Dynamic Online Gradient Descent with Improved Query Complexity: A Theoretical Revisit

Yawei Zhao[1], En Zhu[1], Xinwang Liu[1], and Jianping Yin[2]
[1]National University of Defense Technology, Changsha, 410073, China.
[2]Dongguan University of Technology, Dongguan, Guangdong, 523808, China.
E-mail: {zhaoyawei, enzhu, xinwangliu}@nudt.edu.cn; jpyin@dgut.edu.cn.

Abstract

We provide a new theoretical analysis framework to investigate online gradient descent in the dynamic environment. Comparing with the previous work, the new framework recovers the state-of-the-art dynamic regret, but does not require extra gradient queries for every iteration. Specifically, when functions are α strongly convex and β smooth, to achieve the state-of-the-art dynamic regret, the previous work requires $\mathcal{O}(\kappa)$ with $\kappa = \frac{\beta}{\alpha}$ queries of gradients at every iteration. But, our framework shows that the query complexity can be improved to be $\mathcal{O}(1)$, which does not depend on κ. The improvement is significant for ill-conditioned problems because that their objective function usually has a large κ.

1 Introduction

Online Gradient Descent (OGD) has drawn much attention in the community of machine learning Zhu and Xu [2015], Hazan and Seshadhri [2007], Hall and Willett [2015], Shalev-Shwartz [2012], Garber [2018], Bedi et al. [2018]. It is widely used in various applications such as online recommendation Song et al. [2008], search ranking Moon et al. [2010]. Generally, OGD is formulated as a game between a learner and an adversary. At the t-th round of the game, the learner submits $\mathbf{x}_t$ from the feasible set $\mathcal{X}$, and the adversary selects a function $f_t : \mathcal{X} \mapsto \mathbb{R}$. Then, the function f_t is returned to the learner, and incurs the loss $f_t(\mathbf{x}_t)$.

Recently, there has been a surge of interest in analyzing OGD by using the dynamic regret Zinkevich [2003], Mokhtari et al. [2016], Yang et al. [2016], Lei et al. [2017]. The dynamic regret is usually defined as

$$R_T^* = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^*), \tag{1}$$

where $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$. Unfortunately, it is well-known that a sublinear dynamic regret bound cannot be achieved in the worst case Zinkevich [2003]. The reason is that the functions $f_1, ..., f_T$ may be changed arbitrarily in the dynamic environment. But, it is possible to upper bound the dynamic regret in terms of certain regularity of the comparator sequence. Those regularities are usually defined as the *path length* Mokhtari et al. [2016], Yang et al. [2016]:

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

or *squared path length* Zhang et al. [2017]:

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

Table 1: Our method OGD recovers the state-of-the-art regret with improved query complexity.

Algo.	Obj. type	Dynamic regret	Avg. queries
Mokhtari et al. [2016]	strongly convex	$\mathcal{O}(\mathcal{P}_t^*)$	$\mathcal{O}(1)$
Zhang et al. [2017]	strongly convex	$\mathcal{O}(\min\{\mathcal{S}_T^*, \mathcal{P}_t^*\})$	$\mathcal{O}(\kappa)$
Ours	strongly convex	$\mathcal{O}(\min\{\mathcal{S}_T^*, \mathcal{P}_t^*\})$	$\mathcal{O}(1)$

They capture the cumulative Euclidean norm or the square of Euclidean norm of the difference between successive comparators. When all the functions $f_1, ..., f_T$ are α-strongly convex and β-smooth, the dynamic regret is bounded by $\mathcal{O}(\mathcal{P}_T^*)$ Mokhtari et al. [2016]. When the local variations are small, $\mathcal{S}_T^*$ is much smaller than $\mathcal{P}_T^*$. Thus, the state-of-the-art dynamic regret of OGD is improved to be $\mathcal{O}(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\})$ Zhang et al. [2017].

But, to achieve the state-of-the-art dynamic regret, i.e., $\mathcal{O}(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\})$, the variant of OGD in Zhang et al. [2017] has to query $\mathcal{O}(\kappa)$ gradients for every iteration. Here, $\kappa := \frac{\beta}{\alpha}$ represents the condition number for the β smooth and α strongly convex objective function f_t. For a large κ, the extremely large query complexity makes it not practical in the online setting. In the paper, we investigate the basic online gradient descent, and provide a new theoretical analysis framework. **Using the new analysis framework, we show that the dynamic regret $\mathcal{O}(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\})$ can be achieved with $\mathcal{O}(1)$, instead of $\mathcal{O}(\kappa)$ queries of gradients in Zhang et al. [2017].** Main theoretical results are outlined in Table 1 briefly.

The improvement of the query complexity is vitally important for ill-conditioned[1] problems Tarantola [2004] whose objective function usually has a large condition number, i.e., κ. Let us take the image deblurring problem as an example. Suppose we have a blurred image $\mathbf{y}$, which is modeled by using an unknown real image $\mathbf{x}$ and a blurring matrix $\mathbf{A}$. That is, $\mathbf{y} = \mathbf{A}\mathbf{x}$. Here, $\mathbf{A}$ is usually a non-singular matrix with a large condition number, e.g., $\kappa = 10^6$. We want to recover the real image $\mathbf{x}$ from the blurred image $\mathbf{y}$, that is, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. Comparing with the method in Zhang et al. [2017], our new analysis framework shows that OGD is good enough, and the required queries of gradients can be reduced by multiple orders.

The paper is organized as follows. Section 2 reviews the related work. Section 3 presents the preliminaries. Section 4 presents our theoretical analysis framework. Section 5 presents the improved bounds of regret and query complexity for the strongly convex case. Section 6 concludes the paper.

2 Related work

2.1 Regrets of OGD in the static environment.

Online gradient descent in the static environment has been extensively investigated over the last ten years. The sublinear static regrets for smooth or strongly convex functions have been obtained in many literatures Shalev-Shwartz [2012], Hazan [2016], Duchi et al. [2011], Zinkevich [2003]. Specifically, when $f_t(\cdot)$ is strongly convex, the regret of online gradient descent is $\mathcal{O}(\log T)$ Hazan [2016]. When $f_t(\cdot)$ is convex but not strongly convex, the regret of online gradient descent is $\mathcal{O}(\sqrt{T})$ Hazan [2016].

2.2 Regrets of OGD in the dynamic environment.

When all the functions $f_1, ..., f_T$ are α strongly-convex and β smooth, the dynamic regret of OGD is $\mathcal{O}(\mathcal{P}_T^*)$ Mokhtari et al. [2016], Yang et al. [2016]. If OGD queries $\mathcal{O}(\kappa)$ gradients at every iteration, the dynamic regret of OGD can be improved to be $\mathcal{O}(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\})$ Zhang et al. [2017]. But, our analysis framework shows that the $\mathcal{O}(1)$ gradient queries for every iteration is enough to obtain $\mathcal{O}(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\})$ dynamic regret. Additionally, there are some other regularities including the functional variation Zhu and Xu [2015], Besbes, Omar et al. [2015] and the gradient variation Chiang et al. [2012]. Those regularities measure different aspects of the variation in the dynamic environment. Since they are not comparable directly, some researchers

[1] 'ill-conditioned' may be notated by 'ill-posed' or 'badly posed' in some literatures.

consider to bound the dynamic regret by using the mixed regularity Jadbabaie et al. [2015]. Extending our theoretical framework to different regularities is an interesting avenue for future work.

Besides, the new proposed theoretical analysis framework is inspired by Joulani et al. [2017]. Joulani et al. [2017] provides a theoretical analysis framework in the static environment, but our theoretical analysis framework works in the dynamic environment.

3 Preliminaries

3.1 Notations and assumptions

We use the following notation.

- The bold lower-case letters, e.g., $\mathbf{x}$ represent vectors. The normal letters, e.g., β represent a scalar number.
- η_t represents the learning rate of Algorithm 1 at the t-th iteration, and $\eta_{\min} := \min\{\eta_1, ..., \eta_T\}$.
- The condition number κ is defined by $\kappa := \frac{\beta}{\alpha}$ for any β smooth and α strongly convex function f_t.
- $\|\cdot\|$ represents the l_2 norm of a vector.
- $\Pi_{\mathcal{X}}(\cdot)$ represents the projection to a set $\mathcal{X}$.
- $\mathcal{X}_t^* := \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$ represents the minimizer set at the t-th iteration.
- Bregman divergence $B_f(\mathbf{x}, \mathbf{y})$ is defined by $B_f(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$ for any function f.

In the paper, functions $\{f_t\}_{t=1}^T$ are assumed to be convex and β smooth (defined as follows).

Definition 1 (β smoothness)**.** *A function $f : \mathcal{X} \mapsto \mathbb{R}$ is β smooth, if, for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}$, we have $f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^2$.*

If the function f_t is β smooth, according to the definition of the Bregman divergence, we have $B_{f_t}(\mathbf{x}, \mathbf{y}) \le \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|^2$ holds for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}$. The other assumptions used in the paper are presented as follows.

Assumption 1 (α strong convexity)**.** *For any t, the function $f_t : \mathcal{X} \mapsto \mathbb{R}$ is α strongly convex. That is, for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}$, $f_t(\mathbf{y}) \ge f_t(\mathbf{x}) + \langle \nabla f_t(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|^2$.*

Assumption 2 (Boundedness of gradients)**.** *We assume $\|\nabla f_t(\mathbf{x}_t)\|^2 \le G$ for any t.*

Assumption 3 (Boundedness of the domain of $\mathbf{x}$)**.** *We assume $\|\mathbf{x}_t - \mathbf{x}_t^*\|^2 \le R$ for any t.*

The above assumptions, i.e., Assumptions 1-3, are the basic assumptions, which are used widely in previous researches Shalev-Shwartz [2012], Hazan [2016], Duchi et al. [2011], Zinkevich [2003]. Additionally, we make the following assumption, which is used to model the dynamic environment.

Assumption 4 (Boundedness of variations in the dynamic environment.)**.** *For any $i \in [T]$ and $j \in [T]$, there exists $V \ge 1$ such that $\|\mathbf{x}_{i+1}^* - \mathbf{x}_i^*\| \le V\|\mathbf{x}_{j+1}^* - \mathbf{x}_j^*\|$.*

Assumption 4 models the dynamic environment by using V. A small V means the environment changes mildly. The environment changes significantly with the increase of V.

Algorithm 1 OGD: Online Gradient Descent.

```
Require: The learning rate $\eta_t$ with $1 \leq t \leq T$.
for $t = 1, 2, ..., T$ do
    Submit $\mathbf{x}_t \in \mathcal{X}$ and receive the function $f_t$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$.
    Query the gradient $\nabla f_t(\mathbf{x}_t)$ of $f_t$.
    $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left( \mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$.
return $\mathbf{x}_{T+1}$
```

Algorithm 2 OMGD: Online Multiple Gradient Descent Zhang et al. [2017].

```
Require: The learning rate $\eta_t$ with $1 \leq t \leq T$.
for $t = 1, 2, ..., T$ do
    Submit $\mathbf{x}_t \in \mathcal{X}$ and receive the function $f_t$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$.
    $\mathbf{z}_t^{(1)} = \mathbf{x}_t$, and $K = \frac{\kappa+1}{2}$.
    for $j = 1, 2, ..., K$ do
        Query the gradient $\nabla f_t(\mathbf{z}_t^{(j)})$ of $f_t$.
        $\mathbf{z}_t^{(j+1)} = \Pi_{\mathcal{X}} \left( \mathbf{z}_t^{(j)} - \eta_t \nabla f_t(\mathbf{z}_t^{(j)}) \right)$.
    $\mathbf{x}_{t+1} = \mathbf{z}_t^{(K+1)}$.
        return $\mathbf{x}_{T+1}$
```

3.2 Algorithm

Recall the algorithm of the OGD. At the t-th iteration, it submits $\mathbf{x}_t$, and receives the loss function $f_t(\mathbf{x}_t)$. Querying the gradient of $f_t(\mathbf{x}_t)$, it updates $\mathbf{x}_t$ by using the projected gradient descent method. The details are presented in Algorithm 1.

Comparing with the state-of-the-art method, i.e., Algorithm 2, OGD only requires one query of gradient for every iteration, while Algorithm 2 requires $\frac{\kappa+1}{2}$ queries of gradient. When κ is large, the query complexity of Algorithm 2 is much higher than OGD. Comparing with OMGD, i.e., Algorithm 2, our new theoretical analysis framework shows that **OGD is good enough to recover the state-of-the-art dyanmic regret yielded by OMGD, but it only leads to $\mathcal{O}(1)$ query of gradient, instead of $\mathcal{O}(\kappa)$ queries of gradient required by OMGD.**

4 A new theoretical analysis framework

In the section, we first provide a modular analysis framework, which does not depend on the assumption on the functions. Then, equipped with the strongly convex assumption, it yields specific results.

4.1 High-level thought

Our original goal is equivalent to investigate whether the basic OGD, i.e., Algorithm 1 can obtain the state-of-the-art dynamic regret, i.e., $\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\}$. Using the divide-and-control strategy, we divide the dynamic regret of OGD into two parts.

1. The first part, denoted by R_T^{o}, is caused by the online setting in the dynamic environment. It does not depend on the strongly convex assumption on the function f_t.

2. The second part, denoted by R_T^{m}, is due to the projected gradient descent step in Algorithm 1. It depends on the assumption on the function f_t such as convexity or strong convexity.

In the paper, our first contribution is to provide an upper bound of R_T^{o} without the strongly convex assumption of f_t. Then, benefiting from the rich theoretical tools in the static optimization, we successfully bound R_T^{m} by using the strongly convex assumption of f_t.

4.2 Meta framework

Generally, the dynamic regret of OGD is bounded as follows.

Theorem 1. *For any $\eta_t > 0$ in Algorithm 1, the dynamic regret of OGD defined in* (1) *is bounded by*

$$R_T^* \le R_T^{\mathrm{o}} + R_T^{\mathrm{m}}$$

where

$$R_T^{\mathrm{o}} := \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(- \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 + \left\| \mathbf{x}_t^* - \mathbf{x}_t \right\|^2 \right)$$

and

$$R_T^{\mathrm{m}} := \sum_{t=1}^{T} \frac{1}{\eta_t} \left(-B_{\eta_t f_t}(\mathbf{x}_t^*, \mathbf{x}_t) + \eta_t (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1})) \right) + \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\frac{\beta\eta_t - 1}{2} \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|^2 \right).$$

In Theorem 1, R_T^{o} represents the regret due to the online setting, and R_T^{m} represents the regret due to the projected gradient descent updating step in Algorithm 1.

Remark 1. *Note that the upper bound of R_T^{m} depends on the strongly convex assumption of the function f_t.*

Theorem 2. *Use Assumption 4, and set $\eta_t > 0$ in Algorithm 1. Denote $\mathbf{x}_0^* = \mathbf{x}_1$ and $\eta_{\min} = \min\{\eta_1, ..., \eta_T\}$. For any $0 < \rho \le 1$, the regret due to the online setting, i.e., R_T^{o} is bounded by*

$$R_T^{\mathrm{o}} \le \frac{1 - \rho + 2\rho V}{2\eta_{\min}(1-\rho)} \mathcal{S}_T^* + \frac{1}{2\eta_1} \left\| \mathbf{x}_1^* - \mathbf{x}_1 \right\|^2 + \frac{1}{2} \left(\sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \right).$$

Remark 2. *Note that this upper bound of R_T^{o} does not depend on the strongly convex assumption of the function f_t. It still holds for the convex function f_t.*

Lemma 1 (Appeared in Proposition 2 in Mokhtari et al. [2016])**.** *Use Assumption 1. Let $\mathbf{v}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{v}_t - \eta_t \nabla f_t(\mathbf{v}_t) \right)$ and $\mathcal{X}_t^* := \operatorname{argmin}_{\mathbf{v} \in \mathcal{X}} f_t(\mathbf{v})$. Denote $\kappa = \frac{\beta}{\alpha}$. If $\eta_t \le \frac{1}{\beta}$ and $\rho = \sqrt{\frac{\kappa - 1}{\kappa}}$, we have $\left\| \mathbf{v}_{t+1} - \mathbf{x}_t^* \right\| \le \rho \left\| \mathbf{v}_t - \mathbf{x}_t^* \right\|$.*

According to Lemma 1, when f_t's are strongly convex, $0 < \rho < 1$ (See Lemma 1). When f_t's are just convex, $\rho = 1$ (that is, $\alpha \to 0$). Recall that R_T^{m} depends on the strongly convex assumption of f_t's. Equipped by Lemma 1, we find that as long as R_T^{m} is further bounded, we are able to provide an upper bound for the dynamic regret.

5 Improved regret and query complexity for strongly convex f_t

When all f_t's are smooth and strongly convex, the dynamic regret of our method OGD is upper bounded by the following theorem.

Theorem 3. *Use Assumptions 1, 2, 3 and 4. Setting $\eta_t = \eta = \frac{1}{2(\beta + \beta^2/\alpha)}$ in Algorithm 1, and $\rho = \sqrt{\frac{\kappa - 1}{\kappa}} < 1$, we bound the dynamic regret of OGD as*

$$R_T^* \le \min\{J_1, J_2\},$$

where

$$
\begin{aligned}
J_1 =&\frac{(1-\rho+2\rho V)\left(\beta+\frac{\beta^2}{\alpha}\right)}{1-\rho}\mathcal{S}_T^* + \left(\beta+\frac{\beta^2}{\alpha}\right)\left\|\mathbf{x}_1^*-\mathbf{x}_1\right\|^2 + \frac{1}{2\left(\beta+\frac{\beta^2}{\alpha}\right)}\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 \\
\lesssim&\mathcal{O}\left(\mathcal{S}_T^*+\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2\right),
\end{aligned}
$$

and

$$
J_2 =\frac{G\left\|\mathbf{x}_1-\mathbf{x}_1^*\right\|}{1-\rho}\mathcal{P}_T^* + \frac{G}{1-\rho} \lesssim \mathcal{O}\left(\mathcal{P}_T^*\right).
$$

Corollary 1. *Suppose $\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 = \mathcal{O}\left(\mathcal{S}_T^*\right)$. According to Theorem 3, the dynamic regret of OGD is bounded by*

$$
R_T^* \leq \min\{J_1, J_2\} \lesssim \mathcal{O}\left(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\}\right).
$$

Proof. Recall Assumption 3, and we have $\left\|\mathbf{x}_1^*-\mathbf{x}_1\right\|^2 \leq R$. When $\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 = \mathcal{O}\left(\mathcal{S}_T^*\right)$, we have $J_1 \lesssim \mathcal{O}\left(\mathcal{S}_T^*\right)$. Similarly, we have $J_2 \leq \frac{G\sqrt{R}}{1-\rho}\mathcal{P}_T^* + \frac{G}{1-\rho} \lesssim \mathcal{O}\left(\mathcal{P}_T^*\right)$. Thus, we finally obtain

$$
R_T^* \leq \min\{J_1, J_2\} \lesssim \mathcal{O}\left(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\}\right).
$$

It completes the proof.

□

Recall the previous method, i.e., Algorithm 2. Its dynamic regret has been proved, and we present it as follows.

Lemma 2 (Appeared in Theorem 3 and Corollary 4 in Zhang et al. [2017].)**.** *Use Assumptions 1, 2, and 3, and choose $\eta_t \leq \frac{1}{\beta}$ in Algorithm 2. Denote the dynamic regret of Algorithm 2 by $\tilde{R}_T^*$. Then, for any constant $\sigma > 0$, $\tilde{R}_T^*$ is bounded by*

$$
\tilde{R}_T^* \leq \min\{J_3, J_4\},
$$

where

$$
\begin{aligned}
J_3 =&2G\mathcal{P}_T^* + 2G\left\|\mathbf{x}_1-\mathbf{x}_1^*\right\| \lesssim \mathcal{O}\left(\mathcal{P}_T^*\right), \\
J_4 =&\frac{1}{2\sigma}\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 + (\beta+\sigma)\left(2\mathcal{S}_T^*+\left\|\mathbf{x}_1-\mathbf{x}_1^*\right\|^2\right) \\
\lesssim&\mathcal{O}\left(\mathcal{S}_T^*+\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2\right).
\end{aligned}
$$

Furthermore, suppose $\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 = \mathcal{O}\left(\mathcal{S}_T^*\right)$, and we thus have $\tilde{R}_T^* \lesssim \mathcal{O}\left(\min\{\mathcal{P}_T^*, \mathcal{S}_T^*\}\right)$.

Comparing with Lemma 2, our new result achieves the same bound of the regret. But, OGD, i.e., Algorithm 1, only requires one query of gradient for every iteration, which does not depend on κ, and thus outperforms Algorithm 2 by reduing the query complexity significantly. The following remarks hightlight the advantages of our analysis framework.

Remark 3. *Our analysis framework achieves the state-of-the-art dynamic regret presented in Zhang et al. [2017] with a constant factor, and outperforms the dynamic regret $\mathcal{O}(\mathcal{P}_T^*)$ presented in Mokhtari et al. [2016].*

Remark 4. *Our analysis framework shows that $\mathcal{O}(1)$ queries of gradients for every iteration is enough to achieve the state-of-the-art dynamic regret, but Zhang et al. [2017] requires $\mathcal{O}(\kappa)$ queries of gradients for every iteration.*

6 Conclusion

We provide a new theoretical analysis framework to analyze the regret and query complexity of OGD in the dynamic environment. Comparing with the previous work, our framework achieves the state-of-the-art dynamic regret, and improve the required queries of gradient to be $\mathcal{O}(1)$.

Proof of theorems.

Proof of Theorem 1:

Proof.

$$\begin{aligned} R_T^* =& \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\eta_t f_t(\mathbf{x}_t) - \eta_t f_t(\mathbf{x}_t^*)\right) \\ =& \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\underbrace{\langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle}_{I_1} - B_{\eta_t f_t}(\mathbf{x}_t^*, \mathbf{x}_t) \right) + \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\underbrace{\langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{I_2} \right). \end{aligned} \tag{2}$$

Now, we begin to bound I_1. According to Lemma 4, we obtain

$$I_1 \le \frac{1}{2} \left(- \left\|\mathbf{x}_t^* - \mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t^* - \mathbf{x}_t\right\|^2 - \left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2 \right). \tag{3}$$

After that, we begin to bound I_2.

$$\begin{aligned} I_2 &= \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\ &= \eta_t f_t(\mathbf{x}_t) - \eta_t f_t(\mathbf{x}_{t+1}) + \eta_t B_{f_t}(\mathbf{x}_{t+1}, \mathbf{x}_t) \\ &\le \eta_t (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1})) + \frac{\beta \eta_t}{2} \left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2. \end{aligned} \tag{4}$$

The last inequality holds because that all f_t's are β smooth. Substituting (3) and (4) into (2), we finally complete the proof. □

Proof of Theorem 2:

Proof. According to the cosine theorem, we have

$$- \left\|\mathbf{x}_t^* - \mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_{t+1}\right\|^2 \le 2 \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\| \left\|\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^*\right\| - \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\|^2. \tag{5}$$

According to Lemma 1, if f_t is convex and smooth, $\left\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\right\| \le \rho \left\|\mathbf{x}_t - \mathbf{x}_t^*\right\|$ holds for $0 < \rho \le 1$. Specifically, $0 < \rho < 1$ holds when f_t is strongly convex, and $\rho = 1$ holds when f_t is just convex. We thus have

$$2 \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\| \left\|\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^*\right\| - \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\|^2 \ge -\rho^2 \left\|\mathbf{x}_t - \mathbf{x}_t^*\right\|^2 + \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_{t+1}\right\|^2.$$

Let $A_{t+1} = \left\|\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^*\right\|$, $M_{t+1} = \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\|$, and we thus have

$$2A_{t+1}M_{t+1} - M_{t+1}^2 \ge A_{t+1}^2 - \rho^2 A_t^2,$$

that is, $(A_{t+1} - M_{t+1})^2 \leq \rho^2 A_t^2$. Thus, we have

$$\begin{aligned} A_{t+1} - M_{t+1} &\leq \rho A_t \\ \rho A_t - \rho M_t &\leq \rho^2 A_{t-1} \\ &\cdots \\ \rho^{t-1} A_2 - \rho^{t-1} M_2 &\leq \rho^t A_1. \end{aligned}$$

Summing up, we obtain

$$\begin{aligned} A_{t+1} \leq & \rho^t A_1 + \left(M_{t+1} + \rho M_t + ... + \rho^{t-1} M_2\right) \\ = & \rho^t \left\| \mathbf{x}_1 - \mathbf{x}_1^* \right\| + \sum_{i=2}^{t+1} \rho^{t+1-i} \left\| \mathbf{x}_i^* - \mathbf{x}_{i-1}^* \right\| \\ \overset{\textcircled{1}}{=} & \sum_{i=1}^{t+1} \rho^{t+1-i} \left\| \mathbf{x}_i^* - \mathbf{x}_{i-1}^* \right\| \\ = & \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| + \sum_{i=1}^{t} \rho^i \left\| \mathbf{x}_{t+1-i}^* - \mathbf{x}_{t-i}^* \right\|. \end{aligned} \tag{6}$$

$\textcircled{1}$ holds due to letting $\mathbf{x}_0^* = \mathbf{x}_1$.

Substituting (6) into (5), we obtain,

$$\begin{aligned} & - \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 + \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \\ \leq & 2 \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| A_{t+1} - \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 \\ \leq & \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 + 2 \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| \left(\sum_{i=1}^{t} \rho^i \left\| \mathbf{x}_{t+1-i}^* - \mathbf{x}_{t-i}^* \right\| \right) \\ \leq & \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 + 2V \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 \left(\sum_{i=1}^{t} \rho^i \right) \\ \leq & \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 + \frac{2\rho V}{1-\rho} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 \\ = & \frac{1-\rho+2\rho V}{1-\rho} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2. \end{aligned} \tag{7}$$

Thus, we obtain

$$\begin{aligned} & 2R_T^{\mathrm{o}} = \sum_{t=1}^{T} \frac{1}{\eta_t} \left(- \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 + \left\| \mathbf{x}_t^* - \mathbf{x}_t \right\|^2 \right) \\ = & \sum_{t=1}^{T-1} \frac{1}{\eta_t} \left(- \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 + \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \right) + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \\ & + \frac{1}{\eta_1} \left\| \mathbf{x}_1^* - \mathbf{x}_1 \right\|^2 - \frac{1}{\eta_T} \left\| \mathbf{x}_T^* - \mathbf{x}_{T+1} \right\|^2 \\ \leq & \sum_{t=1}^{T-1} \frac{1}{\eta_t} \left(- \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 + \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \right) + \frac{1}{\eta_1} \left\| \mathbf{x}_1^* - \mathbf{x}_1 \right\|^2 + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \\ \overset{\textcircled{1}}{\leq} & \sum_{t=1}^{T-1} \frac{1}{\eta_t} \left(\frac{1-\rho+2\rho V}{1-\rho} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|^2 \right) + \frac{1}{\eta_1} \left\| \mathbf{x}_1^* - \mathbf{x}_1 \right\|^2 + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_{t+1} \right\|^2 \end{aligned}$$

$$\leq \frac{1-\rho+2\rho V}{\eta_{\min}(1-\rho)}\mathcal{S}_T^* + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\left\|\mathbf{x}_{t+1}^*-\mathbf{x}_{t+1}\right\|^2 + \frac{1}{\eta_1}\left\|\mathbf{x}_1^*-\mathbf{x}_1\right\|^2.$$

Here, $\eta_{\min} = \min\{\eta_1, \eta_2, ..., \eta_T\}$. ① holds due to (7). Dividing $\frac{1}{2}$ on both sides, we complete the proof. □

Proof of Theorem 3:

Proof. When the function f_t is α strongly convex, we have

$$B_{f_t}(\mathbf{x}_t^*, \mathbf{x}_t) \geq \frac{\alpha}{2}\left\|\mathbf{x}_t^*-\mathbf{x}_t\right\|^2. \tag{8}$$

Substituting (8) into Theorem 1, we obtain

$$\begin{aligned}
&R_T^* \\
\leq &\sum_{t=1}^{T}\frac{1}{\eta_t}\left(-\frac{1}{2}\left\|\mathbf{x}_t^*-\mathbf{x}_{t+1}\right\|^2 + \frac{1-\alpha\eta_t}{2}\left\|\mathbf{x}_t^*-\mathbf{x}_t\right\|^2\right) + \sum_{t=1}^{T}\frac{1}{\eta_t}\left(\frac{\beta\eta_t-1}{2}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2\right) \\
&+\sum_{t=1}^{T}\frac{1}{\eta_t}\left(\eta_t(f_t(\mathbf{x}_t)-f_t(\mathbf{x}_{t+1}))\right) \\
\overset{\textcircled{1}}{\leq} &\sum_{t=1}^{T}\frac{1}{2\eta_t}\left(-\left\|\mathbf{x}_t^*-\mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t^*-\mathbf{x}_t\right\|^2\right) \\
&+\sum_{t=1}^{T}\frac{\eta_t\left(\beta+\frac{1}{2\eta_t}+\frac{\beta^2}{\alpha}\right)-1}{2\eta_t}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2 + \sum_{t=1}^{T}\eta_t\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 \\
\overset{\textcircled{2}}{\leq} &\sum_{t=1}^{T}\frac{1}{2\eta_t}\left(-\left\|\mathbf{x}_t^*-\mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t^*-\mathbf{x}_t\right\|^2\right) + \sum_{t=1}^{T}\eta_t\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2 \\
\leq &\frac{(1-\rho+2\rho V)\left(\beta+\frac{\beta^2}{\alpha}\right)}{1-\rho}\mathcal{S}_T^* + \left(\beta+\frac{\beta^2}{\alpha}\right)\left\|\mathbf{x}_1^*-\mathbf{x}_1\right\|^2 + \frac{1}{2\left(\beta+\frac{\beta^2}{\alpha}\right)}\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t^*)\right\|^2.
\end{aligned}$$

① holds due to (13) in Lemma 5 by setting $\theta_1 = \alpha$ and $\theta_2 = 2\eta_t$. ② holds because of $\eta_t = \frac{1}{2\left(\beta+\frac{\beta^2}{\alpha}\right)}$ for $1 \leq t \leq T$. The last inequality holds due to Theorem 2.

Combining Lemma 6, we finally complete the proof.

□

Proof of lemmas.

Lemma 3. *Denote $h(\mathbf{x}) = \langle\eta_t\nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{x}-\mathbf{x}_t\right\|^2$. If $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \eta_t\nabla f_t(\mathbf{x}_t)\right)$, we have*

$$\mathbf{x}_{t+1} \in \operatorname*{Argmin}_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x}).$$

Proof. Consider the following convex optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{X}}\langle\eta_t\nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{x}-\mathbf{x}_t\right\|^2 \tag{9}$$

Denote the optimum set is $\mathcal{X}_t^*$, that is, for any $\mathbf{x}^* \in \mathcal{X}_t^*$, $h(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x})$ holds.

According to the first-order optimality condition Boyd and Vandenberghe [2004], we have, for any $\mathbf{z} \in \mathcal{X}$ and $\mathbf{x}^* \in \mathcal{X}_t^*$,

$$\begin{aligned} 0 \le& \langle \nabla h(\mathbf{x}^*), \mathbf{z} - \mathbf{x}^* \rangle \\ =& \langle \eta_t \nabla f_t(\mathbf{x}_t) + \mathbf{x}^* - \mathbf{x}_t, \mathbf{z} - \mathbf{x}^* \rangle . \end{aligned} \tag{10}$$

Recall that $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$. Thus, we have

$$\begin{aligned} & \langle \eta_t \nabla f_t(\mathbf{x}_t) + \mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{z} - \mathbf{x}_{t+1} \rangle \\ =& \langle \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right) - (\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t)), \mathbf{z} - \mathbf{x}_{t+1} \rangle \\ \ge& 0. \end{aligned}$$

That is, $\mathbf{x}_{t+1}$ satisfies the first-order optimality condition of (9). It completes the proof. □

Lemma 4. *Use Assumption 3. For any minimizer $\mathbf{x}_t^* \in \mathcal{X}_t^*$ and $\mathcal{X}_t^* := \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})$, we have*

$$2 \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle \le - \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 - \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|^2 + \left\| \mathbf{x}_t^* - \mathbf{x}_t \right\|^2 . \tag{11}$$

Proof. First, we construct an auxiliary function $h(\cdot) = \langle \eta_t \nabla f_t(\mathbf{x}_t), \cdot \rangle + \frac{1}{2} \left\| \cdot - \mathbf{x}_t \right\|^2$. According to Lemma 3, we have $\mathbf{x}_{t+1} \in \operatorname*{Argmin}_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x})$. Let $\bar{\mathbf{x}} = \mathbf{x}_{t+1} + \tau(\mathbf{x}_t^* - \mathbf{x}_{t+1})$ with $\tau \in (0, 1]$.

$$\begin{aligned} 0 \le& h(\bar{\mathbf{x}}) - h(\mathbf{x}_{t+1}) \\ =& \langle \eta_t \nabla f_t(\mathbf{x}_t), \tau(\mathbf{x}_t^* - \mathbf{x}_{t+1}) \rangle + \frac{1}{2} \left\| \bar{\mathbf{x}} \right\|^2 - \frac{1}{2} \left\| \mathbf{x}_{t+1} \right\|^2 + \tau \langle \mathbf{x}_t, \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle . \end{aligned} \tag{12}$$

Dividing τ on both sides, we obtain

$$\begin{aligned} 0 \le& \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle + \frac{1}{2\tau} \left(\left\| \bar{\mathbf{x}} \right\|^2 - \left\| \mathbf{x}_{t+1} \right\|^2 \right) + \langle \mathbf{x}_t, \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle \\ \overset{\textcircled{1}}{\le}& \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle + \lim_{\tau \to 0^+} \frac{1}{\tau} \left(\frac{1}{2} \left\| \bar{\mathbf{x}} \right\|^2 - \frac{1}{2} \left\| \mathbf{x}_{t+1} \right\|^2 \right) + \langle \mathbf{x}_t, \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle \\ =& \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle + \lim_{\tau \to 0^+} \left(\frac{\tau}{2} \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 + \langle \mathbf{x}_{t+1}, \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle \right) + \langle \mathbf{x}_t, \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle \\ =& \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle + \langle \mathbf{x}_{t+1}, \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle + \langle \mathbf{x}_t, \mathbf{x}_{t+1} - \mathbf{x}_t^* \rangle \\ =& \langle \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}_{t+1} \rangle - \frac{1}{2} \left\| \mathbf{x}_t^* - \mathbf{x}_{t+1} \right\|^2 - \frac{1}{2} \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|^2 + \frac{1}{2} \left\| \mathbf{x}_t^* - \mathbf{x}_t \right\|^2 . \end{aligned}$$

① holds because that (12) holds for any $\tau \in (0, 1]$. Re-arranging the items, we prove the conclusion. □

Lemma 5. *Suppose that all f_t's are β smooth. For any $\theta_1 > 0$, $\theta_2 > 0$ and any minimizer $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})$, we have*

$$\begin{aligned} & f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1}) \\ \le& \frac{\theta_1}{2} \left\| \mathbf{x}_t - \mathbf{x}_t^* \right\|^2 + \left(\frac{\beta^2}{2\theta_1} + \frac{1}{2\theta_2} \right) \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|^2 + \frac{\theta_2}{2} \left\| \nabla f_t(\mathbf{x}_t^*) \right\|^2 . \end{aligned} \tag{13}$$

Proof. For any $\theta_1 > 0$ and $\theta_2 > 0$, we have

$$\begin{aligned} & f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1}) - \frac{\theta_2}{2} \left\| \nabla f_t(\mathbf{x}_t^*) \right\|^2 - \frac{1}{2\theta_2} \left\| \mathbf{x}_t - \mathbf{x}_{t+1} \right\|^2 \\ \le& \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \langle \nabla f_t(\mathbf{x}_t^*), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\theta_1}{2\beta^2} \left\| \nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_t^*) \right\|^2 + \frac{\beta^2}{2\theta_1} \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|^2 \\
&\leq \frac{\theta_1}{2} \left\| \mathbf{x}_t - \mathbf{x}_t^* \right\|^2 + \frac{\beta^2}{2\theta_1} \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|^2 .
\end{aligned}
$$

The last inequality holds because that all f_t's are β smooth. Re-arranging the items, we thus complete the proof. □

Lemma 6 (Appeared in Theorem 1 in Mokhtari et al. [2016])**.** *Suppose that Assumptions 1-3 hold, and all f_t are β smooth. Thus, $\rho = \sqrt{\frac{\kappa-1}{\kappa}} < 1$ with $\kappa := \frac{\beta}{\alpha}$. Set $\eta_t \leq \frac{1}{L}$ in OGD, i.e., Algorithm 1. The dynamic regret of OGD is bounded as*

$$
R_T^* \leq \frac{G \left\| \mathbf{x}_1 - \mathbf{x}_1^* \right\|}{1-\rho} \mathcal{P}_T^* + \frac{G}{1-\rho}.
$$

Acknowledgment

This work was supported by the National Key R & D Program of China 2018YFB1003203 and the National Natural Science Foundation of China (Grant No. 61672528, 61773392, and 61671463).

References

A. S. Bedi, P. Sarma, and K. Rajawat. Tracking moving agents via inexact online gradient descent algorithm. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):202–217, Feb 2018.

Besbes, Omar, Gur, Yonatan, and Zeevi, Assaf J. Non-Stationary Stochastic Optimization. *Operations Research*, 63(5):1227–1244, 2015.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

C. K. Chiang, T. Yang, C. J. Lee, M. Mahdavi, C. J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. *Journal of Machine Learning Research*, 23, 2012.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 12:2121–2159, 2011.

D. Garber. Fast Rates for Online Gradient Descent Without Strong Convexity via Hoffman's Bound. *arXiv.org*, 2018.

E. C. Hall and R. M. Willett. Online Convex Optimization in Dynamic Environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4): 157–325, 2016.

E. Hazan and C. Seshadhri. Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity*, 14, 2007.

A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan. Online Optimization : Competing with Dynamic Comparators. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 398–406, 2015.

P. Joulani, A. György, and C. Szepesvári. A Modular Analysis of Adaptive (Non-)Convex Optimization - Optimism, Composite Objectives, and Variational Bounds. In *Proceedings of International Conference on Algorithmic Learning Theory (ALT)*, 2017.

Y. Lei, L. Shi, and Z.-C. Guo. Convergence of unregularized online learning algorithms. *Journal of Machine Learning Research*, 18(1):6269–6301, Jan. 2017.

A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *Proceedings of IEEE Conference on Decision and Control (CDC)*, pages 7195–7201. IEEE, 2016.

T. Moon, L. Li, W. Chu, C. Liao, Z. Zheng, and Y. Chang. Online learning for recency search ranking using real-time user feedback. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, CIKM '10, pages 1501–1504, New York, NY, USA, 2010. ACM.

S. Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 515–522, New York, NY, USA, 2008.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2004. ISBN 0898715725.

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 34th International Conference on Machine Learning (ICML)*, 2016.

L. Zhang, T. Yang, J. Yi, R. Jin, and Z.-H. Zhou. Improved Dynamic Regret for Non-degenerate Functions. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2017.

C. Zhu and H. Xu. Online Gradient Descent in Function Space. *arXiv.org*, 2015.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 928–935, 2003.