Improved Dynamic Regret for Non-degenerate Functions

Lijun Zhang[*], **Tianbao Yang**[†], **Jinfeng Yi**[‡], **Rong Jin**[§], **and Zhi-Hua Zhou**[*]
[*]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[†]Department of Computer Science, The University of Iowa, Iowa City, USA
[‡]IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA
[§]Alibaba Group, Seattle, USA
zhanglj@lamda.nju.edu.cn, tianbao-yang@uiowa.edu, jinfengy@us.ibm.com
jinrong.jr@alibaba-inc.com, zhouzh@lamda.nju.edu.cn

Abstract

Recently, there has been a growing research interest in the analysis of dynamic regret, which measures the performance of an online learner against a sequence of local minimizers. By exploiting the strong convexity, previous studies have shown that the dynamic regret can be upper bounded by the path-length of the comparator sequence. In this paper, we illustrate that the dynamic regret can be further improved by allowing the learner to query the gradient of the function multiple times, and meanwhile the strong convexity can be weakened to other non-degenerate conditions. Specifically, we introduce the *squared* path-length, which could be much smaller than the path-length, as a new regularity of the comparator sequence. When multiple gradients are accessible to the learner, we first demonstrate that the dynamic regret of strongly convex functions can be upper bounded by the minimum of the path-length and the squared path-length. We then extend our theoretical guarantee to functions that are semi-strongly convex or self-concordant. To the best of our knowledge, this is the first time that semi-strong convexity and self-concordance are utilized to tighten the dynamic regret.

1 Introduction

Online convex optimization is a fundamental tool for solving a wide variety of machine learning problems [Shalev-Shwartz, 2011]. It can be formulated as a repeated game between a learner and an adversary. On the t-th round of the game, the learner selects a point $\mathbf{x}_t$ from a convex set $\mathcal{X}$ and the adversary chooses a convex function $f_t : \mathcal{X} \mapsto \mathbb{R}$. Then, the function is revealed to the learner, who incurs loss $f_t(\mathbf{x}_t)$. The standard performance measure is the regret, defined as the difference between the learner's cumulative loss and the cumulative loss of the optimal fixed vector in hindsight:

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}). \tag{1}$$

Over the past decades, various online algorithms, such as the online gradient descent [Zinkevich, 2003], have been proposed to yield sub-linear regret under different scenarios [Hazan et al., 2007, Shalev-Shwartz et al., 2007].

Though equipped with rich theories, the notion of regret fails to illustrate the performance of online algorithms in dynamic setting, as a *static* comparator is used in (1). To overcome this limitation, there has been a recent surge of interest in analyzing a more stringent metric—*dynamic* regret [Hall and Willett, 2013, Besbes et al., 2015, Jadbabaie et al., 2015, Mokhtari et al., 2016, Yang et al.,

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2016], in which the cumulative loss of the learner is compared against a sequence of local minimizers, i.e.,

$$R_T^* := R(\mathbf{x}_1^*, \ldots, \mathbf{x}_T^*) = \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_t^*) = \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T \min_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x}) \tag{2}$$

where $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})$. A more general definition of dynamic regret is to evaluate the difference of the cumulative loss with respect to any sequence of comparators $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{X}$ [Zinkevich, 2003].

It is well-known that in the worst-case, it is impossible to achieve a sub-linear dynamic regret bound, due to the arbitrary fluctuation in the functions. However, it is possible to upper bound the dynamic regret in terms of certain regularity of the comparator sequence or the function sequence. A natural regularity is the path-length of the comparator sequence, defined as

$$\mathcal{P}_T^* := \mathcal{P}(\mathbf{x}_1^*, \ldots, \mathbf{x}_T^*) = \sum_{t=2}^T \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\| \tag{3}$$

that captures the cumulative Euclidean norm of the difference between successive comparators. For convex functions, the dynamic regret of online gradient descent can be upper bounded by $O(\sqrt{T}\mathcal{P}_T^*)$ [Zinkevich, 2003]. And when all the functions are strongly convex and smooth, the upper bound can be improved to $O(\mathcal{P}_T^*)$ [Mokhtari et al., 2016].

In the aforementioned results, the learner uses the gradient of each function only *once*, and performs one step of gradient descent to update the intermediate solution. In this paper, we examine an interesting question: is it possible to improve the dynamic regret when the learner is allowed to query the gradient *multiple* times? Note that the answer to this question is no if one aims to promote the static regret in (1), according to the results on the minimax regret bound [Abernethy et al., 2008a]. We however show that when coming to the dynamic regret, multiple gradients can reduce the upper bound significantly. To this end, we introduce a new regularity—the *squared* path-length:

$$\mathcal{S}_T^* := \mathcal{S}(\mathbf{x}_1^*, \ldots, \mathbf{x}_T^*) = \sum_{t=2}^T \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|^2 \tag{4}$$

which could be much smaller than $\mathcal{P}_T^*$ when the local variations are small. For example, when $\|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\| = \Omega(1/\sqrt{T})$ for all $t \in [T]$, we have $\mathcal{P}_T^* = \Omega(\sqrt{T})$ but $\mathcal{S}_T^* = \Omega(1)$. We advance the analysis of dynamic regret in the following aspects.

- When all the functions are strongly convex and smooth, we propose to apply gradient descent multiple times in each round, and demonstrate that the dynamic regret is reduced from $O(\mathcal{P}_T^*)$ to $O(\min(\mathcal{P}_T^*, \mathcal{S}_T^*))$, provided the gradients of minimizers are small. We further present a matching lower bound which implies our result cannot be improved in general.
- When all the functions are semi-strongly convex and smooth, we show that the standard online gradient descent still achieves the $O(\mathcal{P}_T^*)$ dynamic regret. And if we apply gradient descent multiple times in each round, the upper bound can also be improved to $O(\min(\mathcal{P}_T^*, \mathcal{S}_T^*))$, under the same condition as strongly convex functions.
- When all the functions are self-concordant, we establish a similar guarantee if both the gradient and Hessian of the function can be queried multiple times. Specifically, we propose to apply the damped Newton method multiple times in each round, and prove an $O(\min(\mathcal{P}_T^*, \mathcal{S}_T^*))$ bound of the dynamic regret under appropriate conditions.[1]

Application to Statistical Learning Most studies of dynamic regret, including this paper do not make stochastic assumptions on the function sequence. In the following, we discuss how to interpret our results when facing the problem of statistical learning. In this case, the learner receives a sequence of losses $\ell(\mathbf{x}^\top \mathbf{z}_1, y_1), \ell(\mathbf{x}^\top \mathbf{z}_2, y_2), \ldots$, where $(\mathbf{z}_i, y_i)$'s are instance-label pairs sampled from a unknown distribution, and $\ell(\cdot, \cdot)$ measures the prediction error. To avoid the random fluctuation caused by sampling, we can set f_t as the loss averaged over a mini-batch of instance-label pairs. As a result, when the underlying distribution is stationary or drifts slowly, successive functions will be close to each other, and thus the path-length and the squared path-length are expected to be small.

[1] $\mathcal{P}_T^*$ and $\mathcal{S}_T^*$ are modified slightly when functions are semi-strongly convex or self-concordant.

2 Related Work

The static regret in (1) has been extensively studied in the literature [Shalev-Shwartz, 2011]. It has been established that the static regret can be upper bounded by $O(\sqrt{T})$, $O(\log T)$, and $O(\log T)$ for convex functions, strongly convex functions, and exponentially concave functions, respectively [Zinkevich, 2003, Hazan et al., 2007]. Furthermore, those upper bounds are proved to be minimax optimal [Abernethy et al., 2008a, Hazan and Kale, 2011].

The notion of dynamic regret is introduced by Zinkevich [2003]. If we choose the online gradient descent as the learner, the dynamic regret with respect to any comparator sequence $\mathbf{u}_1, \ldots, \mathbf{u}_T$, i.e., $R(\mathbf{u}_1, \ldots, \mathbf{u}_T)$, is on the order of $\sqrt{T}\mathcal{P}(\mathbf{u}_1, \ldots, \mathbf{u}_T)$. When a prior knowledge of $\mathcal{P}_T^*$ is available, the dynamic regret R_T^* can be upper bounded by $O(\sqrt{T\mathcal{P}_T^*})$ [Yang et al., 2016]. If all the functions are strongly convex and smooth, the upper bound of R_T^* can be improved to $O(\mathcal{P}_T^*)$ [Mokhtari et al., 2016]. The $O(\mathcal{P}_T^*)$ rate is also achievable when all the functions are convex and smooth, and all the minimizers $\mathbf{x}_t^*$'s lie in the interior of $\mathcal{X}$ [Yang et al., 2016].

Another regularity of the comparator sequence, which is similar to the path-length, is defined as

$$\mathcal{P}'(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \sum_{t=2}^{T} \|\mathbf{u}_t - \Phi_t(\mathbf{u}_{t-1})\|$$

where $\Phi_t(\cdot)$ is a dynamic model that predicts a reference point for the t-th round. The advantage of this measure is that when the comparator sequence follows the dynamical model closely, it can be much smaller than the path-length $\mathcal{P}(\mathbf{u}_1, \ldots, \mathbf{u}_T)$. A novel algorithm named dynamic mirror descent is proposed to take $\Phi_t(\mathbf{u}_{t-1})$ into account, and the dynamic regret $R(\mathbf{u}_1, \ldots, \mathbf{u}_T)$ is on the order of $\sqrt{T}\mathcal{P}'(\mathbf{u}_1, \ldots, \mathbf{u}_T)$ [Hall and Willett, 2013]. There are also some regularities defined in terms of the function sequence, such as the functional variation [Besbes et al., 2015]

$$\mathcal{F}_T := \mathcal{F}(f_1, \ldots, f_T) = \sum_{t=2}^{T} \max_{\mathbf{x} \in \mathcal{X}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| \tag{5}$$

or the gradient variation [Chiang et al., 2012]

$$\mathcal{G}_T := \mathcal{G}(f_1, \ldots, f_T) = \sum_{t=2}^{T} \max_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2. \tag{6}$$

Under the condition that $\mathcal{F}_T \leq F_T$ and F_t is given beforehand, a restarted online gradient descent is developed by Besbes et al. [2015], and the dynamic regret is upper bounded by $O(T^{2/3}F_T^{1/3})$ and $O(\log T\sqrt{TF_T})$ for convex functions and strongly convex functions, respectively.

The regularities mentioned above reflect different aspects of the learning problem, and are not directly comparable in general. Thus, it is appealing to develop an algorithm that adapts to the smaller regularity of the problem. Jadbabaie et al. [2015] propose an adaptive algorithm based on the optimistic mirror descent [Rakhlin and Sridharan, 2013], such that the dynamic regret is given in terms of all the three regularities ($\mathcal{P}_T^*$, $\mathcal{F}_T$, and $\mathcal{G}_T$). However, it relies on the assumption that the learner can calculate each regularity incrementally.

In the setting of prediction with expert advice, the dynamic regret is also referred to as tracking regret or shifting regret [Herbster and Warmuth, 1998, Cesa-bianchi et al., 2012]. The path-length of the comparator sequence is named as shift, which is just the number of times the expert changes. Another related performance measure is the adaptive regret, which aims to minimize the static regret over any interval [Hazan and Seshadhri, 2007, Daniely et al., 2015]. Finally, we note that the study of dynamic regret is similar to the competitive analysis in the sense that both of them compete against an optimal offline policy, but with significant differences in their assumptions and techniques [Buchbinder et al., 2012].

3 Online Learning with Multiple Gradients

In this section, we discuss how to improve the dynamic regret by allowing the learner to query the gradient multiple times. We start with strongly convex functions, and then proceed to semi-strongly convex functions, and finally investigate self-concordant functions.

Algorithm 1 Online Multiple Gradient Descent (OMGD)

Require: The number of inner iterations K and the step size η

Let $\mathbf{x}_1$ be any point in $\mathcal{X}$
for $t = 1, \ldots, T$ **do**
  Submit $\mathbf{x}_t \in \mathcal{X}$ and the receive loss $f_t : \mathcal{X} \mapsto \mathbb{R}$
  $\mathbf{z}_t^1 = \mathbf{x}_t$
  for $j = 1, \ldots, K$ **do**
    $$\mathbf{z}_t^{j+1} = \Pi_{\mathcal{X}} \left(\mathbf{z}_t^j - \eta \nabla f_t(\mathbf{z}_t^j) \right)$$
  end for
  $\mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$
end for

3.1 Strongly Convex and Smooth Functions

To be self-contained, we provide the definitions of strong convexity and smoothness.

Definition 1. *A function $f : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex, if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Definition 2. *A function $f : \mathcal{X} \mapsto \mathbb{R}$ is L-smooth, if*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Example 1. *The following functions are both strongly convex and smooth.*

1. *A quadratic form $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} - 2\mathbf{b}^\top \mathbf{x} + c$ where $aI \preceq \mathcal{A} \preceq bI$, $a > 0$ and $b < \infty$;*
2. *The regularized logistic loss $f(\mathbf{x}) = \log(1 + \exp(\mathbf{b}^\top \mathbf{x})) + \frac{\lambda}{2}\|\mathbf{x}\|^2$, where $\lambda > 0$.*

Following previous studies [Mokhtari et al., 2016], we make the following assumptions.

Assumption 1. *Suppose the following conditions hold for each $f_t : \mathcal{X} \mapsto \mathbb{R}$.*

1. *f_t is λ-strongly convex and L-smooth over $\mathcal{X}$;*
2. *$\|\nabla f_t(\mathbf{x})\| \leq G$, $\forall \mathbf{x} \in \mathcal{X}$.*

When the learner can query the gradient of each function only once, the most popular learning algorithm is the online gradient descent:

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta \nabla f_t(\mathbf{x}_t) \right)$$

where $\Pi_{\mathcal{X}}(\cdot)$ denotes the projection onto the nearest point in $\mathcal{X}$. Mokhtari et al. [2016] have established an $O(\mathcal{P}_T^*)$ bound of dynamic regret, as stated below.

Theorem 1. *Suppose Assumption 1 is true. By setting $\eta \leq 1/L$ in online gradient descent, we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \frac{1}{1-\gamma} G \mathcal{P}_T^* + \frac{1}{1-\gamma} G \|\mathbf{x}_1 - \mathbf{x}_1^*\|$$

where $\gamma = \sqrt{1 - \frac{2\lambda}{1/\eta + \lambda}}$.

We now consider the setting that the learner can access the gradient of each function multiple times. The algorithm is a natural extension of online gradient descent by performing gradient descent multiple times in each round. Specifically, in the t-th round, given the current solution $\mathbf{x}_t$, we generate a sequence of solutions, denoted by $\mathbf{z}_t^1, \ldots, \mathbf{z}_t^{K+1}$, where K is a constant independent from T, as follows:

$$\mathbf{z}_t^1 = \mathbf{x}_t, \quad \mathbf{z}_t^{j+1} = \Pi_{\mathcal{X}} \left(\mathbf{z}_t^j - \eta \nabla f_t(\mathbf{z}_t^j) \right), \ j = 1, \ldots, K.$$

Then, we set $\mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$. The procedure is named as Online Multiple Gradient Descent (OMGD) and is summarized in Algorithm 1.

By applying gradient descent multiple times, we are able to extract more information from each function and therefore are more likely to obtain a tight bound for the dynamic regret. The following theorem shows that the multiple accesses of the gradient indeed help improve the dynamic regret.

Theorem 2. *Suppose Assumption 1 is true. By setting $\eta \leq 1/L$ and $K = \lceil \frac{1/\eta+\lambda}{2\lambda} \ln 4 \rceil$ in Algorithm 1, for any constant $\alpha > 0$, we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \min \begin{cases} 2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, \\ \frac{\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2}{2\alpha} + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2. \end{cases}$$

When $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2$ is small, Theorem 2 can be simplified as follows.

Corollary 3. *Suppose $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2 = O(\mathcal{S}_T^*)$, from Theorem 2, we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) = O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right).$$

In particular, if $\mathbf{x}_t^*$ belongs to the relative interior of $\mathcal{X}$ (i.e., $\nabla f_t(\mathbf{x}_t^*) = 0$) for all $t \in [T]$, Theorem 2, as $\alpha \to 0$, implies

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \min\left(2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, 2L\mathcal{S}_T^* + L\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2\right).$$

Compared to Theorem 1, the proposed OMGD improves the dynamic regret from $O(\mathcal{P}_T^*)$ to $O\left(\min\left(\mathcal{P}_T^*, \mathcal{S}_T^*\right)\right)$, when the gradients of minimizers are small. Recall the definitions of $\mathcal{P}_T^*$ and $\mathcal{S}_T^*$ in (3) and (4), respectively. We can see that $\mathcal{S}_T^*$ introduces a square when measuring the difference between $\mathbf{x}_t^*$ and $\mathbf{x}_{t-1}^*$. In this way, if the local variations ($\|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|$'s) are small, $\mathcal{S}_T^*$ can be significantly smaller than $\mathcal{P}_T^*$, as indicated below.

Example 2. *Suppose $\|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\| = T^{-\tau}$ for all $t \geq 1$ and $\tau > 0$, we have*

$$\mathcal{S}_{T+1}^* = T^{1-2\tau} \ll \mathcal{P}_{T+1}^* = T^{1-\tau}.$$

In particular, when $\tau = 1/2$, we have $\mathcal{S}_{T+1}^* = 1 \ll \mathcal{P}_{T+1}^* = \sqrt{T}$.

$\mathcal{S}_T^*$ is also closely related to the gradient variation in (6). When all the $\mathbf{x}_t^*$'s belong to the relative interior of $\mathcal{X}$, we have $\nabla f_t(\mathbf{x}_t^*) = 0$ for all $t \in [T]$ and therefore

$$\mathcal{G}_T \geq \sum_{t=2}^{T} \|\nabla f_t(\mathbf{x}_{t-1}^*) - \nabla f_{t-1}(\mathbf{x}_{t-1}^*)\|^2 = \sum_{t=2}^{T} \|\nabla f_t(\mathbf{x}_{t-1}^*) - \nabla f_t(\mathbf{x}_t^*)\|^2 \geq \lambda^2 \mathcal{S}_T^* \tag{7}$$

where the last inequality follows from the property of strongly convex functions [Nesterov, 2004]. The following corollary is an immediate consequence of Theorem 2 and the inequality in (7).

Corollary 4. *Suppose Assumption 1 is true, and further assume all the $\mathbf{x}_t^*$'s belong to the relative interior of $\mathcal{X}$. By setting $\eta \leq 1/L$ and $K = \lceil \frac{1/\eta+\lambda}{2\lambda} \ln 4 \rceil$ in Algorithm 1, we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \min\left(2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, \frac{2L\mathcal{G}_T}{\lambda^2} + L\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2\right).$$

In Theorem 2, the number of accesses of gradients K is set to be a constant depending on the condition number of the function. One may ask whether we can obtain a tighter bound by using a larger K. Unfortunately, according to our analysis, even if we take $K = \infty$, which means $f_t(\cdot)$ is minimized exactly, the upper bound can only be improved by a constant factor and the order remains the same. A related question is whether we can reduce the value of K by adopting more advanced optimization techniques, such as the accelerated gradient descent [Nesterov, 2004]. This is an open problem to us, and will be investigated as a future work.

Finally, we prove that the $O(\mathcal{S}_T^*)$ bound is optimal for strongly convex and smooth functions.

Theorem 5. *For any online learning algorithm $\mathcal{A}$, there always exists a sequence of strongly convex and smooth functions $f_1, \ldots, f_T$, such that*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) = \Omega(\mathcal{S}_T^*)$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_T$ is the solutions generated by $\mathcal{A}$.

Thus, the upper bound in Theorem 2 cannot be improved in general.

3.2 Semi-strongly Convex and Smooth Functions

During the analysis of Theorems 1 and 2, we realize that the proof is built upon the fact that "when the function is strongly convex and smooth, gradient descent can reduce the distance to the optimal solution by a constant factor" [Mokhtari et al., 2016, Proposition 2]. From the recent developments in convex optimization, we know that a similar behavior also happens when the function is semi-strongly convex and smooth [Necoara et al., 2015, Theorem 5.2], which motivates the study in this section.

We first introduce the definition of semi-strong convexity [Gong and Ye, 2014].

Definition 3. *A function $f : \mathcal{X} \mapsto \mathbb{R}$ is semi-strongly convex over $\mathcal{X}$, if there exists a constant $\beta > 0$ such that for any $\mathbf{x} \in \mathcal{X}$*

$$f(\mathbf{x}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \geq \frac{\beta}{2} \left\| \mathbf{x} - \Pi_{\mathcal{X}^*}(\mathbf{x}) \right\|^2 \tag{8}$$

where $\mathcal{X}^* = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})\}$ is the set of minimizers of f over $\mathcal{X}$.

The semi-strong convexity generalizes several non-strongly convex conditions, such as the quadratic approximation property and the error bound property [Wang and Lin, 2014, Necoara et al., 2015]. A class of functions that satisfy the semi-strongly convexity is provided below [Gong and Ye, 2014].

Example 3. *Consider the following constrained optimization problem*

$$\min_{\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d} \; f(\mathbf{x}) = g(E\mathbf{x}) + \mathbf{b}^\top \mathbf{x}$$

where $g(\cdot)$ is strongly convex and smooth, and $\mathcal{X}$ is either $\mathbb{R}^d$ or a polyhedral set. Then, $f : \mathcal{X} \mapsto \mathbb{R}$ is semi-strongly convex over $\mathcal{X}$ with some constant $\beta > 0$.

Based on the semi-strong convexity, we assume the functions satisfy the following conditions.

Assumption 2. *Suppose the following conditions hold for each $f_t : \mathcal{X} \mapsto \mathbb{R}$.*

1. *f_t is semi-strongly convex over $\mathcal{X}$ with parameter $\beta > 0$, and L-smooth;*
2. *$\|\nabla f_t(\mathbf{x})\| \leq G$, $\forall \mathbf{x} \in \mathcal{X}$.*

When the function is semi-strongly convex, the optimal solution may not be unique. Thus, we need to redefine P_T^* and $\mathcal{S}_T^*$ to account for this freedom. We define

$$\mathcal{P}_T^* := \sum_{t=2}^{T} \max_{\mathbf{x} \in \mathcal{X}} \left\| \Pi_{\mathcal{X}_t^*}(\mathbf{x}) - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}) \right\|, \text{ and } \mathcal{S}_T^* := \sum_{t=2}^{T} \max_{\mathbf{x} \in \mathcal{X}} \left\| \Pi_{\mathcal{X}_t^*}(\mathbf{x}) - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}) \right\|^2$$

where $\mathcal{X}_t^* = \{\mathbf{x} \in \mathcal{X} : f_t(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})\}$ is the set of minimizers of f_t over $\mathcal{X}$.

In this case, we will use the standard online gradient descent when the learner can query the gradient only once, and apply the online multiple gradient descent (OMGD) in Algorithm 1, when the learner can access the gradient multiple times. Using similar analysis as Theorems 1 and 2, we obtain the following dynamic regret bounds for functions that are semi-strongly convex and smooth.

Theorem 6. *Suppose Assumption 2 is true. By setting $\eta \leq 1/L$ in online gradient descent, we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x}) \leq \frac{G\mathcal{P}_T^*}{1-\gamma} + \frac{G\|\mathbf{x}_1 - \bar{\mathbf{x}}_1\|}{1-\gamma}$$

where $\gamma = \sqrt{1 - \frac{\beta}{1/\eta + \beta}}$, and $\bar{\mathbf{x}}_1 = \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)$.

Thus, online gradient descent still achieves an $O(\mathcal{P}_T^*)$ bound of the dynamic regret.

Theorem 7. *Suppose Assumption 2 is true. By setting $\eta \leq 1/L$ and $K = \lceil \frac{1/\eta+\beta}{\beta} \ln 4 \rceil$ in Algorithm 1, for any constant $\alpha > 0$, we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) \leq \min \begin{cases} 2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \bar{\mathbf{x}}_1\| \\ \dfrac{G_T^*}{2\alpha} + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\|\mathbf{x}_1 - \bar{\mathbf{x}}_1\|^2 \end{cases}$$

where $G_T^* = \max_{\{\mathbf{x}_t^* \in \mathcal{X}_t^*\}_{t=1}^T} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t^*)\|^2$, and $\bar{\mathbf{x}}_1 = \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)$.

Again, when the gradients of minimizers are small, in other words, $G_T^* = O(\mathcal{S}_T^*)$, the proposed OMGD improves the dynamic regret form $O(\mathcal{P}_T^*)$ to $O(\min(\mathcal{P}_T^*, \mathcal{S}_T^*))$.

3.3 Self-concordant Functions

We extend our previous results to self-concordant functions, which could be non-strongly convex and even non-smooth. Self-concordant functions play an important role in interior-point methods for solving convex optimization problems. We note that in the study of bandit linear optimization [Abernethy et al., 2008b], self-concordant functions have been used as barriers for constraints. However, to the best of our knowledge, this is the first time that losses themselves are self-concordant.

The definition of self-concordant functions is given below [Nemirovski, 2004].

Definition 4. *Let $\mathcal{X}$ be a nonempty open convex set in $\mathbb{R}^d$ and f be a C^3 convex function defined on $\mathcal{X}$. f is called self-concordant on $\mathcal{X}$, if it possesses the following two properties:*

1. *$f(\mathbf{x}_i) \to \infty$ along every sequence $\{\mathbf{x}_i \in \mathcal{X}\}$ converging, as $i \to \infty$, to a boundary point of $\mathcal{X}$;*
2. *f satisfies the differential inequality*

$$|D^3 f(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]| \leq 2\left(\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h}\right)^{3/2}$$

for all $\mathbf{x} \in \mathcal{X}$ and all $\mathbf{h} \in \mathbb{R}^d$, where

$$D^3 f(x)[\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3] = \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3} |_{t_1=t_2=t_3=0} f(\mathbf{x} + t_1\mathbf{h}_1 + t_2\mathbf{h}_2 + t_3\mathbf{h}_3) \ .$$

Example 4. *We provide some examples of self-concordant functions below [Boyd and Vandenberghe, 2004, Nemirovski, 2004].*

1. *The function $f(x) = -\log x$ is self-concordant on $(0,\infty)$.*
2. *A convex quadratic form $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} - 2\mathbf{b}^\top\mathbf{x} + c$ where $A \in \mathbb{R}^{d\times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $c \in \mathbb{R}$, is self-concordant on $\mathbb{R}^d$.*
3. *If $f : \mathbb{R}^d \mapsto \mathbb{R}$ is self-concordant, and $A \in \mathbb{R}^{d\times k}$, $\mathbf{b} \in \mathbb{R}^d$, then $f(A\mathbf{x}+\mathbf{b})$ is self-concordant.*

Using the concept of self-concordance, we make the following assumptions.

Assumption 3. *Suppose the following conditions hold for each $f_t : \mathcal{X}_t \mapsto \mathbb{R}$.*

1. *f_t is self-concordant on domain $\mathcal{X}_t$;*
2. *f_t is non-degenerate on $\mathcal{X}_t$, i.e., $\nabla^2 f_t(\mathbf{x}) \succ 0$, $\forall x \in \mathcal{X}_t$;*
3. *f_t attains its minimum on $\mathcal{X}_t$, and denote $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}_t} f_t(\mathbf{x})$.*

Our approach is similar to previous cases except for the updating rule of $\mathbf{x}_t$. Since we do not assume functions are strongly convex, we need to take into account the second order structure when updating the current solution $\mathbf{x}_t$. Thus, we assume the learner can query both the gradient and Hessian of each function multiple times. Specifically, we apply the damped Newton method [Nemirovski, 2004] to update $\mathbf{x}_t$, as follows:

$$\mathbf{z}_t^1 = \mathbf{x}_t, \quad \mathbf{z}_t^{j+1} = \mathbf{z}_t^j - \frac{1}{1+\lambda_t(\mathbf{z}_t^j)} \left[\nabla^2 f_t(\mathbf{z}_t^j)\right]^{-1} \nabla f_t(\mathbf{z}_t^j), \ j = 1,\ldots,K$$

where

$$\lambda_t(\mathbf{z}_t^j) = \sqrt{\nabla f_t(\mathbf{z}_t^j)^\top \left[\nabla^2 f_t(\mathbf{z}_t^j)\right]^{-1} \nabla f_t(\mathbf{z}_t^j)}. \tag{9}$$

Algorithm 2 Online Multiple Newton Update (OMNU)

Require: The number of inner iterations K in each round

Let $\mathbf{x}_1$ be any point in $\mathcal{X}_1$
for $t = 1, \ldots, T$ **do**
  Submit $\mathbf{x}_t \in \mathcal{X}$ and the receive loss $f_t : \mathcal{X} \mapsto \mathbb{R}$
  $\mathbf{z}_t^1 = \mathbf{x}_t$
  for $j = 1, \ldots, K$ **do**

$$\mathbf{z}_t^{j+1} = \mathbf{z}_t^j - \frac{1}{1 + \lambda_t(\mathbf{z}_t^j)} \left[\nabla^2 f_t(\mathbf{z}_t^j)\right]^{-1} \nabla f_t(\mathbf{z}_t^j)$$

    where $\lambda_t(\mathbf{z}_t^j)$ is given in (9)
  end for
  $\mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$
end for

Then, we set $\mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$. Since the damped Newton method needs to calculate the inverse of the Hessian matrix, its complexity is higher than gradient descent. The procedure is named as Online Multiple Newton Update (OMNU) and is summarized in Algorithm 2.

To analyze the dynamic regret of OMNU, we redefine the two regularities $\mathcal{P}_T^*$ and $\mathcal{S}_T^*$ as follows:

$$\mathcal{P}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|_t = \sum_{t=2}^{T} \sqrt{(\mathbf{x}_t^* - \mathbf{x}_{t-1}^*)^\top \nabla^2 f_t(\mathbf{x}_t^*)(\mathbf{x}_t^* - \mathbf{x}_{t-1}^*)}$$

$$\mathcal{S}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|_t^2 = \sum_{t=2}^{T} (\mathbf{x}_t^* - \mathbf{x}_{t-1}^*)^\top \nabla^2 f_t(\mathbf{x}_t^*)(\mathbf{x}_t^* - \mathbf{x}_{t-1}^*)$$

where $\|\mathbf{h}\|_t = \sqrt{\mathbf{h}^\top \nabla^2 f_t(\mathbf{x}_t^*)\mathbf{h}}$. Compared to the definitions in (3) and (4), we introduce $\nabla^2 f_t(\mathbf{x}_t^*)$ when measuring the distance between $\mathbf{x}_t^*$ and $\mathbf{x}_{t-1}^*$. When functions are strongly convex and smooth, these definitions are equivalent up to constant factors. We then define a quantity to compare the second order structure of consecutive functions:

$$\mu = \max_{t=2,\ldots,T} \left\{ \lambda_{\max} \left(\left[\nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\right]^{-1/2} \nabla^2 f_t(\mathbf{x}_t^*) \left[\nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\right]^{-1/2} \right) \right\} \tag{10}$$

where $\lambda_{\max}(\cdot)$ computes the maximum eigenvalue of its argument. When all the functions are λ-strongly convex and L-smooth, $\mu \leq L/\lambda$. Then, we have the following theorem regarding the dynamic regret of the proposed OMNU algorithm.

Theorem 8. *Suppose Assumption 3 is true, and further assume*

$$\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_t^2 \leq \frac{1}{144}, \ \forall t \geq 2. \tag{11}$$

When $t = 1$, we choose $K = O(1)(f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*) + \log\log\mu)$ in OMNU such that

$$\|\mathbf{x}_2 - \mathbf{x}_1^*\|_1^2 \leq \frac{1}{144\mu}. \tag{12}$$

For $t \geq 2$, we set $K = \lceil \log_4(16\mu) \rceil$ in OMNU, then

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \min\left(\frac{1}{6}\mathcal{P}_T^*, 4\mathcal{S}_T^* + \frac{1}{36}\right) + f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*).$$

The above theorem again implies the dynamic regret can be upper bounded by $O(\min(\mathcal{P}_T^*, \mathcal{S}_T^*))$ when the learner can access the gradient and Hessian multiple times. From the first property of self-concordant functions in Definition 4, we know that $\mathbf{x}_t^*$ must lie in the interior of $\mathcal{X}_t$, and thus $\nabla f_t(\mathbf{x}_t^*) = 0$ for all $t \in [T]$. As a result, we do not need the additional assumption that the gradients of minimizers are small, which has been used before to simplify Theorems 2 and 7.

Compared to Theorems 2 and 7, Theorem 8 introduces an additional condition in (11). This condition is required to ensure that $\mathbf{x}_t$ lies in the feasible region of $f_t(\cdot)$, otherwise, $f_t(\mathbf{x}_t)$ can be infinity

and it is impossible to bound the dynamic regret. The multiple applications of damped Newton method can enforce $\mathbf{x}_t$ to be close to $\mathbf{x}_{t-1}^*$. Combined with (11), we conclude that $\mathbf{x}_t$ is also close to $\mathbf{x}_t^*$. Then, based on the property of the Dikin ellipsoid of self-concordant functions [Nemirovski, 2004], we can guarantee that $\mathbf{x}_t$ is feasible for $f_t(\cdot)$.

4 Conclusion and Future Work

In this paper, we discuss how to reduce the dynamic regret of online learning by allowing the learner to query the gradient/Hessian of each function multiple times. By applying gradient descent multiple times in each round, we show that the dynamic regret can be upper bounded by the minimum of the path-length and the squared path-length, when functions are strongly convex and smooth. We then extend this theoretical guarantee to functions that are semi-strongly convex and smooth. We finally demonstrate that for self-concordant functions, applying the damped Newton method multiple times achieves a similar result.

In the current study, we upper bound the dynamic regret in terms of the path-length or the squared path-length of the comparator sequence. As we mentioned before, there also exist some regularities defined in terms of the function sequence, e.g., the functional variation [Besbes et al., 2015]. In the future, we will investigate whether multiple accesses of gradient/Hessian can improve the dynamic regret when measured by certain regularities of the function sequence. Another future work is to extend our results to the more general dynamic regret

$$R(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{X}$ is an arbitrary sequence of comparators [Zinkevich, 2003].

Acknowledgments

This work was partially supported by the NSFC (61603177, 61333014), JiangsuSF (BK20160658), NSF (IIS-1545995), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal stragies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008a.

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 263–274, 2008b.

O. Besbes, Y. Gur, and A. Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63 (5):1227–1244, 2015.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

N. Buchbinder, S. Chen, J. S. Naor, and O. Shamir. Unified algorithms for online learning and competitive analysis. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

N. Cesa-bianchi, P. Gaillard, G. Lugosi, and G. Stoltz. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25*, pages 980–988, 2012.

C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

A. Daniely, A. Gonen, and S. Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of The 32nd International Conference on Machine Learning*, 2015.

P. Gong and J. Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *ArXiv e-prints*, arXiv:1406.1102, 2014.

E. C. Hall and R. M. Willett. Dynamical models and tracking regret in online convex programming. In *Proceedings of the 30th International Conference on Machine Learning*, pages 579–587, 2013.

E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

E. Hazan and C. Seshadhri. Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity*, 88, 2007.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan. Online optimization: Competing with dynamic comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. *ArXiv e-prints*, arXiv:1603.04954, 2016.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *ArXiv e-prints*, arXiv:1504.06298, 2015.

A. Nemirovski. Interior point polynomial time methods in convex programming. Lecture notes, Technion – Israel Institute of Technology, 2004.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization*. Kluwer Academic Publishers, 2004.

S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26*, pages 3066–3074, 2013.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, pages 807–814, 2007.

P.-W. Wang and C.-J. Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15:1523–1548, 2014.

T. Yang, L. Zhang, R. Jin, and J. Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Supplementary Material: Improved Dynamic Regret for Non-degenerate Functions

Lijun Zhang[*], **Tianbao Yang**[†], **Jinfeng Yi**[‡], **Rong Jin**[§], **and Zhi-Hua Zhou**[*]
[*]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[†]Department of Computer Science, The University of Iowa, Iowa City, USA
[‡]IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA
[§]Alibaba Group, Seattle, USA
zhanglj@lamda.nju.edu.cn, tianbao-yang@uiowa.edu, jinfengy@us.ibm.com
jinrong.jr@alibaba-inc.com, zhouzh@lamda.nju.edu.cn

A Proof of Theorem 1

For the sake of completeness, we include the proof of Theorem 1, which was proved by Mokhtari et al. [2016]. We need the following property of gradient descent.

Lemma 1. *Assume that $f : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex and L-smooth, and $\mathbf{x}_* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. Let $\mathbf{v} = \Pi_{\mathcal{X}}(\mathbf{u} - \eta\nabla f(\mathbf{u}))$, where $\eta \leq 1/L$. We have*

$$\|\mathbf{v} - \mathbf{x}_*\| \leq \sqrt{1 - \frac{2\lambda}{1/\eta + \lambda}}\|\mathbf{u} - \mathbf{x}_*\|.$$

The constant in the above lemma is better than that in Proposition 2 of Mokhtari et al. [2016].

Since $\|\nabla f_t(\mathbf{x})\| \leq G$ for any $t \in [T]$ and any $\mathbf{x} \in \mathcal{X}$, we have

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq G\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|. \tag{13}$$

We now proceed to bound $\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|$. By the triangle inequality, we have

$$\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\| \leq \|\mathbf{x}_1 - \mathbf{x}_1^*\| + \sum_{t=2}^{T}\left(\|\mathbf{x}_t - \mathbf{x}_{t-1}^*\| + \|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|\right). \tag{14}$$

Since

$$\mathbf{x}_t = \Pi_{\mathcal{X}}\left(\mathbf{x}_{t-1} - \eta\nabla f_{t-1}(\mathbf{x}_{t-1})\right)$$

using Lemma 1, we have

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}^*\| \leq \gamma\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|. \tag{15}$$

From (14) and (15), we have

$$\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\| \leq \|\mathbf{x}_1 - \mathbf{x}_1^*\| + \gamma\sum_{t=2}^{T}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\| + \mathcal{P}_T^* \leq \|\mathbf{x}_1 - \mathbf{x}_1^*\| + \gamma\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\| + \mathcal{P}_T^*$$

implying

$$\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\| \leq \frac{1}{1-\gamma}\mathcal{P}_T^* + \frac{1}{1-\gamma}\|\mathbf{x}_1 - \mathbf{x}_1^*\|. \tag{16}$$

We complete the proof by substituting (16) into (13).

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

B Proof of Lemma 1

We first introduce the following property of strongly convex functions [Hazan and Kale, 2011].

Lemma 2. *Assume that $f : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex, and $\mathbf{x}_* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. Then, we have*

$$f(\mathbf{x}) - f(\mathbf{x}_*) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_*\|^2, \ \forall \mathbf{x} \in \mathcal{X}. \tag{17}$$

From the updating rule, we have

$$\mathbf{v} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \ f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{x} - \mathbf{u}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{u}\|^2.$$

According to Lemma 2, we have

$$\begin{aligned} &f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle + \frac{1}{2\eta}\|\mathbf{v} - \mathbf{u}\|^2 \\ \leq &f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{x}_* - \mathbf{u}\rangle + \frac{1}{2\eta}\|\mathbf{x}_* - \mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{v} - \mathbf{x}_*\|^2. \end{aligned} \tag{18}$$

Since $f(\mathbf{x})$ is λ-strongly convex, we have

$$f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{x}_* - \mathbf{u}\rangle \leq f(\mathbf{x}_*) - \frac{\lambda}{2}\|\mathbf{x}_* - \mathbf{u}\|^2. \tag{19}$$

On the other hand, the smoothness assumption implies

$$f(\mathbf{v}) \leq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle + \frac{L}{2}\|\mathbf{v} - \mathbf{u}\|^2 \leq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle + \frac{1}{2\eta}\|\mathbf{v} - \mathbf{u}\|^2. \tag{20}$$

Combining (18), (19), and (20), we obtain

$$f(\mathbf{v}) \leq f(\mathbf{x}_*) - \frac{\lambda}{2}\|\mathbf{x}_* - \mathbf{u}\|^2 + \frac{1}{2\eta}\|\mathbf{x}_* - \mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{v} - \mathbf{x}_*\|^2. \tag{21}$$

Applying Lemma 2 again, we have

$$f(\mathbf{v}) - f(\mathbf{x}_*) \geq \frac{\lambda}{2}\|\mathbf{v} - \mathbf{x}_*\|^2. \tag{22}$$

We complete the proof by substituting (22) into (21) and rearranging.

C Proof of Theorem 2

Since $f_t(\cdot)$ is L-smooth, we have

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \langle \nabla f_t(\mathbf{x}_t^*), \mathbf{x}_t - \mathbf{x}_t^*\rangle + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2 \leq \|\nabla f_t(\mathbf{x}_t^*)\|\|\mathbf{x}_t - \mathbf{x}_t^*\| + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2.$$

Combining with the fact

$$\|\nabla f_t(\mathbf{x}_t^*)\|\|\mathbf{x}_t - \mathbf{x}_t^*\| \leq \frac{1}{2\alpha}\|\nabla f_t(\mathbf{x}_t^*)\|^2 + \frac{\alpha}{2}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2$$

for any $\alpha > 0$, we obtain

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \frac{1}{2\alpha}\|\nabla f_t(\mathbf{x}_t^*)\|^2 + \frac{L+\alpha}{2}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2.$$

Summing the above inequality over $t = 1, \ldots, T$, we get

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \frac{1}{2\alpha}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t^*)\|^2 + \frac{L+\alpha}{2}\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2. \tag{23}$$

We now proceed to bound $\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2$. We have

$$\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|^2 \leq \|\mathbf{x}_1 - \mathbf{x}_1^*\|^2 + 2\sum_{t=2}^{T}\left(\|\mathbf{x}_t - \mathbf{x}_{t-1}^*\|^2 + \|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|^2\right). \tag{24}$$

Recall the updating rule

$$\mathbf{z}_{t-1}^{j+1} = \Pi_{\mathcal{X}} \left(\mathbf{z}_{t-1}^{j} - \eta \nabla f_{t-1}(\mathbf{z}_{t-1}^{j}) \right), \ j = 1, \ldots, K.$$

From Lemma 1, we have

$$\|\mathbf{z}_{t-1}^{j+1} - \mathbf{x}_{t-1}^{*}\|^2 \leq \left(1 - \frac{2\lambda}{1/\eta + \lambda}\right) \|\mathbf{z}_{t-1}^{j} - \mathbf{x}_{t-1}^{*}\|^2$$

which implies

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}^{*}\|^2 = \|\mathbf{z}_{t-1}^{K+1} - \mathbf{x}_{t-1}^{*}\|^2 \leq \left(1 - \frac{2\lambda}{1/\eta + \lambda}\right)^K \|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{*}\|^2 \leq \frac{1}{4}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{*}\|^2 \tag{25}$$

where we choose $K = \lceil \frac{1/\eta+\lambda}{2\lambda} \ln 4 \rceil$ such that

$$\left(1 - \frac{2\lambda}{1/\eta + \lambda}\right)^K \leq \exp\left(-\frac{2K\lambda}{1/\eta + \lambda}\right) \leq \frac{1}{4}.$$

From (24) and (25), we have

$$\begin{aligned}
\sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{x}_t^{*}\|^2 \leq & \|\mathbf{x}_1 - \mathbf{x}_1^{*}\|^2 + \frac{1}{2}\sum_{t=2}^{T} \|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{*}\|^2 + 2\mathcal{S}_T^{*} \\
\leq & \|\mathbf{x}_1 - \mathbf{x}_1^{*}\|^2 + \frac{1}{2}\sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{x}_t^{*}\|^2 + 2\mathcal{S}_T^{*}
\end{aligned}$$

implying

$$\sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{x}_t^{*}\|^2 \leq 4\mathcal{S}_T^{*} + 2\|\mathbf{x}_1 - \mathbf{x}_1^{*}\|^2.$$

Substituting the above inequality into (23), we have

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^{*}) \leq \frac{1}{2\alpha}\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^{*})\|^2 + 2(L+\alpha)\mathcal{S}_T^{*} + (L+\alpha)\|\mathbf{x}_1 - \mathbf{x}_1^{*}\|^2$$

for all $\alpha \geq 0$. Finally, we show that the dynamic regret can still be upper bounded by $\mathcal{P}_T^{*}$. From the previous analysis, we have

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}^{*}\|^2 \overset{(25)}{\leq} \frac{1}{4}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{*}\|^2 \Rightarrow \|\mathbf{x}_t - \mathbf{x}_{t-1}^{*}\| \leq \frac{1}{2}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{*}\|.$$

Then, we can set $\gamma = 1/2$ in Theorem 1 and obtain

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^{*}) \leq 2G\mathcal{P}_T^{*} + 2G\|\mathbf{x}_1 - \mathbf{x}_1^{*}\|.$$

D Proof of Theorem 5

We will randomly generate a sequence of functions $f_t : \mathbb{R}^d \mapsto \mathbb{R}, t = 1, \ldots, T$, where each $f_t(\cdot)$ is independently sampled from a distribution $\mathcal{P}$. For any deterministic algorithm $\mathcal{A}$, it generates a sequence of solutions $\mathbf{x}_t \in \mathcal{X}, t = 1, \ldots, T$, we define the expected dynamic regret as

$$\mathrm{E}\left[R_T^{*}\right] = \mathrm{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^{*})\right].$$

We will show that there exists a distribution of strongly convex and smooth functions such that for any fixed algorithm $\mathcal{A}$, we have $\mathrm{E}[R_T^{*}] \geq \mathrm{E}[\mathcal{S}_T^{*}]$.

For each round t, we randomly sample a vector $\varepsilon_t \in \mathbb{R}^d$ from the Gaussian distribution $\mathcal{N}(0, I)$. Using ε_t, we create a function

$$f_t(\mathbf{x}) = 2\left\|\mathbf{x} - \tau\varepsilon_t\right\|^2$$

which is both strongly convex and smooth. Notice that $\mathbf{x}_t$ is independent from ε_t, and thus we can bound the expected dynamic regret as follows:

$$\mathrm{E}\left[R_T^*\right] = \sum_{t=1}^{T} \mathrm{E}\left[f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)\right] = 2\sum_{t=1}^{T} \mathrm{E}\left[\|\mathbf{x}_t\|^2 + d\tau^2\right] \geq 2dT\tau^2.$$

We furthermore bound $\mathcal{S}_T^*$ as follows

$$\mathrm{E}[\mathcal{S}_T^*] = \sum_{t=2}^{T} \mathrm{E}\left[\|\varepsilon_t - \varepsilon_{t-1}\|^2\tau^2\right] = 2d(T-1)\tau^2.$$

Therefore, $\mathrm{E}[R_T^*] \geq \mathrm{E}[\mathcal{S}_T^*]$. Hence, for any given algorithm $\mathcal{A}$, there exists a sequence of functions $f_1, \ldots, f_T$, such that $\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) = \Omega(\mathcal{S}_T^*)$.

E Proof of Theorem 6

The proof is similar to that of Theorem 1.

We need the following property of gradient descent when applied to semi-strongly convex and smooth functions Necoara et al. [2015], which is analogous to Lemma 1 developed for strongly convex functions.

Lemma 3. *Assume that $f(\cdot)$ is L-smooth and satisfies the semi-strong convexity condition in (8). Let $\mathbf{v} = \Pi_{\mathcal{X}}(\mathbf{u} - \eta\nabla f(\mathbf{u}))$, where $\eta \leq 1/L$. We have*

$$\|\mathbf{v} - \Pi_{\mathcal{X}^*}(\mathbf{v})\| \leq \sqrt{1 - \frac{\beta}{1/\eta + \beta}}\|\mathbf{u} - \Pi_{\mathcal{X}_*}(\mathbf{u})\|.$$

Since $\|\nabla f_t(\mathbf{x})\| \leq G$ for any $t \in [T]$ and any $\mathbf{x} \in \mathcal{X}$, we have

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t\left(\Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right) \leq G\sum_{t=1}^{T} \left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\|. \tag{26}$$

We now proceed to bound $\sum_{t=1}^{T} \|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\|$. By the triangle inequality, we have

$$\sum_{t=1}^{T} \left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\| \leq \left\|\mathbf{x}_1 - \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)\right\| + \sum_{t=2}^{T} \left(\left\|\mathbf{x}_t - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_t)\right\| + \left\|\Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_t) - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\|\right). \tag{27}$$

Since

$$\mathbf{x}_t = \Pi_{\mathcal{X}}\left(\mathbf{x}_{t-1} - \eta\nabla f_{t-1}(\mathbf{x}_{t-1})\right)$$

using Lemma 3, we have

$$\left\|\mathbf{x}_t - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_t)\right\| \leq \gamma\left\|\mathbf{x}_{t-1} - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_{t-1})\right\|. \tag{28}$$

From (27) and (28), we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\| \\
&\leq \left\|\mathbf{x}_1 - \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)\right\| + \gamma\sum_{t=2}^{T} \left\|\mathbf{x}_{t-1} - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_{t-1})\right\| + \sum_{t=2}^{T} \left\|\Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_t) - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\| \\
&\leq \left\|\mathbf{x}_1 - \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)\right\| + \gamma\sum_{t=1}^{T} \left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\| + \mathcal{P}_T^*
\end{aligned}$$

implying

$$\sum_{t=1}^{T} \left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\| \leq \frac{1}{1-\gamma}\mathcal{P}_T^* + \frac{1}{1-\gamma}\left\|\mathbf{x}_1 - \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)\right\|. \tag{29}$$

We complete the proof by substituting (29) into (26).

F Proof of Lemma 3

For the sake of completeness, we provide the proof of Lemma 3, which can also be found in the work of Necoara et al. [2015].

The analysis is similar to that of Lemma 1. Define

$$\bar{\mathbf{u}} = \Pi_{\mathcal{X}^*}(\mathbf{u}), \text{ and } \bar{\mathbf{v}} = \Pi_{\mathcal{X}^*}(\mathbf{v}).$$

From the optimality condition of $\mathbf{v}$, we have

$$\begin{aligned} & f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{1}{2\eta}\|\mathbf{v} - \mathbf{u}\|^2 \\ \leq & f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \bar{\mathbf{u}} - \mathbf{u} \rangle + \frac{1}{2\eta}\|\bar{\mathbf{u}} - \mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{v} - \bar{\mathbf{u}}\|^2. \end{aligned} \tag{30}$$

From the convexity of $f(\mathbf{x})$, we have

$$f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \bar{\mathbf{u}} - \mathbf{u} \rangle \leq f(\bar{\mathbf{u}}). \tag{31}$$

Combining (30), (31), and (20), we obtain

$$f(\mathbf{v}) \leq f(\bar{\mathbf{u}}) + \frac{1}{2\eta}\|\bar{\mathbf{u}} - \mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{v} - \bar{\mathbf{u}}\|^2. \tag{32}$$

From the semi-strong convexity of $f(\cdot)$, we further have

$$f(\mathbf{v}) - f(\bar{\mathbf{u}}) \geq \frac{\beta}{2}\left\|\mathbf{v} - \bar{\mathbf{v}}\right\|^2.$$

Substituting the above inequality into (32), we have

$$\frac{1}{2\eta}\|\bar{\mathbf{u}} - \mathbf{u}\|^2 \geq \frac{1}{2\eta}\|\mathbf{v} - \bar{\mathbf{u}}\|^2 + \frac{\beta}{2}\left\|\mathbf{v} - \bar{\mathbf{v}}\right\|^2 \geq \left(\frac{1}{2\eta} + \frac{\beta}{2}\right)\left\|\mathbf{v} - \bar{\mathbf{v}}\right\|^2$$

which completes the proof.

G Proof of Theorem 7

The proof is similar to that of Theorem 2. In the following, we just provide the key differences.

Following the derivation of (23), we get

$$\begin{aligned} \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) \leq & \frac{1}{2\alpha}\sum_{t=1}^{T}\left\|\nabla f_t\left(\Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right)\right\|^2 + \frac{L+\alpha}{2}\sum_{t=1}^{T}\left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\|^2 \\ \leq & \frac{1}{2\alpha}G_T^* + \frac{L+\alpha}{2}\sum_{t=1}^{T}\left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\|^2 \end{aligned} \tag{33}$$

for any $\alpha > 0$.

To bound $\sum_{t=1}^{T}\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\|^2$, we have

$$\sum_{t=1}^{T}\left\|\mathbf{x}_t - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\|^2 \leq \left\|\mathbf{x}_1 - \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)\right\|^2 + 2\sum_{t=2}^{T}\left(\left\|\mathbf{x}_t - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_t)\right\|^2 + \left\|\Pi_{\mathcal{X}_{t-1}^*}(\mathbf{x}_t) - \Pi_{\mathcal{X}_t^*}(\mathbf{x}_t)\right\|^2\right). \tag{34}$$

From Lemma 3 and the updating rule

$$\mathbf{z}_{t-1}^{j+1} = \Pi_{\mathcal{X}}\left(\mathbf{z}_{t-1}^{j} - \eta\nabla f_{t-1}(\mathbf{z}_{t-1}^{j})\right),\ j = 1, \ldots, K$$

we have

$$\left\|\mathbf{z}_{t-1}^{j+1} - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{z}_{t-1}^{j+1})\right\|^2 \leq \left(1 - \frac{\beta}{1/\eta + \beta}\right)\left\|\mathbf{z}_{t-1}^{j} - \Pi_{\mathcal{X}_{t-1}^*}(\mathbf{z}_{t-1}^{j})\right\|^2,\ j = 1, \ldots, K$$

which implies

$$\begin{aligned}&\left\|\mathbf{x}_t - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{x}_t)\right\|^2 = \left\|\mathbf{z}_{t-1}^{K+1} - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{z}_{t-1}^{K+1})\right\|^2 \\ \leq &\left(1 - \frac{\beta}{1/\eta+\beta}\right)^K \left\|\mathbf{x}_{t-1} - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{x}_{t-1})\right\|^2 \leq \frac{1}{4}\left\|\mathbf{x}_{t-1} - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{x}_{t-1})\right\|^2\end{aligned} \tag{35}$$

where we choose $K = \lceil \frac{1/\eta+\beta}{\beta} \ln 4 \rceil$ such that

$$\left(1 - \frac{\beta}{1/\eta+\beta}\right)^K \leq \exp\left(-\frac{K\beta}{1/\eta+\beta}\right) \leq \frac{1}{4}.$$

From (34) and (35), we have

$$\begin{aligned}\sum_{t=1}^T \left\|\mathbf{x}_t - \Pi_{\mathcal{X}^*_t}(\mathbf{x}_t)\right\|^2 &\leq \left\|\mathbf{x}_1 - \Pi_{\mathcal{X}^*_1}(\mathbf{x}_1)\right\|^2 + \frac{1}{2}\sum_{t=2}^T \left\|\mathbf{x}_{t-1} - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{x}_{t-1})\right\|^2 + 2\mathcal{S}_T^* \\ &\leq \left\|\mathbf{x}_1 - \Pi_{\mathcal{X}^*_1}(\mathbf{x}_1)\right\|^2 + \frac{1}{2}\sum_{t=1}^T \left\|\mathbf{x}_t - \Pi_{\mathcal{X}^*_t}(\mathbf{x}_t)\right\|^2 + 2\mathcal{S}_T^*\end{aligned} \tag{36}$$

implying

$$\sum_{t=1}^T \left\|\mathbf{x}_t - \Pi_{\mathcal{X}^*_t}(\mathbf{x}_t)\right\|^2 \leq 4\mathcal{S}_T^* + 2\left\|\mathbf{x}_1 - \Pi_{\mathcal{X}^*_1}(\mathbf{x}_1)\right\|^2.$$

Substituting the above inequality into (33), we have

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) \leq \frac{1}{2\alpha}G_T^* + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\left\|\mathbf{x}_1 - \Pi_{\mathcal{X}^*_1}(\mathbf{x}_1)\right\|^2, \ \forall \alpha \geq 0.$$

Finally, we show that the dynamic regret can still be upper bounded by $\mathcal{P}_T^*$. From the previous analysis, we have

$$\left\|\mathbf{x}_t - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{x}_t)\right\| \overset{(35)}{\leq} \frac{1}{2}\left\|\mathbf{x}_{t-1} - \Pi_{\mathcal{X}^*_{t-1}}(\mathbf{x}_{t-1})\right\|.$$

Then, we can set $\gamma = 1/2$ in Theorem 6 and obtain

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) \leq 2G\mathcal{P}_T^* + 2G\left\|\mathbf{x}_1 - \Pi_{\mathcal{X}^*_1}(\mathbf{x}_1)\right\|.$$

H Proof of Theorem 8

The inequality (12) follows directly from the result in Section 2.2.X.C of Nemirovski [2004]. To prove the rest of this theorem, we will use the following properties of self-concordant functions and the damped Newton method Nemirovski [2004].

Lemma 4. *Let $f(\mathbf{x})$ be a self-concordant function, and $\|\mathbf{h}\|_{\mathbf{x}} = \sqrt{\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h}}$. Then, all points within the Dikin ellipsoid $W_{\mathbf{x}}$ centered at $\mathbf{x}$, defined as $W_{\mathbf{x}} = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_{\mathbf{x}} \leq 1\}$, share similar second order structure. More specifically, for a given point $\mathbf{x}$ and for any $\mathbf{h}$ with $\|\mathbf{h}\|_{\mathbf{x}} \leq 1$, we have*

$$(1 - \|\mathbf{h}\|_{\mathbf{x}})^2 \nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{x}+\mathbf{h}) \preceq \frac{\nabla^2 f(\mathbf{x})}{(1-\|\mathbf{h}\|_{\mathbf{x}})^2}. \tag{37}$$

Define $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x})$. Then, we have

$$\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{x}^*} \leq \frac{\lambda(\mathbf{x})}{1-\lambda(\mathbf{x})} \tag{38}$$

where $\lambda(\mathbf{x}) = \sqrt{\mathbf{x}^\top \left[\nabla^2 f(\mathbf{x})\right]^{-1}\mathbf{x}}$.

Consider the the damped Newton method: $\mathbf{v} = \mathbf{u} - \frac{1}{1+\lambda(\mathbf{u})}\left[\nabla^2 f(\mathbf{u})\right]^{-1}\nabla f(\mathbf{u})$. Then, we have

$$\lambda(\mathbf{v}) \leq 2\lambda^2(\mathbf{u}). \tag{39}$$

We will also use the following inequality frequently

$$\begin{aligned}&\|\mathbf{x}\|_t^2 = \mathbf{x}^\top \nabla^2 f_t(\mathbf{x}_t^*)\mathbf{x} \\ =&\mathbf{x}^\top \left[\nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\right]^{\frac{1}{2}} \left[\nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\right]^{-\frac{1}{2}} \nabla^2 f_t(\mathbf{x}_t^*) \left[\nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\right]^{-\frac{1}{2}} \left[\nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\right]^{\frac{1}{2}} \mathbf{x} \\ \overset{(10)}{\leq}& \mu \mathbf{x}^\top \nabla^2 f_{t-1}(\mathbf{x}_{t-1}^*)\mathbf{x} = \mu \|\mathbf{x}\|_{t-1}^2.\end{aligned} \tag{40}$$

We will assume that for any $t \geq 2$,

$$\|\mathbf{x}_t - \mathbf{x}_t^*\|_t \leq \frac{1}{6} \tag{41}$$

which will be proved at the end of the analysis.

According to the Taylor's theorem, for any $t \geq 2$, we have

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) = \frac{1}{2}(\mathbf{x}_t - \mathbf{x}_t^*)^\top \nabla^2 f_t(\xi_t)(\mathbf{x}_t - \mathbf{x}_t^*)$$

where ξ_t is a point on the line segment between $\mathbf{x}_t$ and $\mathbf{x}_t^*$. Now, using the property of self-concordant functions, we have

$$\nabla^2 f_t(\xi_t) = \nabla^2 f_t(\mathbf{x}_t^* + \xi_t - \mathbf{x}_t^*) \overset{(37)}{\preceq} \frac{1}{(1 - \|\xi_t - \mathbf{x}_t^*\|_t)^2} \nabla^2 f_t(\mathbf{x}_t^*) \preceq \frac{1}{(1 - \|\mathbf{x}_t - \mathbf{x}_t^*\|_t)^2} \nabla^2 f_t(\mathbf{x}_t^*)$$

where we use the inequality in (41) to ensure $\|\mathbf{x}_t - \mathbf{x}_t^*\|_t \leq 1$. We thus have

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \frac{\|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2}{2(1 - \|\mathbf{x}_t - \mathbf{x}_t^*\|_t)^2} \overset{(41)}{\leq} \|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2.$$

As a result

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*) + \sum_{t=2}^T \|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2. \tag{42}$$

From (41) and (42), we immediately have

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*) + \frac{1}{6}\mathcal{P}_T^*.$$

We now proceed to bound $\sum_{t=2}^T \|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2$ by $\mathcal{S}_T^*$. First, we have

$$\sum_{t=2}^T \|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2 \leq \sum_{t=2}^T 2\left(\|\mathbf{x}_t - \mathbf{x}_{t-1}^*\|_t^2 + \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|_t^2\right) \overset{(40)}{\leq} 2\mu \sum_{t=2}^T \|\mathbf{x}_t - \mathbf{x}_{t-1}^*\|_{t-1}^2 + 2\mathcal{S}_T^*. \tag{43}$$

Next, we discuss how to bound $\sum_{t=2}^T \|\mathbf{x}_t - \mathbf{x}_{t-1}^*\|_{t-1}^2$. Since $\mathbf{x}_t$ is derived by applying the damped Newton method multiple times to the initial solution $\mathbf{x}_{t-1}$, we need to first bound $\lambda_{t-1}(\mathbf{x}_{t-1})$. To this end, we establish the following lemma.

Lemma 5. *Let $f(\mathbf{x})$ be a self-concordant function, and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x})$. If $\|\mathbf{u} - \mathbf{x}^*\|_{\mathbf{x}^*} < 1/2$, we have*

$$\lambda(\mathbf{u}) \leq \frac{1}{1 - 2\|\mathbf{u} - \mathbf{x}^*\|_{\mathbf{x}^*}} \|\mathbf{u} - \mathbf{x}^*\|_{\mathbf{x}^*}.$$

The above lemma implies

$$\lambda_{t-1}(\mathbf{x}_{t-1}) \leq \frac{1}{1 - 2\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_{t-1}} \|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_{t-1} \overset{(41)}{\leq} \min\left(\frac{3}{2}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_{t-1}, \frac{1}{4}\right). \tag{44}$$

Recall the updating rule

$$\mathbf{z}_{t-1}^{j+1} = \mathbf{z}_{t-1}^j - \frac{1}{1 + \lambda_{t-1}(\mathbf{z}_{t-1}^j)} \left[\nabla^2 f_{t-1}(\mathbf{z}_{t-1}^j)\right]^{-1} \nabla f_{t-1}(\mathbf{z}_{t-1}^j),\ j = 1, \ldots, K.$$

From Lemma 4, we have

$$\lambda_{t-1}(\mathbf{z}_{t-1}^{j+1}) \overset{(39)}{\leq} 2\lambda_{t-1}^2(\mathbf{z}_{t-1}^{j}),\ j=1,\ldots,K.$$

Since $\lambda_{t-1}(\mathbf{z}_{t-1}^1) = \lambda_{t-1}(\mathbf{x}_{t-1}) \leq 1/4$. By induction, it is easy to verify

$$\lambda_{t-1}(\mathbf{z}_{t-1}^{j}) \leq \frac{1}{4},\ j=1,\ldots,K,K+1. \tag{45}$$

Therefore,

$$\lambda_{t-1}(\mathbf{x}_t) = \lambda_{t-1}(\mathbf{z}_{t-1}^{K+1}) \leq \frac{1}{2}\lambda_{t-1}(\mathbf{z}_{t-1}^{K}) \leq \cdots \leq \frac{1}{2^K}\lambda_{t-1}(\mathbf{z}_{t-1}^{1}) = \frac{1}{2^K}\lambda_{t-1}(\mathbf{x}_{t-1}). \tag{46}$$

Again, using Lemma 4, we have

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}^*\|_{t-1} \overset{(38)}{\leq} \frac{\lambda_{t-1}(\mathbf{x}_t)}{1-\lambda_{t-1}(\mathbf{x}_t)} \overset{(45),(46)}{\leq} \frac{4}{3}\frac{1}{2^K}\lambda_{t-1}(\mathbf{x}_{t-1}) \overset{(44)}{\leq} \frac{2}{2^K}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_{t-1}$$

implying

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}^*\|_{t-1}^2 \leq \frac{4}{4^K}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_{t-1}^2. \tag{47}$$

Combining (43) with (47), we have

$$\begin{aligned}\sum_{t=2}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2 \leq& \frac{8\mu}{4^K}\sum_{t=3}^{T}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_{t-1}^2 + 2\mu\|\mathbf{x}_2 - \mathbf{x}_1^*\|_1^2 + 2\mathcal{S}_T^* \\ \leq& \frac{1}{2}\sum_{t=2}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2 + 2\mu\|\mathbf{x}_2 - \mathbf{x}_1^*\|_1^2 + 2\mathcal{S}_T^*\end{aligned} \tag{48}$$

where we use the fact $\frac{8\mu}{4^K} \leq 1/2$. From (48), we have

$$\sum_{t=2}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|_t^2 \leq 4\mu\|\mathbf{x}_2 - \mathbf{x}_1^*\|_1^2 + 4\mathcal{S}_T^* \overset{(12)}{\leq} \frac{1}{36} + 4\mathcal{S}_T^*. \tag{49}$$

Substituting (49) into (42), we obtain

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq 4\mathcal{S}_T^* + f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*) + \frac{1}{36}.$$

We now proceed to show that the inequality in (41) holds. For $t=2$, we have

$$\|\mathbf{x}_2 - \mathbf{x}_2^*\|_2^2 \leq 2\|\mathbf{x}_2 - \mathbf{x}_1^*\|_2^2 + 2\|\mathbf{x}_1^* - \mathbf{x}_2^*\|_2^2 \overset{(11),(40)}{\leq} 2\mu\|\mathbf{x}_2 - \mathbf{x}_1^*\|_1^2 + \frac{1}{72} \overset{(12)}{\leq} \frac{1}{36}.$$

Now, we suppose (41) is true for $t=2,\ldots,k$. We will prove (41) holds for $t=k+1$. We have

$$\begin{aligned}&\|\mathbf{x}_{k+1} - \mathbf{x}_{k+1}^*\|_{k+1}^2 \leq 2\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|_{k+1}^2 + 2\|\mathbf{x}_k^* - \mathbf{x}_{k+1}^*\|_{k+1}^2 \\ \overset{(11),(40)}{\leq}& 2\mu\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|_k^2 + \frac{1}{72} \overset{(47)}{\leq} \frac{8\mu}{4^K}\|\mathbf{x}_k - \mathbf{x}_k^*\|_k^2 + \frac{1}{72} \leq \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_k^*\|_k^2 + \frac{1}{72} \leq \frac{1}{36}.\end{aligned}$$

I Proof of Lemma 5

By the mean value theorem for vector-valued functions, we have

$$\nabla f(\mathbf{u}) = \nabla f(\mathbf{u}) - \nabla f(\mathbf{x}^*) = \int_0^1 \nabla^2 f\left(\mathbf{x}^* + \tau(\mathbf{u}-\mathbf{x}^*)\right)(\mathbf{u}-\mathbf{x}^*)\,\mathrm{d}\,\tau. \tag{50}$$

Define

$$g(\mathbf{x}) = \mathbf{x}^\top\left[\nabla^2 f(\mathbf{u})\right]^{-1}\mathbf{x}$$

which is a convex function of $\mathbf{x}$. Then, we have

$$\begin{aligned}\lambda^2(\mathbf{u}) &= \left\langle \nabla f(\mathbf{u}), \left[\nabla^2 f(\mathbf{u})\right]^{-1} \nabla f(\mathbf{u}) \right\rangle = g\left(\nabla f(\mathbf{u})\right) \\ &\overset{(50)}{=} g\left(\int_0^1 \nabla^2 f\left(\mathbf{x}^* + \tau(\mathbf{u}-\mathbf{x}^*)\right)(\mathbf{u}-\mathbf{x}^*)\,\mathrm{d}\,\tau\right) \leq \int_0^1 g\left(\nabla^2 f\left(\mathbf{x}^* + \tau(\mathbf{u}-\mathbf{x}^*)\right)(\mathbf{u}-\mathbf{x}^*)\right)\mathrm{d}\,\tau\end{aligned} \tag{51}$$

where the last step follows from Jensen's inequality.

Define $\xi_\tau = \mathbf{x}^* + \tau(\mathbf{u}-\mathbf{x}^*)$ which lies in the line segment between $\mathbf{u}$ and $\mathbf{x}^*$. In the following, we will provide an upper bound for

$$g\left(\nabla^2 f(\xi_\tau)(\mathbf{u}-\mathbf{x}^*)\right) = (\mathbf{u}-\mathbf{x}^*)^\top \nabla^2 f(\xi_\tau)\left[\nabla^2 f(\mathbf{u})\right]^{-1}\nabla^2 f(\xi_\tau)(\mathbf{u}-\mathbf{x}^*).$$

Following Lemma 4, we have

$$\nabla^2 f(\xi_\tau) = \nabla^2 f(\mathbf{x}^* + \xi_\tau - \mathbf{x}^*) \overset{(37)}{\preceq} \frac{1}{(1-\|\xi_\tau - \mathbf{x}^*\|_{\mathbf{x}^*})^2}\nabla^2 f(\mathbf{x}^*) \preceq \frac{1}{(1-\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*})^2}\nabla^2 f(\mathbf{x}^*), \tag{52}$$

$$\|\mathbf{u}-\xi_\tau\|_{\xi_\tau}^2 \overset{(52)}{\leq} \frac{\|\mathbf{u}-\xi_\tau\|_{\mathbf{x}^*}^2}{(1-\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*})^2} \leq \frac{\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*}^2}{(1-\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*})^2} < 1, \tag{53}$$

$$\nabla^2 f(\mathbf{u}) = \nabla^2 f(\xi_\tau + \mathbf{u} - \xi_\tau) \overset{(37)}{\succeq} (1-\|\mathbf{u}-\xi_\tau\|_{\xi_\tau})^2 \nabla^2 f(\xi_\tau) \overset{(53)}{\succeq} \left(\frac{1-2\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*}}{1-\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*}}\right)^2 \nabla^2 f(\xi_\tau). \tag{54}$$

As a result

$$\begin{aligned} g\left(\nabla^2 f(\xi_\tau)(\mathbf{u}-\mathbf{x}^*)\right) &\overset{(54)}{\leq} \left(\frac{1-\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*}}{1-2\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*}}\right)^2 \left\langle (\mathbf{u}-\mathbf{x}^*), \nabla^2 f(\xi_\tau)(\mathbf{u}-\mathbf{x}^*)\right\rangle \\ &\overset{(52)}{\leq} \frac{1}{(1-2\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*})^2}\|\mathbf{u}-\mathbf{x}^*\|_{\mathbf{x}^*}^2. \end{aligned} \tag{55}$$

We complete the proof by substituting (55) into (51).