A Proof of Theorem 4.1

A.1 Preliminaries

A.1.1 Useful concentration

Our proof will require applying the following concentration inequality, derived from Azuma's inequality:

Lemma A.1. *Let $W_1, \ldots, W_\tau$ be random variables in $\mathbb{R}$ such that $|W_t| \leq W_{max}$. Suppose for all $t \in [\tau]$, for all $w_1, \ldots, w_{t-1}$,*

$$\mathbb{E}\left[W_t | W_{t-1} = w_{t-1}, \ldots, W_1 = w_1\right] = 0.$$

Then, with at least $1 - \delta$,

$$\left| \sum_{t=1}^{\tau} W_t \right| \leq W_{max} \sqrt{2\tau \log(2/\delta)}.$$

Proof. This is a reformulated version of Azuma's inequality. To see this, define

$$Z_t = \sum_{i=1}^{t} W_i \quad \forall t,$$

and initialize $Z_0 = 0$. We start by noting that for all $t \in [\tau]$, since

$$Z_t = \sum_{i=1}^{t} W_i = W_t + \sum_{i=1}^{t-1} W_i = W_t + Z_{t-1},$$

we have

$$\begin{aligned} \mathbb{E}\left[Z_t | Z_{t-1}, \ldots, Z_1\right] &= \mathbb{E}\left[W_t | Z_{t-1}, \ldots, Z_1\right] + \mathbb{E}\left[Z_{t-1} | Z_{t-1}, \ldots, Z_1\right] \\ &= \mathbb{E}\left[W_t | Z_{t-1}, \ldots, Z_1\right] + Z_{t-1}. \end{aligned}$$

Further, it is easy to see that $Z_i = z_i \ \forall i \in [t-1]$ if and only if $W_i = z_i - z_{i-1} \ \forall i \in [t-1]$, hence

$$\mathbb{E}\left[W_t | Z_{t-1} = z_{t-1}, \ldots, Z_1 = z_1\right] = \mathbb{E}\left[W_t | W_i = z_i - z_{i-1} \ \forall i \in [t-1]\right] = 0.$$

Combining the last two equations implies that

$$\mathbb{E}\left[Z_t | Z_{t-1}, \ldots, Z_1\right] = Z_{t-1},$$

and the Z_t's define a martingale. Since for all t,

$$|Z_t - Z_{t-1}| = |W_t| \leq W_{max},$$

we can apply Azuma's inequality to show that with probability at least $1 - \delta$,

$$|Z_\tau - Z_0| \geq W_{max} \sqrt{2\tau \log(2/\delta)},$$

which immediately gives the result. ∎

A.1.2 Sub-space decomposition and projection

We will also need to divide $\mathbb{R}^d$ in several sub-spaces, and project our observations to said subspaces.

Sub-space decomposition We focus on the sub-space generated by the non-modified features x_t's and the sub-space generated by the feature modifications Δ_t's. We let r be the rank of Σ, and let $\lambda_r \geq \ldots \geq \lambda_1 > 0$ be the non-zero eigenvalues of Σ. Further, we let $f_1, \ldots, f_r$ be the unit eigenvectors (i.e., such that $\|f_1\|_1 = \ldots = \|f_r\|_1 = 1$) corresponding to eigenvalues $\lambda_1, \ldots, \lambda_r$ of Σ. As Σ is a symmetric matrix, $f_1, \ldots, f_r$ are orthonormal. We abuse notations in the proof of Theorem 4.1 and denote $\Sigma = \text{span}(f_1, \ldots, f_r)$ when clear from context.

For all k, let e_k be the unit vector such that $e_k(k) = 1$ and $e_k(j) = 0 \ \forall j \neq k$. At time τ, we denote $\mathcal{D}_\tau = \text{span}\left(e_k\right)_{k \in D_\tau}$ the sub-space of $\mathbb{R}^d$ spanned by the features in D_τ.

Finally, we let

$$\mathcal{V}_\tau = \Sigma + \mathcal{D}_\tau = \text{span}\left(f_1, \ldots, f_r\right) + \text{span}\left(e_k\right)_{k \in D_\tau}$$

be the Minkowski sum of sub-spaces Σ and $\mathcal{D}_\tau$.

Projection onto sub-spaces For any vector z, sub-space $\mathcal{H}$ of $\mathbb{R}^d$, we write $z = z(\mathcal{H}) + z(\mathcal{H}^\perp)$ where $z(\mathcal{H})$ is the projection of z onto sub-space $\mathcal{H}$, i.e. is uniquely defined as

$$z(\mathcal{H}) = \sum_{q \in B} (z^\top q) q$$

for any orthonormal basis B of $\mathcal{H}$. We also let $z(\mathcal{H}^\perp)$ be the projection on the orthogonal complement $\mathcal{H}^\perp$. In particular, $z(\mathcal{H})$ is orthogonal to $z(\mathcal{H}^\perp)$. Further, we write $\bar{X}_\tau(\mathcal{H})$ the matrix whose rows are given by $\bar{x}_t(\mathcal{H})^\top$ for all $t \in [\tau]$.

A.2 Main Proof

Characterization of the least-square estimate via first-order conditions First, for any least square solution $\hat{\beta}_E$ at time $\tau(E)$, we write the first order conditions solved by $\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)$, the projection of $\hat{\beta}_E$ on sub-space $\mathcal{V}_{\tau(E)}$. We abuse notations to let $\varepsilon_{\tau(E)} \triangleq (\varepsilon_t)_{t \in [\tau(E)]}$ the vector of all ε_t's up until time $\tau(E)$, and state the result as follows:

Lemma A.2 (First-order conditions projected onto $\mathcal{V}_{\tau(E)}$)**.** *Suppose $\hat{\beta}_E \in LSE(\tau(E))$. Then,*

$$\left(\bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) = \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)}.$$

Proof. For simplicity of notations, we drop all $\tau(E)$ indices and subscripts in this proof. Remember that

$$LSE = \underset{\beta}{\operatorname{argmin}} \left(\bar{X}\beta - \bar{Y}\right)^\top \left(\bar{X}\beta - \bar{Y}\right).$$

Since $\hat{\beta}_E \in LSE$, it must satisfy the first order conditions given by

$$2\bar{X}^\top\left(\bar{X}\hat{\beta}_E - \bar{Y}\right) = 0,$$

which can be rewritten as

$$\bar{X}^\top \bar{X}\hat{\beta}_E = \bar{X}^\top \bar{Y}.$$

Second, we note that for all t, $x_t \in \operatorname{span}(f_1, \ldots, f_r)$ and $\Delta_t \in \operatorname{span}\left((e_k)_{k \in D}\right)$ (by definition of D). This immediately implies, in particular, that $\bar{x}_t = x_t + \Delta_t \in \mathcal{V}$. In turn, $\bar{x}_t\left(\mathcal{V}\right) = \bar{x}_t$ for all t, and

$$\bar{X} = \bar{X}\left(\mathcal{V}\right).$$

As such, the first order condition can be written

$$\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)\hat{\beta}_E = \bar{X}\left(\mathcal{V}\right)^\top \bar{Y}.$$

Now, we remark that

$$\begin{aligned}
\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)\hat{\beta}_E &= \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right)\bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E \\
&= \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right)\bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E\left(\mathcal{V}\right) + \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right)\bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E(\mathcal{V}^\perp) \\
&= \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right)\bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E\left(\mathcal{V}\right) \\
&= \bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)\hat{\beta}_E\left(\mathcal{V}\right),
\end{aligned}$$

where the second-to-last equality follows from the fact that $\mathcal{V}$ and $\mathcal{V}^\perp$ are orthogonal, which immediately implies $\bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E(\mathcal{V}^\perp) = 0$ for all t. To conclude the proof, we note that $\bar{Y} = \bar{X}^\top \beta^* + \varepsilon = \bar{X}\left(\mathcal{V}\right)^\top \beta^*\left(\mathcal{V}\right) + \varepsilon$. Plugging this in the above equation, we obtain that

$$\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)\hat{\beta}_E\left(\mathcal{V}\right) = \bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)^\top \beta^*\left(\mathcal{V}\right) + \bar{X}\left(\mathcal{V}\right)^\top \varepsilon.$$

This can be rewritten

$$\left(\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}\right) - \beta^*\left(\mathcal{V}\right)\right) = \bar{X}\left(\mathcal{V}\right)^\top \varepsilon,$$

which completes the proof. □

Upper-bounding the right-hand side of the first order conditions We now use concentration to give an upper bound on a function of the right-hand side of the first order conditions,

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)}.$$

Lemma A.3. *With probability at least* $1-\delta$*,*

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon \\ &\leq \left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \cdot K'\sqrt{d\tau(E)\log(2d/\delta)}.\end{aligned}$$

where K' *is a constant that only depends on the distribution of costs and the bound* σ *on the noise.*

Proof. Pick any $k \in [d]$, and define $W_t = \bar{x}_t(k)\varepsilon_t$. First, we remark that

$$|\bar{x}_t(k)| \leq |x_t(k)| + |\Delta_t(k)| \leq 1 + \max_{k\in[d],\ i\in[l]} \frac{B^i}{c^i(k)}.$$

In turn, $|W_t| \leq K'$ where

$$K' \triangleq \left(1 + \max_{k\in[d],\ i\in[l]} \frac{B^i}{c^i(k)}\right)\sigma.$$

Further, note that both $x_t(k)$ and ε_t are independent of the history of play up through time $t-1$, hence of $W_1,\ldots,W_{t-1}$, and that ε_t is further independent of Δ_t (the distribution of Δ_t is a function of the currently posted $\hat{\beta}_{E-1}$ only, which only depends on the previous time steps). Noting that if A, B, C are random variables, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{A,B}[AB|C=c] &= \sum_a\sum_b ab\Pr[A=a, B=b|C=c] \\ &= \sum_a\sum_b ab\Pr[A=a|B=b,C=c]\Pr[B=b|C=c] \\ &= \sum_b b\left(\sum_a a\Pr[A=a|B=b,C=c]\right)\Pr[B=b|C=c] \\ &= \sum_b b\mathop{\mathbb{E}}_A[A|B=b,C=c]\Pr[B=b|C=c] \\ &= \mathop{\mathbb{E}}_B\left[\mathop{\mathbb{E}}_A[A|B,C=c]\,B|C=c\right],\end{aligned}$$

and applying this with $A=\varepsilon_t$, $B=\Delta_t(k)$, $C = W_1\cap\ldots\cap W_{t-1}$, we obtain

$$\begin{aligned}\mathbb{E}[W_t|W_{t-1},\ldots,W_1] &= \mathbb{E}[\bar{x}_t(k)\varepsilon_t|W_{t-1},\ldots,W_1] \\ &= \mathbb{E}[x_t(k)\varepsilon_t|W_{t-1},\ldots,W_1] + \mathbb{E}[\Delta_t(k)\varepsilon_t|W_{t-1},\ldots,W_1] \\ &= \mathbb{E}[x_t(k)\varepsilon_t] + \mathop{\mathbb{E}}_{\Delta_t}\left[\mathop{\mathbb{E}}_{\varepsilon_t}[\varepsilon_t|\Delta_t(k),W_{t-1},\ldots,W_1]\cdot\Delta_t(k)\middle|W_{t-1},\ldots,W_1\right] \\ &= \mathop{\mathbb{E}}_{x_t}\left[x_t(k)\cdot\mathop{\mathbb{E}}_{\varepsilon}[\varepsilon_t|x_t(k)]\right] + \mathop{\mathbb{E}}_{\Delta_t}\left[\Delta_t(k)\cdot\mathop{\mathbb{E}}_{\varepsilon_t}[\varepsilon_t]\middle|W_{t-1},\ldots,W_1\right] \\ &= 0,\end{aligned}$$

since $\mathbb{E}_{\varepsilon_t}[\varepsilon_t]=0$ and $\mathbb{E}_\varepsilon[\varepsilon_t|x_t(k)]=0$. Hence, we can apply Lemma A.1 and a union bound over all d features to show that with probability at least $1-\delta$,

$$\sum_{t=1}^{\tau(E)}\bar{x}_t(k)\varepsilon_t \geq -K'\sqrt{2\tau(E)\log(2d/\delta)}\ \ \forall k\in[d].$$

By Cauchy-Schwarz, we have

$$
\begin{aligned}
\left(\hat{\beta}_E\left(\mathcal{V}\right)-\beta^*\left(\mathcal{V}\right)\right)^\top \sum_{t=1}^{\tau(E)} \bar{x}_t\varepsilon_t &\leq \left\|\hat{\beta}_E\left(\mathcal{V}\right)-\beta^*\left(\mathcal{V}\right)\right\|_2 \cdot \left\|\sum_{t=1}^{\tau(E)} \bar{x}_t\varepsilon_t\right\|_2 \\
&\leq \left\|\hat{\beta}_E\left(\mathcal{V}\right)-\beta^*\left(\mathcal{V}\right)\right\|_2 \sqrt{\sum_{k=1}^{d}\left(\sum_t \bar{x}_t(k)\varepsilon_t\right)^2} \\
&\leq \left\|\hat{\beta}_E\left(\mathcal{V}\right)-\beta^*\left(\mathcal{V}\right)\right\|_2 \cdot K'\sqrt{2d\tau(E)\log(2d/\delta)}.
\end{aligned}
$$

∎

Strong convexity of the mean-squared error in sub-space $\mathcal{V}(\tau(E))$ We give a lower bound on the eigenvalues of $\bar{X}^\top\bar{X}$ on sub-space $\mathcal{V}(\tau(E))$, so as to show that at time $\tau(E)$, any least square solution $\hat{\beta}_E$ satisfies

$$
\begin{aligned}
&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) \\
&\qquad\qquad \geq \Omega(n)\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2^2.
\end{aligned}
$$

To do so, we will need the following concentration inequalities:

Lemma A.4. *Suppose $\mathbb{E}\left[x_t\right]=0$. Fix $\tau(E)=En$ for some $E\in\mathbb{N}$. With probability at least $1-\delta$, we have that*

$$
\sum_{t=1}^{\tau(E)} z^\top x_t x_t^\top z \geq \left(\lambda_r\tau(E)-2rd\sqrt{\tau(E)\log(6r/\delta)}\right)\|z\|_2^2 \quad \forall z\in\Sigma,
$$

and

$$
\sum_{t=1}^{\tau(E)} z^\top \Delta_t\Delta_t^\top z \geq \left(\min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\}n-\left(\max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\right)^2\sqrt{2n\log(6d/\delta)}\right)\|z\|_2^2 \quad \forall z\in\mathcal{D}_{\tau(E)}
$$

and

$$
\sum_{t=1}^{\tau(E)} z^\top x_t\Delta_t^\top z \geq -2\max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}d\sqrt{\tau(E)\log(6d/\delta)}\|z\|_2^2 \quad \forall z\in\mathbb{R}^d.
$$

Proof. Deferred to Appendix A.2.1. ∎

We will also need the following statement on the norm of the projections of any $z\in\mathcal{V}$ to $\mathcal{D}$ and Σ:

Lemma A.5. *Let*

$$
\begin{aligned}
\lambda(\mathcal{D},\Sigma)=\inf_{z\in\mathcal{D}+\Sigma}\quad & \|z(\mathcal{D})\|_2+\|z(\Sigma)\|_2 \\
s.t.\quad & \|z\|_2=1.
\end{aligned}
$$

Then, $\lambda(\mathcal{D},\Sigma)>0$.

Proof. With respect to the Euclidean metric, the objective function is continuous in z (the orthogonal projection operators are linear hence continuous functions of z and $z\to\|z\|_2$ also is a continuous function), and its feasible set is compact (as it is a sphere in a bounded-dimensional space over real values). By the extreme value theorem, the optimization problem admits an optimal solution, i.e., there exists z^* with $\|z^*\|_2=1$ such that $\lambda(\mathcal{D},\Sigma)=\|z^*(\mathcal{D})\|_2+\|z^*(\Sigma)\|_2$. Now, supposing $\lambda(\mathcal{D},\Sigma)\leq 0$, it must necessarily be the case that $z(\mathcal{D})=0$, $z(\Sigma)=0$. In particular, this means z is orthogonal to both $\mathcal{D}$ and Σ. In turn, z must be orthogonal to every vector in $\mathcal{D}+\Sigma$; since $z\in\mathcal{D}+\Sigma$, this is only possible when $z=0$, contradicting $\|z\|_2=1$. ∎

We can now move onto the proof of our lower bound for

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right).$$

Corollary A.6. *Fix $\tau(E) = En$ for some $E \in \mathbb{N}$. With probability at least $1-\delta$,*

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)\\ &\geq \left(\frac{\lambda n}{2} - \kappa' d^2\sqrt{\tau(E)\log(6d/\delta)}\right)\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2^2,\end{aligned}$$

for some constants κ', λ that only depend on σ, $\mathcal{C}$, and Σ, with $\lambda > 0$.

Proof. Since it is clear from context, we drop all $\tau(E)$ subscripts in the notation of this proof. First, we remark that

$$\begin{aligned}z^\top \bar{X}^\top \bar{X} z &= \sum_t z^\top \bar{x}_t \bar{x}_t^\top z\\ &= \sum_t z^\top x_t x_t^\top z + \sum_t z^\top \Delta_t \Delta_t^\top z + 2\sum_t z^\top \Delta_t z^\top x_t.\end{aligned}$$

We have by Lemma A.5 that for all $z \in \mathcal{V} = \mathcal{D} + \Sigma$,

$$\|z(\mathcal{D})\|_2 + \|z(\Sigma)\|_2 \geq \lambda(\mathcal{D}, \Sigma)\|z\|_2.$$

Let $\lambda(\Sigma) \triangleq \min_{D\subset[d]} \lambda(\mathcal{D}, \Sigma)$. Since there are finitely many subsets D of $[d]$ (and corresponding sub-spaces $\mathcal{D}$) and since for all such subsets, $\lambda(\mathcal{D}, \Sigma) > 0$, we have that $\lambda(\Sigma) > 0$. Further,

$$\|z(\mathcal{D})\|_2 + \|z(\Sigma)\|_2 \geq \lambda(\Sigma)\|z\|_2.$$

Therefore, it must be the case that either $\|z(\mathcal{D})\|_2 \geq \frac{\lambda(\Sigma)}{2}\|z\|_2$ or $\|z(\Sigma)\|_2 \geq \frac{\lambda(\Sigma)}{2}\|z\|_2$. We divide our proof into the corresponding two cases:

1. The first case is when $\|z(\Sigma)\|_2 \geq \frac{\lambda(\Sigma)}{2}\|z\|_2$. Then, note that since $z^\top \Delta_t \Delta_t^\top z \geq 0$ always, we have

$$\begin{aligned}\sum_t z^\top \bar{x}_t \bar{x}_t^\top z &\geq \sum_t z^\top x_t x_t^\top z + 2\sum_t z^\top \Delta_t z^\top x_t\\ &= \sum_t z(\Sigma)^\top x_t x_t^\top z(\Sigma) + 2\sum_t z^\top \Delta_t z^\top x_t,\end{aligned}$$

where the last equality follows from the fact that $x_t \in \Sigma$ and $z = z(\Sigma) + z(\Sigma^\perp)$. By Lemma A.4, we get that for some constant C_1 that depends only on $\mathcal{C}$,

$$\begin{aligned}&\sum_t z^\top \bar{x}_t \bar{x}_t^\top z\\ &\geq \left(\lambda_r \tau(E) - 2rd\sqrt{\tau(E)\log(6r/\delta)}\right)\|z(\Sigma)\|_2^2 - C_1 d\sqrt{\tau(E)\log(6d/\delta)}\|z\|_2^2\\ &\geq \left(\frac{\lambda(\Sigma)\lambda_r}{2}\tau(E) - \lambda(\Sigma)rd\sqrt{\tau(E)\log(6r/\delta)} - C_1 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2\\ &\geq \left(\frac{\lambda(\Sigma)\lambda_r}{2}\tau(E) - \lambda(\Sigma)d^2\sqrt{\tau(E)\log(6d/\delta)} - C_1 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2.\end{aligned}$$

(The second step assumes $\lambda_r \tau(E) - 2rd\sqrt{\tau(E)\log(6r/\delta)} \geq 0$. When this is negative, the bound trivially holds as $\sum_t z^\top \bar{x}_t \bar{x}_t^\top z \geq 0$.)

2. The second case arises when $\|z(\mathcal{D})\|_2 \geq \frac{\lambda(\Sigma)}{2}\|z\|_2$. Note that

$$\begin{aligned}\sum_t z^\top \bar{x}_t \bar{x}_t^\top z &\geq \sum_t z^\top \Delta_t \Delta_t^\top z + 2\sum_t z^\top \Delta_t z^\top x_t \\ &= \sum_t z(\mathcal{D})^\top \Delta_t \Delta_t^\top z(\mathcal{D}) + 2\sum_t z^\top \Delta_t z^\top x_t,\end{aligned}$$

as $\Delta_t \in \mathcal{D}$ and $z = z(\mathcal{D}) + z(\mathcal{D}^\perp)$. By Lemma A.4, it follows that for some constants C_2, C_3 that only depend on $\mathcal{C}$,

$$\begin{aligned}&\sum_t z^\top \bar{x}_t \bar{x}_t^\top z \\ &\geq \left(n \min_{i,k}\left\{\pi^i \left(\frac{B^i}{c^i(k)}\right)^2\right\} - C_2\sqrt{n\log(6d/\delta)}\right)\|z(\mathcal{D})\|_2^2 - C_3 d\sqrt{\tau(E)\log(6d/\delta)}\|z\|_2^2 \\ &\geq \left(\frac{\lambda(\Sigma)n}{2}\min_{i,k}\left\{\pi^i \left(\frac{B^i}{c^i(k)}\right)^2\right\} - \frac{\lambda(\Sigma)C_2}{2}\sqrt{n\log(6d/\delta)} - C_3 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2 \\ &\geq \left(\frac{\lambda(\Sigma)n}{2}\min_{i,k}\left\{\pi^i \left(\frac{B^i}{c^i(k)}\right)^2\right\} - \frac{\lambda(\Sigma)C_2}{2}\sqrt{\tau(E)\log(6d/\delta)} - C_3 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2.\end{aligned}$$

Noting that by definition $\lambda_r > 0$ and $\min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\} > 0$, and picking the worse of the two above bounds on $\sum_t z^\top \bar{x}_t \bar{x}_t^\top z$ concludes the proof with

$$\lambda = \frac{\lambda(\Sigma)}{2}\min\left(\lambda_r, \min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\}\right) > 0.$$

□

We can now prove Theorem 4.1. By Lemma A.2, we have that

$$\left(\bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) = \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)},$$

which immediately yields

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top\left(\bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) \\ &= \left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)}\end{aligned}$$

by performing matrix multiplication with $\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top$ on both sides on the first-order conditions. Further, by Lemma A.3, Corollary A.6, and a union bound, we get that with probability at least $1-\delta$,

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) \\ &\geq \left(\frac{\lambda n}{2} - \kappa' d^2\sqrt{\tau(E)\log(12d/\delta)}\right)\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2^2,\end{aligned}$$

and

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon \\ &\leq \left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \cdot K'\sqrt{d\tau(E)\log(4d/\delta)}.\end{aligned}$$

Combining the two above inequalities with the first-order conditions yields

$$\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \leq \frac{K'\sqrt{d\tau(E)\log(4d/\delta)}}{\frac{\lambda n}{2}-\kappa' d^2\sqrt{\tau(E)\log(12d/\delta)}}.$$

For

$$n \geq \frac{4\kappa' d^2}{\lambda}\sqrt{\tau(E)\log(12d/\delta)},$$

the bound becomes

$$\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \leq \frac{4K'\sqrt{d\tau(E)\log(4d/\delta)}}{\lambda n}.$$

The proof concludes by letting $K \triangleq 4K'$, $\kappa \triangleq 4\kappa'$ and noting that since $\mathcal{D}_{\tau(E)} \subset \mathcal{V}_{\tau(E)}$ by construction, the statement holds true over $\mathcal{D}_{\tau(E)}$ (projecting onto a subspace cannot increase the $\ell 2$-norm).

A.2.1 Proof of Lemma A.4

For the first statement, note that for all $k \neq j \leq r$,

$$\mathbb{E}\left[f_k^\top x_t x_t^\top f_j\right] = f_k^\top \mathbb{E}\left[x_t x_t^\top\right] f_j = \lambda_j f_k^\top f_j,$$

as f_j is (by definition) an eigenvector of $\Sigma = \mathbb{E}\left[x_t x_t^\top\right]$ for eigenvalue λ_j. Note that the $f_j^\top x_t x_t^\top f_k = (f_j^\top x_t)(f_k^\top x_t)$ are random variables that are independent across t. Further, by Cauchy-Schwarz,

$$\left|(f_k^\top x_t)(f_j^\top x_t)\right| \leq \|f_k\|_2\|f_j\|_2\|x_t\|_2^2 = \|x_t\|_2^2 \leq d.$$

Therefore, we can apply Hoeffding with a union bound over the r^2 choices of (f_k, f_j) to show that with probability at least $1-\delta'$,

$$\left|\sum_{t=1}^{\tau(E)} f_k^\top x_t x_t^\top f_j - \lambda_j \tau(E) f_k^\top f_j\right| \leq d\sqrt{2\tau(E)\log(2r^2/\delta')}.$$

Note now that for all $z \in \Sigma$, we can write $z = \sum_{k=1}^r \left(z^\top f_k\right) f_k$, and as such

$$\begin{aligned}
&\left|\sum_{t=1}^{\tau(E)} z^\top x_t x_t^\top z - \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j)\lambda_j \tau(E) f_k^\top f_j\right| \\
&= \left|\sum_{t=1}^{\tau(E)}\sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j) f_k^\top x_t x_t^\top f_j - \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j)\lambda_j \tau(E) f_k^\top f_j\right| \\
&= \left|\sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j)\left(\sum_t f_k^\top x_t x_t^\top f_j - \lambda_j\tau(E) f_k^\top f_j\right)\right| \\
&\leq d\sqrt{2\tau(E)\log(2r^2/\delta')}\sum_{k,j=1}^{r}|z^\top f_k||z^\top f_j| \\
&\leq rd\sqrt{2\tau(E)\log(2r^2/\delta')}\|z\|_2^2,
\end{aligned}$$

where the last step follows from the fact that by Cauchy-Schwarz,

$$\sum_{k=1}^r |z^\top f_k| \leq \sqrt{\sum_{k=1}^r 1^2}\sqrt{\sum_{k=1}^r (z^\top f_k)^2} = \sqrt{r}\|z\|_2.$$

Hence, for $z \in \Sigma$, remembering $f_k^\top f_j = 0$ when $k \neq j$ and $f_k^\top f_k = 1$, and noting $\|z\|_2^2 = \sum_{k=1}^r (z^\top f_k)^2$, we get that

$$\begin{aligned}\sum_{t=1}^{\tau(E)} z^\top x_t x_t^\top z &\geq \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j)\lambda_j \tau(E) f_k^\top f_j - rd\sqrt{2\tau(E)\log(2r^2/\delta')}\|z\|_2^2 \\ &= \sum_{k=1}^{r} \lambda_k \tau(E)(z^\top f_k)^2 - rd\sqrt{2\tau(E)\log(2r^2/\delta')}\|z\|_2^2 \\ &\geq \lambda_r \tau(E) \sum_{k=1}^{r} (z^\top f_k)^2 - rd\sqrt{2\tau(E)\log(2r^2/\delta')}\|z\|_2^2 \\ &= \left(\lambda_r \tau(E) - 2rd\sqrt{\tau(E)\log(2r/\delta')}\right)\|z\|_2^2.\end{aligned}$$

For the second statement, we remind the reader that the costs of modification are such that $\left|\Delta_t(k)^2\right| \leq \left(\max_{i,j}\left\{\frac{B^i}{c^i(j)}\right\}\right)^2$, and that within any epoch ϕ, the Δ_t's are independent of each other. We can therefore apply Hoeffding's inequality and a union bound (over $k \in D_{\tau(E)} \subset [d]$) to show that with probability at least $1-\delta'$, for any $k \in D_{\tau(E)}$, there exists an epoch $\phi(k) \leq E$ (pick any ϕ in which k is modified) such that

$$\begin{aligned}\sum_{t \in \phi(k)} e_k^\top \Delta_t \Delta_t^\top e_k &\geq n\,\mathbb{E}\left[\Delta_t(k)^2\right] - \left(\max_{i,j}\left\{\frac{B^i}{c^i(j)}\right\}\right)^2 \sqrt{2n\log(d/\delta')} \\ &\geq n \min_{i \in [l], j \in [d]} \left\{\pi^i \left(\frac{B^i}{c^i(j)}\right)^2\right\} - \left(\max_{i,j}\left\{\frac{B^i}{c^i(j)}\right\}\right)^2 \sqrt{2n\log(d/\delta')}.\end{aligned}$$

The last inequality holds noting that k can be modified in period $\phi(k)$ only if there exists a cost type i on the support of $\mathcal{C}$ such that k is a best response to $\hat{\beta}_{\phi(k)-1}$; in turn, k is modified with probability π^i by amount $\Delta(k) = B^i/c^i(k)$, leading to

$$\mathbb{E}\left[\Delta_t(k)^2\right] \geq \pi^i \left(\frac{B^i}{c^i(k)}\right)^2.$$

Since $\Delta_t(k)\Delta_t(j) = 0$ when $k \neq j$ as a single direction is modified at a time, note that for all $z \in \mathcal{D}_{\tau(E)}$, we have

$$\begin{aligned}&\sum_{t \leq \tau(E)} z^\top \Delta_t \Delta_t^\top z \\ &= \sum_{t \leq \tau(E)} \sum_{k=1}^{d} \Delta_t(k)^2 z^\top e_k e_k^\top z \\ &= \sum_{k=1}^{d} \sum_{t \leq \tau(E)} \Delta_t(k)^2 (z^\top e_k)^2 \\ &\geq \sum_{k \in D_{\tau(E)}} \sum_{t \in \phi(k)} \Delta_t(k)^2 (z^\top e_k)^2 \\ &\geq \sum_{k \in D_{\tau(E)}} \left(n \min_{i \in [l], j \in [d]} \left\{\pi^i \left(\frac{B^i}{c^i(j)}\right)^2\right\} - \left(\max_{i,j}\left\{\frac{B^i}{c^i(j)}\right\}\right)^2 \sqrt{2n\log(d/\delta')}\right)(z^\top e_k)^2 \\ &= \left(n \min_{i \in [l], j \in [d]} \left\{\pi^i \left(\frac{B^i}{c^i(j)}\right)^2\right\} - \left(\max_{i,j}\left\{\frac{B^i}{c^i(j)}\right\}\right)^2 \sqrt{2n\log(d/\delta')}\right) \sum_{k \in D_{\tau(E)}} (z^\top e_k)^2.\end{aligned}$$

For $z \in \mathcal{D}_{\tau(E)}$, $\sum_{k \in D_{\tau(E)}} (z^\top e_k)^2 = \|z\|_2^2$, and the second inequality immediately holds.

Finally, let us prove the last inequality. Take $(k,j) \in [d]^2$, and let us write $W_t = e_k^\top x_t \Delta_t^\top e_j$. First, note that x_t and Δ_t are independent: in epoch ϕ, the distribution of Δ_t is a function of $\hat{\beta}_{\phi-1}$ (and $\mathcal{C}$) only, which only depends on the realizations of $x,\ \varepsilon,\ \Delta$ in previous time steps. Further, x_t is independent of the history of features and modifications up until time $t-1$ included. Hence, it must be the case that

$$\begin{aligned}\mathbb{E}\left[W_t|W_{t-1},\ldots,W_1\right] &= \mathbb{E}\left[\mathbb{E}\left[e_k^\top x_t\middle|\Delta_t,W_{t-1},\ldots,W_1\right]\Delta_t^\top e_j\middle|W_{t-1},\ldots,W_1\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[e_k^\top x_t\right]\Delta_t^\top e_j\middle|W_{t-1},\ldots,W_1\right]\\ &= \mathbb{E}\left[e_k^\top x_t\right]\cdot\mathbb{E}\left[\Delta_t^\top e_j\middle|W_{t-1},\ldots,W_1\right]\\ &= 0,\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}\left[x_t\right]=0$. Further,

$$\left|e_k^\top x_t\Delta_t^\top e_j\right| = |x_t(k)||\Delta_t(j)| \leq \max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}.$$

We can therefore apply Lemma A.1 and a union bound over all $(k,j)\in[d]^2$ to show that with probability at least $1-\delta'$,

$$\left|\sum_{t=1}^{\tau(E)} e_k^\top x_t\Delta_t^\top e_j\right| \leq \max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\sqrt{2\tau(E)\log(2d^2/\delta')}.$$

In particular, we get that for all $z\in\mathbb{R}^d$,

$$\begin{aligned}\left|\sum_{t\in E} z^\top x_t\Delta_t^\top z\right| &= \left|\sum_{k,j}\sum_{t\in E}(z^\top e_k)(z^\top e_j)e_k^\top x_t\Delta_t^\top e_j\right|\\ &\leq \sum_{k,j}|z^\top e_k||z^\top e_j|\left|\sum_{t\in E}e_k^\top x_t\Delta_t^\top e_j\right|\\ &\leq \max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\sqrt{2\tau(E)\log(2d^2/\delta')}\left(\sum_k|z^\top e_k|\right)^2\\ &\leq 2d\max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\sqrt{\tau(E)\log(2d/\delta')}\|z\|_2^2,\end{aligned}$$

where the last step follows from the fact that by Cauchy-Schwarz,

$$\left(\sum_k|z^\top e_k|\right)^2 = \left(\sum_k|z(k)|\right)^2 \leq \sum_k 1^2\cdot\sum_k z(k)^2 = d\cdot\|z\|_2^2.$$

We conclude the proof with a union bound over all three inequalities, taking $\delta'=3\delta$.

B Proof of Theorem 5.2

We drop the $\tau(E)$ subscripts when clear from context. We first note that $\hat{\beta}_E$ is a least-square solution.

Claim B.1.

$$\hat{\beta}_E \in LSE(\tau(E)).$$

Proof. This follows immediately from noting that

$$\left(\bar{X}\hat{\beta}_E-\bar{Y}\right)^\top\left(\bar{X}\hat{\beta}_E-\bar{Y}\right) = \left(\bar{X}\beta_E-\bar{Y}\right)^\top\left(\bar{X}\beta_E-\bar{Y}\right),$$

as $\bar{X}^\top v = \bar{X}(\mathcal{U})^\top v = 0$ by definition of $\mathcal{U}$, and since $v\in\mathcal{U}^\perp$. □

Second, we show that $\hat{\beta}_E$ has large norm:

Claim B.2.

$$\left\|\hat{\beta}_E\right\|_2 \geq \alpha.$$

Proof. First, we note that necessarily, $\beta_E \in \mathcal{U}_{\tau(E)}$. Suppose not, then we can write

$$\beta_E = \beta_E\left(\mathcal{U}_{\tau(E)}\right) + \beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right),$$

with $\beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right) \neq 0$. By the same argument as in Claim B.1, $\beta_E\left(\mathcal{U}_{\tau(E)}\right)$ is a least-square solution. Using orthogonality of $\mathcal{U}_{\tau(E)}$ and $\mathcal{U}^{\perp}_{\tau(E)}$ and the fact that $\left\|\beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right)\right\|_2 > 0$, we have

$$\|\beta_E\|^2 = \left\|\beta_E\left(\mathcal{U}_{\tau(E)}\right)\right\|_2^2 + \left\|\beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right)\right\|_2^2 > \left\|\beta_E\left(\mathcal{U}_{\tau(E)}\right)\right\|_2^2.$$

This contradicts β_E being a minimum norm least-square solution. Hence, it must be the case that $\beta_E \in \mathcal{U}_{\tau(E)}$. Since $v \in \mathcal{U}^{\perp}_{\tau(E)}$, we have that β_E and v are orthogonal with $\|v\|_2 = 1$, implying

$$\left\|\hat{\beta}_E\right\|_2^2 = \|\beta_E\|_2^2 + \alpha^2\|v\|_2^2 \geq \alpha^2.$$

This concludes the proof. □

We argue that such a solution places a large amount of weight on currently unexplored features:

Lemma B.3. *At time $\tau(E)$, suppose rank $\left(\mathcal{U}_{\tau(E)}\right) \leq [d]$. Suppose $n \geq \frac{\kappa d^2}{\lambda}\sqrt{\tau(E)\log(12d/\delta')}$. Take any α with*

$$\alpha \geq \gamma\left(\sqrt{d} + \frac{Kd\sqrt{T\log(4d/\delta')}}{\lambda n}\right),$$

where γ is a constant that depends only on $\mathcal{C}$. With probability at least $1-\delta'$, there exists $i \in [l]$ and a feature $k \notin D_{\tau(E)}$ with

$$\frac{\left|\hat{\beta}_E(k)\right|}{c^i(k)} > \frac{\left|\hat{\beta}_E(j)\right|}{c^i(j)},\ \forall j \in D_{\tau(E)}.$$

Proof. Since $\hat{\beta}_E \in LSE(\tau(E))$, it must be by Theorem 4.1 that with probability at least $1-\delta'$,

$$\begin{aligned}\sqrt{\sum_{k\in D}\left(\hat{\beta}_E(k) - \beta^*(k)\right)^2} &\leq \frac{K\sqrt{d\tau(E)\log(4d/\delta')}}{\lambda n} \\ &\leq \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.\end{aligned} \tag{4}$$

First, since $z \to \sqrt{\sum_{k\in D} z(k)^2}$ defines a norm (in fact, the $\ell 2$-norm in $\mathbb{R}^{|D|}$), it must be the case that

$$\sqrt{\sum_{k\in D}(z(k) - z'(k))^2} \geq \sqrt{\sum_{k\in D} z(k)^2} - \sqrt{\sum_{k\in D} z'(k)^2}.$$

In turn, plugging this in Equation (4), we obtain

$$\begin{aligned}\sqrt{\sum_{k\in D}\hat{\beta}_E(k)^2} &\leq \sqrt{\sum_{k\in D}\beta^*(k)^2} + \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n} \\ &\leq \|\beta^*\|_2 + \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n} \\ &\leq \sqrt{d} + \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.\end{aligned}$$

By the triangle inequality and the lemma's assumption, we also have that

$$\sqrt{\sum_{k\in D}\hat{\beta}_E(k)^2}+\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2}\geq\|\hat{\beta}_E\|_2\geq\alpha.$$

Combining the last two equations, we obtain

$$\sqrt{d}+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}+\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2},\geq\alpha$$

which implies that for $\alpha\geq\gamma\left(\sqrt{d}+\frac{Kd\sqrt{T\log(4d/\delta')}}{\lambda n}\right)$, we have:

$$\begin{aligned}\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2}&\geq\alpha-\sqrt{d}-\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\\&\geq\alpha-\sqrt{d}-\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\\&\geq\sqrt{d}\,(\gamma-1)\left(1+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\right).\end{aligned}$$

Second, note that Equation (4) implies immediately that for any $j\in D_T$,

$$\left|\hat{\beta}_E(j)-\beta^*(j)\right|\leq\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n},$$

and in turn,

$$\left|\hat{\beta}_E(j)\right|\leq|\beta^*(j)|+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\leq 1+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.$$

Therefore,

$$\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2}\geq\sqrt{d}\,(\gamma-1)\max_{j\in D}\hat{\beta}_E(j).$$

Hence, there must exist feature $k\notin D$ with

$$\left|\hat{\beta}_E(k)\right|\geq(\gamma-1)\max_{j\in D}\hat{\beta}_E(j).$$

Picking γ such that for some $i\in[l]$,

$$\gamma-1\geq\max_{j\in D}\frac{c^i(k)}{c^i(j)}$$

yields the result immediately. □

The proof of Theorem 5.2 follows directly from Lemma B.3 and a union bound over the first d epochs. With probability at least $1-d\delta'$, for every epoch $E\in[d]$, there is a feature $k\notin D_{\tau(E)}$ such that for some $i\in[l]$,

$$\frac{\left|\hat{\beta}_E(k)\right|}{c^i(k)}>\frac{\left|\hat{\beta}_E(j)\right|}{c^i(j)}\ \forall j\in D_{\tau(E)}.$$

This implies that there exists $k\in D_{\tau(E+1)}$ but $k\notin D_{\tau(E)}$. Applying this d times, we have that if $T\geq dn$, necessarily $D_T=[d]$. We can then apply Theorem 4.1 to then show that with probability at least $1-\delta'$

$$\left\|\hat{\beta}_{T/n}-\beta^*\right\|_2\leq\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.$$

Taking a union bound over the two above events and $\delta=2d\delta'$, we get the theorem statement with probability at least $1-\delta'\,(d+1)\geq 1-\delta$.