Online Forgetting Process for Linear Regression Models

Yuantong Li  
Department of Statistics  
Purdue University

Chi-Hua Wang  
Department of Statistics  
Purdue University

Guang Cheng  
Department of Statistics  
Purdue University

Abstract

Motivated by the EU’s "Right To Be Forgotten" regulation, we initiate a study of statistical data deletion problems where users’ data are accessible only for a limited period of time. This setting is formulated as an online supervised learning task with \textit{constant memory limit}. We propose a deletion-aware algorithm FIFD-OLS for the low dimensional case, and witness a catastrophic rank swinging phenomenon due to the data deletion operation, which leads to statistical inefficiency. As a remedy, we propose the FIFD-Adaptive Ridge algorithm with a novel online regularization scheme, that effectively offsets the uncertainty from deletion. In theory, we provide the cumulative regret upper bound for both online forgetting algorithms. In the experiment, we showed FIFD-Adaptive Ridge outperforms the ridge regression algorithm with fixed regularization level, and hopefully sheds some light on more complex statistical models.

1 Introduction

Today many internet companies and organizations are facing the situation that \textit{certain} individual’s data can no longer be used to train their models by legal requirements. Such circumstance forces companies and organizations to delete their database on the demand of users’ willing to be forgotten. On the ground of the ‘Right to be Forgotten’, regularization is established in many countries and states’ laws, including the EU’s General Data Protection Regulation (GDPR) \cite{GDPR2016}, and the recent California Consumer Privacy Act Right (CCPA) \cite{CCPA2018}, which also stipulates users to require companies and organizations such as Google, Facebook, Twitter, etc to forget and delete these personal data to protect their privacy. Besides, users also have the right to request the platform to delete his/her obsolete data at any time or only to authorize the platform to hold his/her personal information such as photos, emails, etc, only for a \textit{limited} period. Unfortunately, given these data are typically incrementally collected online, it is a disaster for the machine learning model to forget these data in chronological order. Such a challenge opens the needs to design and analyze \textit{deletion-aware} online machine learning method.

In this paper, we propose and investigate a class of online learning procedure, termed \textit{online forgetting process}, to adapt users’ requests to delete their data before a specific time bar. To proceed with the discussion, we consider a special deletion practice, termed \textit{First In First Delete} (FIFD), to address the scenario that the users only authorize their data for a limited period. (See Figure 1 for an illustration of the online forgetting process with constant memory limit $s$.) In FIFD, the agent is required to delete the oldest data as soon as receiving the latest data, to meet a \textit{constant memory limit}. The FIFD deletion practice is inspired by the situation that, the system may only use data from the past three months to train their machine learning model to offer service for new customers \cite{Google2020}. The proposed online forgetting process is an \textit{online} extension of recent works that consider offline data deletion \cite{Izzo2020, Ginart2019} or detect data been forgotten or not \cite{Liu2020}.

In such a ‘machine forgetting’ setting, we aim to determine its difference with standard statistical machine learning methods via an online regression framework. To accommodate such limited authority, we provide solutions on designing \textit{deletion-aware} online linear regression algorithms, and discuss the harm due to the "constant memory limit" setting. Such a setting is challenging for a general online statistical learning task.
since the "sample size" never grows to infinity but stays a constant size along the whole learning process. As an evil consequence, statistical efficiency is never improving as the time step grows due to the constant memory limit.

Our Contribution. We first investigate the online forgetting process in the ordinary linear regression. We find a new phenomenon defined as Rank Swinging Phenomenon, which exists in the online forgetting process. If the deleted data can be fully represented by the data memory, then it will not introduce any regret. Otherwise, it will introduce extra regret to make this online forgetting process task not online learnable. The rank swinging phenomenon plays such a role that it indirectly represents the dissimilarity between the deletion data and the new data to affect the instantaneous regret. Besides, if the gram matrix does not have full rank, it will cause the adaptive constant $\zeta$ to be unstable and then the confidence ellipsoid will become wider. Taking both of these effects into consideration, the order of the FIFD-OLS’s regret will become linear in time horizon $T$.

The rank swinging phenomenon affects the regret scale and destabilizes the FIFD-OLS algorithm. Thus, to remedy this problem, we propose the FIFD-Adaptive Ridge to offset this phenomenon because when we add the regularization parameter, the gram matrix will have full rank. Different from using the fixed regularization parameter in the standard online learning model, we use the martingale technique to adaptively select the regularization parameter over time.

Notation. Throughout this paper, we denote $[T]$ as the set $\{1, 2, \ldots, T\}$. $|S|$ denotes the number of elements for any collection $S$. We use $x_p$ to denote the $p$-norm of a vector $x \in \mathbb{R}^d$ and $x_\infty = \sup_i |x_i|$. For any vector $v \in \mathbb{R}^d$, notation $\mathcal{P}(v) \equiv \{i | v_i > 0\}$ denotes the indexes of positive coordinate of $v$ and $\mathcal{N}(v) \equiv \{i | v_i < 0\}$ denotes the indexes of negative coordinate of $v$. $\mathcal{P}_{\min}(v) \equiv \min\{v_i | i \in \mathcal{P}(v)\}$ denotes the minimum value of $v_i$, where $i$ is in the indexes of positive coordinate and $\mathcal{N}_{\max}(v) \equiv \max\{v_i | i \in \mathcal{N}(v)\}$ denotes the maximum value of $v_i$, where $i$ is in the indexes of negative coordinate.

For a positive semi-definite matrix $\Phi \in \mathbb{R}_{\geq 0}^{d \times d}$, $\lambda_{\min}(\Phi)$ denotes the minimum eigenvalue of $\Phi$. We denote $\Phi^\top$ as the generalized inverse of $\Phi$ if it satisfies the condition $\Phi = \Phi \Phi^\top \Phi$. The weighted 2-norm of vector $x \in \mathbb{R}^d$ with respect to positive definite matrix $\Phi$ is defined by $x_\Phi = \sqrt{x^\top \Phi x}$. The inner product is denoted by $\langle \cdot, \cdot \rangle$ and the weighted inner-product is denoted by $x^\top \Phi y = \langle x, y \rangle\Phi$. For any sequence $\{x_t\}_{t=0}^\infty$, we denote matrix $x_{[a,b]} = \{x_a, x_{a+1}, \ldots, x_b\}$ and $x_{[a,b]}_\infty = \sup_{i,j} |x_{[a,b]}(i,j)|$. For any matrix $x_{[a,b]}$, notation $\Phi_{[a,b]} = \sum_{i=a}^b x_i x_i^\top$ represents a gram matrix with constant memory limit $s = b - a + 1$ and notation $\Phi_{\lambda_{[a,b]}} = \sum_{i=a}^b x_i x_i^\top + \lambda I_{d \times d}$ represents a gram matrix with ridge hyperparameter $\lambda$.

2 Statistical Data Deletion Problem

At each time step $t \in [T]$, where $T$ is a finite time horizon, the learner receives a context-response pair $z_t = (x_t, y_t)$, where $x_t \in \mathbb{R}^d$ is a $d$-dimensional context and $y_t \in \mathbb{R}$ is the response. The observed sequence of context $\{x_t\}_{t \geq 1}$ are drawn i.i.d from a distribution of $\mathcal{P}_x$ with a bounded support $\mathcal{X} \subset \mathbb{R}^d$ and $x_{t-2} \leq L$. Let $D_{[t-s,t-1]} = \{z_i\}_{i=t-s}^{t-1}$ denote the data collected at $[t-s, t-1]$ following the FIFD scheme.

We assume that for all $t \in [T]$, the response $y_t$ is a linear combination of context $x_t$; formally,

$$y_t = \langle x_t, \theta_\star \rangle + \epsilon_t, \quad (1)$$

where $\theta_\star \in \mathbb{R}^d$ is the target parameter that summarizes the relation between the context $x_t$ and response $y_t$. The noise $\epsilon_t$’s are drawn independently from a $\sigma$-subgaussian distribution. That is, for every $\alpha \in \mathbb{R}$, it is satisfied that $\mathbb{E}[\exp(\alpha \epsilon_t)] \leq \exp(\alpha^2 \sigma^2 / 2)$. Under the proposed FIFD scheme with a constant memory limit $s$, the algorithm $\mathcal{A}$ at time $t$ only can keep the information of historical data from time step $t-s$ to time step $t-1$ and then to make the prediction,
After prediction, the algorithm \( A \) first receives a context \( x_t \) at time step \( t \), and make a prediction based only on the information in previous \( s \) time steps \( D_{[t-s:t-1]} \) (\(|D_{[t-s:t-1]}|=s\)) and hence the agent forgets the information up to the time step \( t-s-1 \). The predicted value \( \hat{y}_t \) computed by the algorithm \( A \) based on information \( D_{[t-s:t-1]} \) and current the context \( x_t \), which is denoted by
\[
\hat{y}_t = A(x_t, D_{[t-s:t-1]}). \tag{2}
\]
After prediction, the algorithm \( A \) receives the true response \( y_t \) and suffers a pseudo regret, \( r(\hat{y}_t, y_t) = (x_t, \theta_s - \hat{y}_t)^2 \). We define the cumulative (pseudo-)regret of the algorithm \( A \) up to time horizon \( T \) as
\[
R_{T,s}(A) = \sum_{t=s+1}^{T} ((x_t, \theta_s) - \hat{y}_t)^2. \tag{3}
\]

Our theoretical goal is to explore the relationship between constant memory limit \( s \) and the cumulative regret \( R_{T,s}(A) \). We proposed two algorithms FIFD-OLS and FIFD-Adaptive Ridge and studied their cumulative regret, respectively. In particular, we discuss the effect of constant memory limit \( s \), dimension \( d \), subgaussian parameter \( \sigma \), and confidence level \( 1-\delta \) on the obtained cumulative regret \( R_{T,s}(A) \) for these two algorithms. For example, what’s the effect of constant memory \( s \) on the order of the regret \( R_{T,s}(A) \) if other parameters keep constant. In other words, how many data do we need to keep in order to achieve the satisfied performance under the FIFD scheme. Besides, is there any amazing or unexpected phenomenon both in the experiment and theory occurred when the agent used the ordinary least square method under the FIFD scheme and how to improve it?

Remark. The FIFD scheme can also be generalized to the standard online learning paradigm when we add two and delete one data or add more and delete one data. These settings will automatically transfer to the standard online learning paradigm when we add and delete data. These settings will automatically transfer to the standard online learning paradigm when we add and delete data.

3 FIFD OLS

In this section, we present the FIFD - ordinary least square regression (FIFD-OLS) algorithm under “First In First Delete” scheme and the corresponding confidence ellipsoid for the FIFD-OLS estimator and regret analysis. Besides, the rank swinging phenomenon will be discussed in section 3.4.

3.1 FIFD-OLS Algorithm

The FIFD-OLS algorithm uses the least square estimator based on the constant data memory from time window \([t-s,t-1]\), defined as \( \hat{\theta}_{[t-s,t-1]} = \Phi_{[t-s,t-1]}^{-1} \left[ \sum_{i=t-s}^{t-1} y_i x_i \right] \). Then an incremental update formula for \( \hat{\theta} \) from time window \([t-s,t-1]\) to \([t-s+1,t]\) is shown as follows.

**OLS incremental update:** At time step \( t \), the estimator \( \hat{\theta}_{[t-s+1,t]} \) is updated by previous estimator \( \hat{\theta}_{[t-s,t-1]} \) and new data \( z_t \),
\[
\hat{\theta}_{[t-s+1,t]} = f(\Phi_{[t-s,t-1]}^{-1}, x_{t-s}, x_t) \cdot g(\hat{\theta}_{[t-s,t-1]}, \Phi_{[t-s,t-1]}^{-1}, z_{t-s}, z_t), \tag{4}
\]
where \( f(\Phi_{[t-s,t-1]}^{-1}, x_{t-s}, x_t) \) is defined as
\[
\Gamma(\Phi_{[t-s,t-1]}^{-1}) = (x_{t-s}^\top \Gamma(\Phi_{[t-s,t-1]}^{-1}) x_{t-s} + 1)^{-1} \Phi_{[t-s,t-1]}^{-1} x_{t-s}^\top \Phi_{[t-s,t-1]}^{-1} \tag{5}
\]
and \( g(\hat{\theta}_{[t-s,t-1]}, \Phi_{[t-s,t-1]}^{-1}, z_{t-s}, z_t) \) is defined as \( \Phi_{[t-s,t-1]}^{-1} \hat{\theta}_{[t-s,t-1]} + y_t x_t - y_{t-s} x_{t-s} \). The detailed online incremental update is in Appendix F.

The algorithm has two steps. First, \( f \)-step is to update the inverse of gram matrix from \( \Phi_{[t-s,t-1]}^{-1} \) to \( \Phi_{[t-s+1,t]}^{-1} \) and the Woodbury formula [Woodbury (1950)] is applied to reduce the time complexity to \( \mathcal{O}(s^d) \) compared with directly computing the inverse of gram matrix. Secondly, \( g \)-step is a simple algebra and uses the definition of adding and forgetting data to update the least square estimator. The detailed update procedure of FIFD-OLS is displayed in Algorithm 1.

3.2 Confidence Ellipsoid for FIFD-OLS

Our first contribution is to obtain a confidence ellipsoid for the OLS method based on the sample set collected under the FIFD scheme, showed in Lemma 1.

**Lemma 1. (FIFD-OLS Confidence Ellipsoid)** For any \( \delta > 0 \), if the event \( \lambda_{\min}(\Phi_{[t-s,t-1]}/s) > \phi_{[t-s,t-1]}^2 > 0 \) holds, with probability at least \( 1-\delta \), for all \( t \geq s+1 \), \( \theta_s \) lies in the set
\[
C_{[t-s,t-1]} = \left\{ \theta \in \mathbb{R}^d : \Theta_{[t-s,t-1]} - \theta \in \mathbb{R}^d \right\} \tag{6}
\]
where \( q_{[t-s,t-1]} = \sum_{i=t-s}^{t-1} y_i x_i / \phi_{[t-s,t-1]}^2 \) is the adaptive constant and we denote \( \beta_{[t-s,t-1]} \) as the RHS bound.
Algorithm 1: FIFD-OLS

Given parameters: $s, T, \delta$
for $t \in \{s+1, \ldots, T\}$ do
  Observe $x_t$
  $\Phi_{t-s,t-1} = x_{t-s,t-1}^T x_{t-s,t-1}$
  if $t = s+1$ then
    $\hat{\theta}_{[1,s]} = \Phi_{[1,s]}^{-1} x_{[1,s]}^T y_{[1,s]}
  else
    $\hat{\theta}_{[t-s,t-1]} = f(\Phi_{[t-s-1,t-2]}^{-1}, x_{t-s-1}, x_t),
    g(\hat{\theta}_{[t-s-1,t-2]}, \Phi_{[t-s-1,t-2]}^{-1}, z_{t-s-1}, z_{t-1})$
  end
Predict $\hat{y}_t = \hat{\theta}_{[t-s,t-1]} x_t$ and observe $y_t$
Compute loss $r_t = |\hat{y}_t - \langle \theta^*, x_t \rangle|$
Delete data $z_{t-s} = (x_{t-s}, y_{t-s})$
end

Proof. The key step to get the confidence ellipsoid is to use the martingale noise technique presented in Bastani and Bayati [2020] to obtain an adaptive bound. The detailed proof can be obtained in Appendix A. Besides, the confidence ellipsoid in [6] requires the information of minimum eigenvalue of gram matrix $\Phi_{[t-s,t-1]}$ and infinity norm of our observations with constant data memory $s$. Thus, it is an adaptive confidence ellipsoid following the change of data.

3.3 FIFD OLS Regret

The following theorem provides the regret upper bounds for the FIFD-OLS algorithm.

Theorem 1. (Regret Upper Bound of The FIFD-OLS Algorithm). Assume that for all $t \in [s+1, T-s]$ and $X_t$ is i.i.d random variables with distribution $\mathcal{P}_X$. With probability at least $1 - \delta = [0, 1]$, for all $T > s, s > d$, we have an upper bound on the cumulative regret at time $T$:

\[
R_{T,s}(\mathcal{A}_{OLS}) \leq 2\sigma_\zeta \sqrt{(d/s) \log(2d/\delta)}(T-s)(d \log(sL^2/d) + (T-s)),
\]

where the adaptive constant $\zeta = \max_{s+1 \leq t \leq T} q_{[t-s,t-1]}$. \hspace{1cm} (7)

Proof. We provide a roadmap for the proof of Theorem 1. The proof is motivated by Abbasi-Yadkori et al. [2011], Bastani and Bayati [2020]. We first prove Lemma 2 to obtain a confidence ellipsoid holding for the FIFD-OLS estimator. Then we use the confidence ellipsoid, to sum up the regret over time in Lemma 3 and find a key term called FRT (defined below equation 4), which affects the derivation of regret upper bound a lot. The detailed proof of this theorem can be found in the appendix C.

Remarks. We develop the FIFD-OLS algorithm and prove an upper bound of the cumulative regret in order of $O(\sigma_\zeta (d/s) (\log 2d/\delta)^2 T)$. The agent using this algorithm cannot improve the performance of this algorithm because it has a constant memory limit $s$ to update the estimated parameters. What’s more, the adaptive constant $q_{[t-s,t-1]}$ is unstable caused by the forgetting process. The main factor causing the oscillation of gram matrix $\Phi_{[t-s,t-1]}$ is that the minimum eigenvalue is close to zero. In other words, the gram matrix is full rank at some time. Therefore, the adaptive constant $q_{[t-s,t-1]}$ will go to infinity at such time points. So the generalization’s ability of FIFD-OLS is poor.

Following the derivation of the regret upper bound of FIFD-OLS in Theorem 1, we find an interesting phenomenon, which we called Rank Swinging Phenomenon (defined below Definition 1). This phenomenon will unstabilize the FIFD-OLS algorithm and introduce some extreme bad events, which results in a larger value of the regret upper bound of the FIFD-OLS algorithm.

3.4 Rank Swinging Phenomenon

Below we give the definition of the Rank Swinging Phenomenon.

Definition 1. (Rank Swinging Phenomenon) At time $t$, when we delete data $x_{t-s}$ and add data $x_t$, the gram matrix switching from $\Phi_{[t-s,t-1]}$ to $\Phi_{[t-s+1,t]}$ will cause its rank increasing or decreasing by 1 or 0 if $\text{Rank}(\Phi_{[t-s,t-1]}) \leq d$,

\[
\text{Rank}(\Phi_{[t-s+1,t]}) = \begin{cases}
\text{Rank}(\Phi_{[t-s,t-1]}) + 1, & \text{Case 1} \\
\text{Rank}(\Phi_{[t-s,t-1]}), & \text{Case 2} \\
\text{Rank}(\Phi_{[t-s,t-1]}) - 1, & \text{Case 4}
\end{cases}
\]

where four examples are illustrated in Appendix A. The real example can be found in Figure D.

The rank swinging phenomenon results in unstable regret, which is measured by the term called ‘Forgetting Regret Term’ (FRT) caused by deletion at time $t$.

Definition 2. (Forgetting Regret Term)

\[
FRT_{[t-s,t-1]} = x_{t-s}^2 \Phi_{[t-s,t-1]}^{-1}(x_{t-s}^T \Phi_{[t-s,t-1]}^{-1} \sin^2(\theta, \Phi_{[t-s,t-1]}^{-1}) + 1), \hspace{1cm} (9)
\]

where $FRT_{[t-s,t-1]} \in [0, 2]$.

Let’s use $\sin^2(\theta, \Phi_{[t-s,t-1]}^{-1})$ denote the dissimilarity between $x_{t-s}$ and $x_t$ under $\Phi_{[t-s,t-1]}^{-1}$ as follows,

\[
\sin^2(\theta, \Phi_{[t-s,t-1]}^{-1}) = \sin^2(\theta, \Phi_{[t-s,t-1]}^{-1}) + 1 - \cos^2(\theta, \Phi_{[t-s,t-1]}^{-1})
\]
Lemma 2. The cumulative regret of FIFD-OLS is partially determined by FRT at each time step, and the cumulative representation term is
\[
\sum_{t=s+1}^{T} x_t \Phi^{-1}_{[t-s,t-1]} \leq 2\eta_{OLS} + \sum_{t=s+1}^{T} FRT_{[t-s,t-1]} \tag{10}
\]
where \(\eta_{OLS} = \log(\det(\Phi_{[T-s,T]}))\) is a constant based on data time window \([T-s,T]\).

Proof. Let’s first denote \(r_t\) as the instantaneous regret at time \(t\) and decompose the instantaneous regret, \(r_t = (\hat{\theta}_{[t-s,t-1]} - \theta_{[t-s,t-1]} x_t) \leq \sqrt{n_{[t-s,t-1]} (\delta)} x_t \Phi^{-1}_{[t-s,t-1]}\), where the inequality is from Lemma \(\Box\). Thus, with probability at least \(1-\delta\), for all \(T > s\),
\[
R_{t,s}(\mathcal{A}_{OLS}) \leq \sqrt{(T-s) \sum_{t=s+1}^{T} r_t^2} \leq \sqrt{(T-s) \max_{s+1 \leq t \leq T} n_{[t-s,t-1]} (\delta) \sum_{t=s+1}^{T} x_t^2 \Phi^{-1}_{[t-s,t-1]}}.
\]

4 FIFD-Adaptive Ridge

To remedy the rank switching phenomenon, we take advantage of the ridge regression method to avoid this happening in the online forgetting process. In this section, we present the FIFD - adaptive ridge regression (FIFD-Adaptive Ridge) algorithm under “First In First Delete” scheme and the corresponding confidence ellipsoid for the FIFD-Adaptive Ridge estimator and regret analysis.

4.1 FIFD-Adaptive Ridge Algorithm

The FIFD-Adaptive Ridge algorithm use the ridge estimator \(\hat{\Phi}_{\lambda,[t-s,t-1]}\) to predict the response. The definition of it for time window \([t-s,t-1]\) is showed as follows.

Adaptive Ridge update:
\[
\hat{\Phi}_{\lambda,[t-s,t-1]} = \sum_{i=t-s}^{t-1} y_i x_i.
\]

Then we display the adaptive choice of hyperparameter \(\lambda\) for time window \([t-s,t-1]\).

Lemma 3. (Adaptive Ridge Parameter \(\lambda_{\delta,[t-s,t-1]}\)) If the event \(\lambda_{\min}(\Phi_{\lambda,[t-s,t-1]}/s) > \phi_{\lambda,[t-s,t-1]}^2\) holds, with probability \(1-\delta\), for any
\[
\chi(\delta) > \sigma x_{[a,b]} \infty \sqrt{(2d/s) \log(2d/\delta)} / \phi_{\lambda,[a,b]}^2,
\]
we have a control of $L_2$ estimation error that
$$
Pr \left[ \hat{\theta}_{s,t-s-1} - \theta_{s,t-1} \leq (\delta) \right] \geq 1 - \delta,
$$
and to satisfy this condition, we select the adaptive $\lambda_{[s,t-s-1]}$ as follows for the limited time window $[t-s, t-1]$,

$$
\lambda_{[t-s,t-1]} \leq \sigma x_{[t-1,s-1]}^{\infty} \sqrt{2s \log(2d/\delta)} / \theta_{s}^{\infty}.
$$

The detailed derivation can be found in Appendix E.

The detailed update procedure of FIFD-Adaptive Ridge is displayed in Algorithm 2.

**Algorithm 2: FIFD-Adaptive Ridge**

Given parameters: $s$, $T$, $\delta$

for $t \in \{s+1, ..., T\}$ do

- Observe $x_t$

  - $x_{[t-1,s-1]}^{\infty} = \max_{1 \leq i \leq 1, 1 \leq j \leq d} x_{i,j}$
  - $\hat{\sigma}_{[t-s,t-1]} = SD(y_{[t-1,s-1]}^{\infty})$
  - $\lambda_{[t-s,t-1]} = \sqrt{2 \sigma_{[t-s,t-1]}^{2} x_{[t-1,s-1]}^{\infty} \sqrt{\log(1/\delta)}}$
  - $\Phi_{\lambda_{[t-s,t-1]}^{-1}} = x_{[t-1,s-1]}^{\infty} (x_{[t-1,s-1]}^{\infty})^{T}$

- Predict $\hat{y}_t = (\hat{\theta}_{[t-s,t-1]}, x_t)$ and observe $y_t$

- Compute loss $r_t = |\hat{y}_t - (\theta^*, x_t)|$

- Delete data $z_{[t-s]} = (x_{t-s}, y_{t-s})$

end

4.2 Confidence Ellipsoid for FIFD-Adaptive Ridge

Before we moving to the confidence ellipsoid of the adaptive ridge estimator, we define some notions about the true parameter $\theta^*$, where $P_{\min}(\theta^*)$ represents the weakest positive signal and $N_{\max}(\theta^*)$ serves as the weakest negative signal. To make the adaptive ridge confidence ellipsoid more simplified, without loss of generality, we assume the following assumption.

**Assumption 1.** (Weakest Positive to Strongest Signal Ratio) We assume positive coordinate of $\theta^*$ dominates the bad events happening,

$$
WPSSR = \frac{P_{\min}(\theta^*)}{\theta^*} \leq \sqrt{C_3 + C_4 + s^2 \log \frac{2d}{\delta}},
$$

where $C_3 = \log \frac{6d}{\delta} \log \frac{2d}{\delta}$. The WPSSR is monotone increasing in $s$ and $P(\theta^*)$, and is monotone decreasing in $d$. In most cases, the LHS is greater than one, such as $s = 100$, $d = 110$, $\delta = 0.05$, $P(\theta^*) = 30$, then WPS Ratio needs to be less than 1.02, which is satisfied this assumption.

If Assumption 1 holds, a high probability confidence ellipsoid can be obtained for FIFD-Adaptive ridge.

**Lemma 4.** (FIFD-Adaptive Ridge Confidence Ellipsoid) For any $\delta \in [0,1]$, with probability at least $1 - \delta$, for all $t \geq s + 1$, with Assumption 1 and the event $\lambda_{\min}(\Phi_{\lambda_{[t-s,t-1]}}) \geq \delta^2 \lambda_{\max}(\Phi_{\lambda_{[t-s,t-1]}}) > 0$ holds, then $\theta^*$ lies in the set

$$
C_{\lambda_{[t-s,t-1]}} = \left\{ \theta \in \mathbb{R}^d : |\hat{\theta}_{[t-s,t-1]} - \theta_{[t-s,t-1]}| \leq \frac{\alpha d}{\sqrt{T_s, \theta_s}} \right\},
$$

where $\alpha_{[t-s,t-1]} = x_{[t-s,t-1]}^{\infty} / \Phi_{\lambda_{[t-s,t-1]}^{-1}}$, $\eta = \log^2(1/\delta)$, and $\nu = \theta^*/P_{\min}(\theta^*)$.

**Proof.** Key steps. The same technique used in Lemma 4 is applied here to obtain an adaptive confidence ellipsoid for ridge estimator. Here we assume Assumption 1 to make the confidence ellipsoid more simplified. The detailed proof can be obtained in the Appendix E.

**Remark.** The order of constant memory limit $s$ in the confidence ellipsoid for FIFD-OLS and FIFD-Adaptive ridge’s confidence ellipsoids are $O(\sqrt{s})$. The adaptive ridge confidence ellipsoid requires the information of minimum eigenvalue of gram matrix $\Phi_{\lambda_{[t-s,t-1]}^{-1}}$ with hyperparameter $\lambda$ and infinity norm of data $x_{[t-1,s-1]}^{\infty}$, which is similar as FIFD-OLS confidence ellipsoid required. The benefits of introduction of $\lambda$ avoids the singularity of gram matrix $\Phi$, which will stabilize the algorithm FIFD-Adaptive ridge and make the confidence ellipsoid narrow in most times. Besides, it is an adaptive confidence ellipsoid.

4.3 FIFD Adaptive Ridge Regret

In the following theorem, we provide the regret upper bound for the FIFD-Adaptive Ridge method.

**Theorem 2.** (Regret Upper Bound of the FIFD-Adaptive Ridge algorithm) With Assumption 1 and with probability at least $1 - \delta$, the cumulative regret satisfies:

$$
R_{T,s}(A_{\text{Ridge}}) \leq \sigma \kappa \nu \sqrt{\lambda_{s} \log(1/\delta)} (T-s)(1+\eta_{\text{Ridge}}),
$$

where $\lambda_{s} = \max_{s+1 \leq T} \frac{\tilde{\sigma}^{2}_{[t-1,s-1]}^{\infty}}{\Phi_{\lambda_{[t-s,t-1]}^{-1}}}$ is the maximum adaptive constant, $\eta_{\text{Ridge}} = d \log(1/\delta) + \lambda_{[t-s,t-1]}$, $\kappa = \log\left(\frac{d}{\delta}\right)$, and $C_2(\phi)$ is a constant related to the last data memory, $C_2(\phi) = \prod_{t=s+1}^{T} (1 + \tilde{\sigma}_{[t,s]}^{2})$, $\lambda_{[t,s]}^{\Delta} = \lambda_{[t-1,s]}^{\Delta}$, and $\lambda_{[t,s]}^{\Delta} = \lambda_{[t-1,s]}^{\Delta}$ represents the change of $\lambda$ over time steps.

**Proof.** The key step is to use the same technical lemma used in the OLS setting to show a confidence ellipsoid
holds for the adaptive ridge estimator. Lemma 2 is used to compute FRT and then sums up FTRs to get the cumulative regret upper bound. The detailed proof of this theorem can be found in Appendix C.

Remarks. (Relationship between regret upper bound and parameters) We develop the FIFD-Adaptive Ridge algorithm and then provide a regret upper bound in order \(O(\sigma \epsilon \Delta \sqrt{d/s} T)\). This order represents the relationship between the regret, noise \(\sigma\), dimension \(d\), constant memory limit \(s\), signal level \(\nu\), and confidence level \(1 - \delta\). Since the agent always keeps the constant memory limit \(s\) data to update the estimated parameters, thus the agent can’t improve the performance of the algorithm. Therefore, the regret is \(O(T)\) with respect to the decision number. Besides, we find that using the ridge estimator can offset the rank swinging phenomenon since Ridge’s gram matrix is always full rank.

5 Simulation Experiments

We compare the FIFD-Adaptive Ridge method where \(\lambda\) is chosen under Lemma 2 with the baselines with pre-tuned \(\lambda\). For all experiments unless otherwise specified, we choose \(T = 3000, \delta = 0.05, d = 100, L = 1\) for all simulation settings. All results are averaged over 100 runs.

Simulation settings. The setting for constant memory limit \(s\) is 20, 40, 60, 80 and subgaussian parameter \(\sigma\) is 1, 2, 3. Context \(\{x_t\}_{t \in [T]}\) is generated from \(N(0_d, I_{d \times d})\), and then we normalize it. The response \(\{y_t\}_{t \in [T]}\) is generated by \(y_t = \langle x_t, \theta_* \rangle + \epsilon_t\), where \(\theta_* \sim N(0_d, I_{d \times d})\) and then normalize it with \(\theta_{s*} = 1\).

We set \(\epsilon_t \overset{i.i.d.}{\sim} \sigma\)-subgaussian. The adaptive ridge’s noise \(\sigma\) in the simulation is estimated by \(\sigma_{[t-s, t-1]} = sd(y_{[t-s, t-1]}), t \in [s + 1, T]\).

Hyperparameter settings for competing methods. The hyperparameter \(\lambda\) we select for \(\sigma = 1\) setting is \(\{1, 10, 100\}\). Since the relationship of hyperparameter \(\lambda\) and noise level \(\sigma\), we know they are in linear order. Thus, for the \(\sigma = 2\) setting, we set \(\lambda = \{2, 20, 200\}\) and for the \(\sigma = 3\) setting, we set \(\lambda = \{3, 30, 300\}\). The adaptive ridge hyperparameter is automatically calculated according to Lemma 2 and we assume \(\theta_{s\infty} = 1\) since \(\theta_{s*} = 1\).

Results. In figure 3, we present the simulation result of relationship between the cumulative regret \(R_{T,s}(A_{ridge})\), noise level \(\sigma\), hyperparameter \(\lambda\), and constant memory limit \(s\). We find all of the cumulative regret are linear in time horizon \(T\) just with different constant levels.

From three rows, as the memory limit \(s\) increases, we find that the cumulative regret of the adaptive ridge (blue line) decreases. Since we know that as the sample size increases, the regret will decrease in theory. For each column, when we fix the memory limit \(s\), as the noise level \(\sigma\) increases, the regret of FIFD-Adaptive Ridge and Fixed Ridge increase with different hyperparameters.

Overall, we find that the FIFD-Adaptive Ridge method is more robust to the change of noise level \(\sigma\) when we view figure 3 by row. Fixed Ridge methods with different hyperparameters such as green line and red line are sensitive to the change of noise \(\sigma\) and memory limit \(s\) because Fixed Ridge doesn’t have any prior to adaptive the data. Although it performs well in the top left, it doesn’t work well in other settings. The yellow line has relative comparable performance compared with the Adaptive Ridge method since its hyperparameter \(\lambda\) is determined by the knowledge from Adaptive Ridge. The FIFD-Adaptive Ridge is always the best (2nd row and 3rd row) or close to the best (1st row) choice of \(\lambda\) among all of these settings, which means that the Adaptive Ridge method is robust to the large noise.

Besides, the Adaptive Ridge method can save computational cost compared with the fixed Ridge method, which needs cross-validation to select the optimal \(\lambda\). In addition, more results about the choice of hyperpa-
parameter $\lambda$ and $L_2$ estimation error of the estimator can be found in Appendix G.

**Heavy Tail Noise Distribution.** Moreover, we also test the robustness of these methods concerning different noise. Here we assume $\varepsilon_t \sim t_{df}, \forall t \in [T]$. The degree of freedom $df$ is set to be 5, 10, 15. As we know when $df$ is greater than 30, it behaves like the normal distribution. When $df$ is small, $t$-distribution has much heavier tails. Thus, in figure 4, we find that when $df$ increases, the cumulative regret is decreasing since the error is more like the normal distribution, which can be well captured by our algorithm. However, fixed Ridge methods are not robust to the change of noise distribution, especially green line and red line.

**Results for (+k, 1) addition-deletion operation.** Here we test the pattern when we add more than one data point at each time step, such as $k = 2, 3, 4$, and delete one data, denoted as (+$k$, -1) addition-deletion pattern. In figure 5, the 1st, 2nd, 3rd column are cases that delete one data after receiving 2, 3, 4 data, respectively. Each row has different noise level $\sigma = 1, 2, 3$. From the figure, we notice that the cumulative regret in each subplot has an increasing sublinear pattern. For the first row, the noise level is the smallest, which shows the sublinear pattern quickly. For other rows, all subplots show increasing sublinear patterns sooner or later, which satisfied our anticipation that when adding $k > 1$ data points and delete one data point at each time step, it finally will convert the normal online learn regime. However, the noise level will affect the time it moves to the sublinear pattern.

**Switching from Ridge to OLS.** Besides, we also consider when the agent accumulates many data such as $n > 2d$, where $n$ represents the number of data the agent has at some time point $t$, and then transfer the algorithm from the Adaptive ridge method to the OLS method. This method called Switching Adaptive Ridge represented as the cyan line in figure 5. We see that when the noise level is relatively small, it performs well. However, when $\sigma$ is large, it works worse than the green line (fixed Ridge with prior knowledge about $\lambda$) and the blue line (Adaptive Ridge).

### 6 Discussion and Conclusion

In this paper, we have proposed two online forgetting algorithms under the FIFD scheme and provide the theoretical regret upper bound. To our knowledge, this is the first, theoretically well-proved work in the online forgetting process under the FIFD scheme.

Besides, we find the existence of rank swinging phenomenon in the online least square regression and tackle it using the online adaptive ridge regression. In the future, we hope we can provide the lower bound for these two algorithms and design other online forgetting algorithms under the FIFD scheme.

### References


