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

$$ \begin{cases} 
1 \times K^0 + \sum_{j=1}^{k} \beta_j K^j, & 2 \times K^0 + \sum_{j=1}^{k} \beta_j K^j, \ldots, \beta_0 \times K^0 + \sum_{j=1}^{k} \beta_j K^j \\
1 \times K^1 + \sum_{j=2}^{k} \beta_j K^j, & 2 \times K^1 + \sum_{j=2}^{k} \beta_j K^j, \ldots, \beta_1 \times K^1 + \sum_{j=2}^{k} \beta_j K^j \\
\vdots \\
1 \times K^{k-1} + \beta_{k-1} K^k, & 1 \times K^{k-1} + \beta_k K^k, \ldots, \beta_{k-1} \times K^{k-1} + \beta_k K^k \\
1 \times K^k, & 2 \times K^k, \ldots, \beta_k K^k 
\end{cases} $$

The total number of alive integers are upper bounded by

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

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

$$ r = \sum_{j=0}^{k} \beta_j K^j. $$
Dynamic Regret of Strongly Adaptive Methods

We generate a sequence of segments as

\[ I_1 = [t_1, e^{t_1} - 1] = \left[ \sum_{j=0}^{k} \beta_j K^j, (\beta_1 + 1)K^1 + \sum_{j=2}^{k} \beta_j K^j - 1 \right], \]

\[ I_2 = [t_2, e^{t_2} - 1] = \left[ (\beta_1 + 1)K^1 + \sum_{j=2}^{k} \beta_j K^j, (\beta_2 + 1)K^2 + \sum_{j=3}^{k} \beta_j K^j - 1 \right], \]

\[ I_3 = [t_3, e^{t_3} - 1] = \left[ (\beta_2 + 1)K^2 + \sum_{j=3}^{k} \beta_j K^j, (\beta_3 + 1)K^3 + \sum_{j=4}^{k} \beta_j K^j - 1 \right], \]

\[ \ldots \]

\[ I_k = [t_k, e^{t_k} - 1] = \left[ (\beta_{k-1} + 1)K^{k-1} + \beta_k K^k, (\beta_k + 1)K^k - 1 \right], \]

\[ I_{k+1} = [t_{k+1}, e^{t_{k+1}} - 1] = \left[ (\beta_k + 1)K^k, K^{k+1} - 1 \right], \]

\[ I_{k+2} = [t_{k+2}, e^{t_{k+2}} - 1] = \left[ K^{k+1}, K^{k+2} - 1 \right], \]

\[ \ldots \]

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

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

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

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

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

\[ \frac{1}{\alpha} \left( \log t_j + 2 \sum_{t=t_j+1}^{t_j} \frac{1}{t} \right), \quad \forall j \in [m-1] \]

where \( w_{t_j}, \ldots, w_{e^{t_j} - 1} \) is the sequence of solutions generated by the expert \( E^{t_j} \). Similarly, for the last segment, we have

\[ \frac{1}{\alpha} \left( \log t_m + 2 \sum_{t=t_m+1}^{t_m} \frac{1}{t} \right). \]
By adding things together, we have
\[
\sum_{j=1}^{m-1} \left( \sum_{t=t_j}^{t_j-1} f_t(w_t) - f_t(w_{t_j}^i) \right) + \sum_{t=t_m}^{s} f_t(w_t) - f_t(w_{t_m}^i) \\
\leq \frac{1}{\alpha} \sum_{j=1}^{m} \log t_j + \frac{2}{\alpha} \sum_{t=r+1}^{s} \frac{1}{t} \leq \frac{m + 2}{\alpha} \log T.
\] (8)

According to the property of online Newton step (Hazan et al., 2007, Theorem 2), we have, for any \( w \in \Omega \),
\[
\sum_{t=t_j}^{t_j-1} f_t(w_t^i) - f_t(w) \leq 5d \left( \frac{1}{\alpha} + GB \right) \log T, \; \forall j \in [m - 1]
\] (9)
and
\[
\sum_{t=t_m}^{s} f_t(w_{t_m}^i) - f_t(w) \leq 5d \left( \frac{1}{\alpha} + GB \right) \log T.
\] (10)

Combining (8), (9), and (10), we have,
\[
\sum_{t=r}^{s} f_t(w_t) - \sum_{t=r}^{s} f_t(w) \leq \left( \frac{(5d + 1)m + 2}{\alpha} + 5dmGB \right) \log T
\]
for any \( w \in \Omega \).

**C. Proof of Lemma 2**

The gradient of \( \exp(-\alpha f(w)) \) is
\[
\nabla \exp(-\alpha f(w)) = \exp(-\alpha f(w)) - \alpha \nabla f(w) = -\alpha \exp(-\alpha f(w)) \nabla f(w).
\]
and the Hessian is
\[
\nabla^2 \exp(-\alpha f(w)) = -\alpha \exp(-\alpha f(w)) \nabla f(w) \nabla^\top f(w) - \alpha \exp(-\alpha f(w)) \nabla^2 f(w) = -\alpha \exp(-\alpha f(w)) \left( \alpha \nabla f(w) \nabla^\top f(w) - \nabla^2 f(w) \right).
\]
Thus, \( f(\cdot) \) is \( \alpha \)-exp-concave if
\[
\alpha \nabla f(w) \nabla^\top f(w) \preceq \nabla^2 f(w).
\]

We complete the proof by noticing
\[
\frac{\lambda}{G^2} \nabla f(w) \nabla^\top f(w) \preceq \lambda I \preceq \nabla^2 f(w).
\]

**D. Proof of Theorem 2**

Lemma 2 implies that all the \( \lambda \)-strongly convex functions are also \( \frac{1}{G^2} \)-exp-concave. As a result, we can reuse the proof of Theorem 1. Specifically, (8) with \( \alpha = \frac{1}{G^2} \) becomes
\[
\sum_{j=1}^{m-1} \left( \sum_{t=t_j}^{t_j-1} f_t(w_t) - f_t(w_{t_j}^i) \right) + \sum_{t=t_m}^{s} f_t(w_t) - f_t(w_{t_m}^i) \leq \frac{(m + 2)G^2}{\lambda} \log T.
\] (11)

According to the property of online gradient descent (Hazan et al., 2007, Theorem 1), we have, for any \( w \in \Omega \),
\[
\sum_{t=t_j}^{t_j-1} f_t(w_t^i) - f_t(w) \leq \frac{G^2}{2\lambda} (1 + \log T), \; \forall j \in [m - 1]
\] (12)
and
\[ \sum_{t=t_m}^{s} f_t(w_t^m) - f_t(w) \leq \frac{G^2}{2\lambda} (1 + \log T). \]  

Combining (11), (12), and (13), we have,
\[ \sum_{t=r}^{s} f_t(w_t) - \sum_{t=r}^{s} f_t(w) \leq \frac{G^2}{2\lambda} (m + (3m + 4) \log T) \]

for any \( w \in \Omega \).

**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 \( \tau \) 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 \( \tau \), and in this case \( k = T/\tau \). Then, we have
\[
D\text{-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq \min_{1 \leq \tau \leq T} \sum_{i=1}^{k} \left( \text{SA-Regret}(T, \tau) + 2\tau V_T(i) \right)
= \min_{1 \leq \tau \leq T} \left( \frac{\text{SA-Regret}(T, \tau) T}{\tau} + 2\tau \sum_{i=1}^{k} V_T(i) \right)
\leq \min_{1 \leq \tau \leq T} \left( \frac{\text{SA-Regret}(T, \tau) T}{\tau} + 2\tau V_T \right).
\]
Combining with Theorem 4, we have
\[
D\text{-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq \min_{1 \leq \tau \leq T} \left( \frac{(c + 8\sqrt{\tau \log T + 5}) T}{\sqrt{\tau}} + 2\tau V_T \right)
\]
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
\[
D\text{-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq \frac{(c + 8\sqrt{T \log T + 5}) T^{2/3} V_T^{1/3}}{\log^{1/6} T} + 2T^{2/3} V_T^{1/3} \log^{1/3} T
\leq \frac{(c + 8\sqrt{T}) T^{2/3} V_T^{1/3}}{\log^{1/6} T} + (2 + 8\sqrt{T}) T^{2/3} V_T^{1/3} \log^{1/3} T.
\]
Otherwise, we choose \( \tau = T \), and have
\[
D\text{-Regret}(\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*) \leq (c + 8\sqrt{\log T + 5}) \sqrt{T} + 2TV_T
\leq (c + 8\sqrt{T \log T + 5}) \sqrt{T} + 2T \sqrt{\frac{\log T}{T}}
\leq (c + 9\sqrt{T \log T + 5}) \sqrt{T}.
\]
Dynamic Regret of Strongly Adaptive Methods

In summary, we have

\[
\text{D-Regret}(w_1^*, \ldots, w_T^*) \leq \max \left\{ \frac{(c + 9\sqrt{77\log T} + 5)\sqrt{T}}{\log^{1/6} T}, \frac{(c + 8\sqrt{5})T^{2/3}V_T^{1/3}}{\log^{1/6} T} + 24T^{2/3}V_T^{1/3}\log^{1/3} T \right\} = O \left( \max \left\{ \sqrt{T\log T}, T^{2/3}V_T^{1/3}\log^{1/3} T \right\} \right).
\]

G. Proof of Corollary 6

The first part of Corollary 6 is a direct consequence of Theorem 1 by setting \( K = \lceil T^{1/\gamma} \rceil \).

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

\[
\text{D-Regret}(w_1^*, \ldots, w_T^*) \leq \min_{1 \leq \tau \leq T} \left\{ \left( \frac{(5d + 1)(\gamma + 1) + 2}{\alpha} + 5d(\gamma + 1)GB \right) \frac{T\log T}{\tau} + 2\tau V_T \right\}.
\]

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

\[
\text{D-Regret}(w_1^*, \ldots, w_T^*) \leq \left( \frac{(5d + 1)(\gamma + 1) + 2}{\alpha} + 5d(\gamma + 1)GB + 2 \right) \sqrt{T}V_T\log T.
\]

Otherwise, we choose \( \tau = T \), and have

\[
\text{D-Regret}(w_1^*, \ldots, w_T^*) \leq \left( \frac{(5d + 1)(\gamma + 1) + 2}{\alpha} + 5d(\gamma + 1)GB \right) \log T + 2TV_T
\]

\[
\leq \left( \frac{(5d + 1)(\gamma + 1) + 2}{\alpha} + 5d(\gamma + 1)GB \right) \log T + 2T \frac{\log T}{T}
\]

\[
= \left( \frac{(5d + 1)(\gamma + 1) + 2}{\alpha} + 5d(\gamma + 1)GB + 2 \right) \log T.
\]

In summary, we have

\[
\text{D-Regret}(w_1^*, \ldots, w_T^*) \leq \left( \frac{(5d + 1)(\gamma + 1) + 2}{\alpha} + 5d(\gamma + 1)GB + 2 \right) \max \left\{ \log T, \sqrt{T}V_T\log T \right\} = O \left( d \cdot \max \left\{ \log T, \sqrt{T}V_T\log T \right\} \right).
\]

H. Proof of Corollary 7

The first part of Corollary 7 is a direct consequence of Theorem 2 by setting \( K = \lceil T^{1/\gamma} \rceil \).

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

\[
\text{D-Regret}(w_1^*, \ldots, w_T^*) \leq \min_{1 \leq \tau \leq T} \left\{ \frac{G^2}{2\lambda} (\gamma + 1 + (3\gamma + 7) \log T) \frac{T}{\tau} + 2\tau V_T \right\}
\]

\[
\leq \min_{1 \leq \tau \leq T} \left\{ \frac{G^2}{\lambda\tau} (\gamma + 5\gamma \log T) \frac{T}{\tau} + 2\tau V_T \right\}
\]

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

$$D\text{-Regret}(w_1^*, \ldots, w_T^*) \leq \frac{\gamma G^2}{\lambda} \sqrt{T V_T \log T} + \frac{5\gamma G^2}{\lambda} \sqrt{T V_T \log T} + 2 \sqrt{T V_T \log T}$$

$$= \frac{\gamma G^2}{\lambda} \sqrt{T V_T \log T} + \left( \frac{5\gamma G^2}{\lambda} + 2 \right) \sqrt{T V_T \log T}.$$ 

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

$$D\text{-Regret}(w_1^*, \ldots, w_T^*) \leq \frac{(\gamma + 5\gamma \log T)G^2}{\lambda} + 2TV_T$$

$$\leq \frac{(\gamma + 5\gamma \log T)G^2}{\lambda} + 2T \log T$$

$$= \frac{\gamma G^2}{\lambda} + \left( \frac{5\gamma G^2}{\lambda} + 2 \right) \log T.$$ 

In summary, we have

$$D\text{-Regret}(w_1^*, \ldots, w_T^*) \leq \max \left\{ \frac{\gamma G^2}{\lambda} + \left( \frac{5\gamma G^2}{\lambda} + 2 \right) \log T, \right.$$ 

$$\left. \frac{\gamma G^2}{\lambda} \sqrt{T V_T \log T} + \left( \frac{5\gamma G^2}{\lambda} + 2 \right) \sqrt{T V_T \log T} \right\} = O \left( \max \left\{ \log T, \sqrt{T V_T \log T} \right\} \right).$$