Supplementary Material of "Dynamic Regret of Strongly Adaptive Methods"

Lijun Zhang [1] **Tianbao Yang** [2] **Rong Jin** [3] **Zhi-Hua Zhou** [1]

A. Proof of Lemma 1

We first prove the first part of Lemma 1. Let $k = \lfloor \log_K t \rfloor$. Then, integer t can be represented in the base-K number system as

$$t = \sum_{j=0}^{k} \beta_j K^j .$$

From the definition of base-K ending time, integers that are no larger than t and alive at t are

$$\left\{ \begin{array}{l} 1 \quad K^0 + \sum_{j=1}^{k} \beta_j K^j, \ 2 \quad K^0 + \sum_{j=1}^{k} \beta_j K^j, \ \ldots, \ \beta_0 \quad K^0 + \sum_{j=1}^{k} \beta_j K^j \\ 1 \quad K^1 + \sum_{j=2}^{k} \beta_j K^j, \ 2 \quad K^1 + \sum_{j=2}^{k} \beta_j K^j, \ \ldots, \ \beta_1 \quad K^1 + \sum_{j=2}^{k} \beta_j K^j \\ \ldots \\ 1 \quad K^{k-1} + \beta_k K^k, \ 1 \quad K^{k-1} + \beta_k K^k, \ \ldots, \ \beta_{k-1} \quad K^{k-1} + \beta_k K^k \\ 1 \quad K^k, \ 2 \quad K^k, \ \ldots, \ \beta_k K^k \end{array} \right\} .$$

The total number of alive integers are upper bounded by

$$\sum_{i=0}^{k} \beta_i \quad (k+1)(K \quad 1) = (\lfloor \log_K t \rfloor + 1)(K \quad 1).$$

We proceed to prove the second part of Lemma 1. Let $k = \lfloor \log_K r \rfloor$, and the representation of r in the base-K number system be

$$r = \sum_{j=0}^{k} \beta_j K^j .$$

We generate a sequence of segments as

$$
\begin{aligned}
I_1 &= [t_1, e^{t_1} - 1] = \left[\sum_{j=0}^{k} \beta_j K^j, (\beta_1+1)K^1 + \sum_{j=2}^{k} \beta_j K^j - 1\right], \\
I_2 &= [t_2, e^{t_2} - 1] = \left[(\beta_1+1)K^1 + \sum_{j=2}^{k} \beta_j K^j, (\beta_2+1)K^2 + \sum_{j=3}^{k} \beta_j K^j - 1\right], \\
I_3 &= [t_3, e^{t_3} - 1] = \left[(\beta_2+1)K^2 + \sum_{j=3}^{k} \beta_j K^j, (\beta_3+1)K^3 + \sum_{j=4}^{k} \beta_j K^j - 1\right], \\
&\ldots \\
I_k &= [t_k, e^{t_k} - 1] = \left[(\beta_{k-1}+1)K^{k-1} + \beta_k K^k, (\beta_k+1)K^k - 1\right], \\
I_{k+1} &= [t_{k+1}, e^{t_{k+1}} - 1] = \left[(\beta_k+1)K^k, K^{k+1} - 1\right], \\
I_{k+2} &= [t_{k+2}, e^{t_{k+2}} - 1] = \left[K^{k+1}, K^{k+2} - 1\right], \\
&\ldots
\end{aligned}
$$

until s is covered. It is easy to verify that

$$t_{m+1} > t_m + K^{m-1} - 1.$$

Thus, s will be covered by the first m intervals as long as

$$t_m + K^{m-1} - 1 \geq s.$$

A sufficient condition is

$$r + K^{m-1} - 1 \geq s$$

which is satisfied when

$$m = \lceil \log_K (s - r + 1) \rceil + 1.$$

B. Proof of Theorem 1

From the second part of Lemma 1, we know that there exist m segments

$$I_j = [t_j, e^{t_j} - 1],\ j \in [m]$$

with $m \leq \lceil \log_K (s - r + 1) \rceil + 1$, such that

$$t_1 = r,\ e^{t_j} = t_{j+1},\ j \in [m-1], \text{ and } e^{t_m} > s.$$

Furthermore, the expert E^{t_j} is alive during the period $[t_j, e^{t_j} - 1]$.

Using Claim 3.1 of Hazan & Seshadhri (2009), we have

$$\sum_{t=t_j}^{e^{t_j}-1} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_t^{t_j}) \leq \frac{1}{\alpha}\left(\log t_j + 2 \sum_{t=t_j+1}^{e^{t_j}-1} \frac{1}{t}\right),\ \forall j \in [m-1]$$

where $\mathbf{w}_{t_j}^{t_j}, \ldots, \mathbf{w}_{e^{t_j}-1}^{t_j}$ is the sequence of solutions generated by the expert E^{t_j}. Similarly, for the last segment, we have

$$\sum_{t=t_m}^{s} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_t^{t_m}) \leq \frac{1}{\alpha}\left(\log t_m + 2 \sum_{t=t_m+1}^{s} \frac{1}{t}\right).$$

By adding things together, we have

$$
\begin{split}
&\sum_{j=1}^{m-1}\left(\sum_{t=t_j}^{e^{t_j}-1} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_t^{t_j})\right) + \sum_{t=t_m}^{s} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_t^{t_m}) \\
\leq & \frac{1}{\alpha}\sum_{j=1}^{m}\log t_j + \frac{2}{\alpha}\sum_{t=r+1}^{s}\frac{1}{t} \leq \frac{m+2}{\alpha}\log T.
\end{split} \tag{8}
$$

According to the property of online Newton step (Hazan et al., 2007, Theorem 2), we have, for any $\mathbf{w} \in \Omega$,

$$
\sum_{t=t_j}^{e^{t_j}-1} f_t(\mathbf{w}_t^{t_j}) - f_t(\mathbf{w}) \leq 5d\left(\frac{1}{\alpha} + GB\right)\log T, \ \forall j \in [m-1] \tag{9}
$$

and

$$
\sum_{t=t_m}^{s} f_t(\mathbf{w}_t^{t_m}) - f_t(\mathbf{w}) \leq 5d\left(\frac{1}{\alpha} + GB\right)\log T. \tag{10}
$$

Combining (8), (9), and (10), we have,

$$
\sum_{t=r}^{s} f_t(\mathbf{w}_t) - \sum_{t=r}^{s} f_t(\mathbf{w}) \leq \left(\frac{(5d+1)m+2}{\alpha} + 5dmGB\right)\log T
$$

for any $\mathbf{w} \in \Omega$.

C. Proof of Lemma 2

The gradient of $\exp(-\alpha f(\mathbf{w}))$ is

$$
\nabla \exp(-\alpha f(\mathbf{w})) = \exp(-\alpha f(\mathbf{w}))\ -\alpha \nabla f(\mathbf{w}) = -\alpha \exp(-\alpha f(\mathbf{w})) \nabla f(\mathbf{w}).
$$

and the Hessian is

$$
\begin{split}
\nabla^2 \exp(-\alpha f(\mathbf{w})) = & -\alpha \exp(-\alpha f(\mathbf{w}))\ -\alpha \nabla f(\mathbf{w}) \nabla^\top f(\mathbf{w}) - \alpha \exp(-\alpha f(\mathbf{w})) \nabla^2 f(\mathbf{w}) \\
= & \alpha \exp(-\alpha f(\mathbf{w})) \left(\alpha \nabla f(\mathbf{w}) \nabla^\top f(\mathbf{w}) - \nabla^2 f(\mathbf{w})\right).
\end{split}
$$

Thus, $f(\cdot)$ is α-exp-concave if

$$
\alpha \nabla f(\mathbf{w}) \nabla^\top f(\mathbf{w}) \preceq \nabla^2 f(\mathbf{w}).
$$

We complete the proof by noticing

$$
\frac{\lambda}{G^2} \nabla f(\mathbf{w}) \nabla^\top f(\mathbf{w}) \preceq \lambda I \preceq \nabla^2 f(\mathbf{w}).
$$

D. Proof of Theorem 2

Lemma 2 implies that all the λ-strongly convex functions are also $\frac{\lambda}{G^2}$-exp-concave. As a result, we can reuse the proof of Theorem 1. Specifically, (8) with $\alpha = \frac{\lambda}{G^2}$ becomes

$$
\sum_{j=1}^{m-1}\left(\sum_{t=t_j}^{e^{t_j}-1} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_t^{t_j})\right) + \sum_{t=t_m}^{s} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_t^{t_m}) \leq \frac{(m+2)G^2}{\lambda}\log T. \tag{11}
$$

According to the property of online gradient descent (Hazan et al., 2007, Theorem 1), we have, for any $\mathbf{w} \in \Omega$,

$$
\sum_{t=t_j}^{e^{t_j}-1} f_t(\mathbf{w}_t^{t_j}) - f_t(\mathbf{w}) \leq \frac{G^2}{2\lambda}(1+\log T), \ \forall j \in [m-1] \tag{12}
$$

and

$$\sum_{t=t_m}^{s} f_t(\mathbf{w}_t^{t_m}) - f_t(\mathbf{w}) \leq \frac{G^2}{2\lambda}(1+\log T). \tag{13}$$

Combining (11), (12), and (13), we have,

$$\sum_{t=r}^{s} f_t(\mathbf{w}_t) - \sum_{t=r}^{s} f_t(\mathbf{w}) \leq \frac{G^2}{2\lambda}\big(m + (3m+4)\log T\big)$$

for any $\mathbf{w} \in \Omega$.

E. Proof of Theorem 4

As pointed out by Daniely et al. (2015), the static regret of online gradient descent (Zinkevich, 2003) over any interval of length τ is upper bounded by $3BG\sqrt{\tau}$. Combining this fact with Theorem 2 of Jun et al. (2017), we get Theorem 4 in this paper.

F. Proof of Corollary 5

To simplify the upper bound in Theorem 3, we restrict to intervals of the same length τ, and in this case $k = T/\tau$. Then, we have

$$\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) &\leq \min_{1 \leq \tau \leq T} \sum_{i=1}^{k} \big(\text{SA-Regret}(T,\tau) + 2\tau V_T(i)\big) \\
&= \min_{1 \leq \tau \leq T} \left(\frac{\text{SA-Regret}(T,\tau)T}{\tau} + 2\tau \sum_{i=1}^{k} V_T(i) \right) \\
&\leq \min_{1 \leq \tau \leq T} \left(\frac{\text{SA-Regret}(T,\tau)T}{\tau} + 2\tau V_T \right).
\end{aligned}$$

Combining with Theorem 4, we have

$$\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq \min_{1 \leq \tau \leq T} \left(\frac{(c + 8\sqrt{7\log T + 5})T}{\sqrt{\tau}} + 2\tau V_T \right).$$

where $c = 12BG/(\sqrt{2} - 1)$.

In the following, we consider two cases. If $V_T \geq \sqrt{\log T / T}$, we choose

$$\tau = \left(\frac{T\sqrt{\log T}}{V_T} \right)^{2/3} \leq T$$

and have

$$\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) &\leq \frac{(c + 8\sqrt{7\log T + 5})T^{2/3}V_T^{1/3}}{\log^{1/6} T} + 2T^{2/3}V_T^{1/3}\log^{1/3} T \\
&\leq \frac{(c + 8\sqrt{5})T^{2/3}V_T^{1/3}}{\log^{1/6} T} + (2 + 8\sqrt{7})T^{2/3}V_T^{1/3}\log^{1/3} T.
\end{aligned}$$

Otherwise, we choose $\tau = T$, and have

$$\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) &\leq (c + 8\sqrt{7\log T + 5})\sqrt{T} + 2TV_T \\
&\leq (c + 8\sqrt{7\log T + 5})\sqrt{T} + 2T\sqrt{\frac{\log T}{T}} \\
&\leq (c + 9\sqrt{7\log T + 5})\sqrt{T}.
\end{aligned}$$

In summary, we have

$$
\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq & \max \begin{cases} (c + 9\sqrt{7 \log T + 5})\sqrt{T} \\ \dfrac{(c + 8\sqrt{5})T^{2/3}V_T^{1/3}}{\log^{1/6} T} + 24T^{2/3}V_T^{1/3}\log^{1/3} T \end{cases} \\
= & O\left(\max\left\{\sqrt{T \log T}, T^{2/3}V_T^{1/3}\log^{1/3} T\right\}\right).
\end{aligned}
$$

G. Proof of Corollary 6

The first part of Corollary 6 is a direct consequence of Theorem 1 by setting $K = \lceil dT^{1/\alpha} \rceil$.

Now, we prove the second part. Following similar analysis of Corollary 5, we have

$$
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq \min_{1 \leq \tau \leq T} \left\{ \left(\frac{(5d+1)(\gamma+1)+2}{\alpha} + 5d(\gamma+1)GB \right) \frac{T \log T}{\tau} + 2\tau V_T \right\}.
$$

Then, we consider two cases. If $V_T \geq \log T / T$, we choose

$$
\tau = \sqrt{\frac{T \log T}{V_T}} \leq T
$$

and have

$$
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq \left(\frac{(5d+1)(\gamma+1)+2}{\alpha} + 5d(\gamma+1)GB + 2 \right) \sqrt{TV_T \log T}.
$$

Otherwise, we choose $\tau = T$, and have

$$
\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq & \left(\frac{(5d+1)(\gamma+1)+2}{\alpha} + 5d(\gamma+1)GB \right) \log T + 2TV_T \\
\leq & \left(\frac{(5d+1)(\gamma+1)+2}{\alpha} + 5d(\gamma+1)GB \right) \log T + 2T\frac{\log T}{T} \\
= & \left(\frac{(5d+1)(\gamma+1)+2}{\alpha} + 5d(\gamma+1)GB + 2 \right) \log T.
\end{aligned}
$$

In summary, we have

$$
\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq & \left(\frac{(5d+1)(\gamma+1)+2}{\alpha} + 5d(\gamma+1)GB + 2 \right) \max\left\{\log T, \sqrt{TV_T \log T}\right\} \\
= & O\left(d \cdot \max\left\{\log T, \sqrt{TV_T \log T}\right\}\right).
\end{aligned}
$$

H. Proof of Corollary 7

The first part of Corollary 7 is a direct consequence of Theorem 2 by setting $K = \lceil dT^{1/\alpha} \rceil$.

The proof of the second part is similar to that of Corollary 6. First, we have

$$
\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq & \min_{1 \leq \tau \leq T} \left\{ \frac{G^2}{2\lambda}(\gamma + 1 + (3\gamma + 7)\log T)\frac{T}{\tau} + 2\tau V_T \right\} \\
\leq & \min_{1 \leq \tau \leq T} \left\{ \frac{(\gamma + 5\gamma \log T)G^2 T}{\lambda \tau} + 2\tau V_T \right\}
\end{aligned}
$$

where the last inequality is due to the condition $\gamma > 1$.

Then, we consider two cases. If $V_T \quad \log T/T$, we choose

$$\tau = \sqrt{\frac{T \log T}{V_T}} \quad T$$

and have

$$\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \quad & \frac{\gamma G^2}{\lambda}\sqrt{\frac{TV_T}{\log T}} + \frac{5\gamma G^2}{\lambda}\sqrt{TV_T \log T} + 2\sqrt{TV_T \log T} \\
= & \frac{\gamma G^2}{\lambda}\sqrt{\frac{TV_T}{\log T}} + \left(\frac{5\gamma G^2}{\lambda} + 2\right)\sqrt{TV_T \log T}.
\end{aligned}$$

Otherwise, we choose $\tau = T$, and have

$$\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \quad & \frac{(\gamma + 5\gamma \log T)G^2}{\lambda} + 2TV_T \\
& \frac{(\gamma + 5\gamma \log T)G^2}{\lambda} + 2T\frac{\log T}{T} \\
= & \frac{\gamma G^2}{\lambda} + \left(\frac{5\gamma G^2}{\lambda} + 2\right)\log T.
\end{aligned}$$

In summary, we have

$$\begin{aligned}
\text{D-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \quad & \max \begin{cases} \frac{\gamma G^2}{\lambda} + \left(\frac{5\gamma G^2}{\lambda} + 2\right)\log T \\ \frac{\gamma G^2}{\lambda}\sqrt{\frac{TV_T}{\log T}} + \left(\frac{5\gamma G^2}{\lambda} + 2\right)\sqrt{TV_T \log T} \end{cases} \\
= & O\left(\max\left\{\log T, \sqrt{TV_T \log T}\right\}\right).
\end{aligned}$$