim$
	AdaRem-S	0.4	0.999	1e-4,**4e-5**	10,**100**	1e-8

B.2 Other experimental results

Table 6: Top-5 accuracy of various networks on the ImageNet dataset. The bold number indicates the best result.

Model	Top-5 Accuracy(%)	
	SGDM	AdaRem-S
ResNet50	92.9	**92.92**
ResNet18	89.74	**89.87**
MobileNetV2-1.0	89.81	**90.34**
MobileNetV2-0.5	84.66	**85.30**
ShuffleNetV2-1.0	87.68	**87.96**
ShuffleNetV2-0.5	80.23	**81.77**

Table 7: Train loss of various networks on the ImageNet dataset. The bold number indicates the best result.

Model	Train Loss	
	SGDM	AdaRem-S
ResNet50	0.932	**0.823**
ResNet18	1.331	**1.148**
MobileNetV2-1.0	1.462	**1.343**
MobileNetV2-0.5	1.965	**1.887**
ShuffleNetV2-1.0	1.340	**1.337**
ShuffleNetV2-0.5	1.839	**1.744**