A. Proofs

In this section, we describe how to derive the theoretical results presented in the paper.

First, the exponential convergence rate of the estimated value function to the conjugate regularized value function at the root node (Theorem 1) is derived based on induction with respect to the depth D of the tree. When $D = 1$, we derive the concentration of the average reward at the leaf node with respect to the ∞-norm (as shown in Lemma 1) based on the result from Theorem 2.19 in (Wainwright, 2019), and the induction is done over the tree by additionally exploiting the contraction property of the convex regularized value function. Second, based on Theorem 1, we prove the exponential convergence rate of choosing the best action at the root node (Theorem 2). Third, the pseudo-regret analysis of E3W is derived based on the Bregman divergence properties and the contraction properties of the Legendre-Fenchel transform (Proposition 1). Finally, the bias error of estimated value at the root node is derived based on results of Theorem 1, and the boundedness property of the Legendre-Fenchel transform (Proposition 1).

Let $\hat{r}$ and r be respectively the average and the the expected reward at the leaf node, and the reward distribution at the leaf node be σ^2-sub-Gaussian.

Lemma 1 *For the stochastic bandit problem E3W guarantees that, for $t \geq 4$,*

$$\mathbb{P}\big(\parallel r - \hat{r}_t \parallel_\infty \geq \frac{2\sigma}{\log(2+t)}\big) \leq 4|\mathcal{A}| \exp\Big(- \frac{t}{(\log(2+t))^3}\Big).$$

Proof 1 *Let us define $N_t(a)$ as the number of times action a have been chosen until time t, and $\hat{N}_t(a) = \sum_{s=1}^{t} \pi_s(a)$, where $\pi_s(a)$ is the E3W policy at time step s. By choosing $\lambda_s = \frac{|\mathcal{A}|}{\log(1+s)}$, it follows that for all a and $t \geq 4$,*

$$\begin{aligned}\hat{N}_t(a) = \sum_{s=1}^{t} \pi_s(a) &\geq \sum_{s=1}^{t} \frac{1}{\log(1+s)} \geq \sum_{s=1}^{t} \frac{1}{\log(1+s)} - \frac{s/(s+1)}{(\log(1+s))^2} \\ &\geq \int_1^{1+t} \frac{1}{\log(1+s)} - \frac{s/(s+1)}{(\log(1+s))^2} ds = \frac{1+t}{\log(2+t)} - \frac{1}{\log 2} \geq \frac{t}{2\log(2+t)}.\end{aligned}$$

From Theorem 2.19 in (Wainwright, 2019), we have the following concentration inequality:

$$\mathbb{P}(|N_t(a) - \hat{N}_t(a)| > \epsilon) \leq 2\exp\{-\frac{\epsilon^2}{2\sum_{s=1}^{t} \sigma_s^2}\} \leq 2\exp\{-\frac{2\epsilon^2}{t}\},$$

where $\sigma_s^2 \leq 1/4$ is the variance of a Bernoulli distribution with $p = \pi_s(k)$ at time step s. We define the event

$$E_\epsilon = \{\forall a \in \mathcal{A}, |\hat{N}_t(a) - N_t(a)| \leq \epsilon\},$$

and consequently

$$\mathbb{P}(|\hat{N}_t(a) - N_t(a)| \geq \epsilon) \leq 2|\mathcal{A}| \exp(-\frac{2\epsilon^2}{t}). \tag{18}$$

Conditioned on the event E_ϵ, for $\epsilon = \frac{t}{4\log(2+t)}$, we have $N_t(a) \geq \frac{t}{4\log(2+t)}$. For any action a by the definition of sub-gaussian,

$$\mathbb{P}\left(|r(a) - \hat{r}_t(a)| > \sqrt{\frac{8\sigma^2 \log(\frac{2}{\delta}) \log(2+t)}{t}} \right) \leq \mathbb{P}\left(|r(a) - \hat{r}_t(a)| > \sqrt{\frac{2\sigma^2 \log(\frac{2}{\delta})}{N_t(a)}} \right) \leq \delta$$

by choosing a δ satisfying $\log(\frac{2}{\delta}) = \frac{1}{(\log(2+t))^3}$, we have

$$\mathbb{P}\left(|r(a) - \hat{r}_t(a)| > \sqrt{\frac{2\sigma^2 \log(\frac{2}{\delta})}{N_t(a)}} \right) \leq 2\exp\left(- \frac{1}{(\log(2+t))^3} \right).$$

Therefore, for $t \geq 2$

$$\mathbb{P}\left(\parallel r - \hat{r}_t \parallel_\infty > \frac{2\sigma}{\log(2+t)} \right) \leq \mathbb{P}\left(\parallel r - \hat{r}_t \parallel_\infty > \frac{2\sigma}{\log(2+t)} \middle| E_\epsilon \right) + \mathbb{P}(E_\epsilon^C)$$
$$\leq \sum_k \left(\mathbb{P}\left(|r(a) - \hat{r}_t(a)| > \frac{2\sigma}{\log(2+t)} \right) + \mathbb{P}(E_\epsilon^C) \leq 2|\mathcal{A}| \exp\left(- \frac{1}{(\log(2+t))^3} \right) \right)$$
$$+ 2|\mathcal{A}| \exp\left(- \frac{t}{(\log(2+t))^3} \right) = 4|\mathcal{A}| \exp\left(- \frac{t}{(\log(2+t))^3} \right).$$

Lemma 2 *Given two policies $\pi^{(1)} = \nabla\Omega^*(r^{(1)})$ and $\pi^{(2)} = \nabla\Omega^*(r^{(2)}), \exists L$, such that*

$$\parallel \pi^{(1)} - \pi^{(2)} \parallel_p \leq L \parallel r^{(1)} - r^{(2)} \parallel_p .$$

Proof 2 *This comes directly from the fact that $\pi = \nabla\Omega^*(r)$ is Lipschitz continuous with ℓ^p-norm. Note that p has different values according to the choice of regularizer. Refer to (Niculae & Blondel, 2017) for a discussion of each norm using maximum entropy and Tsallis entropy regularizer. Relative entropy shares the same properties with maximum Entropy.*

Lemma 3 *Consider the E3W policy applied to a tree. At any node s of the tree with depth d, Let us define $N_t^*(s,a) = \pi^*(a|s).t$, and $\hat{N}_t(s,a) = \sum_{s=1}^t \pi_s(a|s)$, where $\pi_k(a|s)$ is the policy at time step k. There exists some C and $\hat{C}$ such that*

$$\mathbb{P}\big(|\hat{N}_t(s,a) - N_t^*(s,a)| > \frac{Ct}{\log t}\big) \leq \hat{C}|\mathcal{A}|t\exp\{-\frac{t}{(\log t)^3}\}.$$

Proof 3 *We denote the following event,*

$$E_{r_k} = \{\parallel r(s',\cdot) - \hat{r}_k(s',\cdot) \parallel_\infty < \frac{2\sigma}{\log(2+k)}\}.$$

Thus, conditioned on the event $\bigcap_{i=1}^t E_{r_t}$ and for $t \geq 4$, we bound $|\hat{N}_t(s,a) - N_t^*(s,a)|$ as

$$\begin{aligned}
|\hat{N}_t(s,a) - N_t^*(s,a)| &\leq \sum_{k=1}^t |\hat{\pi}_k(a|s) - \pi^*(a|s)| + \sum_{k=1}^t \lambda_k \\
&\leq \sum_{k=1}^t \parallel \hat{\pi}_k(\cdot|s) - \pi^*(\cdot|s) \parallel_\infty + \sum_{k=1}^t \lambda_k \\
&\leq \sum_{k=1}^t \parallel \hat{\pi}_k(\cdot|s) - \pi^*(\cdot|s) \parallel_p + \sum_{k=1}^t \lambda_k \\
&\leq L \sum_{k=1}^t \parallel \hat{Q}_k(s',\cdot) - Q(s',\cdot) \parallel_p + \sum_{k=1}^t \lambda_k (\textit{Lemma 2}) \\
&\leq L|\mathcal{A}|^{\frac{1}{p}} \sum_{k=1}^t \parallel \hat{Q}_k(s',\cdot) - Q(s',\cdot) \parallel_\infty + \sum_{k=1}^t \lambda_k (\textit{ Property of p-norm}) \\
&\leq L|\mathcal{A}|^{\frac{1}{p}} \gamma^d \sum_{k=1}^t \parallel \hat{r}_k(s'',\cdot) - r(s'',\cdot) \parallel_\infty + \sum_{k=1}^t \lambda_k (\textit{Contraction 3.1}) \\
&\leq L|\mathcal{A}|^{\frac{1}{p}} \gamma^d \sum_{k=1}^t \frac{2\sigma}{\log(2+k)} + \sum_{k=1}^t \lambda_k \\
&\leq L|\mathcal{A}|^{\frac{1}{p}} \gamma^d \int_{k=0}^t \frac{2\sigma}{\log(2+k)} dk + \int_{k=0}^t \frac{|\mathcal{A}|}{\log(1+k)} dk \\
&\leq \frac{Ct}{\log t}.
\end{aligned}$$

for some constant C depending on $|\mathcal{A}|, p, d, \sigma, L$, and γ . Finally,

$$\begin{aligned}\mathbb{P}(|\hat{N}_t(s,a) - N_t^*(s,a)| \geq \frac{Ct}{\log t}) &\leq \sum_{i=1}^{t} \mathbb{P}(E_{r_t}^c) = \sum_{i=1}^{t} 4|\mathcal{A}| \exp(-\frac{t}{(\log(2+t))^3}) \\ &\leq 4|\mathcal{A}| t \exp(-\frac{t}{(\log(2+t))^3}) \\ &= O(t \exp(-\frac{t}{(\log(t))^3})).\end{aligned}$$

Lemma 4 *Consider the E3W policy applied to a tree. At any node s of the tree, Let us define $N_t^*(s,a) = \pi^*(a|s).t$, and $N_t(s,a)$ as the number of times action a have been chosen until time step t. There exists some C and $\hat{C}$ such that*

$$\mathbb{P}\big(|N_t(s,a) - N_t^*(s,a)| > \frac{Ct}{\log t}\big) \leq \hat{C} t \exp\{-\frac{t}{(\log t)^3}\}.$$

Proof 4 *Based on the result from Lemma 3, we have*

$$\begin{aligned}&\mathbb{P}\big(|N_t(s,a) - N_t^*(s,a)| > (1+C)\frac{t}{\log t}\big) \leq Ct \exp\{-\frac{t}{(\log t)^3}\} \\ &\leq \mathbb{P}\big(|\hat{N}_t(s,a) - N_t^*(s,a)| > \frac{Ct}{\log t}\big) + \mathbb{P}\big(|N_t(s,a) - \hat{N}_t(s,a)| > \frac{t}{\log t}\big) \\ &\leq 4|\mathcal{A}| t \exp\{-\frac{t}{(\log(2+t))^3}\} + 2|\mathcal{A}| \exp\{-\frac{t}{(\log(2+t))^2}\} (\textit{Lemma 3 and } (18)) \\ &\leq O(t \exp(-\frac{t}{(\log t)^3})).\end{aligned}$$

Theorem 1 *At the root node s of the tree, defining $N(s)$ as the number of visitations and $V_{\Omega^*}(s)$ as the estimated value at node s, for $\epsilon > 0$, we have*

$$\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) \leq C \exp\{-\frac{N(s)\epsilon}{\hat{C}(\log(2+N(s)))^2}\}.$$

Proof 5 *We prove this concentration inequality by induction. When the depth of the tree is $D = 1$, from Proposition 1, we get*

$$|V_\Omega(s) - V_\Omega^*(s)| = \| \Omega^*(Q_\Omega(s,\cdot)) - \Omega^*(Q_\Omega^*(s,\cdot)) \|_\infty \leq \gamma \| \hat{r} - r^* \|_\infty \ (\textit{Contraction})$$

where $\hat{r}$ is the average rewards and r^* is the mean reward. So that

$$\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) \leq \mathbb{P}(\gamma \| \hat{r} - r^* \|_\infty > \epsilon).$$

From Lemma 1, with $\epsilon = \frac{2\sigma\gamma}{\log(2+N(s))}$, we have

$$\begin{aligned}\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) \leq \mathbb{P}(\gamma \| \hat{r} - r^* \|_\infty > \epsilon) &\leq 4|\mathcal{A}| \exp\{-\frac{N(s)\epsilon}{2\sigma\gamma(\log(2+N(s)))^2}\} \\ &= C \exp\{-\frac{N(s)\epsilon}{\hat{C}(\log(2+N(s)))^2}\}.\end{aligned}$$

Let assume we have the concentration bound at the depth $D-1$, Let us define $V_\Omega(s_a) = Q_\Omega(s,a)$, where s_a is the state reached taking action a from state s. then at depth $D-1$

$$\mathbb{P}(|V_\Omega(s_a) - V_\Omega^*(s_a)| > \epsilon) \leq C \exp\{-\frac{N(s_a)\epsilon}{\hat{C}(\log(2+N(s_a)))^2}\}. \tag{19}$$

Now at the depth D, because of the Contraction Property, we have

$$\begin{aligned}|V_\Omega(s) - V_\Omega^*(s)| &\leq \gamma \parallel Q_\Omega(s,\cdot) - Q_\Omega^*(s,\cdot) \parallel_\infty \\ &= \gamma |Q_\Omega(s,a) - Q_\Omega^*(s,a)|.\end{aligned}$$

So that

$$\begin{aligned}\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) &\leq \mathbb{P}(\gamma \parallel Q_\Omega(s,a) - Q_\Omega^*(s,a) \parallel > \epsilon) \\ &\leq C_a \exp\{-\frac{N(s_a)\epsilon}{\hat{C}_a(\log(2+N(s_a)))^2}\} \\ &\leq C_a \exp\{-\frac{N(s_a)\epsilon}{\hat{C}_a(\log(2+N(s)))^2}\}.\end{aligned}$$

From (19), we can have $\lim_{t\to\infty} N(s_a) = \infty$ *because if* $\exists L, N(s_a) < L$*, we can find* $\epsilon > 0$ *for which (19) is not satisfied. From Lemma 4, when* $N(s)$ *is large enough, we have* $N(s_a) \to \pi^*(a|s)N(s)$ *(for example* $N(s_a) > \frac{1}{2}\pi^*(a|s)N(s)$*), that means we can find* C *and* $\hat{C}$ *that satisfy*

$$\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) \leq C \exp\{-\frac{N(s)\epsilon}{\hat{C}(\log(2+N(s)))^2}\}.$$

Lemma 5 *At any node* s *of the tree,* $N(s)$ *is the number of visitations. We define the event*

$$E_s = \{\forall a \in \mathcal{A}, |N(s,a) - N^*(s,a)| < \frac{N^*(s,a)}{2}\} \text{ where } N^*(s,a) = \pi^*(a|s)N(s),$$

where $\epsilon > 0$ *and* $V_{\Omega^*}(s)$ *is the estimated value at node s. We have*

$$\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon | E_s) \leq C \exp\{-\frac{N(s)\epsilon}{\hat{C}(\log(2+N(s)))^2}\}.$$

Proof 6 *The proof is the same as in Theorem 2. We prove the concentration inequality by induction. When the depth of the tree is* $D = 1$*, from Proposition 1, we get*

$$|V_\Omega(s) - V_\Omega^*(s)| = \parallel \Omega^*(Q_\Omega(s,\cdot)) - \Omega^*(Q_\Omega^*(s,\cdot)) \parallel \leq \gamma \parallel \hat{r} - r^* \parallel_\infty \text{ (Contraction Property)}$$

where $\hat{r}$ *is the average rewards and* r^* *is the mean rewards. So that*

$$\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) \leq \mathbb{P}(\gamma \parallel \hat{r} - r^* \parallel_\infty > \epsilon).$$

From Lemma 1, with $\epsilon = \frac{2\sigma\gamma}{\log(2+N(s))}$ *and given* E_s*, we have*

$$\begin{aligned}\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) &\leq \mathbb{P}(\gamma \parallel \hat{r} - r^* \parallel_\infty > \epsilon) \leq 4|\mathcal{A}| \exp\{-\frac{N(s)\epsilon}{2\sigma\gamma(\log(2+N(s)))^2}\} \\ &= C \exp\{-\frac{N(s)\epsilon}{\hat{C}(\log(2+N(s)))^2}\}.\end{aligned}$$

Let assume we have the concentration bound at the depth $D-1$*, Let us define* $V_\Omega(s_a) = Q_\Omega(s,a)$*, where* s_a *is the state reached taking action* a *from state* s*, then at depth* $D-1$

$$\mathbb{P}(|V_\Omega(s_a) - V_\Omega^*(s_a)| > \epsilon) \leq C \exp\{-\frac{N(s_a)\epsilon}{\hat{C}(\log(2+N(s_a)))^2}\}.$$

Now at depth D, because of the Contraction Property and given E_s*, we have*

$$\begin{aligned}|V_\Omega(s) - V_\Omega^*(s)| &\leq \gamma \parallel Q_\Omega(s,\cdot) - Q_\Omega^*(s,\cdot) \parallel_\infty \\ &= \gamma |Q_\Omega(s,a) - Q_\Omega^*(s,a)| (\exists a, \textit{ satisfied}).\end{aligned}$$

So that

$$\begin{aligned}\mathbb{P}(|V_\Omega(s) - V_\Omega^*(s)| > \epsilon) &\leq \mathbb{P}(\gamma \parallel Q_\Omega(s,a) - Q_\Omega^*(s,a) \parallel > \epsilon) \\ &\leq C_a \exp\{-\frac{N(s_a)\epsilon}{\hat{C}_a(\log(2+N(s_a)))^2}\} \\ &\leq C_a \exp\{-\frac{N(s_a)\epsilon}{\hat{C}_a(\log(2+N(s)))^2}\} \\ &\leq C \exp\{-\frac{N(s)\epsilon}{\hat{C}(\log(2+N(s)))^2}\} (\textit{because of } E_s)\end{aligned}$$

.

Theorem 2 *Let a_t be the action returned by algorithm E3W at iteration t. Then for t large enough, with some constants $C, \hat{C}$,*

$$\mathbb{P}(a_t \neq a^*) \leq Ct \exp\{-\frac{t}{\hat{C}\sigma(\log(t))^3}\}.$$

Proof 7 *Let us define event E_s as in Lemma 5. Let a^* be the action with largest value estimate at the root node state s. The probability that E3W selects a sub-optimal arm at s is*

$$\begin{aligned}&\mathbb{P}(a_t \neq a^*) \leq \sum_a \mathbb{P}(V_\Omega(s_a)) > V_\Omega(s_{a^*})|E_s) + \mathbb{P}(E_s^c) \\ &= \sum_a \mathbb{P}((V_\Omega(s_a) - V_\Omega^*(s_a)) - (V_\Omega(s_{a^*}) - V_\Omega^*(s_{a^*})) \geq V_\Omega^*(s_{a^*}) - V_\Omega^*(s_a)|E_s) + \mathbb{P}(E_s^c).\end{aligned}$$

Let us define $\Delta = V_\Omega^*(s_{a^*}) - V_\Omega^*(s_a)$, therefore for $\Delta > 0$, we have

$$\begin{aligned}&\mathbb{P}(a_t \neq a^*) \leq \sum_a \mathbb{P}((V_\Omega(s_a) - V_\Omega^*(s_a)) - (V_\Omega(s_{a^*}) - V_\Omega^*(s_{a^*})) \geq \Delta|E_s) + +\mathbb{P}(E_s^c) \\ &\leq \sum_a \mathbb{P}(|V_\Omega(s_a) - V_\Omega^*(s_a)| \geq \alpha\Delta|E_s) + \mathbb{P}(|V_\Omega(s_{a^*}) - V_\Omega^*(s_{a^*})| \geq \beta\Delta|E_s) + \mathbb{P}(E_s^c) \\ &\leq \sum_a C_a \exp\{-\frac{N(s)(\alpha\Delta)}{\hat{C}_a(\log(2+N(s)))^2}\} + C_{a^*} \exp\{-\frac{N(s)(\beta\Delta)}{\hat{C}_{a^*}(\log(2+N(s)))^2}\} + \mathbb{P}(E_s^c),\end{aligned}$$

where $\alpha+\beta = 1$, $\alpha > 0$, $\beta > 0$, and $N(s)$ is the number of visitations the root node s. Let us define $\frac{1}{\hat{C}} = \min\{\frac{(\alpha\Delta)}{C_a}, \frac{(\beta\Delta)}{C_{a^*}}\}$, and $C = \frac{1}{|\mathcal{A}|}\max\{C_a, C_{a^*}\}$ we have

$$\mathbb{P}(a \neq a^*) \leq C \exp\{-\frac{t}{\hat{C}\sigma(\log(2+t))^2}\} + \mathbb{P}(E_s^c).$$

From Lemma 4, $\exists C^{'}, \hat{C}^{'}$ for which

$$\mathbb{P}(E_s^c) \leq C^{'} t \exp\{-\frac{t}{\hat{C}^{'}(\log(t))^3}\},$$

so that

$$\mathbb{P}(a \neq a^*) \leq O(t \exp\{-\frac{t}{(\log(t))^3}\}).$$

Theorem 3 *Consider an E3W policy applied to the tree. Let define $\mathcal{D}_{\Omega^*}(x,y) = \Omega^*(x) - \Omega^*(y) - \nabla\Omega^*(y)(x-y)$ as the Bregman divergence between x and y, The expected pseudo regret R_n satisfies*

$$\mathbb{E}[R_n] \leq -\tau\Omega(\hat{\pi}) + \sum_{t=1}^{n} \mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) + \mathcal{O}(\frac{n}{\log n}).$$

Proof 8 *Without loss of generality, we can assume that $V_i \in [-1, 0], \forall i \in [1, |A|]$. as the definition of regret, we have*

$$\mathbb{E}[R_n] = nV^* - \sum_{t=1}^{n} \langle \hat{\pi}_t(\cdot), V(\cdot) \rangle \leq \hat{V}_1(0) - \sum_{t=1}^{n} \langle \hat{\pi}_t(\cdot), V(\cdot) \rangle \leq -\tau\Omega(\hat{\pi}) - \sum_{t=1}^{n} \langle \hat{\pi}_t(\cdot), V(\cdot) \rangle .$$

By the definition of the tree policy, we can obtain

$$\begin{aligned}
-\sum_{t=1}^{n} \langle \hat{\pi}_t(\cdot), V(\cdot) \rangle &= -\sum_{t=1}^{n} \left\langle (1-\lambda_t)\nabla\Omega^*(\hat{V}_t(\cdot)), V(\cdot) \right\rangle - \sum_{t=1}^{n} \left\langle \frac{\lambda_t(\cdot)}{|A|}, V(\cdot) \right\rangle \\
&= -\sum_{t=1}^{n} \left\langle (1-\lambda_t)\nabla\Omega^*(\hat{V}_t(\cdot)), V(\cdot) \right\rangle - \sum_{t=1}^{n} \left\langle \frac{\lambda_t(\cdot)}{|A|}, V(\cdot) \right\rangle \\
&\leq -\sum_{t=1}^{n} \left\langle \nabla\Omega^*(\hat{V}_t(\cdot)), V(\cdot) \right\rangle - \sum_{t=1}^{n} \left\langle \frac{\lambda_t(\cdot)}{|A|}, V(\cdot) \right\rangle .
\end{aligned}$$

with

$$\begin{aligned}
-\sum_{t=1}^{n} \left\langle \nabla\Omega^*(\hat{V}_t(\cdot)), V(\cdot) \right\rangle &= \sum_{t=1}^{n} \Omega^*(\hat{V}_t(\cdot) + V(\cdot)) - \sum_{t=1}^{n} \Omega^*(\hat{V}_t(\cdot)) - \sum_{t=1}^{n} \left\langle \nabla\Omega^*(\hat{V}_t(\cdot)), V(\cdot) \right\rangle \\
&\quad - (\sum_{t=1}^{n} \Omega^*(\hat{V}_t(\cdot) + V(\cdot)) - \sum_{t=1}^{n} \Omega^*(\hat{V}_t(\cdot))) \\
&= \sum_{t=1}^{n} \mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) - (\sum_{t=1}^{n} \Omega^*(\hat{V}_t(\cdot) + V(\cdot)) - \sum_{t=1}^{n} \Omega^*(\hat{V}_t(\cdot))) \\
&\leq \sum_{t=1}^{n} \mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) + n \parallel V(\cdot) \parallel_\infty \text{ (\textit{Contraction property, Proposition 1})} \\
&\leq \sum_{t=1}^{n} \mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) .(\textit{ because } V_i \leq 0)
\end{aligned}$$

And

$$-\sum_{t=1}^{n} \left\langle \frac{\lambda_t(\cdot)}{|A|}, V(\cdot) \right\rangle \leq \mathcal{O}(\frac{n}{\log n}), (\textit{Because} \sum_{k=1}^{n} \frac{1}{\log(k+1)} \to \mathcal{O}(\frac{n}{\log n}))$$

So that

$$\mathbb{E}[R_n] \leq -\tau\Omega(\hat{\pi}) + \sum_{t=1}^{n} \mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) + \mathcal{O}(\frac{n}{\log n}).$$

We consider the generalized Tsallis Entropy $\Omega(\pi) = \mathcal{S}_\alpha(\pi) = \frac{1}{1-\alpha}(1 - \sum_i \pi^\alpha(a_i|s))$.
According to (Abernethy et al., 2015), when $\alpha \in (0, 1)$

$$\begin{aligned}
\mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) &\leq (\tau\alpha)^{-1}|\mathcal{A}|^\alpha \\
-\Omega(\hat{\pi}_n) &\leq \frac{1}{1-\alpha}(|\mathcal{A}|^{1-\alpha} - 1).
\end{aligned}$$

Then, for the generalized Tsallis Entropy, when $\alpha \in (0, 1)$, the regret is

$$\mathbb{E}[R_n] \leq \frac{\tau}{1-\alpha}(|\mathcal{A}|^{1-\alpha} - 1) + n(\tau\alpha)^{-1}|\mathcal{A}|^\alpha + \mathcal{O}(\frac{n}{\log n}),$$

when $\alpha = 2$, which is the Tsallis entropy case we consider, according to (Zimmert & Seldin, 2019), By Taylor's theorem $\exists z \in$ conv$(\hat{V}_t, \hat{V}_t + V)$, we have

$$\mathcal{D}_{\Omega^*}(\hat{V}_t(\cdot) + V(\cdot), \hat{V}_t(\cdot)) \leq \frac{1}{2} \left\langle V(\cdot), \nabla^2\Omega^*(z)V(\cdot) \right\rangle \leq \frac{|\mathcal{K}|}{2}.$$

So that when $\alpha = 2$, we have

$$\mathbb{E}[R_n] \leq \tau(\frac{|\mathcal{A}|-1}{|\mathcal{A}|}) + \frac{n|\mathcal{K}|}{2} + \mathcal{O}(\frac{n}{\log n}).$$

when $\alpha = 1$, which is the maximum entropy case in our paper, we derive.

$$\mathbb{E}[R_n] \leq \tau(\log|\mathcal{A}|) + \frac{n|\mathcal{A}|}{\tau} + \mathcal{O}(\frac{n}{\log n})$$

Finally, when the convex regularizer is relative entropy, One can simply write $KL(\pi_t||\pi_{t-1}) = -H(\pi_t) - \mathbb{E}_{\pi_t}\log\pi_{t-1}$, let $m = \min_a \pi_{t-1}(a|s)$, we have

$$\mathbb{E}[R_n] \leq \tau(\log|\mathcal{A}| - \frac{1}{m}) + \frac{n|\mathcal{A}|}{\tau} + \mathcal{O}(\frac{n}{\log n}).$$

Before derive the next theorem, we state the Theorem 2 in (Geist et al., 2019)

- Boundedness: for two constants L_Ω and U_Ω such that for all $\pi \in \Pi$, we have $L_\Omega \leq \Omega(\pi) \leq U_\Omega$, then

$$V^*(s) - \frac{\tau(U_\Omega - L_\Omega)}{1-\gamma} \leq V^*_\Omega(s) \leq V^*(s). \tag{20}$$

Where τ is the temperature and γ is the discount constant.

Theorem 4 *For any $\delta > 0$, with probability at least $1-\delta$, the ε_Ω satisfies*

$$-\sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} - \frac{\tau(U_\Omega - L_\Omega)}{1-\gamma} \leq \varepsilon_\Omega \leq \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}}.$$

Proof 9 *From Theorem 2, let us define $\delta = C\exp\{-\frac{2N(s)\epsilon^2}{\hat{C}\sigma^2}\}$, so that $\epsilon = \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}}$ then for any $\delta > 0$, we have*

$$\mathbb{P}(|V_\Omega(s) - V^*_\Omega(s)| \leq \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}}) \geq 1-\delta.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&|V_\Omega(s) - V^*_\Omega(s)| \leq \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} \\
&-\sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} \leq V_\Omega(s) - V^*_\Omega(s) \leq \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} \\
&-\sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} + V^*_\Omega(s) \leq V_\Omega(s) \leq \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} + V^*_\Omega(s).
\end{aligned}$$

From Proposition 1, we have

$$-\sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} + V^*(s) - \frac{\tau(U_\Omega - L_\Omega)}{1-\gamma} \leq V_\Omega(s) \leq \sqrt{\frac{\hat{C}\sigma^2\log\frac{C}{\delta}}{2N(s)}} + V^*(s).$$