Supplement to "Online Forgetting Process for Linear Regression Models"

This appendix contains the proof of the theoretical results, rank swinging phenomenon examples, and simulation results to illustrate the FIFD-Adaptive Ridge algorithm's performance. For the sake of simplicity, in the following proof of lemmas and theorems in appendix, we denote $a = t - s, b = t - 1$, and then the *constant memory limit* s equals to $b - a + 1$ for the *limited time window* $[t - s, t - 1] = [a, b]$.

A FIFD-OLS Confidence Ellipsoid

(Bernstein Concentration). *Let $\{D_k, S_k\}_{k=1}^{\infty}$ be a martingale difference, and suppose that D_k is a σ-subgaussian in an adapted sense, i.e., for all $\alpha \in \mathbb{R}$. $\mathbb{E}[e^{\alpha D_k}|S_{k-1}] \leq e^{\frac{\alpha^2\sigma^2}{2}}$ almost surely. Then, for all $t \geq 0$, $Pr[|\sum_{k=1}^{n} D_k| \geq t] \leq 2e^{-\frac{t^2}{2n\sigma^2}}$.* Lemma A is from Theorem 2.3 of Wainwright (2019) (**?**) when $\alpha_* = \alpha_k = 0$ and $\nu_k = \sigma$ for all k.

Define the event

$$\mathcal{F}(\lambda_0(\gamma)) \equiv \{\max_{r\in[d]}(2|\epsilon^{\mathsf{T}} X^{(r)}|/n) \leq \lambda_0(\gamma)\}$$

where $X^{(r)}$ is the r^{th} column of matrix $\mathbf{X}$ and $\lambda_0(\gamma) \equiv 2\sigma x_{max}\sqrt{(\gamma^2 + 2\log d)/n}$. Then, we have $Pr[\mathcal{F}(\lambda_0(\gamma))] \geq 1 - 2\exp[-\gamma^2/2]$.

Proof. Let S_t be the sigma algebra generated by random variables $X_1, \ldots, X_{t-1}$ and $Y_1, \ldots, Y_{t-1}$. First, using a union bound, we can write

$$\Pr[\mathcal{F}(\lambda_0(\gamma))] \geq 1 - \sum_{r=1}^{d} \Pr[|\epsilon^{\mathsf{T}} X^{(r)}| > n\lambda_0(\gamma)/2]$$

Now, for each $r \in [d]$, let $D_{t,r} = \epsilon_t X_{t,r}$ and note that $D_{1,r}, \ldots, D_{n,r}$ is a martingale difference sequence adapted to the filtration $S_1 \subset \ldots \subset S_n$ since $\mathbb{E}[\epsilon_t X_{t,r}|S_t] = 0$. On the other hand, each $D_{t,r}$ is a $(x_{\max}, \sigma)$-subgaussian random variable adapted to $\{S_t\}_{t=1}^{n}$, since

$$\mathbb{E}(e^{\alpha D_{t,r}}|S_{t-1}) \leq \mathbb{E}_{X_t}[e^{\alpha^2 X_{t,r}^2 \sigma^2/2}|S_{t-1}] \leq e^{\alpha^2 (x_{\max}\sigma)^2}.$$

Then, using Lemma A, $\Pr[\mathcal{F}(\lambda_0(\gamma))] \geq 1 - 2d\exp[-(\gamma^2 + 2\log d)/2] = 1 - 2\exp[-\gamma^2/2]$. □

(*FIFD-OLS Confidence Ellipsoid*) *For any $\delta > 0$, if the event $\lambda_{min}(\Phi_{[t-s,t-1]}/s) > \phi^2_{[t-s,t-1]} > 0$ holds, with probability at least $1 - \delta$, for all $t \geq s + 1$, $\theta_\star$ lies in the set*

$$C_{[t-s,t-1]} = \left\{\theta \in \mathbb{R}^d : \hat{\theta}_{[t-s,t-1]} - \theta_{\Phi_{[t-s,t-1]}} \leq \sigma q_{[t-s,t-1]}\sqrt{(2d/s)\log(2d/\delta)}\right\} \tag{1}$$

where $q_{[t-s,t-1]} = \mathbf{x}_{[t-s,t-1]\,\infty}/\phi^2_{[t-s,t-1]}$ is the adaptive constant and we denote $\beta_{[t-s,t-1]}$ as the RHS bound.

Proof. Notation $\hat{\Sigma}(\mathbf{x}_{[a,b]})$ represents the normalized covariance matrix, so $\hat{\Sigma}(\mathbf{x}_{[a,b]}) = \Phi_{[a,b]}/s = \mathbf{x}_{[a,b]}^{\mathsf{T}}\mathbf{x}_{[a,b]}/s$. Note that, if the event $\lambda_{\min}(\hat{\Sigma}(\mathbf{x}_{[a,b]})) > \phi^2_{[a,b]}$ holds,

$$\begin{aligned}
\hat{\theta}_{[a,b]} - \theta_{\star}{}_2 &= (\mathbf{x}_{[a,b]}^{\mathsf{T}}\mathbf{x}_{[a,b]})^{-1}\mathbf{x}_{[a,b]}^{\mathsf{T}}(\mathbf{x}_{[a,b]}\theta_\star + \epsilon) - \theta_{\star}{}_2 \\
&= (\mathbf{x}_{[a,b]}^{\mathsf{T}}\mathbf{x}_{[a,b]})^{-1}\mathbf{x}_{[a,b]}^{\mathsf{T}}\epsilon + \theta_\star - \theta_{\star}{}_2 \\
&= (\mathbf{x}_{[a,b]}^{\mathsf{T}}\mathbf{x}_{[a,b]})^{-1}\mathbf{x}_{[a,b]}^{\mathsf{T}}\epsilon_2 \\
&\leq \frac{1}{s\phi^2_{[a,b]}}\mathbf{x}_{[a,b]}^{\mathsf{T}}\epsilon_2
\end{aligned} \tag{2}$$

Then, for any $\chi > 0$, we can write

$$\begin{aligned}
\Pr\left[\hat{\theta}_{[a,b]} - \theta_{\star}{}_2 \leq \chi\right] &\geq \Pr\left[(\mathbf{x}_{[a,b]}^{\mathsf{T}}\epsilon_2 \leq s\chi\phi^2_{[a,b]}) \cap (\lambda_{\min}(\hat{\Sigma}_{[a,b]}) > \phi^2_{[a,b]})\right] \\
&\geq 1 - \sum_{r=1}^{d} \Pr\left[|\epsilon^{\mathsf{T}}\mathbf{x}^{(r)}_{[a,b]}| > \frac{s\chi\phi^2_{[a,b]}}{\sqrt{d}}\right] - \Pr\left[\lambda_{\min}(\hat{\Sigma}_{[a,b]}) \leq \phi^2_{[a,b]})\right]
\end{aligned} \tag{3}$$

where we have let $\mathbf{x}_{[a,b]}^{(r)}$ denote the r^{th} column of $\mathbf{x}_{[a,b]}$. We can expand $\epsilon^\mathsf{T}\mathbf{x}_{[a,b]}^{(r)} = \sum_{j\in[a,b]}\epsilon(j)\mathbf{x}_j^{(r)}$, where we note that $D_{j,r} \equiv \epsilon(k)\mathbf{x}_j^{(r)}$ is a $x_{\max}\sigma$-subgaussian random variable, where $x_{\max} = \mathbf{x}_{[a,b]_\infty}$, conditioned on the sigma algebra S_{j-1} that is generated by random variable $X_1, ..., X_{j-1}, y_1, ..., y_{j-1}$. Defining $D_{0,r} = 0$, the sequence $D_{0,r}, D_{1,r}, ..., D_{(b,r)}$ is a martingale difference sequence adapted to the filtration $S_1 \subset S_2 \subset ...S_b$ since $E[\epsilon(j)X_j^{(r)}|S_{j-1}] = 0$. Using Lemma A,

$$\begin{aligned}
&\Pr\left[\hat{\theta}_{[a,b]} - \theta_{\star 2} \le \chi\right] \\
&\quad\ge 1 - \sum_{r=1}^{d}\Pr\left[|\epsilon^\mathsf{T}\mathbf{x}_{[a,b]}^{(r)}| > \frac{s\chi\phi_{[a,b]}^2}{\sqrt{d}}\right] - \Pr\left[\lambda_{\min}(\hat{\Sigma}_{[a,b]}) \le \phi_{[a,b]}^2)\right] \\
&\quad\ge 1 - 2d\exp\left[-\frac{s\chi^2\phi_{[a,b]}^4}{2d\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right] - \Pr\left[\lambda_{\min}(\hat{\Sigma}_{[a,b]}) \le \phi_{[a,b]}^2)\right],
\end{aligned} \tag{4}$$

Since the event $\lambda_{\min}(\Phi_{[t-s,t-1]}/s) > \phi_{[t-s,t-1]}^2 > 0$ holds by the requirement of condition, then $\Pr\left[\lambda_{\min}(\hat{\Sigma}_{[a,b]}) \le \phi_{[a,b]}^2)\right] = 0$. With probability $1-\delta$, we have

$$1 - 2d\exp\left[-\frac{s\chi^2\phi_{[a,b]}^4}{2d\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right] \ge 1 - \delta. \tag{5}$$

Hence we have

$$\chi(\delta, s, q_{[a,b],\sigma,d}) \ge \sigma q_{[a,b]}\sqrt{\frac{2d}{s}\log(\frac{2d}{\delta})}. \tag{6}$$

where $q_{[a,b]} = \mathbf{x}_{[a,b]_\infty}/\phi_{[a,b]}^2$ is the adaptive constant for the limited time window $[a,b]$. Besides, we denote $\beta_{[a,b]}$ as the RHS constant. Then with probability $1-\delta$, for the limited time window $[t-s, t-1]$ and all $t \ge s+1$, we have the get the `FIFD-OLS` confidence ellipsoid, $\theta_\star$ lies in the set

$$C_{[t-s,t-1]} = \left\{\theta \in \mathbb{R}^d : \hat{\theta}_{[t-s,t-1]} - \theta_{\Phi_{[t-s,t-1]}} \le \sigma q_{[t-s,t-1]}\sqrt{(2d/s)\log(2d/\delta)}\right\}. \tag{7}$$

□

B FIFD-Adaptive Ridge Confidence Ellipsoid

(*FIFD-Adaptive Ridge Confidence Ellipsoid*) *For any $\delta \in [0,1]$, with probability at least $1-\delta$, for all $t \ge s+1$, if condition equation* (**??**) *is satisfied and the event $\lambda_{min}(\Phi_{\lambda,[t-s,t-1]}/s) > \phi_{\lambda,[t-s,t-1]}^2 > 0$ holds, then $\theta_\star$ lies in the set*

$$C_{\lambda,[t-s,t-1]} = \left\{\theta \in \mathbb{R}^d : \hat{\theta}_{\lambda,[t-s,t-1]} - \theta_{\Phi_{\lambda,[t-s,t-1]}} \le \sigma\kappa\nu q_{\lambda,[t-s,t-1]}\sqrt{d/2s}\right\}, \tag{8}$$

where $q_{\lambda,[t-s,t-1]} = \mathbf{x}_{[t-1,t-s]_\infty}/\phi_{\lambda,[t-s,t-1]}^2$, and $\kappa = \sqrt{\log^2(6|\mathcal{P}(\theta_\star)|/\delta)/\log(2d/\delta)}$. $\nu = \theta_{\star\infty}/\mathcal{P}_{\min}(\theta_\star)$ represents the strongest signal to weakest signal ratio.

Proof. For the sake of simplicity, We first denote $\hat{\Sigma}_\lambda(\mathbf{x}_{[a,b]}) = \Phi_{\lambda,[a,b]}/s$, $\hat{\Sigma}_{\lambda,[a,b]} = \hat{\Sigma}_\lambda(\mathbf{x}_{[a,b]})$, $\lambda_{[a,b]} = \lambda$ for the

limited time window $[t-s, t-1]$. Note that, if the event $\lambda_{\min}(\hat{\Sigma}_\lambda(\mathbf{x}_{[a,b]})) = \lambda_{\min}(\hat{\Sigma}_{\lambda,[a,b]}) > \phi^2_{\lambda,[a,b]} > 0$ holds,

$$
\begin{aligned}
&\hat{\theta}_{\lambda,[a,b]} - \theta_{\star_2} \\
&= (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\mathbf{x}_{[a,b]}^\mathsf{T}(\mathbf{x}_{[a,b]}\theta_\star + \epsilon) - \theta_{\star_2} \\
&= (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\mathbf{x}_{[a,b]}^\mathsf{T}\epsilon + (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}(\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})\theta_\star \\
&\qquad\qquad - \lambda(\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\theta_\star - \theta_{\star_2} \\
&= (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\mathbf{x}^\mathsf{T}\epsilon + \theta^\star - \lambda(\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\theta^\star - \theta_{\star_2} \\
&= (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\mathbf{x}_{[a,b]}^\mathsf{T}\epsilon - \lambda(\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}\theta_{\star_2} \\
&= (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}(\mathbf{x}^\mathsf{T}\epsilon - \lambda\theta_\star)_2 \\
&\leq (\mathbf{x}_{[a,b]}^\mathsf{T}\mathbf{x}_{[a,b]} + \lambda\mathbf{I})^{-1}{}_2\,\mathbf{x}^\mathsf{T}\epsilon - \lambda\theta_{\star 2} \\
&\leq \frac{1}{s\phi^2_{\lambda,[a,b]}}\mathbf{x}_{[a,b]}^\mathsf{T}\epsilon - \lambda\theta_{\star_2}.
\end{aligned}
\tag{9}
$$

We can expand $\epsilon^\mathsf{T}\mathbf{x}^{(r)}_{[a,b]} = \sum_{j\in[a,b]}\epsilon(j)\mathbf{x}_j^{(r)}$, where we let $\mathbf{x}^{(r)}_{[a,b]}$ denote the r^{th} column of $\mathbf{x}_{[a,b]}$ and $D_{j,r} \equiv \epsilon(k)\mathbf{x}_j^{(r)}$ is a $x_{\max}\sigma$-subgaussian random variable, conditioned on the sigma algebra S_{j-1} which is generated by random variables $X_1, ..., X_{j-1}, Y_1, ..., Y_{j-1}$. Defining $D_{0,r} = 0$, the sequence $D_{0,r}, D_{1,r}, ..., D_{(b,r)}$ is a martingale difference sequence adapted to the filtration $S_1 \subset S_2 \subset ...S_b$ since $E[\epsilon(j)X_j^{(r)}|S_{j-1}] = 0$. Using Lemma A,

$$
\begin{aligned}
&\Pr\left[\hat{\theta}_\lambda - \theta_{\star 2} \leq \chi\right] \\
&\geq 1 - \sum_{r=1}^{d}\Pr\left[|\epsilon^\mathsf{T}\mathbf{x}^{(r)}_{[a,b]} - \lambda\theta^{(r)}_\star| > \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}\right] - \Pr\left[\lambda_{\min}(\hat{\Sigma}_{\lambda,[a,b]}) \leq \phi^2_{\lambda,[a,b]}\right] \\
&= 1 - (\sum_{r=1}^{d}\Pr\left[\epsilon^\mathsf{T}\mathbf{x}^{(r)}_{[a,b]} > \underbrace{\lambda\theta^{(r)}_\star + \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}}_{\chi_{1,r}(\phi,\lambda,\theta^\star)}\right] + \Pr\left[\epsilon^\mathsf{T}\mathbf{x}^{(r)}_{[a,b]} < \underbrace{\lambda\theta^{(r)}_\star - \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}}_{\chi_{2,r}(\phi,\lambda,\theta^\star)}\right]),
\end{aligned}
\tag{10}
$$

since the event $\lambda_{\min}(\hat{\Sigma}_\lambda(\mathbf{x}_{[a,b]})) > \phi^2_{\lambda,[a,b]} > 0$ holds, $\Pr\left[\lambda_{\min}(\hat{\Sigma}_{\lambda,[a,b]}) \leq \phi^2_{\lambda,[a,b]}\right] = 0$.

Here we denote $\chi_{1,r}(\phi,\lambda,\theta^\star) = \lambda\theta^{(r)}_\star + \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}$ and $\chi_{2,r}(\phi,\lambda,\theta^\star) = \lambda\theta^{(r)}_\star - \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}$. By Lemma A, we can get the probability of this tail event,

B.1 Bounds the first part of inequality 10

In the following, we give a brief case by case analysis to decompose these two tail events' probabilities.

Case B.1.1: *If $\theta^{(r)}_\star = 0$, then $\chi_{1,r}(\phi,\lambda,\theta^\star) = \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} > 0$. We have,*

$$
Pr\left[\epsilon^\mathsf{T}X^{(r)} > \chi_{1,r}(\phi,\lambda,\theta^\star)\right] \leq \exp\left[-\frac{\chi^2_{1,r}(\phi,\lambda,\theta_\star)}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]. \tag{11}
$$

When λ becomes smaller, $\chi^2_{1,r}(\phi,\lambda,\theta_\star)$ becomes smaller. Then the RHS exponential probability bound of B.4 becomes larger. We always hope 10's probability bound smaller, then we can get a larger confidence ellipsoid. So λ becoming smaller is our choice.

Case B.1.2: *If $\theta^{(r)}_\star > 0$, then $\chi_{1,r}(\phi,\lambda,\theta_\star) = \lambda\theta^{(r)}_\star + \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} > 0$. We have,*

$$
Pr\left[\epsilon^\mathsf{T}X^{(r)} > \chi_{1,r}(\phi,\lambda,\theta_\star)\right] \leq \exp\left[-\frac{\chi^2_{1,r}(\phi,\lambda,\theta_\star)}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]. \tag{12}
$$

When λ becomes smaller, $\chi^2_{1,r}(\phi,\lambda,\theta_\star)$ becomes smaller. Then RHS exponential probability bound of B.5 becomes larger. Then part B.1.2's probability bound becomes smaller. We always hope the B.1.2's probability bound smaller, then we can get a larger confidence ellipsoid. So λ becoming smaller is our choice.

Case B.1.3: *If $\theta^{(r)}_\star < 0$ and $\lambda < -\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta^{(r)}_\star}$, then $\chi_{1,r}(\phi,\lambda,\theta_\star) = \lambda\theta^{(r)}_\star + \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} > 0$. We have*

$$Pr\left[\epsilon^\mathsf{T} X^{(r)} > \chi_{1,r}(\phi,\lambda,\theta_\star)\right] \leq \exp\left[-\frac{\chi^2_{1,r}(\phi,\lambda,\theta^\star)}{2s\mathbf{x}^2_{[a,b]_\infty}\sigma^2}\right]. \tag{13}$$

When λ becomes larger, $\chi^2_{1,r}(\phi,\lambda,\theta_\star)$ becomes smaller. Then RHS B.6's exponential probability bound becomes larger. Then part B.1.3's probability becomes smaller. We always hope the B.1.3's probability bound smaller, then we can get a larger confidence ellipsoid. Then the final probability gets smaller. So λ becoming larger is our choice.

Case B.1.4: *If $\theta^{(r)}_\star < 0$ and $\lambda \geq -\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta^{(r)}_\star}$, then $\chi_{1,r}(\phi,\lambda,\theta_\star) = \lambda\theta^{(r)}_\star + \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} < 0$, which means that this probability is larger than $\frac{1}{2}$ because our ϵ is symmetric random variable. We have*

$$Pr\left[\epsilon^\mathsf{T} X^{(r)} > \chi_{1,r}(\phi,\lambda,\theta_\star)\right]. \tag{14}$$

If we want our confidence ellipsoid having a relative large probability, we need to avoid this case. So the choice for λ is $\lambda < -\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta^{(r)}_\star}$.

When consider bounding the first part of 10, λ should be in the interval $\lambda \in [0, \min_{r\in\mathcal{N}(\theta_\star)} -\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta^{(r)}_\star})$. Then we just consider cases B.1.1, B.1.2 and B.1.3.

B.2 Bounds the second part of inequality 10

In following cases. We analyze the second part of 10 and get the probability upper bound.

Case B.2.1: *If $\theta^{(r)}_\star = 0$, then $\chi_{2,r}(\phi,\lambda,\theta_\star) = -\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} < 0$. We have,*

$$Pr\left[\epsilon^\mathsf{T} X^{(r)} < \chi_{2,r}(\phi,\lambda,\theta_\star)\right] \leq \exp\left[-\frac{\chi^2_{2,r}(\phi,\lambda,\theta_\star)}{2s\mathbf{x}^2_{[a,b]_\infty}\sigma^2}\right] \quad . \tag{15}$$

When λ becomes smaller, $\chi^2_{1,r}(\phi,\lambda,\theta_\star)$ becomes smaller. Then RHS exponential probability bound of B.8 becomes larger. We always hope 10's probability bound smaller, then we can get a larger confidence ellipsoid. So λ becoming smaller is our choice..

Case B.2.2: *If $\theta^{(r)}_\star < 0$, then $\chi_{2,r}(\phi,\lambda,\theta_\star) = \lambda\theta^{(r)}_\star - \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} < 0$. We have,*

$$Pr\left[\epsilon^\mathsf{T} X^{(r)} < \chi_{2,r}(\phi,\lambda,\theta_\star)\right] \leq \exp\left[-\frac{\chi^2_{2,r}(\phi,\lambda,\theta_\star)}{2s\mathbf{x}^2_{[a,b]_\infty}\sigma^2}\right]. \tag{16}$$

When λ becomes larger, $|\chi_{2,r}(\phi,\lambda,\theta_\star)|$ becomes larger. Then RHS exponential probability becomes smaller. Then part B.2.2's probability bound becomes smaller, we can get a larger confidence ellipsoid. So λ becoming larger is our choice.

Case B.2.3: *If $\theta^{(r)}_\star > 0$ and $\lambda < \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta^{(r)}_\star}$, then $\chi_{2,r}(\phi,\lambda,\theta_\star) = \lambda\theta^{(r)}_\star - \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} < 0$. We have*

$$Pr\left[\epsilon^\mathsf{T} X^{(r)} < \chi_{2,r}(\phi,\lambda,\theta_\star)\right] \leq \exp\left[-\frac{\chi^2_{2,r}(\phi,\lambda,\theta_\star)}{2s\mathbf{x}^2_{[a,b]_\infty}\sigma^2}\right]. \tag{17}$$

When λ becomes smaller, $|\chi_{2,r}(\phi,\lambda,\theta_\star)|$ gets larger. Then RHS exponential probability bound becomes smaller. Then part B.2.3's probability bound becomes smaller, we can get a larger confidence ellipsoid. So λ becoming smaller is our choice.

Case B.2.4: *If* $\theta_\star^{(r)} > 0$ *and* $\lambda \geq \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta_\star^{(r)}}$, *then* $\chi_{2,r}(\phi,\lambda,\theta_\star) = \lambda\theta_\star^{(r)} - \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}} > 0$, *which means that this probability is larger than* $\frac{1}{2}$ *because our* ϵ *is symmetric random variable. We have*

$$Pr\left[\epsilon^\mathsf{T} X^{(r)} < \chi_{2,r}(\phi,\lambda,\theta_\star)\right] \geq \frac{1}{2}. \tag{18}$$

If we want our confidence ellipsoid having a relative large probability, we need to avoid this case. So the choice for $\lambda < \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta_\star^{(r)}}$.

When consider bounding the second part of 10, λ should be in the interval $\lambda \in [0, \min\limits_{r\in\mathcal{P}(\theta_\star)} \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta_\star^{(r)}})$. Then we just consider cases B.2.1, B.2.2 and B.2.3.

B.3 Lower bound of inequality 10

Combining λ from subsections of B.2 and B.3, we get one adaptive interval for λ

$$\lambda^{\mathrm{bd}} = \min\{\min_{r\in\mathcal{N}(\theta_\star)} -\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta_\star^{(r)}}, \min_{r\in\mathcal{P}(\theta_\star)} \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}\theta_\star^{(r)}}\} = \frac{s\chi\phi^2_{[a,b]}}{\sqrt{d}}\frac{1}{\theta_{\star\infty}}. \tag{19}$$

Since cases B.1.2, B.1.3 and B.2.2, B.2.3 are two counteractive cases. If we know the number of positive coordinate of $\theta_\star$, $|\mathcal{P}(\theta_\star)|$, is more than the number of negative coordinate of $\theta_\star$, $|\mathcal{N}(\theta_\star)|$. We would prefer cases B.1.3, B.2.3; otherwise, we prefer cases B.1.2, B.2.2. Therefore, it is a trade-off.

Now let $d_0 = d - |\mathcal{P}(\theta_\star)| - |\mathcal{N}(\theta_\star)|$ denotes the number of zero coordinate of $\theta_\star$. Then equation (10) becomes

$$\begin{aligned}
&\Pr\left[\hat{\theta}_{\lambda,[a,b]} - {\theta^\star}_2 \leq \chi\right] \\
&\geq 1 - (\sum_{r_+\in\mathcal{P}(\theta_\star)} \exp\left[-\frac{\chi^2_{1,r_+}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \sum_{r_-\in\mathcal{N}(\theta_\star)} \exp\left[-\frac{\chi^2_{1,r_-}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + d_0\exp\left[-\frac{\chi^2_1(\phi,\lambda,0)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right]) \\
&- (\sum_{r_+\in\mathcal{P}(\theta_\star)} \exp\left[-\frac{\chi^2_{2,r_+}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \sum_{r_-\in\mathcal{N}(\theta_\star)} \exp\left[-\frac{\chi^2_{2,r_-}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + d_0\exp\left[-\frac{\chi^2_2(\phi,\lambda,0)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right]).
\end{aligned} \tag{20}$$

If we assume $\lambda \in [0, \lambda^{\mathrm{bd}}_{[a,b]}]$, then

$$\begin{aligned}
&= 1 - d_0(\exp\left[-\frac{\chi^2_1(\phi,\lambda,0)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \exp\left[-\frac{\chi^2_2(\phi,\lambda,0)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right]) \\
&\quad - (\sum_{r_+\in\mathcal{P}(\theta_\star)} \exp\left[-\frac{\chi^2_{1,r_+}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \exp\left[-\frac{\chi^2_{2,r_+}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right]) \\
&\quad - (\sum_{r_-\in\mathcal{N}(\theta_\star)} \exp\left[-\frac{\chi^2_{1,r_-}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \exp\left[-\frac{\chi^2_{2,r_-}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right]).
\end{aligned} \tag{21}$$

Since we know the $\chi^2_1(\phi,\lambda,0) = \chi^2_2(\phi,\lambda,0) = \frac{s^2\chi^2\phi^4_{\lambda,[a,b]}}{d}$. Thus

$$\exp\left[-\frac{\chi^2_1(\phi,\lambda,0)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \exp\left[-\frac{\chi^2_2(\phi,\lambda,0)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right]) = 2\exp(-\frac{s\chi^2\phi^4_{\lambda,[a,b]}}{2d{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}). \tag{22}$$

Then the second positive coordinate of $\theta_\star$ part becomes

$$\begin{aligned}
&= \sum_{r_+\in\mathcal{P}(\theta_\star)} \exp\left[-\frac{\chi^2_{1,r_+}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \exp\left[-\frac{\chi^2_{2,r_+}(\phi,\lambda,\theta_\star)}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] \\
&= \sum_{r_+\in\mathcal{P}(\theta_\star)} \exp\left[-\frac{(\lambda\theta_\star^{r_+} + \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}})^2}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right] + \exp\left[-\frac{(\lambda\theta_\star^{r_+} - \frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}})^2}{2s{\mathbf{x}_{[a,b]}}^2_\infty\sigma^2}\right].
\end{aligned} \tag{23}$$

Since we want to get upper bound of the above equation and then to get the confidence interval, we maximizes this exponential value by selecting the minimum of χ^2_{1,r_+}andχ^2_{2,r_+}.

Since in the case B.1.2, we know $\chi_{1,r_+}(\phi,\lambda,\theta_\star)=\lambda\theta_\star^{r_+}+\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}>0$ and by selecting the $\min_{r_+\in\mathcal{P}(\theta_\star)}\theta_\star^{r_+}$ given fixed λ. We denote the minimum positive coordinate of $\theta_\star$ as $\mathcal{P}_{\min}(\theta_\star)$, $C_1(\phi_{\lambda,[a,b]},\chi,d)=\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}$; In the second part of the exponential, we know in case B.2.3, $\chi_{2,r_+}(\phi,\lambda,\theta_\star)=\lambda\theta_\star^{r_+}-\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}<0$. So if we want to minimize $|\chi_{2,r_+}|$ given fixed lambda, select the maximum: $\max_{r_+\in\mathcal{P}(\theta_\star)}\theta_\star^{r_+}$. We denote the maximum positive coordinate of $\theta_\star$ as $\mathcal{P}_{\max}(\theta_\star)$. Thus, taking both of these into consideration to get an upper bound of this summation of probability, we can obtain

$$
\begin{aligned}
&\leq \sum_{r_+\in\mathcal{P}(\theta_\star)}\exp\left[-\frac{(\lambda\mathcal{P}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]+\exp\left[-\frac{(\lambda\mathcal{P}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]\\
&=|\mathcal{P}(\theta_\star)|(\exp\left[-\frac{(\lambda\mathcal{P}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]+\exp\left[-\frac{(\lambda\mathcal{P}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right])\\
&\leq 2|\mathcal{P}(\theta_\star)|\max\left\{\exp\left[-\frac{(\lambda\mathcal{P}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right],\exp\left[-\frac{(\lambda\mathcal{P}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]\right\},
\end{aligned}
\tag{24}
$$

where constant $C_1(\phi_{\lambda,[a,b]},\chi,d)=\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}}$.

If $\lambda\leq\frac{2C_1(\phi_{\lambda,[a,b]},\chi,d)}{\mathcal{P}_{\max}(\theta_\star)-\mathcal{N}_{\max}(\theta_\star)}$, then (B.17)

$$
\leq 2|\mathcal{P}(\theta_\star)|\exp\left[-\frac{(\lambda\mathcal{P}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]
\tag{25}
$$

This condition is easily to be satisfied since $-\mathcal{N}_{\max}(\theta_\star)$ should be a very small value.

The third negative part becomes

$$
\begin{aligned}
&=\sum_{r_-\in\mathcal{N}(\theta_\star)}\exp\left[-\frac{\chi^2_{1,r_-}(\phi,\lambda,\theta^\star)}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]+\exp\left[-\frac{\chi^2_{2,r_-}(\phi,\lambda,\theta^\star)}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]\\
&=\sum_{r_-\in\mathcal{N}(\theta_\star)}\exp\left[-\frac{(\lambda\theta_\star^{r_-}+\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}})^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right]+\exp\left[-\frac{(\lambda\theta_\star^{r_-}-\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}})^2}{2s\mathbf{x}_{[a,b]}{}^2_\infty\sigma^2}\right].
\end{aligned}
\tag{26}
$$

Since we want to get an upper bound of the above equation and then to get the confidence interval, we maximizes this exponential value by selecting the minimum of χ^2_{1,r_-}andχ^2_{2,r_-}.

In the case B.1.3, we know $\chi_{1,r_-}(\phi,\lambda,\theta_\star)=\lambda\theta_\star^{r_-}+\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}})>0$ and by selecting $\min_{r_-}\theta_\star^{r_-}$ given fixed λ. We denote the minimum negative coordinate of $\theta_\star$ as $\mathcal{N}_{\min}(\theta_\star)$;

In the second part of the exponential, we know in case B.2.2, $\chi_{2,r_-}(\phi,\lambda,\theta_\star)=(\lambda_{[a,b]}\theta^\star_{r_-}-\frac{s\chi\phi^2_{\lambda,[a,b]}}{\sqrt{d}})<0$, so if we want to minimize $|\chi_{2,r_-}|$, we need to select the maximum: $\max_{r_-\in\mathcal{N}(\theta_\star)}\theta_\star^{r_-}$. We denote the maximum negative coordinate of $\theta_\star$ as $\mathcal{N}_{\max}(\theta_\star)$.

Thus, taking both of these into consideration to get an upper bound of this summation of probability, we can

obtain

$$
\begin{aligned}
&\leq \sum_{r_- \in \mathcal{N}(\theta_\star)} \exp\left[-\frac{(\lambda\mathcal{N}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]+\exp\left[-\frac{(\lambda\mathcal{N}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]\\
&=|\mathcal{N}(\theta_\star)|(\exp\left[-\frac{(\lambda\mathcal{N}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]+\exp\left[-\frac{(\lambda\mathcal{N}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right])\\
&\leq 2|\mathcal{N}(\theta_\star)|\max\left\{\exp\left[-\frac{(\lambda\mathcal{N}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right],\exp\left[-\frac{(\lambda\mathcal{N}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]\right\}.
\end{aligned}
\tag{27}
$$

If $\lambda \leq \frac{2C_1(\phi_{\lambda,[a,b]},\chi,d)}{\mathcal{N}_{\max}(\theta_\star)-\mathcal{N}_{\min}(\theta_\star)}$, then

$$
\leq 2|\mathcal{N}(\theta_\star)|\exp\left[-\frac{(\lambda\mathcal{N}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right] \tag{28}
$$

This is similar to the previous one in equation (25).

Finally, when we combine equations (22), (25), (28) together, we get the estimated probability

$$
\begin{aligned}
&\Pr\left[\hat{\theta}_{\lambda,[a,b]}-\theta^\star{}_2\leq\chi\right]\\
&\quad\geq 1-2d_0\exp\left[-\frac{C_1^2(\phi_{\lambda,[a,b]},\chi,d)}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]-2|\mathcal{P}(\theta_\star)|\exp\left[-\frac{(\lambda\mathcal{P}_{\min}(\theta_\star)+C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]\\
&\qquad-2|\mathcal{N}(\theta_\star)|\exp\left[-\frac{(\lambda\mathcal{N}_{\max}(\theta_\star)-C_1(\phi_{\lambda,[a,b]},\chi,d))^2}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right].
\end{aligned}
\tag{29}
$$

Since we want to control the confidence set with probability at least $1-\delta$, let $\Pr\left[\hat{\theta}_{\lambda,[a,b]}-\theta^\star{}_2\leq\chi\right]\geq 1-\delta$, we have

$$
\begin{aligned}
&\underbrace{d\exp\left[-\frac{C_1^2(\phi_{\lambda,[a,b]},\chi,d)}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]}_{\text{Part I}}+\underbrace{|\mathcal{P}(\theta_\star)|\exp\left[-\frac{\lambda^2(\mathcal{P}_{\min}(\theta_\star))^2+2\lambda\mathcal{P}_{\min}(\theta_\star)C_1(\phi_{\lambda,[a,b]},\chi,d)}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]}_{\text{Part II}}\\
&\qquad+\underbrace{|\mathcal{N}(\theta_\star)|\exp\left[-\frac{\lambda^2(\mathcal{N}_{\max}(\theta_\star))^2-2\lambda\mathcal{N}_{\max}(\theta_\star)C_1(\phi_{\lambda,[a,b]},\chi,d)}{2s\mathbf{x}_{[a,b]}{}_\infty^2\sigma^2}\right]}_{\text{Part III}}\leq\frac{\delta}{2}.
\end{aligned}
\tag{30}
$$

In the following, we present the assumption we need to make the above inequality to have an analytic solution.

Assumption 1. *(Weakest Positive to Strongest Signal Ratio) This is the condition for considering the case that positive coordinate of $\theta_\star$ dominates the bad events happening and without loss of generality, we assume that*

$$
\text{WPSSR}=\frac{\mathcal{P}_{\min}(\theta_\star)}{\theta_{\star\infty}}\leq\frac{-\sqrt{\log\frac{6d}{\delta}\log\frac{2d}{\delta}}+\sqrt{\log\frac{6d}{\delta}\log\frac{2d}{\delta}+s^2\log\frac{2d}{\delta}\log\frac{12|\mathcal{P}(\theta_\star)|}{\delta}}}{s\log\frac{2d}{\delta}}. \tag{31}
$$

Remarks. The WPSSR is monotone increasing in s and $\mathcal{P}(\theta_\star)$, and is monotone decreasing in d. However, as long as $s\geq d$, in most cases, the LHS is greater than one. For example, if $s=100, d=110, \delta=0.05, |\mathcal{P}(\theta_\star)|=30$, then WPSSR needs to be less than 1.02, which is satisfied automatically.

First we have a weak assumption that *Part I* is always less than $\delta/2$, with this assumption, we can have a initial interval for $\chi(\delta)$,

$$
\chi(\delta)\geq\sqrt{2d}\sigma\frac{x_{\max,[a,b]}}{\phi^2_{\lambda,[a,b]}\sqrt{b-a}}\sqrt{\log(\frac{2d}{\delta})}. \tag{32}
$$

Then we plug in this initial $\chi(\delta)$ into the λ^{bd}, we can get the initial interval for λ,

$$\lambda \leq \sigma \mathbf{x}_{[t-1,t-s]_\infty} \sqrt{2s \log(2d/\delta)} / \theta_{\star\infty}. \tag{33}$$

To get an analytical confidence ellipsoid, without loss of generality, we assume *Part II* is greater than *Part III* and *Part I* with assumption 1 and if select λ following Lemma **??**, then

$$3|\mathcal{P}(\theta_\star)| \exp\left[-\frac{\lambda^2(\mathcal{P}_{\min}(\theta_\star))^2 + 2\lambda \mathcal{P}_{\min}(\theta_\star) C_1(\phi_{\lambda,[a,b]}, \chi, d)}{2s\mathbf{x}_{[a,b]_\infty^2}^2 \sigma^2}\right] \leq \frac{\delta}{2}. \tag{34}$$

Thus, with confidence $1-\delta$, we have the following confidence ellipsoid for `FIFD-Adaptive Ridge` method,

$$C_{\lambda,[t-s,t-1]} = \left\{\theta \in \mathbb{R}^d : \hat{\theta}_{[t-s,t-1]} - \theta_{\Phi_{\lambda,[t-s,t-1]}} \leq \sigma\kappa\nu q_{\lambda,[t-s,t-1]} \sqrt{d/2s}\right\}, \tag{35}$$

where $q_{\lambda,[t-s,t-1]} = \mathbf{x}_{[t-1,t-s]_\infty} / \phi^2_{\lambda,[t-s,t-1]}$, $\kappa = \sqrt{\log^2(6|\mathcal{P}(\theta_\star)|/\delta)/\log(2d/\delta)}$, and $\nu = \theta_{\star\infty}/\mathcal{P}_{\min}(\theta_\star)$.

□

C FIFD-OLS Regret

(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 $\mathcal{P}_\mathcal{X}$. With probability at least $1-\delta \in [0,1]$ and Lemma A holds, for all $T > s, s \geq d$, we have an upper bound on the cumulative regret at time T:

$$R_{T,s}(\mathcal{A}_{OLS}) \leq 2\sigma\zeta\sqrt{(d/s)\log(2d/\delta)(T-s)\left(d\log(sL^2/d) + (T-s)\right)}, \tag{36}$$

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

Proof. Here we use l_t to denote the instantaneous *absolute loss* at time t. Let's decompose the instantaneous absolute loss as follows:

$$\begin{aligned}
l_t &= |\langle \hat{\theta}_{[t-s,t-1]}, X_t\rangle - \langle \theta_\star, X_t\rangle| \\
&= |\langle \hat{\theta}_{[t-s,t-1]} - \theta_\star, x_t\rangle| \\
&= |[\hat{\theta}_{[t-s,t-1]} - \theta_\star]^\mathsf{T} x_t| \\
&= |[\hat{\theta}_{[t-s,t-1]} - \theta_\star]^\mathsf{T} \Phi^{\frac{1}{2}}_{[t-s,t-1]} \Phi^{-\frac{1}{2}}_{[t-s,t-1]} x_t| \\
&= |[\hat{\theta}_{[t-s,t-1]} - \theta_\star]^\mathsf{T} \Phi^{\frac{1}{2}}_{[t-s,t-1]}| \times |\Phi^{-\frac{1}{2}}_{[t-s,t-1]} x_t| \\
&\leq \hat{\theta}_{[t-s,t-1]} - \theta_{\star_{\Phi_{[t-s,t-1]}}} x_{t_{\Phi^{-1}_{[t-s,t-1]}}} \\
&\leq \sqrt{\beta_{[t-s,t-1]}(\delta)} x_{t_{\Phi^{-1}_{[t-s,t-1]}}},
\end{aligned} \tag{37}$$

where the last step from Lemma **??**. Thus, with probability at least $1-\delta$, for all $T > s$,

$$\begin{aligned}
R_{T,s}(\mathcal{A}) &= \sqrt{(T-s)\sum_{t=s+1}^{T} r_t^2} \\
&\leq \sqrt{(T-s)\sum_{t=s+1}^{T} \beta_{[t-s,t-1]}(\delta) x^2_{t_{\Phi^{-1}_{[t-s,t-1]}}}},
\end{aligned} \tag{38}$$

where the last step we use Lemma **??** to process the deletion and addition procedure. So we have

$$\begin{aligned}
&\leq \sqrt{2(T-s)\max_{s+1\leq t\leq T} \beta_{[t-s-1,t-1]}(\delta)\left(d\log(\frac{sL^2}{d}) + (T-s)\right)} \\
&\leq 2\sigma\zeta\sqrt{(d/s)\log(2d/\delta)(T-s)\left(d\log(sL^2/d) + (T-s)\right)},
\end{aligned} \tag{39}$$

where the last step uses the confidence ellipsoid from Lemma **??**. □

The cumulative regret of FIFD-OLS is partially determined by FRT at each time step,

$$\sum_{t=s+1}^{T} {x_t}^2_{\Phi^{-1}_{[t-s,t-1]}} \leq 2\eta_{\text{OLS}} + \sum_{t=s+1}^{T} \text{FRT}_{[t-s,t-1]} \leq 2\left[d\log(\frac{sL^2}{d}) + (T-s)\right] \tag{40}$$

where $\eta_{\text{OLS}} = \log(\det(\Phi_{[T-s,T]}))$ is a constant based on data time window$[T-s, T]$.

Proof. Here we use $a = t-s, b = t-1$ for the sake of simplicity. Elementary algebra gives

$$\begin{aligned}
&\det(\Phi_{[a+1,b+1]}) \\
&\quad= \det(\Phi_{\lambda,[a,b]} + x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T}) \\
&\quad= \det(\Phi^{1/2}_{\lambda,[a,b]}(\mathbf{I} + \Phi^{-1/2}_{\lambda,[a,b]}(x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T})\Phi^{-1/2}_{\lambda,[a,b]})\Phi^{1/2}_{\lambda,[a,b]}) \\
&\quad= \det(\Phi_{\lambda,[a,b]})\det(\mathbf{I} + \Phi^{-1/2}_{\lambda,[a,b]}(x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T})\Phi^{-1/2}_{\lambda,[a,b]}) \\
&\quad= \det(\Phi_{\lambda,[a,b]})\det(\mathbf{I} + \Phi^{-1/2}_{\lambda,[a,b]}x_{b+1}x_{b+1}^\mathsf{T}\Phi^{-1/2}_{\lambda,[a,b]} - \Phi^{-1/2}_{\lambda,[a,b]}x_a x_a^\mathsf{T}\Phi^{-1/2}_{\lambda,[a,b]}) \\
&\quad= \det(\Phi_{\lambda,[a,b]})\det\left(\mathbf{I} + (\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})(\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})^\mathsf{T} - (\Phi^{-1/2}_{\lambda,[a,b]}x_a)(\Phi^{-1/2}_{\lambda,[a,b]}x_a)^\mathsf{T}\right) \\
&\quad= \det(\Phi_{\lambda,[a,b]})\left((1 + {\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1}}^2)(1 - {\Phi^{-1/2}_{\lambda,[a,b]}x_a}^2) + \langle\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1}, \Phi^{-1/2}_{\lambda,[a,b]}x_a\rangle^2\right).
\end{aligned} \tag{41}$$

where the last step use lemma C,

$$\begin{aligned}
&= \det(\Phi_{\lambda,[a,b]})\left((1 + {x_{b+1}}^2_{\Phi^{-1}_{\lambda,[a,b]}})(1 - {x_a}^2_{\Phi^{-1}_{\lambda,[a,b]}}) + \langle x_{b+1}, x_a\rangle^2_{\Phi^{-1}_{\lambda,[a,b]}}\right) \\
&= \det(\Phi_{\lambda,[a,b]})\left(1 + {x_{b+1}}^2_{\Phi^{-1}_{\lambda,[a,b]}} - {x_a}^2_{\Phi^{-1}_{\lambda,[a,b]}} + \langle x_{b+1}, x_a\rangle^2_{\Phi^{-1}_{\lambda,[a,b]}} - {x_{b+1}}^2_{\Phi^{-1}_{\lambda,[a,b]}}{x_a}^2_{\Phi^{-1}_{\lambda,[a,b]}}\right) \\
&= \det(\Phi_{\lambda,[a,b]})\left(1 + {x_{b+1}}^2_{\Phi^{-1}_{\lambda,[a,b]}} - {x_a}^2_{\Phi^{-1}_{\lambda,[a,b]}} + {x_{b+1}}^2_{\Phi^{-1}_{\lambda,[a,b]}}{x_a}^2_{\Phi^{-1}_{\lambda,[a,b]}}(\cos^2_{\Phi^{-1}_{\lambda,[a,b]}}\theta - 1)\right).
\end{aligned} \tag{42}$$

where $\cos_{\Phi^{-1}_{\lambda,[a,b]}}\theta = \frac{\langle x_{b+1}, x_a\rangle_{\Phi^{-1}_{\lambda,[a,b]}}}{{x_{b+1}}_{\Phi^{-1}_{\lambda,[a,b]}}{x_a}_{\Phi^{-1}_{\lambda,[a,b]}}}$, which measures the similarity of the vector of x_{b+1} and x_a with respect to $\Phi^{-1}_{\lambda,[a,b]}$. If $\cos^2_{\Phi^{-1}_{\lambda,[a,b]}}\theta = 1$, that means the incoming data x_{b+1} and the deleted data x_a are same. However, if $\cos^2_{\Phi^{-1}_{\lambda,[a,b]}}\theta = 0$, that means the incoming data and the deleted data are totally different with respect to $\Phi^{-1}_{\lambda,[a,b]}$. Now we switch to the notation t, where $t-s = a, t-1 = b$. At time step T and combining equation (42), we have

$$\begin{aligned}
\det(\Phi_{[T-s,T]}) = \prod_{t=s+1}^{T} (1 &+ {x_t}^2_{\Phi^{-1}_{[t-s,t-1]}} - {x_{t-s}}^2_{\Phi^{-1}_{[t-s,t-1]}} \\
&+ {x_t}^2_{\Phi^{-1}_{[t-s,t-1]}}{x_{t-s}}^2_{\Phi^{-1}_{[t-s,t-1]}}(\cos^2_{\Phi^{-1}_{[t-s,t-1]}}\theta - 1)).
\end{aligned} \tag{43}$$

Taking log to both side of equation (43), we get

$$\begin{aligned}
\log(\det(\Phi_{[T-s,T]})) = \sum_{t=s+1}^{T} \log(1 &+ {x_t}^2_{\Phi^{-1}_{[t-s,t-1]}} - {x_{t-s}}^2_{\Phi^{-1}_{[t-s,t-1]}} \\
&+ {x_t}^2_{\Phi^{-1}_{[t-s,t-1]}}{x_{t-s}}^2_{\Phi^{-1}_{[t-s,t-1]}}(\cos^2_{\Phi^{-1}_{[t-s,t-1]}}\theta - 1)).
\end{aligned} \tag{44}$$

Combining $\log(1+x) > \frac{x}{1+x}$ which holds when $x > -1$, we first consider each part of the product and use the

dissimilarity measure $\sin^2_{\Phi^{-1}_{\lambda,[a,b]}}\theta = 1 - \cos^2_{\Phi^{-1}_{\lambda,[a,b]}}\theta$. So we get

$$\begin{aligned}
&\log\left(1 + x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2_{\Phi^{-1}_{[t-s,t-1]}}\theta\right) \\
&> \frac{x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2_{\Phi^{-1}_{[t-s,t-1]}}\theta}{1 + x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2_{\Phi^{-1}_{[t-s,t-1]}}\theta} \\
&> \frac{1}{2}\left[x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} - x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2_{\Phi^{-1}_{[t-s,t-1]}}\theta\right].
\end{aligned} \tag{45}$$

Therefore, we can give a bound of $\sum_{t=s+1}^{T} x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}}$,

$$\begin{aligned}
&\sum_{t=s+1}^{T} x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \\
&\leq 2\log(\det(\Phi_{[T-s,T]})) + \sum_{t=s+1}^{T} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} + \sum_{t=1}^{T-s} x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2_{\Phi^{-1}_{[t-s,t-1]}}\theta.
\end{aligned} \tag{46}$$

Thus by combining the last two terms and extracting $x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}}$, we can get

$$\sum_{t=s+1}^{T} x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \leq 2\log(\det(\Phi_{[T-s,T]})) + \sum_{t=1}^{T-s} x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} (x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2_{\Phi^{-1}_{[t-s,t-1]}}\theta + 1). \tag{47}$$

To make the formula simpler, we define *'Forgetting Regret Term' (FRT)* term, at time window $[t-s, t-1]$ or called at time step t as follows,

$$\mathrm{FRT}_{[t-s,t-1]} = x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} (x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) + 1), \tag{48}$$

where $\mathrm{FRT}_{[t-s,t-1]} \in [0, 2]$. The detailed explanation and examples of *'Forgetting Regret Term' (FRT)* can be found in D.

So equation (72) becomes

$$\sum_{t=s+1}^{T} x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \leq 2\eta_{\mathrm{OLS}} + \sum_{t=s+1}^{T} \mathrm{FRT}_{[t-s,t-1]}, \tag{49}$$

where $\eta_{\mathrm{OLS}} = \log(\det(\Phi_{[T-s,T]}))$ is a constant based on data time window$[T-s, T]$. By Lemma C, we get

$$\sum_{t=s+1}^{T} x_{t}{}^2_{\Phi^{-1}_{[t-s,t-1]}} \leq 2\left[d\log(\frac{sL^2}{d}) + (T-s)\right]. \tag{50}$$

□

$$\det(\mathbf{I} - aa^{\mathsf{T}} + bb^{\mathsf{T}}) = (1 + b^2)(1 - a^2) + \langle a, b\rangle^2 \tag{51}$$

Proof. By Woodbury matrix identity

$$(\mathbf{I} - aa^{\mathsf{T}})^{-1} = \mathbf{I} + \frac{aa^{\mathsf{T}}}{1 - a^{\mathsf{T}}a} \tag{52}$$

and by Matrix determinant lemma, suppose $\mathbf{B}$ is an invertible square matrix and u, v are column vectors. Then the matrix determinant lemma states that

$$\det(\mathbf{B} + uv^{\mathsf{T}}) = (1 + v^{\mathsf{T}}\mathbf{B}^{-1}u)\det(\mathbf{B}). \tag{53}$$

So combining above two equations, we get

$$\begin{aligned}\det(\mathbf{I}-aa^{\mathsf{T}}+bb^{\mathsf{T}}) &= \left(1+b^{\mathsf{T}}(\mathbf{I}-aa^{\mathsf{T}})^{-1}b\right)\det(\mathbf{I}-aa^{\mathsf{T}}) \\ &= \left(1+b^{\mathsf{T}}(\mathbf{I}+\frac{aa^{\mathsf{T}}}{1-a^{\mathsf{T}}a})b\right)\det(\mathbf{I}-aa^{\mathsf{T}}) \\ &= (1+b^{\mathsf{T}}b+\frac{b^{\mathsf{T}}aa^{\mathsf{T}}b}{1-a^{\mathsf{T}}a})\det(\mathbf{I}-aa^{\mathsf{T}}) \\ &= (1+b^{\mathsf{T}}b+\frac{\langle a,b\rangle^2}{1-a^{\mathsf{T}}a})(1-a^{\mathsf{T}}a) \\ &= (1+b^2)(1-a^2)+\langle a,b\rangle^2.\end{aligned} \tag{54}$$

□

(Determinant-Trace Inequality). Suppose $x_1, x_2, ..., x_b \in \mathbb{R}^d$ and for any $t \in [b]$, $x_{t2} \leq L$. Let $\Phi_{[1,b]} = \sum_{t=1}^{b} x_t x_t^{\mathsf{T}}$. Then we have

$$det(\Phi_{[1,b]}) \leq (\frac{bL^2}{d})^d. \tag{55}$$

If Rank$(\mathbf{x}_{[1,b]}) = d$, then det$(\Phi_{[1,b]}) = det(\mathbf{x}_{[1,b]})^2 \leq (\frac{bL^2}{d})^d$; else Rank$(\mathbf{x}_{[1,b]}) < d$, then det$(\Phi_{[1,b]}) = 0$.

D Rank Swinging Phenomenon Examples

- Case 1: Introduce the minimum regret, then FRT = 0.
- Case 2: Introduce the maximum regret, FRT = 2.
- Case 3: Introduce medium regret scenario I, FRT = 1.
- Case 4: Introduce medium regret scenario II, FRT = 1.

$$\mathcal{M}_1=\begin{pmatrix}0&0&0&1\\0&1&0&0\\0&0&1&0\\0&0&0&1\\v_1&v_2&v_3&v_4\end{pmatrix}\mathcal{M}_2=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\0&1&0&0\end{pmatrix}\mathcal{M}_3=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&1&0&0\\0&0&0&1\\1&0&0&0\end{pmatrix}\mathcal{M}_4=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&1&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix} \tag{56}$$

Case 1. (Minimum Regret) *Rank$(\Phi_{[t-s,t-1]}) < d$ and the delete term $x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} = 0$. So it won't introduce any extra regret no matter what the new data x_t is. So FRT = 0.*

$x_{t-s}{}^2_{\Phi^{-1}_{[t-s,t-1]}} = 0$ means that the old data x_{t-s} can be fully represented by the original data memory, so delete it won't influence the representation ability of data memory. For example, $\mathcal{M}_1$ in (56), if $s = 4, d = 4$, rank$(\mathbf{x}_{[1,4]}) = 3$, since x_1 and x_4 are linearly correlated. By the FIFD scheme, the forgetting data is $x_1 = (0, 0, 0, 1)^{\mathsf{T}}$, so $x_1{}^2_{\Phi^{-1}_{[1,4]}} = 0$

Case 2. (Maximum Regret) *If the old data and the new data are totally dissimilar $\sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) = 1$ and $x_t{}^2_{\Phi^{-1}_{[t-s,t-1]}} = 1$, then the weight difference term $x_t{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) + 1 = 2$. So FRT = 2.*

which means that x_t is totally dissimilar with x_{t-s} with respect to $\Phi^{-1}_{[t-s,t-1]}$, but x_t can be represented by the rest of data memory from time window $[t - s + 1, t - 1]$, and the gram matrix's rank will decrease by 1. For example, $\mathcal{M}_2$ in (56), $x_1 = (1, 0, 0, 0)^{\mathsf{T}}$ and $x_5 = (0, 1, 0, 0)^{\mathsf{T}}$. So Rank$(\Phi_{[2,5]})$ = Rank$(\Phi_{[1,4]}) - 1$. $x_5{}^2_{\Phi^{-1}_{[1,4]}} \sin^2(\theta, \Phi^{-1}_{[1,4]}) = 1$.

Case 3. (Medium Regret I) *If the old data and the new data are perfectly similar $\sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) = 0$, then $x_t{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) = 0$. So the weight difference term $x_t{}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) + 1 = 1$.So FRT = 1.*

which means that x_t is perfectly simliar with x_{t-s} with respect to $\Phi^{-1}_{[t-s,t-1]}$. For example, $\mathcal{M}_3$ in (56), $x_1 = (1,0,0,0)^\mathsf{T}$ and $x_5 = (1,0,0,0)^\mathsf{T}$, $\text{Rank}(\Phi_{[2,5]}) = \text{Rank}(\Phi_{[1,4]})$, $\sin^2_{\Phi^{-1}_{[1,4]}}\theta = 0$, ${x_5}^2_{\Phi^{-1}_{[1,4]}} \sin^2(\theta, \Phi^{-1}_{[1,4]}) = 0$. Then the weight difference term ${x_5}^2_{\Phi^{-1}_{[1,4]}} \sin^2(\theta, \Phi^{-1}_{[1,4]}) + 1 = 1$ no matter how large ${x_t}^2_{\Phi^{-1}_{[1,4]}}$ is.

Case 4. (Medium Regret II) *If the new data does not lie in the space generated by $\mathbf{x}_{[t-s,t-1]}$, which means that ${x_t}^2_{\Phi^{-1}_{[t-s,t-1]}} = 0$, then ${x_t}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) = 0$. So the weight difference term ${x_t}^2_{\Phi^{-1}_{[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{[t-s,t-1]}) + 1 = 1$. So FRT = 1.*

For example, $\mathcal{M}_4$ in (56), $x_1 = (1,0,0,0)^\mathsf{T}$ and $x_5 = (0,0,1,0)^\mathsf{T}$, $\text{Rank}(\Phi_{[2,5]}) = \text{Rank}(\Phi_{[1,4]}) + 1$, ${x_5}^2_{\Phi^{-1}_{[1,4]}} = 0$, ${x_5}^2_{\Phi^{-1}_{[1,4]}} \sin^2(\theta, \Phi^{-1}_{[1,4]}) = 1$. Then the weight difference term ${x_5}^2_{\Phi^{-1}_{[1,4]}} \sin^2(\theta, \Phi^{-1}_{[1,4]}) + 1 = 1$.

E FIFD-Adaptive Ridge Regret

(Regret Upper Bound of The FIFD-Adaptive Ridge algorithm) *The same assumption in Theorem **??** and if Lemma **??** holds, with probability at least $1-\delta$, , the cumulative regret satisfies:*

$$R_{T,s}(\mathcal{A}_{Ridge}) \leq \sigma\kappa\nu\zeta_\lambda \sqrt{(d/s)(T-s)[\eta_{Ridge} + (T-s)]} \tag{57}$$

where $\zeta_\lambda = \max_{s+1\leq t\leq T} \frac{\mathbf{x}_{[t-1,t-s]\infty}}{\phi^2_{\lambda,[t-s,t-1]}}$ is the maximum adaptive constant over time, $\eta_{Ridge} = d\log(sL^2/d + \lambda_{[T-s,T-1]}) - \log C_2(\boldsymbol{\phi})$ is a constant related to the last data memory, $C_2(\boldsymbol{\phi}) = \prod_{t=s+1}^{T}(1 + \frac{s}{\phi^2_{\lambda,[t-s+1,t]}+\lambda_{\Delta,[t-s+1,t]}}\lambda_{\Delta,[t-s+1,t]})$ is a constant close to 1, and $\lambda_{\Delta,[t-s+1,t]} = \lambda_{[t-s+1,t]} - \lambda_{[t-s,t-1]}$ represents the fluctuation of λ over time steps.

Proof. Here we use $l_{\lambda,t}$ denote the instantaneous *absolute loss* at time t using `FIFD-Adaptive Redge` algorithm. Let's decompose the instantaneous absolute loss as follows:

$$\begin{aligned}
l_{\lambda,t} &= |\langle \hat{\theta}_{\lambda,[t-s,t-1]}, x_t\rangle - \langle \theta_\star, x_t\rangle| \\
&= |\langle \hat{\theta}_{\lambda,[t-s,t-1]} - \theta_\star, x_t\rangle| \\
&= |[\hat{\theta}_{\lambda,[t-s,t-1]} - \theta_\star]^\mathsf{T} x_t| \\
&= |[\hat{\theta}_{\lambda,[t-s,t-1]} - \theta_\star]^\mathsf{T} \Phi^{\frac{1}{2}}_{\lambda,[t-s,t-1]} \Phi^{-\frac{1}{2}}_{\lambda,[t-s,t-1]} x_t| \\
&= |[\hat{\theta}_{\lambda,[t-s,t-1]} - \theta_\star]^\mathsf{T} \Phi^{\frac{1}{2}}_{\lambda,[t-s,t-1]}| \times |\Phi^{-\frac{1}{2}}_{\lambda,[t-s,t-1]} x_t| \\
&\leq \hat{\theta}_{\lambda,[t-s,t-1]} - \theta_{\star_{\Phi_{\lambda,[t-s,t-1]}}} x_{t_{\Phi^{-1}_{\lambda,[t-s,t-1]}}} \\
&\leq \sqrt{\beta_{\lambda,[t-s,t-1]}(\delta)} x_{t_{\Phi^{-1}_{\lambda,[t-s,t-1]}}}
\end{aligned} \tag{58}$$

where the last step from Lemma B. To compute the regret, we first denote $r_{\lambda,t}$ as the instantaneous regret at time t. Let's decompose the instantaneous regret as follows, $r_{\lambda,t} = \langle \hat{\theta}_{\lambda,[t-s,t-1]}, x_t\rangle - \langle \theta_\star, x_t\rangle \leq \sqrt{\beta_{\lambda,[t-s,t-1]}(\delta)} x_{t_{\Phi^{-1}_{\lambda,[t-s,t-1]}}}$, where the inequality is from Lemma B.

Thus, with probability at least $1-\delta$, for all $T > s$,

$$\begin{aligned}
R_{T,s}(\mathcal{A}) &= \sqrt{(T-s)\sum_{t=s+1}^{T} r^2_{\lambda,t}} \\
&\leq \sqrt{(T-s)\sum_{t=s+1}^{T} \beta_{\lambda,[t-s,t-1]}(\delta) {x_t}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}}} \\
&\leq \sqrt{2(T-s)\max_{s+1\leq t\leq T}\beta_{\lambda,[t-s,t-1]}(\delta)\left(2\left[d\log(\frac{sL^2}{d} + \lambda_{[T-s,T-1]}) - \log C_2(\boldsymbol{\phi}) + (T-s)\right]\right)}
\end{aligned} \tag{59}$$

where the last step we use Lemma E to deal with the online forgetting process to get the summation of term $x_{t\,\Phi^{-1}_{\lambda,[t-s,t-1]}}^2$. for each time step. So we have the cumulative regret upper bound for the algorithm `FIFD-Adaptive Ridge` as follows,

$$R_{T,s}(\mathcal{A}_{\text{Ridge}}) = \sigma\kappa\nu\zeta_\lambda\sqrt{(d/s)(T-s)[\eta_{\text{Ridge}} + (T-s)]} \tag{60}$$

where we use the Lemma B to get the maximum ellipsoid confidence $\max_{t\in[s+1,T]} \beta_{\lambda,[t-s,t-1]}(\delta)$. □

The cumulative regret of FIFD-Adaptive Ridge is partially determined by FRT-Ridge at each time step,

$$\sum_{t=s+1}^{T} x_{t\,\Phi^{-1}_{\lambda,[t-s,t-1]}}^2 \leq 2\eta_{Ridge} + \sum_{t=s+1}^{T} FRT_{\lambda,[t-s,t-1]} \leq 2\left[d\log(\frac{sL^2}{d} + \lambda_{[T-s,T-1]}) - \log C_2(\boldsymbol{\phi}) + (T-s)\right] \tag{61}$$

where $\eta_{Ridge} = d\log(sL^2/d + \lambda_{[T-s,T-1]}) - \log C_2(\boldsymbol{\phi})$ *is a constant based on the limited data time window*$[T-s,T]$.

Proof. Here we still use $a = t-s, b = t-1$ for the sake of simplicity. Elementary algebra gives

$$\det(\Phi_{\lambda,[a+1,b+1]}) = \det(\Phi_{\lambda,[a,b]} + x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T} + \lambda_{\Delta,[a+1,b+1]}), \tag{62}$$

where $\lambda_{\Delta,[a+1,b+1]} = \lambda_{[a+1,b+1]} - \lambda_{[a,b]}$ and the choice of $\lambda_{[a+1,b+1]}$ $\lambda_{[a,b]}$ is given by Lemma **??**. Then $\det(\Phi_{\lambda,[a+1,b+1]})$ equals to

$$\begin{aligned}
&= \det(\Phi^{1/2}_{\lambda,[a,b]}(\mathbf{I} + \Phi^{-1/2}_{\lambda,[a,b]}(x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T} + \lambda_{\Delta,[a+1,b+1]}\Phi^{-1}_{\lambda,[a,b]})\Phi^{-1/2}_{\lambda,[a,b]})\Phi^{1/2}_{\lambda,[a,b]}) \\
&= \det(\Phi_{\lambda,[a,b]})\det(\mathbf{I} + \Phi^{-1/2}_{\lambda,[a,b]}(x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T})\Phi^{-1/2}_{\lambda,[a,b]} + \lambda_{\Delta,[a+1,b+1]}\Phi^{-1}_{\lambda,[a,b]}) \\
&= \det(\Phi_{\lambda,[a,b]})\det(\mathbf{I} + \Phi^{-1/2}_{\lambda,[a,b]}x_{b+1}x_{b+1}^\mathsf{T}\Phi^{-1/2}_{\lambda,[a,b]} - \Phi^{-1/2}_{\lambda,[a,b]}x_a x_a^\mathsf{T}\Phi^{-1/2}_{\lambda,[a,b]} + \lambda_{\Delta,[a+1,b+1]}\Phi^{-1}_{\lambda,[a,b]}) \\
&= \det(\Phi_{\lambda,[a,b]})\det\left(\mathbf{I} + (\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})(\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})^\mathsf{T} - (\Phi^{-1/2}_{\lambda,[a,b]}x_a)(\Phi^{-1/2}_{\lambda,[a,b]}x_a)^\mathsf{T} + \lambda_{\Delta,[a+1,b+1]}\Phi^{-1}_{\lambda,[a,b]}\right).
\end{aligned} \tag{63}$$

We first compute the second determinant of equation (63) and denote $\mathbf{B} = \mathbf{I} + (\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})(\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})^\mathsf{T} - (\Phi^{-1/2}_{\lambda,[a,b]}x_a)(\Phi^{-1/2}_{\lambda,[a,b]}x_a)^\mathsf{T}$ and use the matrix technique $\det(\mathbf{A}+\epsilon\mathbf{X}) \approx \det(\mathbf{A}) + \det(\mathbf{A})\text{tr}(\mathbf{A}^{-1}\mathbf{X})\epsilon + \mathcal{O}(\epsilon^2)$. So we have

$$\det(\mathbf{B} + \lambda_{\Delta,[a+1,b+1]}\Phi^{-1}_{\lambda,[a,b]}) \approx \det(\mathbf{B}) + \det(\mathbf{B})\text{tr}(A^{-1}\Phi^{-1}_{\lambda,[a,b]})\lambda_{\Delta,[a+1,b+1]} + \mathcal{O}(\lambda^2_{\Delta,[a+1,b+1]}) \tag{64}$$

So the second determinant part of equation becomes

$$\begin{aligned}
&\det(\mathbf{B} + \lambda_{\Delta,[a+1,b+1]}\Phi^{-1}_{\lambda,[a,b]}) \\
&\approx \det(\mathbf{B})\left[1 + \text{tr}(\mathbf{B}^{-1}\Phi^{-1}_{\lambda,[a,b]})\lambda_{\Delta,[a+1,b+1]}\right] \\
&= \det(\mathbf{B})\left[1 + \text{tr}[(\Phi_{\lambda,[a,b]}\mathbf{B})^{-1}]\lambda_{\Delta,[a+1,b+1]}\right] \\
&= \det(\mathbf{B})\left[1 + \text{tr}[(\Phi_{\lambda,[a,b]} + x_{b+1}x_{b+1}^\mathsf{T} - x_a x_a^\mathsf{T})^{-1}]\lambda_{\Delta,[a+1,b+1]}\right] \\
&= \det(\mathbf{B})\left[1 + \text{tr}[(\Phi_{\lambda,[a+1,b+1]} - \lambda_{\Delta,[a+1,b+1]}\mathbf{I})^{-1}]\lambda_{\Delta,[a+1,b+1]}\right] \\
&\geq \det(\mathbf{B})\left[1 + \frac{d}{\phi^2_{\lambda,[a+1,b+1]} - \lambda_{\Delta,[a+1,b+1]}}\lambda_{\Delta,[a+1,b+1]}\right]
\end{aligned} \tag{65}$$

where the last step we use the minimum eigenvalue of $\Phi_{\lambda,[a+1,b+1]}$ to get the inequality and we denote $c_{2,[a+1,b+1]} = 1 + \frac{d}{\phi^2_{\lambda,[a+1,b+1]} - \lambda_{\Delta,[a+1,b+1]}}\lambda_{\Delta,[a+1,b+1]}$. So when we go back to equation (63), $\det(\Phi_{\lambda,[a+1,b+1]})$ equals to

$$\begin{aligned}
&= \det(\Phi_{\lambda,[a,b]})\det\left(\mathbf{I} + (\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})(\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1})^\mathsf{T} - (\Phi^{-1/2}_{\lambda,[a,b]}x_a)(\Phi^{-1/2}_{\lambda,[a,b]}x_a)^\mathsf{T}\right)c_{2,[a+1,b+1]} \\
&= \det(\Phi_{\lambda,[a,b]})\left[(1 + \Phi^{-1/2}_{\lambda,[a,b]}x_{b+1}{}^2)(1 - \Phi^{-1/2}_{\lambda,[a,b]}x_a{}^2) + \langle\Phi^{-1/2}_{\lambda,[a,b]}x_{b+1}, \Phi^{-1/2}_{\lambda,[a,b]}x_a\rangle^2\right]c_{2,[a+1,b+1]}.
\end{aligned} \tag{66}$$

where the last step use lemma C, same as the technique using in obtaining the regret upper bound of FIFD-OLS. So we have

$$\begin{aligned}
&= \det(\Phi_{\lambda,[a,b]}) \left((1 + x_{b+1}{}^2_{\Phi^{-1}_{\lambda,[a,b]}})(1 - x_a{}^2_{\Phi^{-1}_{\lambda,[a,b]}}) + \langle x_{b+1}, x_a \rangle^2_{\Phi^{-1}_{\lambda,[a,b]}} \right) c_{2,[a+1,b+1]} \\
&= \det(\Phi_{\lambda,[a,b]}) \left(1 + x_{b+1}{}^2_{\Phi^{-1}_{\lambda,[a,b]}} - x_a{}^2_{\Phi^{-1}_{\lambda,[a,b]}} + \langle x_{b+1}, x_a \rangle^2_{\Phi^{-1}_{\lambda,[a,b]}} - x_{b+1}{}^2_{\Phi^{-1}_{\lambda,[a,b]}} x_a{}^2_{\Phi^{-1}_{\lambda,[a,b]}} \right) c_{2,[a+1,b+1]} \\
&= \det(\Phi_{\lambda,[a,b]}) \left(1 + x_{b+1}{}^2_{\Phi^{-1}_{\lambda,[a,b]}} - x_a{}^2_{\Phi^{-1}_{\lambda,[a,b]}} + x_{b+1}{}^2_{\Phi^{-1}_{\lambda,[a,b]}} x_a{}^2_{\Phi^{-1}_{\lambda,[a,b]}} (\cos^2_{\Phi^{-1}_{\lambda,[a,b]}} \theta - 1) \right) c_{2,[a+1,b+1]}.
\end{aligned} \tag{67}$$

where $\cos_{\Phi^{-1}_{\lambda,[a,b]}} \theta = \frac{\langle x_{b+1}, x_a \rangle_{\Phi^{-1}_{\lambda,[a,b]}}}{x_{b+1}{}_{\Phi^{-1}_{\lambda,[a,b]}} x_a{}_{\Phi^{-1}_{\lambda,[a,b]}}}$, which measures the similarity of the vector of x_{b+1} and x_a with respect to $\Phi^{-1}_{\lambda,[a,b]}$. If $\cos^2_{\Phi^{-1}_{\lambda,[a,b]}} \theta = 1$, that means the incoming data x_{b+1} and the deleted data x_a are same. If $\cos^2_{\Phi^{-1}_{\lambda,[a,b]}} \theta = 0$, that means the incoming data and the deleted data are totally different with respect to $\Phi^{-1}_{\lambda,[a,b]}$.

Now we switch back to the notation time step t and $t - s = a, t - 1 = b$. At time step T, and we use equation (42), we have

$$\begin{aligned}
\det(\Phi_{\lambda,[T-s,T]}) = \prod_{t=s+1}^{T} (1 &+ x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \\
&+ x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} (\cos^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta - 1)) c_{2,[t-s+1,t]}.
\end{aligned} \tag{68}$$

Taking log to both sides of equation (68), we can get

$$\begin{aligned}
&\log(\det(\Phi_{\lambda,[T-s,T]})) \\
&= \log(C_2(\phi)) + \sum_{t=s+1}^{T} \log(1 + x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \\
&\qquad + x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} (\cos^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta - 1)),
\end{aligned} \tag{69}$$

where $C_2(\phi) = \prod_{t=s+1}^{T} c_{2,[t-s+1,t]}$. Combining the inequality of $\log(1 + x) > \frac{x}{1+x}$ which holds when $x > -1$, we first consider each part of the product and use the dissimilarity measure $\sin^2_{\Phi^{-1}_{\lambda,[a,b]}} \theta = 1 - \cos^2_{\Phi^{-1}_{\lambda,[a,b]}} \theta$. So we get

$$\begin{aligned}
&\log \left(1 + x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta \right) \\
&> \frac{x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta}{1 + x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta} \\
&> \frac{1}{2} \left[x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} - x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta \right].
\end{aligned} \tag{70}$$

Therefore, we can provide a upper bound of the term $\sum_{t=s+1}^{T} x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}}$ combing equation (68)(69)(70),

$$\begin{aligned}
\sum_{t=s+1}^{T} x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \leq 2 \left[\log(\det(\Phi_{\lambda,[T-s,T]})) - \log(C_2(\phi)) \right] &+ \sum_{t=s+1}^{T} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \\
&+ \sum_{t=1}^{T-s} x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \theta.
\end{aligned} \tag{71}$$

Thus by combining the last two terms and extracting $x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}}$, we can get

$$\begin{aligned}&\sum_{t=s+1}^{T} x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \\ \leq& 2\left[\log(\det(\Phi_{\lambda,[T-s,T]})) - \log(C_2(\phi))\right] + \sum_{t=1}^{T-s} x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}}(x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}}\theta + 1).\end{aligned} \tag{72}$$

To make the formula simpler, we define *'Forgetting Regret Term' (FRT)* term with respect to adaptive ridge parameter λ, at time window $[t-s, t-1]$ or called at time step t as follows,

$$\mathrm{FRT}_{\lambda,[t-s,t-1]} = x_{t-s}{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}}(x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \sin^2(\theta, \Phi^{-1}_{\lambda,[t-s,t-1]}) + 1), \tag{73}$$

where $\mathrm{FRT}_{\lambda,[t-s,t-1]} \in [0, 2]$. The detailed explanation and examples of *'Forgetting Regret Term' (FRT)* can be found in D.

So equation (72) becomes

$$\sum_{t=s+1}^{T} x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \leq 2\eta_{\mathrm{Ridge}} + \sum_{t=s+1}^{T} \mathrm{FRT}_{\lambda,[t-s,t-1]}, \tag{74}$$

where $\eta_{\mathrm{Ridge}} = \log(\det(\Phi_{\lambda,[T-s,T]})) - \log(C_2(\phi))$ is a constant based on data time window$[T-s, T]$. By Lemma C , we get

$$\sum_{t=s+1}^{T} x_t{}^2_{\Phi^{-1}_{\lambda,[t-s,t-1]}} \leq 2\left[d\log(\frac{sL^2}{d} + \lambda_{[T-s,T-1]}) - \log C_2(\boldsymbol{\phi}) + (T-s)\right]. \tag{75}$$

□

F Online Incremental Update for FIFD-OLS

The `FIFD-OLS` $\hat{\theta}_{[a,b]}$ estimator based on data from decision point a to b is defined as

$$\hat{\theta}_{[a,b]} = \Phi_{[a,b]}^{-1}\big[\sum_{i=a}^{b} y_i x_i\big]. \tag{76}$$

We present an incremental update formula from $\hat{\theta}_{[a,b]}$ to $\hat{\theta}_{[a+1,b+1]}$: [Incremental Update for length s `FIFD`-least square estimator]

$$\hat{\theta}_{[a+1,b+1]} = f(\Phi_{[a,b]}^{-1}) \cdot g(\hat{\theta}_{[a,b]}, \Phi_{[a,b]}), \tag{77}$$

where $f(A)$ is defined as

$$f(A) = \Gamma(A) - (x_a^\top \Gamma - 1)^{-1}\big[\Gamma(A) x_a x_a^\top \Gamma(A)\big] \tag{78}$$

with $\Gamma(A) \equiv A - (x_{b+1}^\top A x_{b+1} + 1)^{-1}\big[A x_{b+1} x_{b+1}^\top A\big]$ and $g(\theta)$ is defined as

$$g(\theta, \Phi) = \Phi\theta + y_{b+1}x_{b+1} - y_a x_a. \tag{79}$$

Proof. We break the proof into 2 steps. The first step is to update the inverse of sample covariance matrix from $\Phi_{[a,b]}^{-1}$ to $\Phi_{[a+1,b+1]}^{-1}$. The second step is simple algebra and the definition of least square estimator (76).

Step 01. Bases on Lemmas F and F, we can do incremental update on the inverse of sample covariance matrix from $\Phi_{[a,b]}^{-1}$ to $\Phi_{[a+1,b+1]}^{-1}$ as

$$\begin{cases} \Phi_{[a+1,b+1]}^{-1} & = \Gamma(\Phi_{[a,b]}^{-1}) - (x_a^\top \Gamma(\Phi_{[a,b]}^{-1}) - 1)^{-1}\big[\Gamma(\Phi_{[a,b]}^{-1}) x_a x_a^\top \Gamma(\Phi_{[a,b]}^{-1})\big] \\ \Gamma(\Phi_{[a,b]}^{-1}) & = \Phi_{[a,b]}^{-1} - (x_{b+1}^\top \Phi_{[a,b]}^{-1} x_{b+1} + 1)^{-1}\big[\Phi_{[a,b]}^{-1} x_{b+1} x_{b+1}^\top \Phi_{[a,b]}^{-1}\big] \end{cases}. \tag{80}$$

We write $\Phi_{[a+1,b+1]}^{-1} = f(\Phi_{[a,b]}^{-1})$, where the function $f(\cdot)$ is defined by the update scheme (80).

Step 02. Put into (76) that $\sum_{i=a+1}^{b+1} y_i x_i = \Phi_{[a,b]}\hat{\theta}_{[a,b]} + y_{b+1}x_{b+1} - y_a x_a$. □

[Matrix Inversion Formula]

$$[A + BCD]^{-1} = A^{-1} - A^{-1}B[DA^{-1}B + C^{-1}]^{-1}DA^{-1}. \tag{81}$$

[Update Scheme-Step 01]

$$\Phi_{[a,b+1]}^{-1} = \Phi_{[a,b]}^{-1} - (x_{b+1}^\top \Phi_{[a,b]}^{-1} x_{b+1} + 1)^{-1}\big[\Phi_{[a,b]}^{-1} x_{b+1} x_{b+1}^\top \Phi_{[a,b]}^{-1}\big]. \tag{82}$$

Proof. Note $\Phi_{[a,b+1]} = \Phi_{[a,b]} + x_{b+1}x_{b+1}^\top$. Take $A = \Phi_{[a,b]}, B = x_{b+1}, C = 1, D = x_{b+1}^\top$ in Lemma F. □

[Update Scheme-Step 02]

$$\Phi_{[a+1,b+1]}^{-1} = \Phi_{[a,b+1]}^{-1} - (x_a^\top \Phi_{[a,b+1]}^{-1} x_a - 1)^{-1}\big[\Phi_{[a,b+1]}^{-1} x_a x_a^\top \Phi_{[a,b+1]}^{-1}\big]. \tag{83}$$

Proof. Note $\Phi_{[a+1,b+1]} = \Phi_{[a,b+1]} - x_a x_a^\top$. Take $A = \Phi_{[a,b+1]}, B = x_a, C = -1, D = x_a^\top$ in Lemma F. □

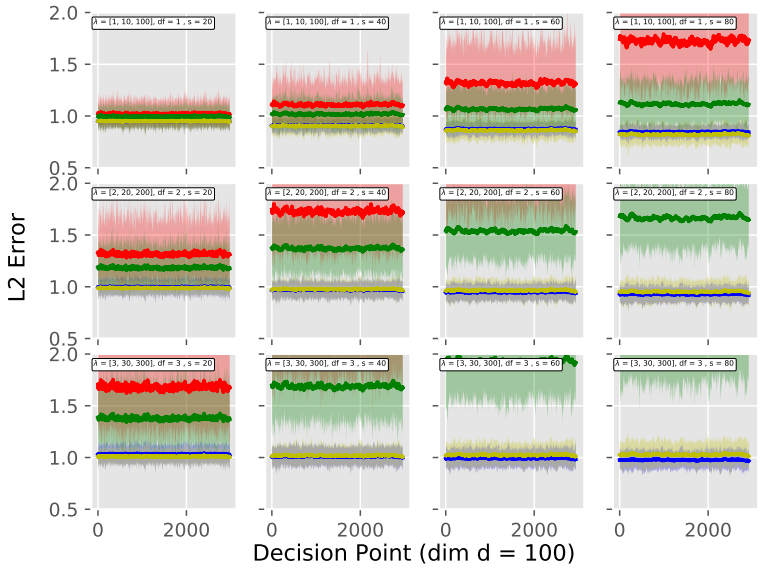

Figure 1: Comparison of L_2 error between `FIFD-Adaptive Ridge` method and Fixed Ridge method. The error bars represent the standard error of the mean regret over 100 runs.

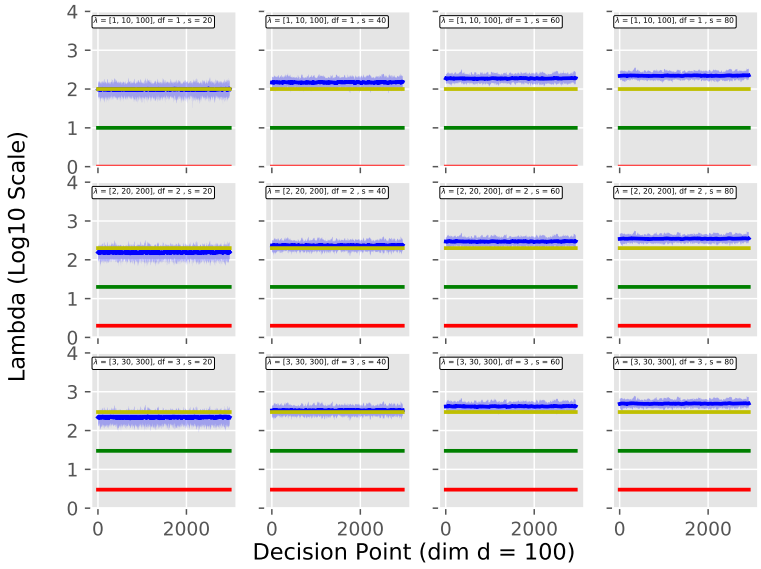

Figure 2: The choice of `FIFD-Adaptive Ridge` and Fixed Ridge λ.

G Additional Simulation Results

In Figure 1, we show the L_2 error of the `FIFD-Adaptive Ridge` method and the Fixed ridge method. From all of the twelve subplots, we can see that L_2 error of the adaptive ridge method can be bounded by 1. As we can see, as the noise level σ increases, the L_2 error increases. If we increase the constant memory limit s, the L_2 error will decrease. Besides, we can see the error bar of the `FIFD-Adaptive Ridge` is relative narrower than the Fixed ridge over all settings. Without any prior knowledge, we can achieve the best or close to the best result compared with the Fixed ridge method with prior knowledge of λ.

In Figure 2, we show the choice of hyperparameter λ over different time steps. Since the hyperparameter λ is calculated over each time interval $[t-s, t-1], \forall t \in [s+1, T]$. Thus, it is adaptive to the incoming data compared with Fixed ridge method with pre-defined hyperparameter λ.