Trading-Off Static and Dynamic Regret in Online Least-Squares and Beyond

Abstract

Recursive least-squares algorithms often use forgetting factors as a heuristic to adapt to non-stationary data streams. The first contribution of this paper rigorously characterizes the effect of forgetting factors for a class of online Newton algorithms. For exp-concave and strongly convex objectives, the algorithms achieve a dynamic regret of $\max\{O(\log T), O(\sqrt{TV})\}$, where V is a bound on the path length of the comparison sequence. In particular, we show how classic recursive least-squares with forgetting factor achieves this dynamic regret bound. By varying V, we obtain a trade-off between static and dynamic regret. In order to obtain more computationally efficient algorithms, our second contribution is a novel gradient descent step-size rule for smooth, strongly convex functions. Here, we obtain static regret of $O(T^{1-\beta})$ and dynamic regret of $O(T^{\beta}V^*)$, where $\beta \in (0,1)$ and V^* is the path length of the sequence of minimizers. By varying β, we obtain a trade-off between static and dynamic regret. Finally, we characterize the strongly convex problem and obtain the dynamic regret of $\max\{O(\log T), O(\sqrt{TV})\}$.

Introduction

Online learning algorithms are designed to solve prediction and learning problems for streaming data or batch data whose volume is too large to be processed all at once. Applications include online routing (Hazan 2016), online auctions (Blum et al. 2004), online classification and regression (Crammer et al. 2006), as well as online resource allocation (Yuan and Lamperski 2018).

The general procedure for online learning algorithms is as follows: at each time t, before the true time-dependent objective function $f_t(\theta)$ is revealed, we need to make the prediction, θ_t, based on the history of the observations $f_i(\theta)$, $i < t$. Then the value of $f_t(\theta_t)$ is the loss suffered due to the lack of the knowledge for the true objective function $f_t(\theta)$. Our prediction of θ is then updated to include the information of $f_t(\theta)$. This whole process is repeated until termination. The functions, $f_t(\theta)$, can be chosen from a function class in an arbitrary, possibly adversarial manner.

An import class of online learning problems is online convex optimization, (Zinkevich 2003), which focuses on the case of convex objective functions. The most basic performance metric in online convex optimization is static regret $\mathcal{R}_s$, which measures the difference between the algorithm's cumulative loss and the cumulative loss of the best fixed decision in hindsight (Cesa-Bianchi and Lugosi 2006). Formally, the static regret is defined by:

$$\mathcal{R}_s = \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta \in \mathcal{S}} \sum_{t=1}^{T} f_t(\theta)$$

where T is the time horizon, $\mathcal{S}$ is a compact convex constraint set with $\|x\| \leq D, \forall x \in \mathcal{S}$. Without loss of generality, we assume throughout the paper that $D \geq 1$.

Another performance metric is called the dynamic regret $\mathcal{R}_d(z_1^T)$ (Zinkevich 2003) which is defined as

$$\mathcal{R}_d(z_1^T) = \sum_{t=1}^{T} f_t(\theta_t) - \sum_{t=1}^{T} f_t(z_t)$$

where $\theta_t, z_t \in \mathcal{S}$, and z_1^T is an arbitrary comparator sequence. (Besbes, Gur, and Zeevi 2015) uses a specific sequence of z_1^T, which is $z_t = \theta_t^*$, the optimal solution of the current $f_t(\theta)$.

For the static regret $\mathcal{R}_s$, a number of algorithms are proposed to upper bound it in terms of the time horizon T under different properties of the convex function $f_t(\theta)$. For the general convex one, (Zinkevich 2003) showed that $\mathcal{R}_s$ can be upper bounded in the order of $O(\sqrt{T})$. For the case when $f_t(\theta)$ is either strongly convex or exp-concave over the convex set S, (Hazan, Agarwal, and Kale 2007) showed that we could upper bound the $\mathcal{R}_s$ in the order of $O(\log T)$. These two upper bounds were shown to be minimax optimal by (Abernethy et al. 2008). For the smooth convex $f_t(\theta)$, (Srebro, Sridharan, and Tewari 2010) proved that it can be upper bounded in terms of the cumulative loss of the fixed optimal solution, which is preferable when it is much smaller than T. Such sub-linear regret upper bounds guarantee that on average the predicted variable θ_t will converge to the global optimal solution θ^* as $T \to \infty$.

For the dynamic regret $\mathcal{R}_d$, it is usually not upper bounded

merely in terms of T. One notion is in terms of the path-length V (Zinkevich 2003), which is defined as

$$V = \sum_{t=2}^{T} \|z_t - z_{t-1}\| \tag{1}$$

And we use V^* when $z_t = \theta_t^*$.

According to (Zinkevich 2003), $\mathcal{R}_d$ can be upper bounded by $O(\sqrt{T}(1+V))$ when $f_t(\theta)$ is convex. Such upper bound is later improved by (Zhang, Lu, and Zhou 2018) to $O(\sqrt{T(1+V)})$ by running an order of $O(\log T)$ algorithms in parallel in order to cover the domain of the possible stepsizes. For the strongly convex and smooth function $f_t(\theta)$, $\mathcal{R}_d$ is improved to the order of $O(V^*)$ by (Mokhtari et al. 2016).

Other notions of the comparator sequence for the dynamic regret include the variant of path-length (Hall and Willett 2013), functional variation (Besbes, Gur, and Zeevi 2015), as well as gradient variation (Chiang et al. 2012).

The contributions of this paper are the following four folds:

1. For the α-exp-concave problem, we propose the Discounted Online Newton Step (D-ONS), which has improved performance of the dynamic regret upper bound $\max\{O(\log T), O(\sqrt{TV})\}$ as compared to the previous result $O(\sqrt{T(1+V)})$ in (Zhang, Lu, and Zhou 2018). Furthermore, it solves the open question of how to achieve the trade-off between dynamic and static regrets by a simple user-determined discounted factor $\beta \in (0,1)$.
2. Although the analysis in the α-exp-concave problem applies to some strongly convex and smooth problems, it is not computationally efficient. To circumvent such obstacle and further improve the regret trade-off performance in the strongly convex and smooth case, we start from the online least-squares problem. For the first time we can not only make the connection between the discounted recursive least-squares algorithm (Fabre and Gueguen 1986) and the regret guarantees it can achieve, but show that the two regrets' trade-off can be achieved by tuning the discounted factor $\beta \in (0,1)$ to achieve both $\mathcal{R}_s \leq O(T^{1-\beta})$ and $\mathcal{R}_d \leq O(T^{\beta}(1+V^*))$, which is an improved result compared to the result from the α-exp-concave case.
3. For the strongly convex and smooth case, we propose a new online gradient descent update inspired by the analysis in the online least-squares problem. This new update rule is not only computationally efficient but enjoys all the same improvements in the special least-squares problem. These user-determined regret trade-off results can provide the flexibility in the priority of the dynamic or static regret minimization while maintaining the other one in a reasonable order.
4. Inspired by the proposed new step-size rule, for the strongly convex case, we obtain $\max\{O(\log T), O(\sqrt{TV})\}$ dynamic regret bound for the general V, which is better than $O(\sqrt{T(1+V)})$ in (Zhang, Lu, and Zhou 2018).

Notation. For the n dimensional vector $\theta \in \mathbb{R}^n$, we use $\|\theta\|$ to denote the ℓ_2-norm. The gradient of the function f_t at time step t in terms of the θ is denoted as $\nabla f_t(\theta)$.

For the matrix $A \in \mathbb{R}^{m\times n}$, its transpose is denoted by $A^\top$ and $A^\top A$ denotes the matrix multiplication. The inverse of A is denoted as A^{-1}. When $m = n$, we use $\|A\|_2$ to represent the induced 2 norm of the square matrix. For the two square matrix $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$, $A \preceq B$ means $A - B$ is negative semi-definite, while $A \succeq B$ means $A - B$ is positive semi-definite. For a positive definite matrix, M, let $\|x\|_M^2 = x^\top M x$. The standard inner product between matrices is given by $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$. The determinant of a square matrix, A is denoted by $|A|$. We use I to represent the identity matrix.

Problem Statement and Motivation

In this section, we discuss the problems that we consider and the motivations behind them.

First, let us see the definition of the f_t being α-exp-concave (Cesa-Bianchi and Lugosi 2006):

Definition 1. *A convex function f_t: $\mathcal{S} \to \mathbb{R}$ is α-exp-concave if $\exp(-\alpha f_t)$ is concave over $\mathcal{S}$ and $\alpha > 0$.*

For the general convex problem, (Zhang, Lu, and Zhou 2018) shows that $\mathcal{R}_d(z_1^T) = \sum_{t=1}^{T} f_t(\theta_t) - \sum_{t=1}^{T} f_t(z_t) \leq O(\sqrt{T(1+V)})$.

But there are no results on the dynamic regret for the cases when f_t is either α-exp-concave or strongly convex. Considering the better static regret for these cases, the dynamic regret is also possible to be better than $O(\sqrt{T(1+V)})$.

As a result, we first consider the α-exp-concave case, in which we are interested in not only upper bounding the dynamic regret, but obtaining the trade-off between dynamic and static regret. To the best of our knowledge, there is only one previous work concerning the dynamic regret in this case (Zhang et al. 2018) when $V = V^*$, while no prior work has been done to do the general analysis of V and the trade-off between dynamic and static regrets as has been done in convex case (Zinkevich 2003; Zhang, Lu, and Zhou 2018).

Then we move to the strongly convex and smooth problem setup in order to further improve the existing results done in the α-exp-concave case.

As a special case of strongly convex and smooth setup, we first consider the online least-squares problem:

$$f_t(\theta) = \frac{1}{2}\|y_t - A_t\theta\|^2$$

where $A_t \in \mathbb{R}^{m\times n}$, $A_t^T A_t$ has full rank with $lI \preceq A_t^T A_t \preceq uI$, and $y_t \in \mathbb{R}^m$ comes from a bounded set with $\|y_t\| \leq D$.

In this setting, previous works consider to upper bound either the static regret using the so-called "Follow-The-Leader" update rule (Shalev-Shwartz and others 2012) or the dynamic regret in terms of V^* (Mokhtari et al. 2016). Besides these regret minimization based approach, the discounted recursive least-squares update (Fabre and Gueguen

1986) is also commonly used in order to track the changes of the environment.

However, there is no result that connects such discounted recursive update with the analysis of both static and dynamic regret guarantees, which is necessary considering the successful applications such as (Duong 2017). Moreover, we want to provide the user with both the flexibility of the trade-off on the two regrets and the user-determined improvement on either static or restricted dynamic regret in not only this online least-squares setting, but the strongly convex and smooth setting.

Although the dynamic and the static regret can be upper bounded by the first-order method with different stepsize choices in (Mokhtari et al. 2016) in terms of V^* and (Hazan, Agarwal, and Kale 2007), respectively, no trade-off between these two metrics for this specific problem setting has been shown in the literature. This open question is solved by our update rule to take both regret metrics into consideration.

Last but not least, we consider the strongly convex case to obtain the general dynamic regret in terms of the path-length V, whose step-size rule is inspired by the strongly convex and smooth case.

Discounted Online Newton Step

In this section, we propose the Discounted Online Newton Step algorithm to achieve the static and dynamic regrets' trade-off.

The original Online Newton Step was proposed in (Hazan, Agarwal, and Kale 2007), which could upper bound the static regret of the α-exp-concave problems in the order of $O(\log T)$. However, there is no result on how to arrive the regret trade-off as shown in the general convex problem using online gradient descent (Zinkevich 2003; Zhang, Lu, and Zhou 2018).

For the α-exp-concave function f_t, Lemma 4.2 of (Hazan 2016) implies that for all $\rho \leq \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$, the following bound holds for all x and y in $\mathcal{S}$:

$$f_t(y) \geq f_t(x) + \nabla f_t(x)^\top (y-x) + \frac{\rho}{2}(x-y)^\top \nabla f_t(x) \nabla f_t(x)^\top (x-y). \quad (2a)$$

Also, if f_t are twice differentiable, then f_t is α-exp-concave if and only if

$$\nabla^2 f_t(x) \succeq \alpha \nabla f_t(x) \nabla f_t(x)^\top \quad (2b)$$

for all $x \in \mathcal{S}$.

In some variations on the algorithm, we will require extra conditions on the function, f_t. In particular, in one variation we will require ℓ-strong convexity. which means that there is a number $\ell > 0$ such that

$$f_t(y) \geq f_t(x) + \nabla f_t(x)^\top (y-x) + \frac{\ell}{2}\|x-y\|^2 \quad (2c)$$

for all x and y in $\mathcal{S}$. For twice-differentiable functions, strong convexity implies α-exp-concavity for $\alpha \leq \ell/G^2$ on $\mathcal{S}$.

In another variant, we will require that the following bound holds for all x and y in $\mathcal{S}$:

$$f_t(y) \geq f_t(x) + \nabla f_t(x)^\top (y-x) + \frac{1}{2}\|x-y\|^2_{\nabla^2 f_t(x)}. \quad (2d)$$

This bound does not correspond to a commonly used convexity class, but it does hold for the important special case of quadratic functions: $f_t(x) = \frac{1}{2}\|y_t - A_t x\|^2$. This fact will be important for analyzing the classic discounted recursive least-squares algorithm. Note that if y_t and A_t are restricted to compact sets, α can be chosen so that f_t is α-exp-concave.

Additionally, the algorithms for strongly convex functions and those satisfying (2d) will require that the gradients $\nabla f_t(x)$ are u-Lipschitz for all $x \in \mathcal{S}$ (equivalently, $f_t(x)$ is u-smoothness), which means the gradient $\nabla f_t(x)$ satisfies the relation

$$\|\nabla f_t(x) - \nabla f_t(y)\| \leq u \|x-y\|, \forall t$$

which is equivalent to $f_t(y) \leq f_t(x) + \nabla f_t(x)^T (y-x) + \frac{u}{2}\|y-x\|^2$. This implies, in particular, that $\nabla^2 f_t(x) \preceq uI$.

Algorithm 1 Discounted Online Newton Step

Given constants $\epsilon > 0$, $\eta > 0$, and $\gamma \in (0,1)$.
Let $\theta_1 \in \mathcal{S}$ and $P_0 = \epsilon I$.
for t=1,...,T **do**
  Play θ_t and incur loss $f_t(\theta_t)$
  Observe $\nabla_t = \nabla f_t(\theta_t)$ and $H_t = \nabla^2 f_t(\theta_t)$ (if needed)
  Update P_t:

$$P_t = \gamma P_{t-1} + \nabla_t \nabla_t^\top \quad \text{(Quasi-Newton)} \quad (3a)$$
$$P_t = \gamma P_{t-1} + H_t \quad \text{(Full-Newton)} \quad (3b)$$

  Update θ_t: $\theta_{t+1} = \Pi_{\mathcal{S}}^{P_t}\left(\theta_t - \frac{1}{\eta} P_t^{-1} \nabla_t\right)$
end for

To accommodate these three different cases, we propose Algorithm 1, in which $\Pi_{\mathcal{S}}^{P_t}(y) = \operatorname{argmin}_{z \in \mathcal{S}} \|z-y\|^2_{P_t}$ is the projection onto $\mathcal{S}$ with respect to the norm induced by P_t.

By using Algorithm 1, the following theorem can be obtained:

Theorem 1. *Consider the following three cases of Algorithm 1:*

1. *f_t is α-exp-concave. The algorithm uses $\eta \leq \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$, $\epsilon = 1$, and* (3a).
2. *f_t is α-exp-concave and ℓ-strongly convex while $\nabla f_t(x)$ is u-Lipschitz. The algorithm uses $\eta \leq \ell/u$, $\epsilon = 1$, and* (3b).
3. *f_t is α-exp-concave and satisfy (2d) while $\nabla f_t(x)$ is u-Lipschitz. The algorithm uses $\eta \leq 1$, $\epsilon = 1$, and* (3b).

For each of these cases, there are positive constants $a_1, \dots a_4$ such that

$$\sum_{t=1}^T (f_t(\theta_t) - f_t(z_t)) \leq -a_1 T \log\gamma - a_2 \log(1-\gamma) + \frac{a_3}{1-\gamma} V + a_4$$

for all $z_1, \dots, z_T \in \mathcal{S}$ such that $\sum_{t=2}^T \|z_t - z_{t-1}\| \leq V$.

Due to space limit, all the omitted proofs are moved to Appendix. Next, we will describe some consequences of Theorem 1.

Corollary 1. *Setting $\gamma = 1 - T^{-\beta}$ with $\beta \in (0,1)$ leads to the following form:*

$$\sum_{t=1}^{T}(f_t(\theta_t) - f_t(z_t)) \leq O(T^{1-\beta} + \beta \log T + T^{\beta} V)$$

Proof. The first term is bounded as:

$$-T\log\gamma = -T\log(1-T^{-\beta}) \leq \frac{T^{1-\beta}}{1-T^{-\beta}} = O(T^{1-\beta}),$$

where the inequality follows from $-\log(1-x) \leq \frac{x}{1-x}$ for $0 \leq x < 1$.

The other terms follow by direct calculation. □

This corollary guarantees that the static regret is bounded in the order of $O(T^{1-\beta})$ since $V = 0$ in that case. The dynamic regret is of order $O(T^{1-\beta} + T^{\beta}V)$. By choosing $\beta \in (0,1)$, we are guaranteed that both the static and dynamic regrets are both sublinear in T as long as $V < O(T)$. Also, small static regret can be obtained by setting β near 1.

In the setting of Corollary 1, the algorithm parameters do not depend on the path length V. Thus, the bounds hold for any path length, whether or not it is known a priori. The next corollary shows how tighter bounds could be obtained if knowledge of V were exploited in choosing the discount factor, γ.

Corollary 2. *Setting $\gamma = 1 - \frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}}$ leads to the form:*

$$\sum_{t=1}^{T}(f_t(\theta_t) - f_t(z_t)) \leq \max\{O(\log T), O(\sqrt{TV})\}$$

The proof is similar to the proof of Corollary 1.

Note that Corollary 2 implies that the discounted Newton method achieves logarithmic static regret by setting $V = 0$. This matches the bounds obtained in (Hazan, Agarwal, and Kale 2007). For positive path lengths bounded by V, we improve the $O(\sqrt{T(1+V)})$ dynamic bounds from (Zhang, Lu, and Zhou 2018). However, the algorithm above current requires knowing a bound on the path length, whereas (Zhang, Lu, and Zhou 2018) achieves its bound without knowing the path length, a priori.

If we view V as the variation budget that $z_1^T = z_1, \ldots, z_T$ can vary over $\mathcal{S}$ like in (Besbes, Gur, and Zeevi 2015), and use this as a pre-fixed value to allow the comparator sequence to vary arbitrarily over the set of admissible comparator sequence $\{z_1^T \in \mathcal{S} : \sum_{t=2}^{T} \|z_t - z_{t-1}\| \leq V\}$, we can tune γ in terms of V.

In order to bound the dynamic regret without knowing a bound on the path length, the method of (Zhang, Lu, and Zhou 2018) runs a collection of gradient descent algorithms in parallel with different step sizes and then uses a meta-optimization (Cesa-Bianchi and Lugosi 2006) to weight their solutions. In a later section, we will show how a related meta-optimization over the discount factor leads to $\max\{O(\log T), O(\sqrt{TV})\}$ dynamic regret bounds for unknown V.

For the Algorithm 1, we need to invert P_t, which can be achieved in time $O(n^2)$ for the Quasi-Newton case in (3a) by utilizing the matrix inversion lemma. However, for the Full-Newton step (3b), the inversion requires $O(n^3)$ time.

In the next two sections, we will use different methods to achieve the static/dynamic regret trade-off for the strongly convex and smooth case considered in the Full-Newton update to both avoid the high computation cost and improve the trade-off performance.

From Forgetting Factors to a Step Size Rule

In this section, we analyze recursive least squares for the special case of quadratic functions of the form:

$$f_t(\theta) = \frac{1}{2}\|\theta - y_t\|^2, \tag{4}$$

where $y_t \in \mathcal{S}$.

In this case, we will see that discounted recursive least squares can be interpreted as online gradient descent method with a special step size rule. We will show how this step size rule achieves a trade-off between static regret and dynamic regret with the specific comparison sequence $\theta_t^* = y_t = \operatorname{argmin}_{\theta \in \mathcal{S}} f_t(\theta)$. In the next section, we will see how this step size rule can achieve similar trade-offs on smooth, strongly convex functions. For a related analysis of more general quadratic functions, $f_t(\theta) = \frac{1}{2}\|A_t\theta - y_t\|^2$, please see the appendix.

Note that the previous section focused on dynamic regret for arbitrary comparison sequences, $z_t \in \mathcal{S}$. The analysis techniques in this and the next section are specialized to comparisons against $\theta_t^* = \operatorname{argmin}_{\theta \in \mathcal{S}} f_t(\theta)$, as studied in works such as (Mokhtari et al. 2016; Yang et al. 2016).

Classic discounted recursive least squares corresponds to Alg. 1 run with full Newton steps, $\eta = 1$, and initial matrix $P_0 = 0$. When f_t is defined as in (4), we have that $P_t = \sum_{k=0}^{t-1}\gamma^k I$. Thus, the update rule can be expressed in the following equivalent ways:

$$\theta_{t+1} = \operatorname*{argmin}_{\theta \in \mathcal{S}} \sum_{i=1}^{t} \gamma^{i-1} f_{t+1-i}(\theta) \tag{5a}$$

$$= \frac{\gamma - \gamma^t}{1-\gamma^t}\theta_t + \frac{1-\gamma}{1-\gamma^t}y_t \tag{5b}$$

$$= \theta_t - P_t^{-1}\nabla f_t(\theta_t) \tag{5c}$$

$$= \theta_t - \eta_t \nabla f_t(\theta_t), \tag{5d}$$

where $\eta_t = \frac{1-\gamma}{1-\gamma^t}$. Note that since $y_t \in \mathcal{S}$, no projection steps are needed.

The above update is the ubiquitous gradient descent with a changing stepsize. The only difference between standard methods is the choice of η_t, which will lead to the useful trade-off between dynamic and static regret.

By using the above update, we can get the relationship between $\theta_{t+1} - \theta_t^*$ and $\theta_t - \theta_t^*$ as the following result:

Lemma 1. *Let $\theta_t^* = \operatorname{argmin}_{\theta \mathcal{S}} f_t(\theta)$ in Eq.*(4)*. When using the discounted recursive least-squares update in Eq.*(5)*, we have the following relation:*

$$\theta_{t+1} - \theta_t^* = \frac{\gamma - \gamma^t}{1 - \gamma^t}(\theta_t - \theta_t^*)$$

Proof. Since $\theta_t^* = \operatorname{argmin} f_t(\theta) = y_t$, for $\theta_{t+1} - \theta_t^*$, we have:

$$\begin{aligned}\theta_{t+1} - \theta_t^* &= \theta_{t+1} - y_t \\ &= \tfrac{\gamma-\gamma^t}{1-\gamma^t}\theta_t + \tfrac{1-\gamma}{1-\gamma^t}y_t - y_t \\ &= \tfrac{\gamma-\gamma^t}{1-\gamma^t}(\theta_t - y_t) \\ &= \tfrac{\gamma-\gamma^t}{1-\gamma^t}(\theta_t - \theta_t^*)\end{aligned}$$

□

Recall from (1) that the path length of optimizer sequence is denoted by V^*. With the help of Lemma 1, we can upper bound the dynamic regret in the next theorem:

Theorem 2. *Let θ_t^* be the solution to $f_t(\theta)$ in Eq.*(4)*. When using the discounted recursive least-squares update in Eq.*(5) *with $1 - \gamma = 1/T^\beta, \beta \in (0, 1)$, we can upper bound the dynamic regret as:*

$$\mathcal{R}_d \leq 2DT^\beta\big(\|\theta_1 - \theta_1^*\| + V^*\big)$$

Proof. According to the Mean Value Theorem, there exists a vector $x \in \{v | v = \delta\theta_t + (1-\delta)\theta_t^*, \delta \in [0,1]\}$ such that $f_t(\theta_t) - f_t(\theta_t^*) = \nabla f_t(x)^T(\theta_t - \theta_t^*) \leq \|\nabla f_t(x)\| \|\theta_t - \theta_t^*\|$. For our problem, $\|\nabla f_t(x)\| = \|x - y_t\| \leq \|x\| + \|y_t\|$. For $\|x\|$, we have:

$$\begin{aligned}\|x\| &= \|\delta\theta_t + (1-\delta)\theta_t^*\| \\ &\leq \delta\|\theta_t\| + (1-\delta)\|y_t\| \\ &= \delta\left\|\frac{\sum_{i=1}^{t-1}\gamma^{i-1}y_{t-i}}{\sum_{i=1}^{t-1}\gamma^{i-1}}\right\| + (1-\delta)\|y_t\| \\ &\leq D\end{aligned}$$

where the second inequality is due to $\|y_i\| \leq D, \forall i$.

As a result, the norm of the gradient can be upper bounded as $\|\nabla f_t(x)\| \leq 2D$. Then we have $\mathcal{R}_d = \sum_{t=1}^{T}\Big(f_t(\theta_t) - f_t(\theta_t^*)\Big) \leq 2D\sum_{t=1}^{T}\|\theta_t - \theta_t^*\|$. Now we could instead upper bound $\sum_{t=1}^{T}\|\theta_t - \theta_t^*\|$, which can be achieved as follows:

$$\begin{aligned}&\sum_{t=1}^{T}\|\theta_t - \theta_t^*\| \\ &= \|\theta_1 - \theta_1^*\| + \sum_{t=2}^{T}\big\|\theta_t - \theta_{t-1}^* + \theta_{t-1}^* - \theta_t^*\big\| \\ &\leq \|\theta_1 - \theta_1^*\| + \sum_{t=1}^{T-1}\|\theta_{t+1} - \theta_t^*\| + \sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\| \\ &= \|\theta_1 - \theta_1^*\| + \sum_{t=1}^{T-1}\tfrac{\gamma-\gamma^t}{1-\gamma^t}\|\theta_t - \theta_t^*\| + \sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\| \\ &\leq \|\theta_1 - \theta_1^*\| + \sum_{t=1}^{T}\tfrac{\gamma-\gamma^t}{1-\gamma^t}\|\theta_t - \theta_t^*\| + \sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\|\end{aligned}$$

where in the second equality, we substitute the result from Lemma 1.

From the above inequality, we get

$$\sum_{t=1}^{T}\left(1 - \frac{\gamma-\gamma^t}{1-\gamma^t}\right)\|\theta_t - \theta_t^*\| \leq \|\theta_1 - \theta_1^*\| + \sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\|$$

Since $\left(1 - \frac{\gamma-\gamma^t}{1-\gamma^t}\right) = \frac{1-\gamma}{1-\gamma^t} \geq 1 - \gamma$, we get

$$\begin{aligned}\sum_{t=1}^{T}\|\theta_t - \theta_t^*\| &\leq \tfrac{1}{1-\gamma}\|\theta_1 - \theta_1^*\| + \tfrac{1}{1-\gamma}\sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\| \\ &= T^\beta(\|\theta_1 - \theta_1^*\| + \sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\|)\end{aligned}$$

Thus, $\mathcal{R}_d \leq 2D\sum_{t=1}^{T}\|\theta_t - \theta_t^*\| \leq 2DT^\beta(\|\theta_1 - \theta_1^*\| + \sum_{t=2}^{T}\big\|\theta_t^* - \theta_{t-1}^*\big\|)$. □

Theorem 2 shows that if we choose the discounted factor $\gamma = 1 - T^{-\beta}$ we obtain a dynamic regret of $O(T^\beta(1+V^*))$. This is a refinement of the Corollary 1 since the bound no longer has the $T^{1-\beta}$ term. Thus, the dynamic regret can be made small by choosing a small β.

In the next theorem, we will show that this carefully chosen γ can also lead to useful static regret, which can give us a trade-off between them.

Theorem 3. *Let θ^* be the solution to $\min\sum_{t=1}^{T} f_t(\theta)$. When using the discounted recursive least-squares update in Eq.*(5) *with $1 - \gamma = 1/T^\beta, \beta \in (0, 1)$, we can upper bound the static regret as:*

$$\mathcal{R}_s \leq O(T^{1-\beta})$$

Recall that the algorithm of this section can be interpreted both as a discounted recursive least squares method, and as a gradient descent method. As a result, this theorem is actually a direct consequence of Corollary 1, by setting $V = 0$. However, we will give a separate proof in the Appendix, since the techniques extend naturally to the analysis of more general work on gradient descent methods of the next sections.

Our Theorems 2 and 3 build a trade-off between dynamic and static regret by the carefully chosen discounted factor γ. Compared with the result from the last section, there are two improvements: 1. The two regrets are decoupled so that we could reduce the β to make the dynamic regret result smaller than the previous section's one. 2. The update is the first-order gradient descent, which is computationally efficient.

In the next section, we will consider the strongly convex and smooth case, whose result is inspired by this section's analysis.

Online Gradient Descent for Smooth, Strongly Convex Problems

In this section, we generalize the previous section idea to the functions with ℓ-strong-convexity and u-smoothness. We will see that similar bounds on $\mathcal{R}_s$ and $\mathcal{R}_d^*$ can be obtained.

The assumption we use is the upper bound of the norm of the gradient, which is $\|\nabla f_t(\theta)\| \leq G, \forall \theta \in \mathcal{S}, \forall t$.

Our proposed update rule for the prediction θ_{t+1} at time step $t+1$ is:

$$\theta_{t+1} = \underset{\theta \in \mathcal{S}}{\operatorname{argmin}} \|\theta - (\theta_t - \eta_t \nabla f_t(\theta_t))\|^2 \tag{6}$$

where $\eta_t = \frac{1-\gamma}{\ell(\gamma - \gamma^t) + u(1-\gamma)}$ and $\gamma \in (0,1)$.

This update rule generalizes the step size rule from the last section.

Before getting to the dynamic regret, we will first derive the relation between $\|\theta_{t+1} - \theta_t^*\|$ and $\|\theta_t - \theta_t^*\|$ to try to mimic the result in Lemma 1 of the quadratic case:

Lemma 2. *Let $\theta_t^* \in \mathcal{S}$ be the solution to $f_t(\theta)$ which is strongly convex and smooth. When we use the update in Eq.(6), the following relation is obtained:*

$$\|\theta_{t+1} - \theta_t^*\| \leq \sqrt{1 - \frac{l(1-\gamma)}{u(1-\gamma) + l\gamma}} \|\theta_t - \theta_t^*\|$$

Now we are ready to present the dynamic regret result:

Theorem 4. *Let θ_t^* be the solution to $f_t(\theta), \theta \in \mathrm{S}$. When using the update in Eq.(6) with $1 - \gamma = 1/T^\beta, \beta \in (0,1)$, we can upper bound the dynamic regret:*

$$\mathcal{R}_d \leq G\big(2(T^\beta - 1) + u/l\big)(\|\theta_1 - \theta_1^*\| + V^*)$$

Theorem 4's result seems promising in achieving the trade-off, since it has the similar formula as the previous successful case of quadratic problem in Theorem 2. Next, we will present the static regret result, which assures such conjecture.

Theorem 5. *Let θ^* be the solution to $\min_{\theta \in \mathcal{S}} \sum_{t=1}^{T} f_t(\theta)$. When using the update in Eq.(6) with $1 - \gamma = 1/T^\beta, \beta \in (0,1)$, we can upper bound the static regret:*

$$\mathcal{R}_s \leq O(T^{1-\beta})$$

The above two theorems' results have the similar bounds as the last section, which will give us the same improvements discussed in the previous section over the strongly convex and smooth problem.

In the next section, we will consider the strongly convex problem.

Online Gradient Descent for Strongly Convex Problems

In this section, we generalize the step-size idea from previous section to consider the problem with ℓ-strong-convexity. The assumption is the same as the previous section's.

The update rule is still oneline gradient descent:

$$\theta_{t+1} = \underset{\theta \in \mathcal{S}}{\operatorname{argmin}} \|\theta - (\theta_t - \eta_t \nabla f_t(\theta_t))\|^2 \tag{7}$$

where $\eta_t = \frac{1-\gamma}{\ell(1-\gamma^t)}$, and $\gamma \in (0,1)$.

We can see that the update rule is the same as the one in Eq.(6) while the stepsize η_t is replaced with $\frac{1-\gamma}{\ell(1-\gamma^t)}$.

By using the new step-size with the update rule in Eq.(7), we can get:

Theorem 6. *If using the update rule in Eq.(7) with $\eta_t = \frac{1-\gamma}{\ell(1-\gamma^t)}$ and $\gamma \in (0,1)$, the following dynamic regret can be obtained:*

$$\sum_{t=1}^{T} \Big(f_t(\theta_t) - f_t(z_t)\Big) \leq 2D\ell \frac{1}{1-\gamma} V + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t$$

We can view V as the variation budget that z_1^T can vary over $\mathcal{S}$ like discussed in previous section. By further restricting $V = T^\beta$, where $\beta \in [0,1)$ like in (Besbes, Gur, and Zeevi 2015), we can get:

Corollary 3. *By setting $\gamma = 1 - \sqrt{\frac{V}{T}}$, $V = T^\beta$, and $\beta \in [0,1)$, for large enough T such that $T^{1-\beta} \geq 4$, the following bound can be obtained:*

$$\sum_{t=1}^{T} \Big(f_t(\theta_t) - f_t(z_t)\Big) \leq O(\sqrt{TV})$$

where $\{z_1, z_2, \ldots, z_T \in \mathcal{S} : \sum_{t=2}^{T} \|z_t - z_{t-1}\| \leq V\}$.

When $V = V^*$, The above corollary's result meets the lower bound in (Besbes, Gur, and Zeevi 2015), which is of the order optimal and requires only online gradient descent as opposed to the complex restarting procedure in (Besbes, Gur, and Zeevi 2015).

If we are instead concerned with the general V, which can vary from 0 up to $2DT$, as opposed to $[1, 2DT)$ in Corollary 3, the corollary below gives a general result:

Corollary 4. *By setting $\gamma = 1 - \frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}}$, the following bound can be obtained:*

$$\sum_{t=1}^{T} \Big(f_t(\theta_t) - f_t(z_t)\Big) \leq \max\{O(\log T), O(\sqrt{TV})\}.$$

The above result characterizes the general dynamic regret, which is never shown before to the best of our knowledge. For the special case when $V = V^*$, $\mathcal{R}_d \leq O(\max\{\log T, \sqrt{TV^*}\})$, which is an improved result compared to $O(\max\{\log T, \sqrt{TV^* \log T}\})$ in (Zhang et al. 2018). According to the result in (Besbes, Gur, and Zeevi 2015), our result is minimax optimal when $V^* \geq \log^2 T/T$, and only up to polylogarithmic factor larger when $V^* < \log^2 T/T$.

Similar problem about the unknown value V arises in the step-size setup of Corollary 4 as in Corollary 2, which will be solved in the next section.

Meta-algorithm

In previous sections, we discussed the results on dynamic regret for both α-exp-concave and ℓ strongly convex cases. But in order to obtain the general results, we need to know the value of V in order to setup the step-size correctly, which may be difficult to know or approximate. In this section, we use the so-called 'Meta-algorithm' to solve this issue by running multiple algorithms in parallel with different step-sizes.

For the online convex optimization, the 'Meta-algorithm' has been used by (Zhang, Lu, and Zhou 2018) to solve the similar issue in the convex case. But it cannot be used directly in either the α-exp-concave or ℓ strongly convex case due to the added $O(\sqrt{T})$ regret from running multiple algorithms.

In this section, we will show that by using appropriate parameters and analysis designed specifically for our cases, the Meta-algorithm can be used to solve our issues.

Algorithm 2 Meta-algorithm

Given step-size λ, and a set $\mathcal{H}$ containing step-sizes for each algorithm.
Activate a set of algorithms $\{A^\gamma|\gamma \in \mathcal{H}\}$ by calling Algorithm 1 (exp-concave case) or the update in Eq.(7) (strongly convex case) for each parameter $\gamma \in \mathcal{H}$.
Sort γ in descending order $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_N$, and set $w_1^{\gamma_i} = \frac{C}{i(i+1)}$ with $C = 1 + 1/|\mathcal{H}|$.
for t=1,...,T **do**
Obtain θ_t^γ from each algorithm A^γ.
Play $\theta_t = \sum_{\gamma\in\mathcal{H}} w_t^\gamma \theta_t^\gamma$, and incur loss $f_t(\theta_t^\gamma)$ for each θ_t^γ.
Update w_t^γ by

$$w_{t+1}^\gamma = \frac{w_t^\gamma \exp(-\lambda f_t(\theta_t^\gamma))}{\sum_{\mu\in\mathcal{H}} w_t^\mu \exp(-\lambda f_t(\theta_t^\mu))}.$$

Send back the gradient $\nabla f_t(\theta_t^\gamma)$ for each algorithm A^γ.
end for

Exp-concave case

Before showing the regret result, we first show that the cumulative loss of the meta-algorithm is comparable to all $A^\gamma \in \mathcal{H}$:

Lemma 3. *If f_t is α-exp-concave and $\lambda = \alpha$, the cumulative loss difference of Algorithm 2 for any $\gamma \in \mathcal{H}$ is bounded as:*

$$\sum_{t=1}^{T}(f_t(\theta_t) - f_t(\theta_t^\gamma)) \leq \frac{1}{\alpha}\log\frac{1}{w_1^\gamma}$$

Based on the above result, if we can show that there exists an algorithm A^γ, which can bound the regret $\sum_{t=1}^{T}(f_t(\theta_t^\gamma) - f_t(z_t)) \leq \max\{O(\log T), O(\sqrt{TV})\}$, then we can combine these two results and show that the regret holds for $\theta_t, t = 1, \ldots, T$ as well:

Theorem 7. *For any comparator sequence $z_1, \ldots, z_T \in \mathcal{S}$, setting $\mathcal{H} = \left\{\gamma_i = 1-\eta_i \middle| i = 1, \ldots, N\right\}$ with $T \geq 2$ where $\eta_i = \frac{1}{2}\frac{\log T}{T\sqrt{2D}}2^{i-1}$, $N = \lceil \frac{1}{2}\log_2(\frac{2DT^2}{\log^2 T})\rceil + 1$, and $\lambda = \alpha$ leads to the result:*

$$\sum_{t=1}^{T}(f_t(\theta_t) - f_t(z_t)) \leq O(\max\{\log T, \sqrt{TV}\})$$

Strongly convex case

For the strongly convex problem, since the parameter γ used in Corollary 4 is the same as the one in Corollary 2, the meta-algorithm may also work with the same setup in Theorem 7 except the parameter $\lambda = \alpha$, which comes from the α-exp-concavity.

To proceed, we first show that the ℓ-strongly convex function with bounded gradient (e.g.,$\|\nabla f_t\| \leq G$) is also ℓ/G^2-exp-concave. Previous works also pointed out this, but their statement only works when f_t is second-order differentiable, while our result is true when f_t is first-order differentiable.

Lemma 4. *For the ℓ-strongly convex function f_t with $\|\nabla f_t\| \leq G$, it is also α-exp-concave with $\alpha = \ell/G^2$.*

Lemma 4 indicates that running Algorithm 2 with strongly convex function leads to the same result as in Lemma 3. Thus, using the similar idea as discussed in the case of α-exp-concavity and Algorithm 2, the theorem below can be obtained:

Theorem 8. *For any comparator sequence $z_1, \ldots, z_T \in \mathcal{S}$, setting $\mathcal{H} = \left\{\gamma_i = 1-\eta_i \middle| i = 1, \ldots, N\right\}$ with $T \geq 2$ where $\eta_i = \frac{1}{2}\frac{\log T}{T\sqrt{2D}}2^{i-1}$, $N = \lceil \frac{1}{2}\log_2(\frac{2DT^2}{\log^2 T})\rceil + 1$, and $\lambda = \ell/G^2$ leads to the result:*

$$\sum_{t=1}^{T}(f_t(\theta_t) - f_t(z_t)) \leq O(\max\{\log T, \sqrt{TV}\})$$

Conclusion

In this paper, we propose the Discounted Online Newton Step (D-ONS) for the α-exp-concave setup to not only improve the dynamic regret result to the order of $\max\{O(\log T), O(\sqrt{TV})\}$ but solve the open question of how to achieve the static/dynamic regret trade-off in this setting.

Faced with the high computational cost when using this algorithm to solve strongly convex and smooth case, we propose a new online gradient descent update to further improve the trade-off performance, which is inspired by the analysis of the connection between discounted recursive least-squares algorithm and the regret guarantees. This new update is generalized to the strongly convex case with improved dynamic regret $\max\{O(\log T), O(\sqrt{TV})\}$.

References

Abernethy, J.; Bartlett, P. L.; Rakhlin, A.; and Tewari, A. 2008. Optimal strategies and minimax lower bounds for online convex games.

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

Blum, A.; Kumar, V.; Rudra, A.; and Wu, F. 2004. Online learning in online auctions. *Theoretical Computer Science* 324(2-3):137–146.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Chiang, C.-K.; Yang, T.; Lee, C.-J.; Mahdavi, M.; Lu, C.-J.; Jin, R.; and Zhu, S. 2012. Online optimization with gradual variations. In *Conference on Learning Theory*, 6–1.

Crammer, K.; Dekel, O.; Keshet, J.; Shalev-Shwartz, S.; and Singer, Y. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7(Mar):551–585.

Duong, V. H. 2017. Adaptive and robust algorithm for lithium-ion battery states estimation for application in electric vehicles.

Fabre, P., and Gueguen, C. 1986. Improvement of the fast recursive least-squares algorithms via normalization: A comparative study. *IEEE transactions on acoustics, speech, and signal processing* 34(2):296–308.

Hall, E. C., and Willett, R. M. 2013. Dynamical models and tracking regret in online convex programming. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, I–579. JMLR. org.

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

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

Mokhtari, A.; Shahrampour, S.; Jadbabaie, A.; and Ribeiro, A. 2016. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 7195–7201. IEEE.

Nesterov, Y. 2013. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Shalev-Shwartz, S., et al. 2012. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning* 4(2):107–194.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, 2199–2207.

Yang, T.; Zhang, L.; Jin, R.; and Yi, J. 2016. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, 449–457.

Yuan, J., and Lamperski, A. 2018. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, 6137–6146.

Zhang, L.; Yang, T.; Zhou, Z.-H.; et al. 2018. Dynamic regret of strongly adaptive methods. In *International Conference on Machine Learning*, 5877–5886.

Zhang, L.; Lu, S.; and Zhou, Z.-H. 2018. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems*, 1323–1333.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 928–936.

Appendix:

The supplementary material contains proofs of the some results of the paper along with supporting results.

Proof of Theorem 1: Before proving the theorem, the following observation is helpful.

Lemma 5. *If P_t is updated via* (3a) *then* $\|P_t\| \leq \epsilon + \frac{G^2}{1-\gamma}$, *while if P_t is updated via* (3b)*, then* $\|P_t\| \leq \epsilon + \frac{u}{1-\gamma}$.

Proof. First consider the quasi-Newton case. The bound holds at $P_0 = \epsilon I$, so assume that it holds at time $t-1$ for $t \geq 1$. Then, by induction we have

$$\begin{aligned}\|P_t\| &= \|\gamma P_{t-1} + \nabla_t \nabla_t\| \\ &\leq \gamma \|P_{t-1}\| + G^2 \\ &\leq \gamma\epsilon + \frac{G^2}{1-\gamma} \\ &\leq \epsilon + \frac{G^2}{1-\gamma}.\end{aligned}$$

The full-Newton case is identical, except it uses the bound $\|H_t\| \leq u$.

□

The generalized Pythagorean theorem implies that

$$\begin{aligned}\|\theta_{t+1} - z_t\|_{P_t}^2 &\leq \left\|\theta_t - \frac{1}{\eta} P_t^{-1}\nabla_t - z_t\right\|_{P_t}^2 \\ &= \|\theta_t - z_t\|_{P_t}^2 + \frac{1}{\eta^2}\nabla_t^\top P_t^{-1}\nabla_t \\ &\quad - \frac{2}{\eta}\nabla_t^\top(\theta_t - z_t).\end{aligned}$$

Re-arranging shows that

$$\begin{aligned}\nabla_t^\top(\theta_t - z_t) \leq \frac{1}{2\eta}\nabla_t^\top P_t^{-1}\nabla_t + \frac{\eta}{2}\Big(&\|\theta_t - z_t\|_{P_t}^2 \\ &- \|\theta_{t+1} - z_t\|_{P_t}^2\Big)\end{aligned} \tag{8}$$

Let c_1 be the upper bound on $\|P_t\|$ from Lemma 5. Then we can lower bound $\|\theta_{t+1} - z_t\|_{P_t}^2$ by

$$\begin{aligned}\|\theta_{t+1} - z_t\|_{P_t}^2 &= \|\theta_{t+1} - z_{t+1}\|_{P_t}^2 + \|z_{t+1} - z_t\|_{P_t}^2 \\ &\quad + 2(\theta_{t+1} - z_{t+1})^\top P_t(z_{t+1} - z_t) \\ &\geq \|\theta_{t+1} - z_{t+1}\|_{P_t}^2 - 4Dc_1\|z_{t+1} - z_t\|\end{aligned} \tag{9}$$

Combining (8) and (9) gives

$$\begin{aligned}\nabla_t^\top(\theta_t - z_t) \leq \frac{1}{2\eta}\nabla_t^\top P_t^{-1}\nabla_t + 2Dc_1\eta\|z_{t+1} - z_t\| \\ \frac{\eta}{2}\left(\|\theta_t - z_t\|_{P_t}^2 - \|\theta_{t+1} - z_{t+1}\|_{P_t}^2\right)\end{aligned}$$

Summing over t, dropping the term $-\|\theta_{T+1} - z_{T+1}\|_{P_T}^2$, setting $z_{T+1} = z_T$, and re-arranging gives

$$\begin{aligned}\sum_{t=1}^T \nabla_t^\top(\theta_t - z_t) &\leq \sum_{t=1}^T \frac{1}{2\eta}\nabla_t^\top P_t^{-1}\nabla_t + 2Dc_1\eta V \\ &+ \frac{\eta}{2}\epsilon\|\theta_1 - z_1\|^2 + \frac{\eta}{2}\sum_{t=1}^T(\theta_t - z_t)^\top(P_t - P_{t-1})(\theta_t - z_t)\end{aligned} \tag{10}$$

Now we will see how the choices of η enable the final sum from (10) to cancel the terms from (2). In Case 1, we have that $\eta(P_t - P_{t-1}) \preceq \eta\nabla_t\nabla_t^\top$ and the bound from (2a) holds for $\rho = \eta$. In Case 2, $\eta(P_t - P_{t-1}) \preceq \eta H_t \preceq \ell I$. In Case 3, $\eta(P_t - P_{t-1}) \preceq \eta H_t \preceq H_t$. Thus in all cases, η has been chosen so that combining the appropriate term of (2) with (10) gives

$$\begin{aligned}\textstyle\sum_{t=1}^T(f_t(\theta_t) - f_t(z_t)) \quad &\leq \textstyle\sum_{t=1}^T \frac{1}{2\eta}\nabla_t^\top P_t^{-1}\nabla_t \\ &\quad + 2Dc_1\eta V + 2\eta\epsilon D^2\end{aligned} \tag{11}$$

Now we will bound the first sum of (11). Note that $\nabla_t^\top P_t^{-1}\nabla_t = \langle P_t^{-1}, \nabla_t\nabla_t^\top\rangle$. In Case 1, we have that $\nabla_t\nabla_t^\top = P_t - \gamma P_{t-1}$, while in Cases 2 and 3, we have that $\nabla_t\nabla_t^\top \preceq \frac{1}{\alpha}H_t = \frac{1}{\alpha}(P_t - \gamma P_{t-1})$. So, in Case 1, let $c_2 = 1$ and in Cases 2 and 3, let $c_2 = 1/\alpha$. Then in all cases, we have that

$$\nabla_t^\top P_t^{-1}\nabla_t \leq c_2\langle P_t^{-1}, P_t - \gamma P_{t-1}\rangle. \tag{12}$$

Lemma 4.5 of (Hazan 2016) shows that

$$\langle P_t^{-1}, P_t - \gamma P_{t-1}\rangle \leq \log\frac{|P_t|}{|\gamma P_{t-1}|} = \log\frac{|P_t|}{|P_{t-1}|} - n\log\gamma, \tag{13}$$

where n is the dimension of x_t.

Combining (12) with (13), summing, and then using the bound that $\|P_T\| \leq c_1$ gives,

$$\begin{aligned}\sum_{t=1}^T \nabla_t^\top P_t^{-1}\nabla_t &\leq c_2\log|P_T| - c_2 n\log\epsilon - nT\log\gamma \\ &\leq c_2 n\log\frac{c_1}{\epsilon} - c_2 nT\log\gamma\end{aligned} \tag{14}$$

Recall that $c_1 = \epsilon + \frac{c_3}{1-\gamma}$, where $c_3 = G^2$ or $c_3 = u$, depending on the case. Then a more explicit upper bound on (14) is given by:

$$\sum_{t=1}^t \nabla_t^\top P_t^{-1}\nabla_t \leq c_2 n\log\left(1 + \frac{c_3}{\epsilon(1-\gamma)}\right) - c_2 nT\log\gamma. \tag{15}$$

Combining (11) and (15) gives the bound:

$$\begin{aligned}&\textstyle\sum_{t=1}^T(f_t(\theta_t) - f_t(z_t)) \leq -\frac{c_2 nT}{2\eta}\log\gamma + \\ &\textstyle\frac{c_2 n}{2\eta}\log\left(1 + \frac{c_3}{\epsilon(1-\gamma)}\right) + 2D\eta\left(\epsilon + \frac{c_3}{1-\gamma}\right)V + 2\eta\epsilon D^2\end{aligned}$$

The desired regret bound can now be found by simplifying the expression on the right, using the fact that $\frac{1}{1-\gamma} > 1$. □

Proof of Theorem 3:

Proof. To proceed, recall that the update in Eq.(5) is

$$\begin{aligned}\theta_{t+1} &= \tfrac{\gamma-\gamma^t}{1-\gamma^t}\theta_t + \tfrac{1-\gamma}{1-\gamma^t}y_t \\ &= \theta_t - \eta_t \nabla f_t(\theta_t)\end{aligned}$$

where $\eta_t = \frac{1-\gamma}{1-\gamma^t}$.

Then we get the relationship between $\nabla f_t(\theta_t)^T(\theta_t - \theta^*)$ and $\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2$ as:

$$\begin{aligned}\|\theta_{t+1} - \theta^*\|^2 &= \|\theta_t - \eta_t \nabla f_t(\theta_t) - \theta^*\|^2 \\ &= \|\theta_t - \theta^*\|^2 - 2\eta_t \nabla f_t(\theta_t)^T(\theta_t - \theta^*) \\ &\quad + \eta_t^2 \|\nabla f_t(\theta_t)\|^2\end{aligned}$$

$$\begin{aligned}\nabla f_t(\theta_t)^T(\theta_t - \theta^*) &= \tfrac{1}{2\eta_t}\big(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\big) \\ &\quad + \tfrac{\eta_t}{2}\|\nabla f_t(\theta_t)\|^2\end{aligned}$$

Moreover, we write $f_t(\theta^*)$ as $f_t(\theta^*) = f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta^* - \theta_t) + \frac{1}{2}\|\theta^* - \theta_t\|^2$, which combined with the previous equation gives us the following equation:

$$\begin{aligned}f_t(\theta_t) - f_t(\theta^*) &= \tfrac{1}{2\eta_t}\big(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\big) \\ &\quad + \tfrac{\eta_t}{2}\|\nabla f_t(\theta_t)\|^2 - \tfrac{1}{2}\|\theta^* - \theta_t\|^2 \\ &\le 2D^2\eta_t + \tfrac{1}{2\eta_t}\big(\|\theta_t - \theta^*\|^2 - \\ &\quad \|\theta_{t+1} - \theta^*\|^2\big) - \tfrac{1}{2}\|\theta^* - \theta_t\|^2\end{aligned}$$

where the inequality is due to $\|\nabla f_t(\theta_t)\| \le 2D$ as shown in Theorem 2.

Sum the above inequality from $t = 1$ to T, we get:

$$\begin{aligned}&\sum_{t=1}^{T}\Big(f_t(\theta_t) - f_t(\theta^*)\Big) \\ &\le 2D^2\sum_{t=1}^{T}\eta_t + \tfrac{1/\eta_1 - 1}{2}\|\theta_1 - \theta^*\|^2 + \tfrac{1}{2}\sum_{t=2}^{T}\big[(\tfrac{1}{\eta_t} \\ &\quad - \tfrac{1}{\eta_{t-1}} - 1)\|\theta^* - \theta_t\|^2\big] - \tfrac{1}{2\eta_T}\|\theta_{T+1} - \theta^*\|^2\end{aligned}$$

Since $\eta_t = \frac{1-\gamma}{1-\gamma^t}$, $\eta_1 = 1$, $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - 1 < 0$. Then for the static regret, we have:

$$\begin{aligned}\mathcal{R}_s &= \sum_{t=1}^{T}\Big(f_t(\theta_t) - f_t(\theta^*)\Big) \\ &\le 2D^2\sum_{t=1}^{T}\eta_t = 2D^2(1-\gamma)\sum_{t=1}^{T}\tfrac{1}{1-\gamma^t}\end{aligned} \tag{16}$$

Since $\gamma \in (0,1)$, $\sum_{t=1}^{T}\frac{1}{1-\gamma^t} \le \frac{1}{1-\gamma} + \int_1^T \frac{1}{1-\gamma^t}\mathrm{d}t = \frac{1}{1-\gamma} + \Big(t - \frac{\log(1-\gamma^t)}{\log(\gamma)}\Big)\Big|_1^T = \frac{1}{1-\gamma} + T - 1 + \frac{\log(1-\gamma)}{\log\gamma} - \frac{\log(1-\gamma^T)}{\log\gamma} \le \frac{1}{1-\gamma} + T - 1 + \frac{\log(1-\gamma)}{\log\gamma}$.

Since $1 - \gamma = 1/T^\beta$, $\frac{\log(1-\gamma)}{\log\gamma} = \frac{\beta\log T}{\log(1+\frac{1}{T^\beta - 1})}$. Since $\log(1+x) \ge \frac{1}{2}x, x \in (0,1)$, $\log(1 + \frac{1}{T^\beta - 1}) \ge \frac{1}{2}\frac{1}{T^\beta - 1}$. Thus, we have $\frac{\log(1-\gamma)}{\log\gamma} \le 2\beta(T^\beta - 1)\log T$. Then $(1-\gamma)\sum_{t=1}^{T}\frac{1}{1-\gamma^t} = O(T^{1-\beta})$, which results in $\mathcal{R}_s \le O(T^{1-\beta})$. □

Proof of Lemma 2:

Proof. The proof follows the analysis in Chapter 2 of (Nesterov 2013).

From the strong convexity of $f_t(\theta)$, we have

$$\begin{aligned}f_t(\theta) &\ge f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta - \theta_t) + \tfrac{\ell}{2}\|\theta - \theta_t\|^2 \\ &= f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta - \theta_t) + \nabla f_t(\theta_t)^T(\theta_{t+1} - \theta_t) \\ &\quad - \nabla f_t(\theta_t)^T(\theta_{t+1} - \theta_t) + \tfrac{\ell}{2}\|\theta - \theta_t\|^2 \\ &= f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta_{t+1} - \theta_t) \\ &\quad + \nabla f_t(\theta_t)^T(\theta - \theta_{t+1}) + \tfrac{\ell}{2}\|\theta - \theta_t\|^2\end{aligned} \tag{17}$$

According to the optimality condition of the update rule in Eq.(6), we have $\big(\nabla f_t(\theta_t) + \frac{1}{\eta_t}(\theta_{t+1} - \theta_t)\big)^T(\theta - \theta_{t+1}) \ge 0, \forall\theta \in \mathcal{S}$, which is $\nabla f_t(\theta_t)^T(\theta - \theta_{t+1}) \ge \frac{1}{\eta_t}(\theta_t - \theta_{t+1})^T(\theta - \theta_{t+1})$. Then combine with Eq.(17), we have

$$\begin{aligned}f_t(\theta) &\ge f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta_{t+1} - \theta_t) \\ &\quad + \tfrac{1}{\eta_t}(\theta_t - \theta_{t+1})^T(\theta - \theta_{t+1}) + \tfrac{\ell}{2}\|\theta - \theta_t\|^2\end{aligned} \tag{18}$$

From the smoothness of $f_t(\theta)$, we have $f_t(\theta_{t+1}) \le f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta_{t+1} - \theta_t) + \frac{u}{2}\|\theta_{t+1} - \theta_t\|^2$. Since $\frac{1}{\eta_t} = \frac{\ell(\gamma - \gamma^t) + u(1-\gamma)}{1-\gamma} \ge u$, we have $f_t(\theta_t) + \nabla f_t(\theta_t)^T(\theta_{t+1} - \theta_t) \ge f_t(\theta_{t+1}) - \frac{1}{2\eta_t}\|\theta_{t+1} - \theta_t\|^2$. Then combined with inequality (18), we have

$$\begin{aligned}f_t(\theta) &\ge f_t(\theta_{t+1}) - \tfrac{1}{2\eta_t}\|\theta_{t+1} - \theta_t\|^2 \\ &\quad + \tfrac{1}{\eta_t}(\theta_t - \theta_{t+1})^T(\theta - \theta_{t+1}) + \tfrac{\ell}{2}\|\theta - \theta_t\|^2 \\ &= f_t(\theta_{t+1}) + \tfrac{1}{2\eta_t}\|\theta_{t+1} - \theta_t\|^2 \\ &\quad + \tfrac{1}{\eta_t}(\theta_t - \theta_{t+1})^T(\theta - \theta_t) + \tfrac{\ell}{2}\|\theta - \theta_t\|^2\end{aligned} \tag{19}$$

By setting $\theta = \theta_t^*$ and using the fact $f_t(\theta_t^*) \le f_t(\theta_{t+1})$, we reformulate the above inequality as:

$$\begin{aligned}&(\theta_t - \theta_{t+1})^T(\theta_t^* - \theta_t) \\ &\le -\tfrac{\ell(1-\gamma)}{2\ell(\gamma - \gamma^t) + 2u(1-\gamma)}\|\theta_t^* - \theta_t\|^2 - \tfrac{1}{2}\|\theta_{t+1} - \theta_t\|^2\end{aligned} \tag{20}$$

Since $\|\theta_{t+1} - \theta_t^*\|^2 = \|\theta_{t+1} - \theta_t + \theta_t - \theta_t^*\|^2$, we have

$$\begin{aligned}\|\theta_{t+1} - \theta_t^*\|^2 &= \|\theta_{t+1} - \theta_t\|^2 + \|\theta_t - \theta_t^*\|^2 \\ &\quad + 2(\theta_t - \theta_{t+1})^T(\theta_t^* - \theta_t) \\ &\le \big(1 - \tfrac{\ell(1-\gamma)}{\ell(\gamma - \gamma^t) + u(1-\gamma)}\big)\|\theta_t - \theta_t^*\|^2 \\ &\le \big(1 - \tfrac{\ell(1-\gamma)}{\ell\gamma + u(1-\gamma)}\big)\|\theta_t - \theta_t^*\|^2\end{aligned} \tag{21}$$

□

Proof of Theorem 4:

Proof. We use the same steps as in the previous section. First, according to the Mean Value Theorem, we have $f_t(\theta_t) - f_t(\theta_t^*) = \nabla f_t(x)^T(\theta_t - \theta_t^*) \le \|\nabla f_t(x)\|\,\|\theta_t - \theta_t^*\|$, where $x \in \{v | v = \delta\theta_t + (1-\delta)\theta_t^*, \delta \in [0,1]\}$. Due to the assumption on the upper bound of the norm of the gradient, we have $f_t(\theta_t) - f_t(\theta_t^*) \le G\|\theta_t - \theta_t^*\|$. As a result, $\sum_{t=1}^{T}\big(f_t(\theta_t) - f_t(\theta_t^*)\big) \le G\sum_{t=1}^{T}\|\theta_t - \theta_t^*\|$.

Now we need to upper bound the term $\sum_{t=1}^{T}\|\theta_t-\theta_t^*\|$. $\sum_{t=1}^{T}\|\theta_t-\theta_t^*\|$ is equal to $\|\theta_1-\theta_1^*\|+\sum_{t=2}^{T}\big\|\theta_t-\theta_{t-1}^*+\theta_{t-1}^*-\theta_t^*\big\|$, which is less than $\|\theta_1-\theta_1^*\|+\sum_{t=1}^{T}\|\theta_{t+1}-\theta_t^*\|+\sum_{t=2}^{T}\big\|\theta_t^*-\theta_{t-1}^*\big\|$. According to Lemma 2, we have $\sum_{t=1}^{T}\|\theta_{t+1}-\theta_t^*\|\leq\rho\sum_{t=1}^{T}\|\theta_t-\theta_t^*\|$ with $\rho=\sqrt{1-\frac{l(1-\gamma)}{u(1-\gamma)+l\gamma}}$. Then we have $\sum_{t=1}^{T}\|\theta_t-\theta_t^*\|\leq\|\theta_1-\theta_1^*\|+\rho\sum_{t=1}^{T}\|\theta_t-\theta_t^*\|+\sum_{t=2}^{T}\big\|\theta_t^*-\theta_{t-1}^*\big\|$, which can be reformulated as $\sum_{t=1}^{T}\|\theta_t-\theta_t^*\|\leq\frac{1}{1-\rho}(\|\theta_1-\theta_1^*\|++\sum_{t=2}^{T}\big\|\theta_t^*-\theta_{t-1}^*\big\|)$.

$1-\rho=1-\sqrt{1-\frac{a_0}{b_0}}=\frac{\sqrt{b_0}-\sqrt{b_0-a_0}}{\sqrt{b_0}}$, where $a_0=\ell$ and $b_0=\frac{\ell\gamma+u(1-\gamma)}{1-\gamma}$. Thus, $1/(1-\rho)=\frac{\sqrt{b_0}}{\sqrt{b_0}-\sqrt{b_0-a_0}}=\frac{\sqrt{b_0}(\sqrt{b_0}+\sqrt{b_0-a_0})}{a_0}$. After plugging in the expression of $1-\gamma=1/T^\beta$, $1/(1-\rho)=\frac{\sqrt{\ell(T^\beta-1)+u}\big(\sqrt{\ell(T^\beta-1)+u}+\sqrt{\ell(T^\beta-1)+u-\ell}\big)}{\ell}\leq\frac{2\big(\ell(T^\beta-1)+u\big)}{\ell}=2(T^\beta-1)+u/\ell$

Then $\mathcal{R}_d=\sum_{t=1}^{T}\big(f_t(\theta_t)-f_t(\theta_t^*)\big)\leq G\frac{1}{1-\rho}\big(\|\theta_1-\theta_1^*\|+\sum_{t=2}^{T}\big\|\theta_t^*-\theta_{t-1}^*\big\|\big)\leq G\big(2(T^\beta-1)+u/\ell\big)\big(\|\theta_1-\theta_1^*\|+\sum_{t=2}^{T}\big\|\theta_t^*-\theta_{t-1}^*\big\|\big)$.

□

Proof of Theorem 5:

Proof. The proof follows the similar steps in the proof of Theorem 3.

According to the non-expansive property of the projection operator and the update rule in Eq.(6), we have

$$\begin{aligned}\|\theta_{t+1}-\theta^*\|^2&\leq\|\theta_t-\eta_t\nabla f_t(\theta_t)-\theta^*\|^2\\&=\|\theta_t-\theta^*\|^2-2\eta_t\nabla f_t(\theta_t)^T(\theta_t-\theta^*)\\&\quad+\eta_t^2\|\nabla f_t(\theta_t)\|^2\end{aligned}$$

The reformulation gives us

$$\begin{aligned}\nabla f_t(\theta_t)^T(\theta_t-\theta^*)&\leq\tfrac{1}{2\eta_t}\big(\|\theta_t-\theta^*\|^2-\|\theta_{t+1}-\theta^*\|^2\big)\\&\quad+\tfrac{\eta_t}{2}\|\nabla f_t(\theta_t)\|^2\end{aligned}\tag{22}$$

Moreover, from the strong convexity, we have $f_t(\theta^*)\geq f_t(\theta_t)+\nabla f_t(\theta_t)^T(\theta^*-\theta_t)+\frac{\ell}{2}\|\theta^*-\theta_t\|^2$, which is equivalent to $\nabla f_t(\theta_t)^T(\theta_t-\theta^*)\geq f_t(\theta_t)-f_t(\theta^*)+\frac{\ell}{2}\|\theta^*-\theta_t\|^2$. Combined with Eq.(22), we have

$$\begin{aligned}f_t(\theta_t)-f_t(\theta^*)&\leq\tfrac{1}{2\eta_t}\big(\|\theta_t-\theta^*\|^2-\|\theta_{t+1}-\theta^*\|^2\big)\\&\quad+\tfrac{\eta_t}{2}\|\nabla f_t(\theta_t)\|^2-\tfrac{\ell}{2}\|\theta^*-\theta_t\|^2\end{aligned}$$

Summing up from $t=1$ to T with $\|\nabla f_t(\theta_t)\|^2\leq G^2$, we get

$$\begin{aligned}&\sum_{t=1}^{T}\big(f_t(\theta_t)-f_t(\theta^*)\big)\\&\leq\sum_{t=1}^{T}\tfrac{1}{2\eta_t}\big(\|\theta_t-\theta^*\|^2-\|\theta_{t+1}-\theta^*\|^2\big)\\&\quad+\sum_{t=1}^{T}\tfrac{\eta_t}{2}G^2-\sum_{t=1}^{T}\tfrac{\ell}{2}\|\theta^*-\theta_t\|^2\\&\leq G^2/2\sum_{t=1}^{T}\eta_t+\tfrac{1/\eta_1-\ell}{2}\|\theta_1-\theta^*\|^2\\&\quad+\tfrac{1}{2}\sum_{t=2}^{T}\Big[\big(\tfrac{1}{\eta_t}-\tfrac{1}{\eta_{t-1}}-\ell\big)\|\theta^*-\theta_t\|^2\Big]\end{aligned}\tag{23}$$

Since $\eta_t=\frac{1-\gamma}{\ell(\gamma-\gamma^t)+u(1-\gamma)}$, $1/\eta_1=u$ and $\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-\ell=\frac{\ell(\gamma^{t-1}-1)(1-\gamma)}{1-\gamma}\leq 0$.

For the term $\sum_{t=1}^{T}\eta_t=\sum_{t=1}^{T}\frac{1-\gamma}{\ell(\gamma-\gamma^t)+u(1-\gamma)}$, it can be reformulated as $\frac{1}{u}\sum_{t=1}^{T}\frac{\frac{u(1-\gamma)}{\ell(\gamma-\gamma^t)}}{1+\frac{u(1-\gamma)}{\ell(\gamma-\gamma^t)}}=\frac{1}{u}+\frac{1}{u}\sum_{t=2}^{T}\frac{\frac{u(1-\gamma)}{\ell(\gamma-\gamma^t)}}{1+\frac{u(1-\gamma)}{\ell(\gamma-\gamma^t)}}\leq\frac{1}{u}+\frac{1}{u}\sum_{t=2}^{T}\frac{u(1-\gamma)}{\ell(\gamma-\gamma^t)}=\frac{1}{u}+\frac{1-\gamma}{\ell\gamma}\sum_{t=2}^{T}\frac{1}{1-\gamma^{t-1}}=\frac{1}{u}+\frac{1-\gamma}{\ell\gamma}\sum_{t=1}^{T-1}\frac{1}{1-\gamma^t}$. For $\sum_{t=1}^{T-1}\frac{1}{1-\gamma^t}$, we know that $\sum_{t=1}^{T-1}\frac{1}{1-\gamma^t}\leq O(T)$ as shown in the proof of Theorem 3. For the term $\frac{1-\gamma}{\ell\gamma}$, $\frac{1-\gamma}{\ell\gamma}=\frac{1}{\ell(T^\beta-1)}$. Combining these two terms' inequalities, we get that $\sum_{t=1}^{T}\eta_t\leq O(T^{1-\beta})$.

As a result, the inequality (23) can be reduced to

$$\sum_{t=1}^{T}\big(f_t(\theta_t)-f_t(\theta^*)\big)\leq O(T^{1-\beta})$$

□

Proof of Theorem 6:

Proof. According to the non-expansive property of the projection operator and the update rule in Eq.(7), we have

$$\begin{aligned}\|\theta_{t+1}-z_t\|^2&\leq\|\theta_t-\eta_t\nabla f_t(\theta_t)-z_t\|^2\\&=\|\theta_t-z_t\|^2-2\eta_t\nabla f_t(\theta_t)^T(\theta_t-z_t)\\&\quad+\eta_t^2\|\nabla f_t(\theta_t)\|^2\end{aligned}$$

The reformulation gives us

$$\begin{aligned}\nabla f_t(\theta_t)^T(\theta_t-z_t)&\leq\tfrac{1}{2\eta_t}\big(\|\theta_t-z_t\|^2-\|\theta_{t+1}-z_t\|^2\big)\\&\quad+\tfrac{\eta_t}{2}\|\nabla f_t(\theta_t)\|^2\end{aligned}\tag{24}$$

Moreover, from the strong convexity, we have $f_t(z_t)\geq f_t(\theta_t)+\nabla f_t(\theta_t)^T(z_t-\theta_t)+\frac{\ell}{2}\|z_t-\theta_t\|^2$, which is equivalent to $\nabla f_t(\theta_t)^T(\theta_t-z_t)\geq f_t(\theta_t)-f_t(z_t)+\frac{\ell}{2}\|z_t-\theta_t\|^2$.

Combined with Eq.(24), we have

$$\begin{aligned} f_t(\theta_t) - f_t(z_t) \quad &\leq \tfrac{1}{2\eta_t}\big(\|\theta_t - z_t\|^2 - \|\theta_{t+1} - z_t\|^2 \big) \\ &\quad + \tfrac{\eta_t}{2} \|\nabla f_t(\theta_t)\|^2 - \tfrac{\ell}{2} \|z_t - \theta_t\|^2 \end{aligned} \tag{25}$$

Then we can lower bound $\|\theta_{t+1} - z_t\|^2$ by

$$\begin{aligned} \|\theta_{t+1} - z_t\|^2 \quad &= \|\theta_{t+1} - z_{t+1}\|^2 + \|z_{t+1} - z_t\|^2 \\ &\quad + 2(\theta_{t+1} - z_{t+1})^\top (z_{t+1} - z_t) \\ &\geq \|\theta_{t+1} - z_{t+1}\|^2 - 4D\|z_{t+1} - z_t\| \end{aligned} \tag{26}$$

Combining (25) and (26) gives

$$\begin{aligned} & f_t(\theta_t) - f_t(z_t) \\ & \leq \tfrac{1}{2\eta_t}\big(\|\theta_t - z_t\|^2 - \|\theta_{t+1} - z_{t+1}\|^2 \big) + \tfrac{2D}{\eta_t}\|z_{t+1} - z_t\| \\ & \quad + \tfrac{\eta_t}{2} \|\nabla f_t(\theta_t)\|^2 - \tfrac{\ell}{2} \|z_t - \theta_t\|^2 \end{aligned}$$

Summing over t from 1 to T, dropping the term $-\frac{1}{2\eta_T}\|\theta_{T+1} - z_{T+1}\|^2$, setting $z_{T+1} = z_T$, using the inequality $\|\nabla f_t(\theta_t)\|^2 \leq G^2$, and re-arranging gives

$$\begin{aligned} & \sum_{t=1}^{T} \Big(f_t(\theta_t) - f_t(z_t) \Big) \\ & \leq \tfrac{1}{2}(\tfrac{1}{\eta_1} - \ell)\|\theta_1 - z_1\|^2 + \tfrac{1}{2}\sum_{t=1}^{T}(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} - \ell)\|\theta_t - z_t\|^2 \\ & \quad + 2D \sum_{t=1}^{T-1} \tfrac{1}{\eta_t}\|z_{t+1} - z_t\| + \tfrac{G^2}{2} \sum_{t=1}^{T} \eta_t \\ & \leq 2D\ell \tfrac{1}{1-\gamma} V + \tfrac{G^2}{2} \sum_{t=1}^{T} \eta_t \end{aligned}$$

where for the second inequality, we use the following results: $\frac{1}{\eta_1} - \ell = 0$, $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \ell = \frac{\ell(1-\gamma)(\gamma^{t-1}-1)}{1-\gamma} \leq 0$, $\frac{1}{\eta_t} = \frac{\ell(1-\gamma^t)}{1-\gamma} \leq \frac{\ell}{1-\gamma}$, and the definition of V.

□

Proof of Corollary 3:

Proof. Since $\gamma = 1 - \sqrt{\frac{V}{T}}$, $V = T^\beta$, and $\beta \in [0, 1)$, $1 - \gamma = T^{\frac{1}{2}(\beta-1)}$. Theorem 6 leads to

$$\sum_{t=1}^{T} \big(f_t(\theta_t) - f_t(z_t) \big) \leq \frac{2D\ell}{1-\gamma} V + \frac{G^2}{\ell}(1-\gamma) \sum_{t=1}^{T} \frac{1}{1-\gamma^t}$$

The first term $\frac{1}{1-\gamma} V = \sqrt{TV}$. Then second term $\sum_{t=1}^{T} \frac{1}{1-\gamma^t} \leq \frac{1}{1-\gamma} + \int_1^T \frac{1}{1-\gamma^t} \mathrm{d}t = \frac{1}{1-\gamma} + \Big(t - \frac{\ln(1-\gamma^t)}{\ln(\gamma)} \Big) \Big|_1^T = \frac{1}{1-\gamma} + T - 1 + \frac{\ln(1-\gamma)}{\ln \gamma} - \frac{\ln(1-\gamma^T)}{\ln \gamma} \leq \frac{1}{1-\gamma} + T - 1 + \frac{\ln(1-\gamma)}{\ln \gamma}$. $\ln(1-\gamma) = -\frac{1}{2}(1-\beta)\ln T$. $-\ln \gamma = \ln \frac{1}{1 - T^{\frac{1}{2}(\beta-1)}} = \ln \frac{T^{\frac{1}{2}(1-\beta)}}{T^{\frac{1}{2}(1-\beta)} - 1} = \ln(1 + \frac{1}{T^{\frac{1}{2}(1-\beta)} - 1}) \geq \frac{1}{2} \frac{1}{T^{\frac{1}{2}(1-\beta)} - 1}$, where the inequality is due to $\ln(1+x) \geq \frac{1}{2}x, x \in [0, 1]$. Then $\frac{\ln(1-\gamma)}{\ln \gamma} \leq O(T^{\frac{1}{2}(1-\beta)} \ln T)$, which leads to $\sum_{t=1}^{T} \frac{1}{1-\gamma^t} \leq O(T)$. Then the second term $(1-\gamma) \sum_{t=1}^{T} \frac{1}{1-\gamma^t} \leq O(\sqrt{TV})$, which completes the proof. □

Proof of Corollary 4:

Proof. Since $\gamma = 1 - \frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}}$ and $V \in [0, 2DT]$, $1/2 \leq \gamma < 1$. Then the integral bound in the proof of Corollary 3 can be used, which is $\sum_{t=1}^{T} \frac{1}{1-\gamma^t} \leq \frac{1}{1-\gamma} + T + \frac{\ln(1-\gamma)}{\ln \gamma}$.

Next, we upper bound each term on the right-hand-side of Theorem 6 individually. $\frac{1}{1-\gamma} V = 2\sqrt{\frac{2DT}{\max\{V, \log^2 T/T\}}} V \leq O(\sqrt{TV})$. $(1-\gamma) \sum_{t=1}^{T} \frac{1}{1-\gamma^t} \leq 1 + (1-\gamma)(T + \frac{\ln(1-\gamma)}{\ln \gamma})$.

$$\begin{aligned} & \tfrac{\ln(1-\gamma)}{\ln \gamma} \\ & = \frac{-\ln(\frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}})}{-\ln(1 - \frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}})} \\ & = \frac{-\ln(\frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}})}{\ln\Big(1 + \frac{\frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}}}{1 - \frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}}}\Big)} \\ & \leq \ln(2\sqrt{\tfrac{2DT}{\max\{V, \log^2 T/T\}}}) 4 \sqrt{\tfrac{2DT}{\max\{V, \log^2 T/T\}}} \\ & \leq O(\ln(T/\log T)\tfrac{T}{\log T}) \\ & \leq O(T) \end{aligned}$$

where the first inequality follows by using $\ln(1+x) \geq \frac{1}{2}x, x \in [0, 1]$, and $1 - \frac{1}{2}\sqrt{\frac{\max\{V, \log^2 T/T\}}{2DT}} < 1$.

Thus, $(1-\gamma) \sum_{t=1}^{T} \frac{1}{1-\gamma^t} \leq \max\{O(\log T), O(\sqrt{TV})\}$. The final result follows by combining the two terms' results. □

Proof of Lemma 3:

Proof. The first part of the proof is the same as the first part of the result in the Proof of Lemma 1 in (Zhang, Lu, and Zhou 2018), which follows the result in (Cesa-Bianchi and Lugosi 2006). We define $L_t^\gamma = \sum_{i=1}^{t} f_i(\theta_i^\gamma)$, and $W_t = \sum_{\gamma \in \mathcal{H}} w_1^\gamma \exp(-\alpha L_t^\gamma)$.

The following update is equivalent to the update rule in Algorithm 2:

$$w_t^\gamma = \frac{w_1^\gamma \exp(-\alpha L_{t-1}^\gamma)}{\sum_{\mu \in \mathcal{H}} w_1^\mu \exp(-\alpha L_{t-1}^\mu)}, \quad t \geq 2. \tag{27}$$

First, we have

$$\begin{aligned} \log W_T \quad &= \log\big(\sum_{\gamma \in \mathcal{H}} w_1^\gamma \exp(-\alpha L_T^\gamma) \big) \\ &\geq \log\big(\max_{\gamma \in \mathcal{H}} w_1^\gamma \exp(-\alpha L_T^\gamma) \big) \\ &= -\alpha \min_{\gamma \in \mathcal{H}} \big(L_T^\gamma + \tfrac{1}{\alpha} \log \tfrac{1}{w_1^\gamma} \big). \end{aligned} \tag{28}$$

Then we bound the quantity $\log(W_t/W_{t-1})$. For $t \geq 2$, we get

$$\begin{aligned}
&\log\Big(\tfrac{W_t}{W_{t-1}}\Big) \\
&= \log\Big(\tfrac{\sum_{\gamma\in\mathcal{H}} w_1^\gamma \exp(-\alpha L_t^\gamma)}{\sum_{\gamma\in\mathcal{H}} w_1^\gamma \exp(-\alpha L_{t-1}^\gamma)}\Big) \\
&= \log\Big(\tfrac{\sum_{\gamma\in\mathcal{H}} w_1^\gamma \exp(-\alpha L_{t-1}^\gamma)\exp(-\alpha f_t(\theta_t^\gamma))}{\sum_{\gamma\in\mathcal{H}} w_1^\gamma \exp(-\alpha L_{t-1}^\gamma)}\Big) \\
&= \log\Big(\sum_{\gamma\in\mathcal{H}} w_t^\gamma \exp(-\alpha f_t(\theta_t^\gamma))\Big)
\end{aligned} \tag{29}$$

where the last equality is due to Eq.(27).

When $t = 1$, $\log W_1 = \log\Big(\sum_{\gamma\in\mathcal{H}} w_1^\gamma \exp(-\alpha f_1(\theta_1^\gamma))\Big)$. Then $\log W_T$ can be expressed as:

$$\begin{aligned}
\log W_T &= \log W_1 + \sum_{t=2}^{T} \log\Big(\tfrac{W_t}{W_{t-1}}\Big) \\
&= \sum_{t=1}^{T} \log\Big(\sum_{\gamma\in\mathcal{H}} w_t^\gamma \exp(-\alpha f_t(\theta_t^\gamma))\Big).
\end{aligned} \tag{30}$$

The rest of the proof is new.

Due to the α-exp-concavity, $\exp(-\alpha f_t(\sum_{\gamma\in\mathcal{H}} w_t^\gamma \theta_t^\gamma)) \geq \sum_{\gamma\in\mathcal{H}} w_t^\gamma \exp(-\alpha f_t(\theta_t^\gamma))$, which is equivalent to

$$\begin{aligned}
\log\Big(\textstyle\sum_{\gamma\in\mathcal{H}} w_t^\gamma \exp(-\alpha f_t(\theta_t^\gamma))\Big) &\leq -\alpha f_t\Big(\textstyle\sum_{\gamma\in\mathcal{H}} w_t^\gamma \theta_t^\gamma\Big) \\
&= -\alpha f_t(\theta_t)
\end{aligned} \tag{31}$$

Combining the Inequalities (28), (30), and (31), we get

$$-\alpha \min_{\gamma\in\mathcal{H}} \big(L_T^\gamma + \frac{1}{\alpha}\log\frac{1}{w_1^\gamma}\big) \leq -\alpha \sum_{t=1}^{T} f_t(\theta_t)$$

which can be reformulated as

$$\sum_{t=1}^{T} f_t(\theta_t) \leq \min_{\gamma\in\mathcal{H}} \Big(\sum_{t=1}^{T} f_t(\theta_t^\gamma) + \frac{1}{\alpha}\log\frac{1}{w_1^\gamma}\Big)$$

Since it holds for the minimum value, it is true for all $\gamma \in \mathcal{H}$, which completes the proof.

□

Proof of Theorem 7:

Proof. When $\gamma = \gamma^* = 1 - \frac{1}{2}\frac{\log T}{T}\sqrt{\frac{\max\{\frac{T}{\log^2 T}V, 1\}}{2D}} = 1 - \eta^*$, we have $\sum_{t=1}^{T}(f_t(\theta_t^{\gamma^*}) - f_t(z_t)) \leq \max\{O(\log T), O(\sqrt{TV})\}$ based on the Corollary 2.

Since $0 \leq V \leq 2TD$, $\frac{1}{2}\frac{\log T}{T\sqrt{2D}} \leq \eta^* \leq \frac{1}{2}$.

According to our definition of η_i, $\min \eta_i = \frac{1}{2}\frac{\log T}{T\sqrt{2D}}$ and $\frac{1}{2} \leq \max \eta_i < 1$, which means for any value of V, there always exists a η_k such that

$$\eta_k = \frac{1}{2}\frac{\log T}{T\sqrt{2D}} 2^{k-1} \leq \eta^* \leq 2\eta_k = \eta_{k+1}$$

where $k = \lfloor \frac{1}{2}\log_2(\max\{\frac{T}{\log^2 T}V, 1\})\rfloor + 1$.

Since $0 < \eta_k \leq \frac{1}{2}$, $\frac{1}{2} \leq \gamma_k = 1 - \eta_k < 1$ and $\gamma_k \geq \gamma^*$.

According to Theorem 1, we have

$$\begin{aligned}
\textstyle\sum_{t=1}^{T}(f_t(\theta_t^{\gamma_k}) - f_t(z_t)) \quad &\leq -a_1 T\log\gamma_k - a_2\log(1-\gamma_k) \\
&\quad + \tfrac{a_3}{1-\gamma_k}V + a_4.
\end{aligned}$$

For the first term on the RHS, $-T\log\gamma_k = T\log\frac{1}{\gamma_k} \leq T\log\frac{1}{\gamma^*}$.

For the second one, $-\log(1-\gamma_k) = -\log\frac{1}{2}(2-2\gamma_k) = -\log\frac{1}{2}2\eta_k$. Since $1 \geq 2\eta_k \geq \eta^*$, $\frac{1}{2}2\eta_k \geq \frac{1}{2}\eta^*$, which leads to $-\log\frac{1}{2}2\eta_k \leq -\log\frac{1}{2}\eta^*$ and $-\log(1-\gamma_k) \leq -\log\frac{1}{2}\eta^* = \log 2 - \log(1-\gamma^*)$.

For the third one, $\frac{1}{1-\gamma_k}V = \frac{1}{\eta_k}V = \frac{2}{2\eta_k}V \leq \frac{2}{\eta^*}V = \frac{2}{1-\gamma^*}V$.

Since all the terms can be expressed in terms of γ^* in the original form without adding the order, based on the Corollary 2, we get:

$$\sum_{t=1}^{T}(f_t(\theta_t^{\gamma_k}) - f_t(z_t)) \leq \max\{O(\log T), O(\sqrt{TV})\} \tag{32}$$

What's more, from Lemma 3 we get

$$\begin{aligned}
\sum_{t=1}^{T}(f_t(\theta_t) - f_t(\theta_t^{\gamma_k})) &\leq \tfrac{1}{\alpha}\log\tfrac{1}{w_1^{\gamma_k}} \\
&\leq \tfrac{1}{\alpha}\log(k(k+1)) \\
&\leq 2\tfrac{1}{\alpha}\log(k+1) \\
&\leq O(\log(\log T))
\end{aligned} \tag{33}$$

Combining the above inequalities (32) and (33) completes the proof. □

Proof of Lemma 4:

Proof. Let $g(x) = \exp(-\alpha f(x))$. To prove the concavity of $g(x)$, it is equivalent to show $\langle \nabla g(x) - \nabla g(y), x - y\rangle \leq 0, x, y \in \mathcal{S}$. Since $\nabla g(x) = \exp(-\alpha f(x))(-\alpha)\nabla f(x)$, it is equivalent to prove that $\langle \exp(-\alpha f(x))\nabla f(x) - \exp(-\alpha f(y))\nabla f(y), x - y\rangle \geq 0$, which can be reformulated as

$$\exp(-\alpha f(x))\langle \nabla f(x), x - y\rangle \geq \exp(-\alpha f(y))\langle \nabla f(y), x - y\rangle \tag{34}$$

Without loss of generality, let us assume $f(x) \geq f(y)$. Due to ℓ-strong convexity, $f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle + \frac{\ell}{2}\|x - y\|^2$, which leads to

$$\langle \nabla f(y), x - y\rangle \leq f(x) - f(y) - \frac{\ell}{2}\|x - y\|^2 \tag{35}$$

What's more, $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\ell}{2}\|x - y\|^2$, which leads to

$$\langle \nabla f(x), x - y\rangle \geq f(x) - f(y) + \frac{\ell}{2}\|x - y\|^2 \tag{36}$$

Combining inequalities (34), (35), and (36), it is enough to prove that $\exp(-\alpha f(x))(f(x) - f(y) + \frac{\ell}{2}\|x - y\|^2) \geq \exp(-\alpha f(y))(f(x) - f(y) - \frac{\ell}{2}\|x - y\|^2)$, which can be reformulated as $\frac{\ell}{2}\|x - y\|^2(\exp(-\alpha f(x)) + \exp(-\alpha f(y))) \geq (f(x) - f(y))(\exp(-\alpha f(y)) - \exp(-\alpha f(x)))$. When $x - y = 0$, it is always true. Let us consider the case when $\|x -$

$y\| > 0$. Then we need to show that $\frac{\ell}{2}\Big(1+\exp\big(\alpha\big(f(x)-f(y)\big)\big)\Big) \geq \frac{f(x)-f(y)}{\|x-y\|}\frac{\exp\Big(\alpha\big(f(x)-f(y)\big)\Big)-1}{\|x-y\|}$. Due to bounded gradient and Mean value theorem, $\frac{f(x)-f(y)}{\|x-y\|} \leq G$, which means it is enough to show that

$$\frac{\ell}{2G}\Big(1+\exp\big(\alpha\big(f(x)-f(y)\big)\big)\Big) \geq \frac{\exp\Big(\alpha\big(f(x)-f(y)\big)\Big)-1}{\|x-y\|} \tag{37}$$

According to the Taylor series, $\exp\Big(\alpha\big(f(x)-f(y)\big)\Big) = 1+\alpha\big(f(x)-f(y)\big)+\frac{1}{2!}\alpha^2\big(f(x)-f(y)\big)^2+\cdots+\frac{1}{n!}\alpha^n\big(f(x)-f(y)\big)^n, n\to\infty$. Thus, $\frac{\exp\Big(\alpha\big(f(x)-f(y)\big)\Big)-1}{\|x-y\|} = \alpha\frac{f(x)-f(y)}{\|x-y\|}+\frac{1}{2}\alpha^2(f(x)-f(y))\frac{f(x)-f(y)}{\|x-y\|}+\cdots+\frac{1}{n!}\alpha^n\big(f(x)-f(y)\big)^{n-1}\frac{f(x)-f(y)}{\|x-y\|}, n\to\infty$. Since $\frac{f(x)-f(y)}{\|x-y\|}\leq G$, we have

$$\begin{aligned}&\frac{\exp\Big(\alpha\big(f(x)-f(y)\big)\Big)-1}{\|x-y\|}\\ &\leq \alpha G+\tfrac{1}{2}\alpha^2(f(x)-f(y))G+\ldots\\ &\quad+\tfrac{1}{n!}\alpha^n\big(f(x)-f(y)\big)^{n-1}G\end{aligned} \tag{38}$$

For the LHS of inequality (37), it is equal to

$$\begin{aligned}&\tfrac{\ell}{G}+\alpha\tfrac{\ell}{2G}(f(x)-f(y))+\tfrac{1}{2!}\alpha^2\tfrac{\ell}{2G}(f(x)-f(y))^2\\ &+\cdots+\tfrac{1}{n!}\alpha^n\tfrac{\ell}{2G}(f(x)-f(y))^n, n\to\infty\end{aligned} \tag{39}$$

If we compare the coefficients of the RHS from the inequality (38) with the one in (39) and plug in $\alpha=\ell/G^2$, we see that it is always smaller or equal, which completes the proof.

□

Proof of Theorem 8:

Proof. As in the proof of Theorem 7, all we need to show is that there exists an algorithm A^γ, which can bound the regret $\sum_{t=1}^T(f_t(\theta_t^\gamma)-f_t(z_t))\leq O(\max\{\log T,\sqrt{TV}\})$.

When $\gamma=\gamma^*=1-\frac{1}{2}\frac{\log T}{T}\sqrt{\frac{\max\{\frac{T}{\log^2 T}V,1\}}{2D}}=1-\eta^*$, we have $\sum_{t=1}^T(f_t(\theta_t^{\gamma^*})-f_t(z_t))\leq O(\max\{\log T,\sqrt{TV}\})$ based on the Corollary 4.

Since $0\leq V\leq 2TD$, $\frac{1}{2}\frac{\log T}{T\sqrt{2D}}\leq\eta^*\leq\frac{1}{2}$.

According to our definition of η_i, $\min\eta_i=\frac{1}{2}\frac{\log T}{T\sqrt{2D}}$ and $\frac{1}{2}\leq\max\eta_i<1$, which means for any value of V, there always exists a η_k such that

$$\eta_k=\frac{1}{2}\frac{\log T}{T\sqrt{2D}}2^{k-1}\leq\eta^*\leq 2\eta_k=\eta_{k+1}$$

where $k=\lfloor\frac{1}{2}\log_2(\max\{\frac{T}{\log^2 T}V,1\})\rfloor+1$.

Since $0<\eta_k\leq\frac{1}{2}$, $\frac{1}{2}\leq\gamma_k=1-\eta_k<1$ and $\gamma_k\geq\gamma^*$.

According to Theorem 6, we have

$$\sum_{t=1}^T\big(f_t(\theta_t^{\gamma_k})-f_t(z_t)\big)\leq\frac{2D\ell}{1-\gamma_k}V+\frac{G^2}{\ell}(1-\gamma_k)\sum_{t=1}^T\frac{1}{1-\gamma_k^t}$$

For the first term on the RHS, $\frac{1}{1-\gamma_k}V=\frac{1}{\eta_k}V=\frac{2}{2\eta_k}V\leq\frac{2}{\eta^*}V=\frac{2}{1-\gamma^*}V$.

For the second one, $1-\gamma_k\leq 1-\gamma^*$. According to the proof in Corollary 4, $\sum_{t=1}^T\frac{1}{1-\gamma_k^t}\leq\frac{1}{1-\gamma_k}+T+\frac{\log(1-\gamma_k)}{\log\gamma_k}$.

$$\frac{\log(1-\gamma_k)}{\log\gamma_k}=\frac{\log\eta_k}{\log(1-\eta_k)}=\frac{-\log\eta_k}{-\log(1-\eta_k)}. \tag{40}$$

Since $\eta_k\geq\frac{1}{2}\eta^*$, $\log\eta_k\geq\log\frac{1}{2}\eta^*$ and

$$0<-\log\eta_k\leq-\log\frac{1}{2}\eta^*=\log 2-\log\eta^*. \tag{41}$$

Since $\eta_k\geq\frac{1}{2}\eta^*$, $1-\eta_k\leq 1-\frac{1}{2}\eta^*$. Then $\log(1-\eta_k)\leq\log(1-\frac{1}{2}\eta^*)$, which results in

$$-\log(1-\eta_k)\geq-\log(1-\frac{1}{2}\eta^*)>0. \tag{42}$$

Combining inequalities (41) and (42) with Eq.(40), we get

$$\begin{aligned}\tfrac{\log(1-\gamma_k)}{\log\gamma_k}&\leq\tfrac{\log 2-\log\eta^*}{-\log(1-\frac{1}{2}\eta^*)}\\ &=\tfrac{\log 2}{-\log(1-\frac{1}{2}\eta^*)}+\tfrac{-\log\eta^*}{-\log(1-\frac{1}{2}\eta^*)}\end{aligned} \tag{43}$$

For the first term on the RHS,

$$\begin{aligned}-\log(1-\tfrac{1}{2}\eta^*)&=\log\Big(\tfrac{1}{1-\frac{1}{4}\sqrt{\frac{\max\{V,\log^2 T/T\}}{2DT}}}\Big)\\ &=\log\Big(1+\tfrac{\frac{1}{4}\sqrt{\frac{\max\{V,\log^2 T/T\}}{2DT}}}{1-\frac{1}{4}\sqrt{\frac{\max\{V,\log^2 T/T\}}{2DT}}}\Big)\\ &\geq\tfrac{1}{2}\tfrac{\frac{1}{4}\sqrt{\frac{\max\{V,\log^2 T/T\}}{2DT}}}{1-\frac{1}{4}\sqrt{\frac{\max\{V,\log^2 T/T\}}{2DT}}}\\ &\geq\tfrac{1}{8}\sqrt{\tfrac{\max\{V,\log^2 T/T\}}{2DT}}\end{aligned}$$

where the first inequality is due to $\log(1+x)\geq\frac{1}{2}x, x\in[0,1]$ and the second one is due to $\sqrt{\frac{\max\{V,\log^2 T/T\}}{2DT}}>0$. As a result,

$$\begin{aligned}\tfrac{\log 2}{-\log(1-\frac{1}{2}\eta^*)}&\leq 8\sqrt{\tfrac{2DT}{\max\{V,\log^2 T/T\}}}\log 2\\ &\leq 8\tfrac{T}{\log T}\sqrt{2D}\log 2<O(T).\end{aligned}$$

For the second term on the RHS of Eq.(43),

$$\begin{aligned}-\log\eta^*&=\log\Big(2\sqrt{\tfrac{2DT}{\max\{V,\log^2 T/T\}}}\Big)\\ &\leq\log 2+\tfrac{1}{2}\log 2D+\tfrac{1}{2}\log\tfrac{T}{\log T}\end{aligned}$$

Combining the inequalities for $-\log\eta^*$ and $-\log(1-\frac{1}{2}\eta^*)$, we get $\frac{-\log\eta^*}{-\log(1-\frac{1}{2}\eta^*)}\leq(\log 2+\frac{1}{2}\log 2D+\frac{1}{2}\log\frac{T}{\log T})8\frac{T}{\log T}\sqrt{2D}\leq O(T)$.

As a result, $\frac{\log(1-\gamma_k)}{\log\gamma_k}\leq O(T)$ and $\sum_{t=1}^T\frac{1}{1-\gamma_k^t}\leq O(T)$.

Since using γ_k does not increase the order when replacing with γ^*, we get

$$\sum_{t=1}^{T}\Big(f_t(\theta_t^{\gamma_k})-f_t(z_t)\Big)\leq O(\max\{\log T,\sqrt{TV}\})$$

which combining with the result of Lemma 3 completes the proof.

□

Online Least-Squares Optimization Consider the online least-squares problem with:

$$f_t(\theta)=\frac{1}{2}\left\|y_t-A_t\theta\right\|^2 \tag{44}$$

where $A_t\in\mathbb{R}^{m\times n}$, $A_t^TA_t$ has full rank with $lI\preceq A_t^TA_t\preceq uI$, and $y_t\in\mathbb{R}^m$ comes from a bounded set with $\|y_t\|\leq D$.

In the main paper, we analyzed the dynamic regret of discounted recursive least squares against comparison sequences $z_1,\ldots,z_T$ with a path length constraint $\sum_{t=2}^{T}\|z_t-z_{t-1}\|\leq V$. Additionally, we analyzed the trade-off between static and dynamic regret of a gradient descent rule with comparison sequence $\theta_t^*=\operatorname{argmin}_{\theta\in\mathcal{S}}f_t(\theta)$. In this appendix, we analyze the trade-off between static regret and dynamic regret with comparison sequence θ_t^* achieved by discounted recursive least squares. We will see that the discounted recursive least squares achieves trade-offs depend on the condition number, $\delta=u/l$. In particular, low dynamic regret is only guaranteed for low condition numbers.

Recall that discounted recursive least squares corresponds to Alg. 1 run with a full Newton step and $\eta=1$. In this case, $P_t=\sum_{i=1}^{t}\gamma^{i-1}A_{t+1-i}^TA_{t+1-i}=\gamma P_{t-1}+A_t^TA_t$, and the update rule can be written more explicitly as

$$\theta_{t+1}=\Big(\sum_{i=1}^{t}\gamma^{i-1}A_{t+1-i}^TA_{t+1-i}\Big)^{-1}\Big(\sum_{i=1}^{t}\gamma^{i-1}A_{t+1-i}^Ty_{t+1-i}\Big) \tag{45}$$

The above update rule can be reformulated as:

$$\theta_{t+1}=\theta_t-P_t^{-1}\nabla f_t(\theta_t). \tag{46}$$

Before we analyze dynamic and static regret for the update (46), we first show some supporting results for $\|y_t-A_tx\|$ and $\|\nabla f_t(x)\|$, where $x\in\{v|v=\beta\theta_t+(1-\beta)\theta_t^*,\beta\in[0,1]\}$.

Lemma 6. *Let θ_t be the result of Eq.*(46)*, and $\theta_t^*=\operatorname{argmin}f_t(\theta)$. For $x\in\{v|v=\beta\theta_t+(1-\beta)\theta_t^*,\beta\in[0,1]\}$, If $\|y_t\|\leq D$, then $\|y_t-A_tx\|\leq(u/l+1)D$.*

Proof. $\|y_t-A_tx\|\leq\|A_t\|_2\|x\|+\|y_t\|$, and $\|A_t\|_2=\sqrt{\sigma_1(A_t^TA_t)}\leq\sqrt{u}$. For $\|x\|$, we have $\|x\|=\|\beta\theta_t+(1-\beta)\theta_t^*\|\leq\beta\|\theta_t\|+(1-\beta)\|\theta_t^*\|$.

For the term $\|\theta_t\|$, $\|\theta_t\|=\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i}\Big)^{-1}\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^Ty_{t-i}\Big)\right\|$, which can be upper bounded by $\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i}\Big)^{-1}\right\|_2\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^Ty_{t-i}\Big)\right\|$. Then we upper bound these two terms individually.

$\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i}\Big)^{-1}\right\|_2=\frac{1}{\sigma_n(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i})}$. Since $lI\preceq A_{t-i}^TA_{t-i}\preceq uI$, $\frac{1-\gamma^{t-1}}{1-\gamma}lI\preceq\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i})\preceq\frac{1-\gamma^{t-1}}{1-\gamma}uI$. Thus, $\sigma_n(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i})\geq l\frac{1-\gamma^{t-1}}{1-\gamma}$, which results in $\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^TA_{t-i}\Big)^{-1}\right\|_2\leq\frac{1-\gamma}{l(1-\gamma^{t-1})}$.

For the term $\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^Ty_{t-i}\Big)\right\|$, we have $\left\|\Big(\sum_{i=1}^{t-1}\gamma^{i-1}A_{t-i}^Ty_{t-i}\Big)\right\|\leq\sum_{i=1}^{t-1}\gamma^{i-1}\left\|A_{t-i}^Ty_{t-i}\right\|\leq\sum_{i=1}^{t-1}\gamma^{i-1}\left\|A_{t-i}^T\right\|_2\|y_{t-i}\|\leq\frac{1-\gamma^{t-1}}{1-\gamma}\sqrt{u}D$. Then we have $\|\theta_t\|\leq\frac{\sqrt{u}}{l}D$.

For $\|\theta_t^*\|$, we have $\|\theta_t^*\|=\left\|(A_t^TA_t)^{-1}A_t^Ty_t\right\|\leq\left\|(A_t^TA_t)^{-1}\right\|_2\left\|A_t^T\right\|_2\|y_t\|\leq\frac{\sqrt{u}}{l}D$. Thus, $\|x\|\leq\frac{\sqrt{u}}{l}D$ and $\|y_t-A_tx\|\leq\|A_t\|_2\|x\|+\|y_t\|\leq(u/l+1)D$.

□

Corollary 5. *Let θ_t be the result of Eq.*(46) *and $\theta_t^*=\operatorname{argmin}f_t(\theta)$. For $x\in\{v|v=\beta\theta_t+(1-\beta)\theta_t^*,\beta\in[0,1]\}$, we have $\|\nabla f_t(x)\|\leq\sqrt{u}(u/l+1)D$.*

Proof. For $\|\nabla f_t(x)\|$, we have $\|\nabla f_t(x)\|=\left\|A_t^TA_tx-A_t^Ty_t\right\|\leq\left\|A_t^T\right\|_2\|A_tx-y_t\|\leq\sqrt{u}(u/l+1)D$, where the second inequality is due to Lemma 6 and the assumption of $A_t^TA_t\preceq uI$.

□

Moreover, we need to obtain the relationship between $\theta_{t+1}-\theta_t^*$ and $\theta_t-\theta_t^*$ as another necessary step to get the dynamic regret.

Lemma 7. *Let θ_t^* be the solution to $f_t(\theta)$ in Eq.*(44)*. When we use the discounted recursive least-squares update in Eq.*(46)*, the following relationship is obtained:*

$$\begin{aligned}&\theta_{t+1}-\theta_t^*\\&=\big(I-\gamma^{-1}P_{t-1}^{-1}A_t^T(I+A_t\gamma^{-1}P_{t-1}^{-1}A_t^T)^{-1}A_t\big)(\theta_t-\theta_t^*)\\&=\Big(I+\gamma^{-1}P_{t-1}^{-1}A_t^TA_t\Big)^{-1}(\theta_t-\theta_t^*)\end{aligned}$$

Proof. If we set $\Phi_t=\sum_{i=1}^{t}\gamma^{i-1}A_{t+1-i}^Ty_{t+1-i}=\gamma\Phi_{t-1}+A_t^Ty_t$, then according to the update of θ_{t+1} in Eq.(45), we have $\theta_{t+1}=(A_t^TA_t+\gamma P_{t-1})^{-1}(A_t^Ty_t+\gamma\Phi_{t-1})$, which by the use of inverse lemma can be further reformulated as:

$$\begin{aligned}\theta_{t+1}\;&=\Big(\gamma^{-1}P_{t-1}^{-1}-\gamma^{-2}P_{t-1}^{-1}A_t^T(I+\\&\quad A_t\gamma^{-1}P_{t-1}^{-1}A_t^T)^{-1}A_tP_{t-1}^{-1}\Big)\big(A_t^Ty_t+\gamma\Phi_{t-1}\big)\end{aligned} \tag{47}$$

Then for $\theta_{t+1} - \theta_t^* = \theta_{t+1} - (A_t^T A_t)^{-1} A_t^T y_t$, we have:

$$\begin{aligned}
&\theta_{t+1} - \theta_t^* \\
&= \underbrace{\left(I - \gamma^{-1} P_{t-1}^{-1} A_t^T (I + A_t \gamma^{-1} P_{t-1}^{-1} A_t^T)^{-1} A_t\right)}_{①} \theta_t + \underbrace{\gamma^{-1} P_{t-1}^{-1} A_t^T y_t}_{②.1} \\
&\underbrace{- \left(\gamma^{-2} P_{t-1}^{-1} A_t^T (I + A_t \gamma^{-1} P_{t-1}^{-1} A_t^T)^{-1} A_t P_{t-1}^{-1} - (A_t^T A_t)^{-1}\right) A_t^T y_t}_{②.2}
\end{aligned} \tag{48}$$

We want to prove ②.1 + ②.2 = ①$(-\theta_t^*)$ = ①$(-(A_t^T A_t)^{-1} A_t^T y_t)$ = ③.

Since $A(I + BA)^{-1}B = AB(I + AB)^{-1} = (I + AB)^{-1}AB$, for any compatible matrix A and B, we have:

$$\begin{aligned}
&③ \\
&= -\left[I - \gamma^{-1} P_{t-1}^{-1} A_t^T (I + A_t \gamma^{-1} P_{t-1}^{-1} A_t^T)^{-1} A_t\right] (A_t^T A_t)^{-1} A_t^T y_t \\
&= -\left[I - (I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t)^{-1} \gamma^{-1} P_{t-1}^{-1} A_t^T A_t\right] (A_t^T A_t)^{-1} A_t^T y_t \\
&= -\left[(A_t^T A_t)^{-1} - (I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t)^{-1} \gamma^{-1} P_{t-1}^{-1}\right] A_t^T y_t
\end{aligned} \tag{49}$$

Also, for any compatible P, we have $(I + P)^{-1} = I - (I + P)^{-1}P$. Then $(I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t)^{-1} = I - (I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t)^{-1} \gamma^{-1} P_{t-1}^{-1} A_t^T A_t$. Then ③ $= -\left[(A_t^T A_t)^{-1} - \gamma^{-1} P_{t-1}^{-1} + (I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t)^{-1} \gamma^{-2} P_{t-1}^{-1} A_t^T A_t P_{t-1}^{-1}\right] A_t^T y_t$. Compared with ②.1 + ②.2, we are left to prove $(I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t)^{-1} \gamma^{-2} P_{t-1}^{-1} A_t^T A_t P_{t-1}^{-1} = \gamma^{-2} P_{t-1}^{-1} A_t^T (I + A_t \gamma^{-1} P_{t-1}^{-1} A_t^T)^{-1} A_t P_{t-1}^{-1}$, which is always true.

As a result, we have $\theta_{t+1} - \theta_t^* = \left(I - \gamma^{-1} P_{t-1}^{-1} A_t^T (I + A_t \gamma^{-1} P_{t-1}^{-1} A_t^T)^{-1} A_t\right)(\theta_t - \theta_t^*)$, which can be simplified as $\theta_{t+1} - \theta_t^* = \left(I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t\right)^{-1} (\theta_t - \theta_t^*)$.

□

Corollary 6. *Let θ_t^* be the solution to $f_t(\theta)$ in Eq.*(44)*. When we use the discounted recursive least-squares update in Eq.*(46)*, the following relation is obtained:*

$$\|\theta_{t+1} - \theta_t^*\| \quad \leq \sqrt{\tfrac{u}{l}} \tfrac{u\gamma}{u\gamma + l(1-\gamma)} \|\theta_t - \theta_t^*\|$$

Proof. From Lemma 7 we know that

$$\theta_{t+1} - \theta_t^* = \left(I + \gamma^{-1} P_{t-1}^{-1} A_t^T A_t\right)^{-1} (\theta_t - \theta_t^*)$$

which can be reformulated as:

$$\theta_{t+1} - \theta_t^* = P_{t-1}^{-1/2} (I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})^{-1} P_{t-1}^{1/2} (\theta_t - \theta_t^*)$$

which gives us the following inequality:

$$\begin{aligned}
&\|\theta_{t+1} - \theta_t^*\| \\
&\leq \left\|P_{t-1}^{-1/2}\right\|_2 \left\|(I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})^{-1}\right\|_2 \\
&\quad \left\|P_{t-1}^{1/2}\right\|_2 \|\theta_t - \theta_t^*\|
\end{aligned}$$

Then we will upper bound the terms on the right-hand side individually.

Since $lI \preceq A_{t-i}^T A_{t-i} \preceq uI$, $\frac{1-\gamma^{t-1}}{1-\gamma} lI \preceq P_{t-1} = \sum_{i=1}^{t-1} \gamma^{i-1} A_{t-i}^T A_{t-i} \preceq \frac{1-\gamma^{t-1}}{1-\gamma} uI$.

For the term $\left\|P_{t-1}^{-1/2}\right\|_2$, we have $\left\|P_{t-1}^{-1/2}\right\|_2 = \frac{1}{\sqrt{\sigma_n(P_{t-1})}}$. Since $\sigma_n(P_{t-1}) \geq \frac{1-\gamma^{t-1}}{1-\gamma} l$, $\left\|P_{t-1}^{-1/2}\right\|_2 \leq \frac{1}{\sqrt{l}} \sqrt{\frac{1-\gamma}{1-\gamma^{t-1}}}$.

For the term $\left\|P_{t-1}^{1/2}\right\|_2$, we have $\left\|P_{t-1}^{1/2}\right\|_2 = \sqrt{\sigma_1(P_{t-1})}$. Since $\sigma_1(P_{t-1}) \leq \frac{1-\gamma^{t-1}}{1-\gamma} u$, $\left\|P_{t-1}^{1/2}\right\|_2 \leq \sqrt{u} \sqrt{\frac{1-\gamma^{t-1}}{1-\gamma}}$.

For the term $\left\|(I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})^{-1}\right\|_2$, we have $\left\|(I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})^{-1}\right\|_2 = 1/\sigma_n(I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})$. For the term $\sigma_n(I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})$, it is equal to $1 + \sigma_n(\gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})$, which is lower bounded by $1 + \gamma^{-1} \sigma_n(P_{t-1}^{-1/2}) \sigma_n(A_t^T A_t) \sigma_n(P_{t-1}^{-1/2})$.

Since $\sigma_n(P_{t-1}^{-1/2}) = \frac{1}{\sqrt{\sigma_1(P_{t-1})}}$ and $\sigma_1(P_{t-1}) \leq \frac{1-\gamma^{t-1}}{1-\gamma} u$, we have $\sigma_n(P_{t-1}^{-1/2}) \geq \frac{1}{\sqrt{u}} \sqrt{\frac{1-\gamma}{1-\gamma^{t-1}}}$. Together with $\sigma_n(A_t^T A_t) \geq l$, we have $\sigma_n(P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2}) \geq \frac{l}{u} \frac{1-\gamma}{1-\gamma^{t-1}}$, which results in $\left\|(I + \gamma^{-1} P_{t-1}^{-1/2} A_t^T A_t P_{t-1}^{-1/2})^{-1}\right\|_2 \leq \frac{1}{1 + \gamma^{-1} \frac{l}{u} \frac{1-\gamma}{1-\gamma^{t-1}}}$.

Combining the above three terms' inequalities, we have $\|\theta_{t+1} - \theta_t^*\| \leq \sqrt{\frac{u}{l}} \frac{u(\gamma - \gamma^t)}{u(\gamma - \gamma^t) + l(1-\gamma)} \|\theta_t - \theta_t^*\| \leq \sqrt{\frac{u}{l}} \frac{u\gamma}{u\gamma + l(1-\gamma)} \|\theta_t - \theta_t^*\|$.

□

Now we are ready to present the dynamic regret for the general recursive least-squares update:

Theorem 9. *Let θ_t^* be the solution to $f_t(\theta)$ in Eq.*(44) *and $\delta = u/l \geq 1$ be the condition number. When using the discounted recursive least-squares update in Eq.*(46) *with $\gamma < \frac{1}{\delta^{3/2} - \delta + 1}$ and $\rho = \sqrt{\frac{u}{l}} \frac{u\gamma}{u\gamma + l(1-\gamma)} < 1$, we can upper bound the dynamic regret:*

$$\mathcal{R}_d \leq \sqrt{u}(u/l + 1) D \frac{1}{1 - \rho} \left(\|\theta_1 - \theta_1^*\| + + \sum_{t=2}^{T} \|\theta_t^* - \theta_{t-1}^*\| \right)$$

Proof. The proof follows the similar steps in the proof of Theorem 2. First, we use the Mean Value Theorem to get $f_t(\theta_t) - f_t(\theta_t^*) = \nabla f_t(x)^T (\theta_t - \theta_t^*) \leq \|\nabla f_t(x)\| \|\theta_t - \theta_t^*\|$, where $x \in \{v | v = \beta\theta_t + (1-\beta)\theta_t^*, \beta \in [0,1]\}$. According to Corollary 5, $\|\nabla f_t(x)\| \leq \sqrt{u}(u/l + 1)D$. As a result, $\sum_{t=1}^{T} \left(f_t(\theta_t) - f_t(\theta_t^*)\right) \leq \sqrt{u}(u/l + 1) D \sum_{t=1}^{T} \|\theta_t - \theta_t^*\|$.

Now we need to upper bound the term $\sum_{t=1}^{T} \|\theta_t - \theta_t^*\|$. $\sum_{t=1}^{T} \|\theta_t - \theta_t^*\| = \|\theta_1 - \theta_1^*\| + \sum_{t=2}^{T} \|\theta_t - \theta_{t-1}^* + \theta_{t-1}^* - \theta_t^*\| \leq \|\theta_1 - \theta_1^*\| +$

$\sum_{t=1}^{T-1}\|\theta_{t+1}-\theta_t^*\| + \sum_{t=2}^{T}\left\|\theta_t^*-\theta_{t-1}^*\right\| \leq \|\theta_1-\theta_1^*\| + \sum_{t=1}^{T}\|\theta_{t+1}-\theta_t^*\| + \sum_{t=2}^{T}\left\|\theta_t^*-\theta_{t-1}^*\right\|$. According to Corollary 6, $\|\theta_{t+1}-\theta_t^*\| \leq \rho\|\theta_t-\theta_t^*\|$. $\sum_{t=1}^{T}\|\theta_t-\theta_t^*\| \leq \|\theta_1-\theta_1^*\| + \rho\sum_{t=1}^{T}\|\theta_t-\theta_t^*\| + \sum_{t=2}^{T}\left\|\theta_t^*-\theta_{t-1}^*\right\|$, which can be reformulated as $\sum_{t=1}^{T}\|\theta_t-\theta_t^*\| \leq \frac{1}{1-\rho}(\|\theta_1-\theta_1^*\| + + \sum_{t=2}^{T}\left\|\theta_t^*-\theta_{t-1}^*\right\|)$. Then $\mathcal{R}_d = \sum_{t=1}^{T}\left(f_t(\theta_t)-f_t(\theta_t^*)\right) \leq \sqrt{u}(u/l+1)D\frac{1}{1-\rho}(\|\theta_1-\theta_1^*\| + + \sum_{t=2}^{T}\left\|\theta_t^*-\theta_{t-1}^*\right\|)$.

□

In the above Theorem 9, the valid range of γ is in $(0, 1/(\delta^{3/2}-\delta+1))$. Let us now examine the requirement of γ to achieve the sub-linear static regret:

Theorem 10. *Let θ^* be the solution to $\min\sum_{t=1}^{T} f_t(\theta)$. When using the discounted recursive least-squares update in Eq.(46) with $1-\gamma = 1/T^\alpha, \alpha \in (0,1)$, we can upper bound the static regret:*

$$\mathcal{R}_s \leq O(T^{1-\alpha})$$

Proof. The proof follows the analysis of the online Newton method (Hazan, Agarwal, and Kale 2007). From the update in Eq.(46), we have $\theta_{t+1}-\theta^* = \theta_t-\theta^*-P_t^{-1}\nabla f_t(\theta_t)$ and $P_t(\theta_{t+1}-\theta^*) = P_t(\theta_t-\theta^*)-\nabla f_t(\theta_t)$. Multiplying the two equalities, we have $(\theta_{t+1}-\theta^*)^T P_t(\theta_{t+1}-\theta^*) = (\theta_t-\theta^*)^T P_t(\theta_t-\theta^*) - 2\nabla f_t(\theta_t)^T(\theta_t-\theta^*) + \nabla f_t(\theta_t)^T P_t^{-1}\nabla f_t(\theta_t)$.

After the reformulation, we have $\nabla f_t(\theta_t)^T(\theta_t-\theta^*) = \frac{1}{2}\nabla f_t(\theta_t)^T P_t^{-1}\nabla f_t(\theta_t) + \frac{1}{2}(\theta_t-\theta^*)^T P_t(\theta_t-\theta^*) - \frac{1}{2}(\theta_{t+1}-\theta^*)^T P_t(\theta_{t+1}-\theta^*) \leq \frac{1}{2}\nabla f_t(\theta_t)^T P_t^{-1}\nabla f_t(\theta_t) + \frac{1}{2}(\theta_t-\theta^*)^T P_t(\theta_t-\theta^*) - \frac{1}{2}(\theta_{t+1}-\theta^*)^T\gamma P_t(\theta_{t+1}-\theta^*)$.

Summing the above inequality from $t=1$ to T, we have: $\sum_{t=1}^{T}\nabla f_t(\theta_t)^T(\theta_t-\theta^*) \leq \sum_{t=1}^{T}\frac{1}{2}\nabla f_t(\theta_t)^T P_t^{-1}\nabla f_t(\theta_t) + \frac{1}{2}(\theta_1-\theta^*)^T P_1(\theta_1-\theta^*) + \sum_{t=2}^{T}\frac{1}{2}(\theta_t-\theta^*)^T(P_t-\gamma P_{t-1})(\theta_t-\theta^*) - \frac{1}{2}(\theta_{T+1}-\theta^*)^T\gamma P_T(\theta_{T+1}-\theta^*) \leq \sum_{t=1}^{T}\frac{1}{2}\nabla f_t(\theta_t)^T P_t^{-1}\nabla f_t(\theta_t) + \frac{1}{2}(\theta_1-\theta^*)^T(P_1-A_1^T A_1)(\theta_1-\theta^*) + \sum_{t=1}^{T}\frac{1}{2}(\theta_t-\theta^*)^T A_t^T A_t(\theta_t-\theta^*)$.

Since $P_1 = A_1^T A_1$ and $f_t(\theta_t)-f_t(\theta^*) = \nabla f_t(\theta_t)^T(\theta_t-\theta^*) - \frac{1}{2}(\theta_t-\theta^*)^T A_t^T A_t(\theta_t-\theta^*)$, we reformulate the above inequality as:

$$\begin{aligned}
&\sum_{t=1}^{T}\left(f_t(\theta_t)-f_t(\theta^*)\right)\\
&=\sum_{t=1}^{T}\left(\nabla f_t(\theta_t)^T(\theta_t-\theta^*) - \tfrac{1}{2}(\theta_t-\theta^*)^T A_t^T A_t(\theta_t-\theta^*)\right)\\
&\leq \sum_{t=1}^{T}\tfrac{1}{2}\nabla f_t(\theta_t)^T P_t^{-1}\nabla f_t(\theta_t)\\
&=\sum_{t=1}^{T}\tfrac{1}{2}(A_t\theta_t-y_t)^T A_t P_t^{-1} A_t^T(A_t\theta_t-y_t)\\
&\leq \sum_{t=1}^{T}\tfrac{1}{2}\sigma_1(P_t^{-1/2}A_t^T A_t P_t^{-1/2})\left\|A_t\theta_t-y_t\right\|^2
\end{aligned} \tag{50}$$

Since $\sigma_1(P_t^{-1/2}A_t^T A_t P_t^{-1/2}) \leq \sigma_1(P_t^{-1})\sigma_1(A_t^T A_t) = \frac{1}{\sigma_n(P_t)}\sigma_1(A_t^T A_t)$. From the proof of Corollary 6 we know that $\sigma_n(P_t) \geq \frac{1-\gamma^t}{1-\gamma}l$ and $\sigma_1(A_t^T A_t) \leq u$. Then $\sigma_1(P_t^{-1/2}A_t^T A_t P_t^{-1/2}) \leq \frac{u}{l}\frac{1-\gamma}{1-\gamma^t}$. As a result, we have

$$\begin{aligned}
\sum_{t=1}^{T}\left(f_t(\theta_t)-f_t(\theta^*)\right) &\leq \sum_{t=1}^{T}\tfrac{1}{2}\tfrac{u}{l}\tfrac{1-\gamma}{1-\gamma^t}\left\|A_t\theta_t-y_t\right\|^2\\
&\leq \sum_{t=1}^{T}\tfrac{1}{2}\tfrac{u}{l}\tfrac{1-\gamma}{1-\gamma^t}(u/l+1)^2D^2\\
&\leq O(T^{1-\alpha})
\end{aligned} \tag{51}$$

where the second inequality is due to Lemma 6 and the third inequality is due to the fact that $\sum_{t=1}^{T} 1/(1-\gamma^t) \leq O(T)$ as shown in the proof of Theorem 3.

□

Recall that the valid range of γ in Theorem 9 is $(0, 1/(\delta^{3/2}-\delta+1))$, while having sub-linear static regret requires $\gamma = \frac{T^\alpha-1}{T^\alpha}$. Although for some specific T, there might be some intersection. In general, these two are contradictory. However, as discussed in the main body of the paper, more flexible trade-offs between static and dynamic regret can be achieved via the gradient descent rule.