# Remaining Proofs

## EC.1. Proof of Proposition 6

PROPOSITION 6. *For all $\vec{\Delta}, \vec{g} \in \mathbb{R}_+^K$ such that $\vec{g} \neq 0$, the function $\pi \mapsto \left(\pi^\top \vec{\Delta}\right)^2 / \pi^\top \vec{g}$ is convex on $\left\{\pi \in \mathbb{R}^K : \pi^\top \vec{g} > 0\right\}$. Moreover, this function is minimized over $\mathcal{S}_K$ by some $\pi^*$ for which $|\{k : \pi_k^* > 0\}| \leq 2$.*

*Proof.* First, we show the function $\Psi : \pi \mapsto (\pi^T \Delta)^2 / \pi^T g$ is convex on $\{\pi \in \mathbb{R}^K | \pi^T g > 0\}$. As shown in Chapter 3 of Boyd and Vandenberghe (2004), $f : (x, y) \mapsto x^2/y$ is convex over $\{(x, y) \in \mathbb{R}^2 : y > 0\}$. The function $h : \pi \mapsto (\pi^T \Delta, \pi^T g) \in \mathbb{R}^2$ is affine. Since convexity is preserved under composition with an affine function, the function $\Psi = g \circ h$ is convex.

We now prove the second claim. Consider the optimization problems

$$\text{minimize } \Psi(\pi) \text{ subject to } \pi^T e = 1, \pi \geq 0 \tag{EC.1}$$

$$\text{minimize } \rho(\pi) \text{ subject to } \pi^T e = 1, \pi \geq 0 \tag{EC.2}$$

where

$$\rho(\pi) := \left(\pi^T \Delta\right)^2 - \left(\pi^T g\right) \Psi^*,$$

and $\Psi^* \in \mathbb{R}$ denotes the optimal objective value for the minimization problem (EC.1). The set of optimal solutions to (EC.1) and (EC.2) correspond. Note that

$$\Psi(\pi) = \Psi^* \implies \rho(\pi) = 0$$

but for any feasible $\pi$, $\rho(\pi) \geq 0$ since $\Delta(\pi)^2 \geq \Psi^* g(\pi)$. Therefore, any optimal solution $\pi_0$ to (EC.1) is an optimal solutions to (EC.2) and satisfies $\rho(\pi_0) = 0$. Similarly, if $\rho(\pi) = 0$ then simple algebra shows that $\Psi(\pi) = \Psi^*$ and hence that $\pi$ is an optimal solution to (EC.1)

We will now show that there is a minimizer of $\rho(\cdot)$ with at most two nonzero components, which implies the same is true of $\Psi(\cdot)$. Fix a minimizer $\pi^*$ of $\rho(\cdot)$. Differentiating $\rho(\pi)$ with respect to $\pi$ at $\pi = \pi^*$ yields

$$\begin{aligned}\frac{\partial}{\partial \pi} \rho(\pi^*) &= 2\left(\Delta^T \pi^*\right)\Delta - \Psi^* g \\ &= 2L^* \Delta - \Psi^* g\end{aligned}$$

where $L^* = \Delta^T \pi^*$ is the expected instantaneous regret of the sampling distribution $\pi^*$. Let $d^* = \min_i \frac{\partial}{\partial \pi_i} \rho(\pi^*)$ denote the smallest partial derivative of $\rho$ at $\pi^*$. It must be the case that any $i$ with $\pi_i^* > 0$ satisfies $d^* = \frac{\partial}{\partial \pi_i} \rho(\pi^*)$, as otherwise transferring probability from action $a_i$ could lead to strictly lower cost. This shows that

$$\pi_i^* > 0 \implies g_i = \frac{-d^*}{\Psi^*} + \frac{2L^*}{\Psi^*} \Delta_i. \tag{EC.3}$$

Let $i_1, .., i_m$ be the indices such that $\pi_{i_k}^* > 0$ ordered so that $g_{i_1} \geq g_{i_2} \geq \cdots \geq g_{i_m}$. Then we can choose a $\beta \in [0,1]$ so that

$$\sum_{k=1}^{m} \pi_{i_k}^* g_{i_k} = \beta g_{i_1} + (1-\beta) g_{i_m}.$$

By equation (EC.3), this implies as well that $\sum_{k=1}^{m} \pi_{i_k}^* \Delta_{i_k} = \beta \Delta_{i_1} + (1-\beta)\Delta_{i_m}$, and hence that the sampling distribution that plays $a_{i_1}$ with probability $\beta$ and $a_{i_m}$ otherwise has the same instantaneous expected regret and the same expected information gain as $\pi^*$. That is, starting with a general sampling distribution $\pi^*$ that maximizes $\rho(\pi)$, we showed there is a sampling distribution with support over at most two actions attains the same objective value and hence that also maximizes $\rho(\pi)$. $\square$

## EC.2. Proof of Proposition 1

The following fact expresses the mutual information between $A^*$ and $Y_{t,a}$ as the as the expected reduction in the entropy of $A^*$ due to observing $Y_{t,a}$.

Fact (Lemma 5.5.6 of Gray (2011)).

$$I_t\left(A^*; Y_{t,a}\right) = \mathbb{E}\left[H(\alpha_t) - H(\alpha_{t+1}) | A_t = a, \mathcal{F}_t\right]$$

Proposition 1. *For any policy $\pi = (\pi_1, \pi_2, \pi_3, \ldots)$ and time $T \in \mathbb{N}$,*

$$\mathbb{E}\left[\text{Regret}\left(T, \pi\right)\right] \leq \sqrt{\overline{\Psi}_T(\pi) H(\alpha_1) T}.$$

*where*

$$\overline{\Psi}_T(\pi) \equiv \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_\pi[\Psi_t(\pi_t)]$$

*is the average expected information ratio under $\pi$.*

*Proof.* Since the policy $\pi$ is fixed throughout, we will simplify notation and write $\Psi_t \equiv \Psi_t(\pi_t)$, $\Delta_t \equiv \Delta_t(\pi_t)$ and $g_t = g_t(\pi_t)$ throughout this proof. First observe that entropy bounds expected cumulative information gain:

$$\mathbb{E}\sum_{t=1}^{T} g_t = \mathbb{E}\sum_{t=1}^{T} \mathbb{E}\left[H(\alpha_t) - H(\alpha_{t+1})|\mathcal{F}_t\right] = \mathbb{E}\sum_{t=1}^{T}\left(H(\alpha_t) - H(\alpha_{t+1})\right) = H(\alpha_1) - H(\alpha_{T+1}) \leq H(\alpha_1),$$

where the first equality relies on Fact 1 and the tower property of conditional expectation and the final inequality follows from the non-negativity of entropy. Then,

$$\begin{aligned}
\mathbb{E}\left[\text{Regret}\left(T,\pi\right)\right] = \mathbb{E}\sum_{t=1}^{T}\Delta_t = \mathbb{E}\sum_{t=1}^{T}\sqrt{\Psi_t}\sqrt{g_t\left(\pi_t^{\text{IDS}}\right)} &\leq \sqrt{\mathbb{E}\sum_{t=1}^{T}\Psi_t}\sqrt{\mathbb{E}\sum_{t=1}^{T} g_t} \\
&\leq \sqrt{H(\alpha_1)}\sqrt{\mathbb{E}\sum_{t=1}^{T}\Psi_t} \\
&= \sqrt{\left(\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\Psi_t\right)H(\alpha_1)T},
\end{aligned}$$

where the first inequality follows from Holder's inequality. □

## EC.3. Proof of Proposition 7

PROPOSITION 7. *Suppose* $\sup_y R(y) - \inf_y R(y) \leq 1$ *and*

$$\pi_t \in \arg\min_{\pi\in\mathcal{S}_K} \frac{\Delta_t(\pi)^2}{v_t(\pi)},$$

*Then the following hold:*

1. $\Psi_t(\pi_t) \leq |\mathcal{A}|/2$.
2. $\Psi_t(\pi_t) \leq d/2$ *when* $\mathcal{A}\subset\mathbb{R}^d$, $\Theta\subset\mathbb{R}^d$, *and* $\mathbb{E}\left[R_{t,a}|\theta\right] = a^T\theta$ *for each action* $a\in\mathcal{A}$.

The proof of this proposition essentially reduces to techniques in Russo and Van Roy (2016), but some new analysis is required to show the results in that paper apply to variance-based IDS. A full proof is provided below.

We will make use of the following fact, which is a matrix-analogue of the Cauchy-Schwartz inequality. For any rank $r$ matrix $M\in\mathbb{R}^{n\times n}$ with singular values $\sigma_1,\ldots,\sigma_r$, let

$$\|M\|_* := \sum_{i=1}^{r}\sigma_i, \qquad \|M\|_F := \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n} M_{i,j}^2} = \sqrt{\sum_{i=1}^{r}\sigma_i^2}, \qquad \text{Trace}(M) := \sum_{i=1}^{n} M_{ii},$$

denote respectively the Nuclear norm, Frobenius norm and trace of $M$.

FACT 2. For any matrix $M \in \mathbb{R}^{k\times k}$,

$$\text{Trace}\,(M) \le \sqrt{\text{Rank}(M)}\|M\|_{\text{F}}.$$

We now prove Proposition 7

*Proof.*

**Preliminaries:** As noted in Section 5.3, $g_t(a) \ge 2v_t(a)$ for all $t$ and $a$. Therefore for any $\pi \in \mathcal{D}(\mathcal{A})$

$$\Psi_t(\pi) = \frac{\Delta_t(\pi)^2}{g_t(\pi)} \le \frac{\Delta_t(\pi)^2}{2v_t(\pi)}.$$

Therefore, if

$$\pi_t = \arg\min_{\pi\in\mathcal{D}(\mathcal{A})} \frac{\Delta_t(\pi)^2}{v_t(\pi)}$$

is the action-sampling distribution chosen by variance based IDS, then

$$\Psi_t(\pi_t) \le \frac{\Delta_t(\pi_t)^2}{2v_t(\pi_t)} \le \frac{\Delta_t(\pi_t^{\text{TS}})^2}{2v_t(\pi_t^{\text{TS}})},$$

where $\pi_t^{\text{TS}}$ is the action-sampling distribution of Thompson sampling at time $t$.

As a result, to show $\Psi_t(\pi_t) \le \lambda/2$, it's enough to show $\Delta_t(\pi_t^{\text{TS}})^2 \le \lambda v_t(\pi_t^{\text{TS}})$. We show that this holds always for $\lambda = |\mathcal{A}|$, and then show it holds for $\lambda = d$ when $\mathcal{A} \subset \mathbb{R}^d$, $\Theta \subset \mathbb{R}^d$, and $\mathbb{E}\left[R_{t,a}|\theta\right] = a^T\theta$ for all $a \in \mathcal{A}$.

Recall that by definition, $\pi_t^{\text{TS}}(a) = \mathbb{P}_t(A^* = a)$ for each $a \in \mathcal{A}$. Therefore

$$\begin{aligned}
\Delta_t(\pi_t^{\text{TS}}) &= \mathbb{E}_t[R_{t,A^*}] - \sum_{a\in\mathcal{A}} \pi_t^{\text{TS}}(a)\mathbb{E}_t[R_{t,a}] \\
&= \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(A^* = a^*)\mathbb{E}[R_{t,a^*}|A^* = a^*] - \sum_{a\in\mathcal{A}} \mathbb{P}_t(A^* = a)\mathbb{E}_t[R_{t,a}] \\
&= \sum_{a\in\mathcal{A}} \mathbb{P}_t(A^* = a)\left(\mathbb{E}_t[R_{t,a}|A^* = a] - \mathbb{E}_t[R_{t,a}]\right) \qquad \text{(EC.4)}
\end{aligned}$$

and

$$\begin{aligned}
v_t(\pi_t^{\text{TS}}) &= \sum_{a\in\mathcal{A}} \pi_t^{\text{TS}}(a)\text{Var}_t(\mathbb{E}[R_{t,a}|A^*]) \\
&= \sum_{a\in\mathcal{A}} \pi_t^{\text{TS}}(a) \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(A^* = a^*)\left(\mathbb{E}_t[R_{t,a}|A^* = a^*] - \mathbb{E}_t[R_{t,a}]\right)^2 \\
&= \sum_{a,a^*\in\mathcal{A}} \mathbb{P}_t(A^* = a)\mathbb{P}_t(A^* = a^*)\left(\mathbb{E}_t[R_{t,a}|A^* = a^*] - \mathbb{E}_t[R_{t,a}]\right)^2. \qquad \text{(EC.5)}
\end{aligned}$$

**Proof part 1:** By the Cauchy-Schwartz inequality, we conclude

$$\begin{aligned}
\Delta_t(\pi_t^{\mathrm{TS}})^2 &= \left(\sum_{a\in\mathcal{A}} \mathbb{P}_t(A^*=a)\left(\mathbb{E}_t[R_{t,a}|A^*=a]-\mathbb{E}_t[R_{t,a}]\right)\right)^2 \\
&\leq |\mathcal{A}|\sum_{a\in\mathcal{A}} \mathbb{P}_t(A^*=a)^2\left(\mathbb{E}_t[R_{t,a}|A^*=a]-\mathbb{E}_t[R_{t,a}]\right)^2 \\
&\leq |\mathcal{A}|\sum_{a,a'\in\mathcal{A}} \mathbb{P}_t(A^*=a)\mathbb{P}_t(A^*=a')\left(\mathbb{E}_t[R_{t,a}|A^*=a']-\mathbb{E}_t[R_{t,a}]\right)^2 \\
&= |\mathcal{A}| v_t(\pi_t^{\mathrm{TS}}).
\end{aligned}$$

As argued above, this implies $\Psi_t(\pi_t) \leq |\mathcal{A}|/2$.

**Proof of part 2:** This argument can be extended to provide a tighter bound under a linearity assumption. Now assume $\mathcal{A}\subset\mathbb{R}^d$, $\Theta\subset\mathbb{R}^d$, and $\mathbb{E}\left[R_{t,a}|\theta\right]=a^T\theta$. Write $\mathcal{A}=\{a_1,\ldots,a_K\}$ and define $M\in\mathbb{R}^{K\times K}$ by

$$\begin{aligned}
M_{i,j} &= \sqrt{\mathbb{P}_t(A^*=a_i)\mathbb{P}_t(A^*=a_j)}\left(\mathbb{E}_t[R_{t,a_i}|A^*=a_j]-\mathbb{E}_t[R_{t,a_i}]\right) \\
&= \sqrt{\alpha_t(a_i)\alpha_t(a_j)}\left(\mathbb{E}_t[R_{t,a_i}|A^*=a_j]-\mathbb{E}_t[R_{t,a_i}]\right)
\end{aligned}$$

for all $i,j\in\{1,..,K\}$. Then, by (EC.4) and (EC.5),

$$\Delta_t(\pi_t^{\mathrm{TS}}) = \mathrm{Trace(M)},$$

and

$$v_t(\pi_t^{\mathrm{TS}}) = \|M\|_{\mathrm{F}}^2.$$

This shows, by Fact 2 that

$$\Delta_t(\pi_t^{\mathrm{TS}})^2 \leq \mathrm{Rank}(M) v_t(\pi_t^{\mathrm{TS}})$$

We now show $\mathrm{Rank}(M)\leq d$. Define

$$\mu = \mathbb{E}\left[\theta|\mathcal{F}_t\right]$$

$$\mu^j = \mathbb{E}\left[\theta|\mathcal{F}_t, A^*=a_j\right].$$

Then, by the linearity of the expectation operator, $\mathbb{E}_t[R_{t,a_i}|A^* = a_j] - \mathbb{E}_t[R_{t,a_i}] = (\mu^j - \mu)^T a_i$. Therefore, $M_{i,j} = \sqrt{\alpha_t(a_i)\alpha_t(a_j)}((\mu^j - \mu)^T a_i)$ and

$$M = \begin{bmatrix} \sqrt{\alpha_t(a_1)}\,(\mu^1 - \mu)^T \\ \vdots \\ \vdots \\ \sqrt{\alpha_t(a_K))}\,(\mu^k - \mu)^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha_t(a_1)}a_1 \cdots \cdots \sqrt{\alpha_t(a_K)}a_K \end{bmatrix}.$$

Since $M$ is the product of a $K$ by $d$ matrix and a $d$ by $K$ matrix, it has rank at most $d$. $\square$