A. Proof of Lemma 10

Notice that when $n \geq c_0 \ell \ln(5/\delta)$ where $\ell \geq d$, we have

$$\mathrm{err}_{\lambda^*}(x, c_0\ell) \leq \sqrt{\frac{4L^2\|x\|^2_{M(\lambda^*)^{-1}}}{c_0\ell} + \frac{\frac{8}{9}L^2(\|x\|_2^2 + 2\|x\|_2\|x\|_{M(\lambda^*)^{-1}}\sqrt{d} + \|x\|^2_{M(\lambda^*)^{-1}}d)}{c_0^2\ell^2}}$$
$$+ \sqrt{2}\kappa\sqrt{\frac{\|x\|^2_{M(\lambda^*)^{-1}}}{c_0\ell}\sqrt{\frac{3d}{c_0\ell}} + \frac{\|x\|^2_{M(\lambda^*)^{-1}}}{c_0\ell}}$$
$$\leq \sqrt{\frac{\|x\|^2_{M(\lambda^*)^{-1}} + \|x\|_2^2 + \|x\|_2\|x\|_{M(\lambda^*)^{-1}}\sqrt{d}}{\ell} + \frac{\|x\|^2_{M(\lambda^*)^{-1}}d}{3\ell^2}} + \sqrt{\frac{2\|x\|^2_{M(\lambda^*)^{-1}}\sqrt{d/\ell}}{3\ell} + \frac{\|x\|^2_{M(\lambda^*)^{-1}}}{\ell}}$$
$$\leq \sqrt{\frac{2\|x\|_2^2 + 2\sqrt{d}\|x\|_2\|x\|_{M(\lambda^*)^{-1}} + (4 + 2\sqrt{d/\ell})\|x\|^2_{M(\lambda^*)^{-1}}}{\ell}},$$

and so the result follows from Lemma 7.

B. Proof of Theorem 12

Let $\ell = \frac{2+6d+\epsilon d}{\epsilon^2}$. Lemma 11 implies that with probability at least $1 - \delta$, for every $x \in \mathcal{X}$,

$$\begin{aligned}
|x^{\mathrm{T}}\theta - x^{\mathrm{T}}\widehat{\boldsymbol{\theta}}| &\leq \sqrt{\frac{2 + 6d + 2d\sqrt{d/\ell}}{\ell}} \\
&= \sqrt{\frac{2+6d}{2+6d+\epsilon d}\epsilon^2 + \frac{2d^{\frac{3}{2}}}{(2+6d+\epsilon d)^{\frac{3}{2}}}\epsilon^3} \\
&= \epsilon\sqrt{1 - \frac{\epsilon d}{2+6d+\epsilon d} + \sqrt{\frac{4d}{2+6d+\epsilon d}} \cdot \frac{\epsilon d}{2+6d+\epsilon d}} \\
&\leq \epsilon.
\end{aligned}$$

C. Proof of Theorem 15

Let $A^{-1} = \sum_{i=1}^d \lambda_i v_i v_i^{\mathrm{T}}$ be its eigen-decomposition (so that v_i's form an orthonormal basis). Note that $\|v_i\| = 1$. By Lemma 10, for each v_i, with probability at least $(1 - \delta/d)$, we have

$$|v_i^{\mathrm{T}}(\theta - \widehat{\boldsymbol{\theta}})| \leq \sqrt{\frac{2 + (4 + 2\sqrt{d/\ell})\|x\|^2_{M(\lambda^*)^{-1}} + 2\sqrt{d}\|x\|_{M(\lambda^*)^{-1}}}{\ell}}.$$

Let $\lambda_i' = \|v_i\|^2_{M(\lambda^*)^{-1}} = n\lambda_i$ for every $i \in \{1, 2, 3, \ldots, d\}$. We have

$$(v_i^{\mathrm{T}}(\theta - \widehat{\boldsymbol{\theta}}))^2 \leq \frac{2 + (4 + 2\sqrt{d/\ell})\lambda_i' + 2\sqrt{\lambda_i' d}}{\ell}. \tag{17}$$

Via a union bound, we know that with probability at least $(1 - \delta)$, (17) holds for every eigenvector v_i. When this event happens, for every vector x such that $\|x\|_2 \leq 1$, let us write $x = \sum_{i=1}^d a_i v_i$. We have $\sum_{i=1}^d a_i^2 \leq 1$ and we have

$$\left|x^{\mathrm{T}}(\theta - \widehat{\boldsymbol{\theta}})\right| = \left|\sum_{i=1}^d a_i v_i^{\mathrm{T}}(\theta - \widehat{\boldsymbol{\theta}})\right| \leq d^{1/2}\left(\sum_{i=1}^d a_i^2 (v_i^{\mathrm{T}}(\theta - \widehat{\boldsymbol{\theta}}))^2\right)^{1/2}$$
$$\leq \left(\frac{d}{\ell}\right)^{1/2}\left(\sum_{i=1}^d a_i^2 (2 + (4 + 2\sqrt{d/\ell})\lambda_i' + 2\sqrt{\lambda_i' d})\right)^{1/2}, \tag{18}$$

where the first inequality is due to Cauchy-Schwartz inequality. Since

$$\sum_{i=1}^{d} a_i^2(2+(4+2\sqrt{d/\ell})\lambda_i'+2\sqrt{\lambda_i' d}) \le \sum_{i=1}^{d} a_i^2(2+(4+2\sqrt{d/\ell})\lambda_i'+\lambda_i'+d)$$
$$= (2+d)\sum_{i=1}^{d} a_i^2 + (5+2\sqrt{d/\ell})\sum_{i=1}^{d} a_i^2\lambda_i' \le (2+d)+(5+2\sqrt{d/\ell})\|x\|^2_{M(\lambda^*)^{-1}}, \quad (19)$$

continuing with (18) we have

$$\left|x^{\mathrm{T}}(\theta-\widehat{\boldsymbol{\theta}})\right| \le \sqrt{\frac{(2+d)d+(5d+2d\sqrt{d/\ell})\|x\|^2_{M(\lambda^*)^{-1}}}{\ell}}$$

holds for every x such that $\|x\|_2 \le 1$.

D. Proof of Lemma 17

Let $\mathcal{E}_r$ denote the event that $|x^{\mathrm{T}}\widehat{\theta}_r - x^{\mathrm{T}}\theta| \le \epsilon_r/2, \forall x \in S$. By Theorem 12, we have $\Pr[\mathcal{E}_r] \ge 1-\delta_r$. Let $\mathcal{E}$ denote the event $\bigwedge_{r=1}^{+\infty}\mathcal{E}_r$. Via a union bound, we get $\Pr[\mathcal{E}] \ge 1-\sum_{r=1}^{+\infty}\delta_r = 1-(6/\pi^2)\delta\cdot\sum_{r=1}^{+\infty}1/r^2 \ge 1-\delta$. We condition the rest of the proof upon event $\mathcal{E}$.

(i) We show that $S_{[1]} \in S_r$ (so the best arm is in the output set) by induction on r. The base case follows since $S_{[1]} \in S = S_1$. Moreover, if $S_{[1]} \in S_k$ for some $k \ge 1$, we have that $S_{[1]}^{\mathrm{T}}\widehat{\theta}_k+\epsilon_k \ge S_{[1]}^{\mathrm{T}}\theta+\epsilon_k/2 \ge x_{a_k}^{\mathrm{T}}\theta+\epsilon_k/2 \ge x_{a_k}^{\mathrm{T}}\widehat{\theta}$, and so $S_{[1]} \in S_{k+1}$.

To show that $\textsc{ElimTil}_p$ outputs at most p arms, let t_i be the smallest index such that $\Delta_i > \epsilon_{t_i-1}$ (so $\Delta_i \in (\epsilon_{t_i-1}, \epsilon_{t_i-2}]$), with ϵ_0 defined to be 1. We prove that if $i \ne 1$, then $S_{[i]} \notin S_{t_i+1}$. Indeed since $S_{[1]} \in S_{t_i}$, if $S_{[i]} \in S_{t_i}$ then $S_{[i]}^{\mathrm{T}}\widehat{\theta}_{t_i} \le S_{[i]}^{\mathrm{T}}\theta+\epsilon_{t_i}/2 < S_{[1]}^{\mathrm{T}}\theta-\epsilon_{t_i-1}+\epsilon_{t_i}/2 = (S_{[1]}^{\mathrm{T}}\theta-\epsilon_{t_i}/2)-\epsilon_{t_i} \le S_{[1]}^{\mathrm{T}}\widehat{\theta}_{t_i}-\epsilon_{t_i} \le x_{a_{t_i}}^{\mathrm{T}}\widehat{\theta}_{t_i}-\epsilon_{t_i}$, and so $S_{[i]} \notin S_{t_i+1}$.

Therefore $\{S_{[1]}\} \subseteq S_{t_{p+1}+1} \subseteq \{S_{[1]},\ldots,S_{[p]}\}$ and hence the algorithm stops after t_{p+1} rounds. Thus, the first part of this lemma is proved.

(ii) Note that the sample complexity of line 5 is $O\left(\frac{c_0 d}{\epsilon_r^2}\ln\frac{|S|}{\delta_r}\right)$. Also, the algorithm stops after t_{p+1} rounds. Therefore, the total number of samples consumed is bounded by

$$O\left(\frac{c_0 d}{\epsilon_{t_{p+1}}^2}\ln\frac{|S|}{\delta_{t_{p+1}}}\right) = O\left(\frac{c_0 d}{\epsilon_{t_{p+1}}^2}(\ln\delta^{-1}+\ln|S|+\ln(t_{p+1}))\right)$$
$$= O\left(\frac{c_0 d}{\Delta_{p+1}^2}(\ln\delta^{-1}+\ln|S|+\ln\ln\Delta_{p+1}^{-1})\right),$$

where the last equality follows from $\epsilon_{t_{p+1}} = \Theta(\Delta_{p+1})$ and $t_{p+1} = \Theta(\ln\Delta_{p+1}^{-1})$. Therefore, the proof of this lemma is complete.

E. Proof of Lemma 20

Let $\epsilon_r = 1/2^r$. Set $Y = \{y = x - x' \mid x, x' \in S\}$. Let $\mathcal{E}_r^{(1)}$ be the event

$$|x^{\mathrm{T}}\widehat{\theta}_r - x^{\mathrm{T}}\theta| \le \epsilon_r/2, \forall x \in T, \quad (20)$$

and let $\mathcal{E}_r^{(2)}$ be the event

$$|y^{\mathrm{T}}\widehat{\theta}_r - y^{\mathrm{T}}\theta| \le \mathrm{Err}_{\lambda_T^*}(y,\ell_r,\theta) \le \mathrm{err}_{\lambda_T^*}(y,\ell_r), \forall y \in Y. \quad (21)$$

By Theorem 12 and a union bound, we have $\Pr[\mathcal{E}_r^{(1)}] \ge 1-\frac{\delta_r}{|S|}$. By Lemma 7 and a union bound, we have $\Pr[\mathcal{E}_r^{(2)}] \ge 1-\frac{|S|-1}{|S|}\cdot\delta_r$. Let $\mathcal{E}_r = \mathcal{E}_r^{(1)} \wedge \mathcal{E}_r^{(2)}$ and $\mathcal{E} = \bigwedge_{r=1}^{+\infty}\mathcal{E}_r$. Hence via a union bound, we have $\Pr[\mathcal{E}] \ge 1-\sum_{r=1}^{+\infty}\delta_r = 1-(6/\pi^2)\delta\cdot\sum_{r=1}^{+\infty}1/r^2 \ge 1-\delta$.

We now condition on the event $\mathcal{E}$ till the end of the proof. Therefore, for all $x', x \in S$ and $r \geq 1$, we have

$$|(x'-x)^{\mathrm{T}}\widehat{\theta}_r - (x'-x)^{\mathrm{T}}\theta| \leq \min\{\mathrm{err}_{\lambda_T^*}(x'-x,\ell_r), \widehat{\mathrm{Err}}_{\lambda_T^*}(x'-x,\ell_r)\}.$$

(i) We prove that $S_{[1]} \in S_r$ by induction on r. The base case follows since $S_{[1]} \in S = S_1$. Furthermore, if $S_{[1]} \in S_k$, we have $(x_{a_k} - S_{[1]})^{\mathrm{T}}\widehat{\theta}_k \leq (x_{a_k} - S_{[1]})^{\mathrm{T}}\theta + \min\{\mathrm{err}_{\lambda_T^*}(x_{a_k} - S_{[1]}, \ell_r), \widehat{\mathrm{Err}}_{\lambda_T^*}(x_{a_k} - S_{[1]}, \ell_r)\} \leq \min\{\mathrm{err}_{\lambda_T^*}(x_{a_k} - S_{[1]}, \ell_r), \widehat{\mathrm{Err}}_{\lambda_T^*}(x_{a_k} - S_{[1]}, \ell_r)\}$. Hence $S_{[1]} \in S_{k+1}$.

To show that $\mathcal{Y}$-ElimTil$_p$ outputs at most p arms, let t_i be the smallest index such that $\Delta_i > \epsilon_{t_i-1}$ (so $\Delta_i \in (\epsilon_{t_i-1}, \epsilon_{t_i-2}]$), with ϵ_0 defined to be 1. We prove that if $i \neq 1$, then $S_{[i]} \notin S_{t_i+1}$. Indeed since $S_{[1]} \in S_{t_i}$, if $S_{[i]} \in S_{t_i}$ then $(x_{a_{t_i}} - S_{[i]})^{\mathrm{T}}\widehat{\theta}_{t_i} \geq (S_{[1]} - S_{[i]})^{\mathrm{T}}\widehat{\theta}_{t_i} \geq (S_{[1]} - S_{[i]})^{\mathrm{T}}\theta - \epsilon_r = \Delta_i - \epsilon_r > \epsilon_r \geq \mathrm{err}_{\lambda_T^*}(x_{a_r} - S_{[i]}, \ell_r) \geq \min\{\mathrm{err}_{\lambda_T^*}(x_{a_r} - S_{[i]}, \ell_r), \widehat{\mathrm{Err}}_{\lambda_T^*}(x_{a_r} - S_{[i]}, \ell_r)\}$, and so $S_{[i]} \notin S_{t_i+1}$.

Therefore, $\{S_{[1]}\} \subseteq S_{t_{p+1}+1} \subseteq \{S_{[1]}, \ldots, S_{[p]}\}$ and hence the algorithm stops after t_{p+1} rounds. Thus, the first part of this lemma is proved.

(ii) Note that $c_1 \leq 2$, the sample complexity of Line 5 is $O\left(\frac{c_0 d}{\epsilon_r^2} \ln \frac{|S|}{\delta_r}\right)$. Also, the algorithm stops after t_{p+1} rounds. Therefore, using the same proof, mutatis mutandis, as that of Lemma 17 (ii), total number of samples consumed is bounded by

$$O\left(\frac{c_0 d}{\Delta_{p+1}^2}(\ln \delta^{-1} + \ln |S| + \ln\ln \Delta_{p+1}^{-1})\right),$$

and the proof of this lemma is now complete.

F. Proof of Theorem 22

Let $\mathcal{E}_r, r \geq 0$ be the event that algorithm $\mathcal{Y}$-ElimTil$_{\lfloor d/2^r \rfloor}(S_r, \mathcal{X} \cap \mathrm{span}(S_r), \delta_r)$ outputs a set of at most $\lfloor d/2^r \rfloor$ arms with the best arm included, and the sample complexity is $O\left(\frac{c_0 \cdot \lfloor d/2^r \rfloor}{\Delta_{\lfloor d/2^r \rfloor+1}^2}(\ln \delta_r^{-1} + \ln |\mathcal{X}| + \ln\ln \Delta_{\lfloor d/2^r \rfloor+1}^{-1})\right)$. By Lemma 20, we have $\Pr[\mathcal{E}_r] \geq 1 - \delta_r$. Let $\mathcal{E}$ be the event $\bigwedge_{r=0}^{+\infty} \mathcal{E}_r$. Via a union bound, we see that $\Pr[\mathcal{E}] \geq 1 - \sum_{r=0}^{+\infty} \delta_r = 1 - 6/\pi^2 \cdot \delta \cdot \sum_{r=1}^{+\infty} 1/r^2 \geq 1 - \delta$. The proof is conditioned upon event $\mathcal{E}$ occurring.

(i) We first claim that the final output is the best arm. It suffices to prove that $\mathcal{X}_{[1]} \in S_r$ for all $r \geq 0$. We prove this claim by induction on r. The base case follows since $\mathcal{X}_{[1]} \in \mathcal{X} = S_0$. If $\mathcal{X}_{[1]} \in S_k$ holds, for some $k \geq 0$, since $\mathcal{E}_k$ holds, we know that S_{k+1} contains the best arm of S_k which is $\mathcal{X}_{[1]}$ by assumption. Therefore, $\mathcal{X}_{[1]} \in S_{k+1}$, and (i) is proved.

(ii) Let $r_0 = \lfloor \log_2 d \rfloor$. Since $\mathcal{E}$ is true, the total sample complexity is bounded by

$$\begin{aligned}
&\sum_{r=0}^{r_0} O\left(\frac{c_0 \cdot \lfloor d/2^r \rfloor}{\Delta_{\lfloor d/2^r \rfloor+1}^2}(\ln \delta_r^{-1} + \ln |\mathcal{X}| + \ln\ln \Delta_{\lfloor d/2^r \rfloor+1}^{-1})\right) \\
&= \sum_{r=0}^{r_0} O\left(\frac{c_0 \cdot (\lfloor d/2^r \rfloor - \lfloor d/2^{r+1} \rfloor)}{\Delta_{\lfloor d/2^r \rfloor+1}^2}(\ln \delta_r^{-1} + \ln |\mathcal{X}| + \ln\ln \Delta_{\lfloor d/2^r \rfloor+1}^{-1})\right) \\
&= \sum_{r=0}^{r_0-1} \sum_{i=\lfloor d/2^{r+1} \rfloor+1}^{\lfloor d/2^r \rfloor} O\left(\frac{c_0}{\Delta_i^2}(\ln \delta^{-1} + \ln |\mathcal{X}| + \ln\ln \Delta_i^{-1})\right) \\
&= O\left(\sum_{i=2}^{d} \frac{c_0}{\Delta_i^2}(\ln \delta^{-1} + \ln |\mathcal{X}| + \ln\ln \Delta_i^{-1})\right).
\end{aligned}$$

G. Details for computing $\lambda_{\mathcal{X}}^*$ and $M(\lambda_{\mathcal{X}}^*)$

Given a distribution λ over $\mathcal{X}$, $M(\lambda)$ is computed by $\sum_{i=1}^{N} \lambda_i x_i x_i^{\mathrm{T}}$. To compute $\lambda_{\mathcal{X}}^*$, we use entropic mirror descent introduced in (Beck & Teboulle, 2003). The details are included in Algorithm 5. For our experiments, we used $\epsilon = 0.1$ and $\eta_t = 0.001$.

Algorithm 5 The entropic mirror descent algorithm for computing $\lambda^*_{\mathcal{X}}$.

Input: Arms set $\mathcal{X}$, dimension d, Lipschitz constant L_f of function $\log\det M(\lambda)$ and tolerance ϵ.
Initialize $t \leftarrow 1$ and $\lambda^{(1)} \leftarrow (1/N, \ldots, 1/N)$.
while $|\max_{x\in\mathcal{X}} x^{\mathrm{T}} M(\lambda^{(t)})^{-1}x - d| \geq \epsilon$ **do**
    $\eta_t \leftarrow \frac{\sqrt{2\ln N}}{L_f}\frac{1}{\sqrt{t}}$.
    Compute gradient $g_i^{(t)} \leftarrow \mathrm{Tr}(M(\lambda^{(t)})^{-1}(x_i x_i^T))$.
    Update $\lambda_i^{(t+1)} \leftarrow \frac{\lambda_i^{(t)}\exp(\eta_t g_i^{(t)})}{\sum_{i=1}^N \lambda_i^{(t)}\exp(\eta_t g_i^{(t)})}$.
    $t \leftarrow t+1$.
Output: $\lambda^{(t)}$.