Learning in Online MDPs: Is there a Price for Handling the Communicating Case? (Supplementary Material)

Gautam Chandrasekaran[1] **Ambuj Tewari**[2]

[1]Department of Computer Science, University of Texas at Austin, Austin, Texas, USA
[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

1 ANALYSIS OF ALGORITHM 1

Before analysing the FPL algorithm described above, we first introduce some notation and definitions. We define the loss of a cycle c at time t as $\ell_t(s_t(a), a_t(a))$. For any cycle c with start state s, let L^c denote the total cumulative loss that we would have received if we followed the cycle c from the start to the end of the interaction. We use $\tilde{L}^c$ to denote the total perturbed cumulative loss received by cycle c. Let the cycle with lowest total cumulative loss be c^*. Also, let the cycle with lowest perturbed cumulative loss be $\tilde{c}^*$. We use $\tilde{L}_t^c$ to denote the total perturbed cumulative loss incurred by cycle c after t steps. We use $\tilde{c}_t^*$ to denote the cycle with lowest perturbed cumulative loss after t steps. Let C_t be the cycle chosen by the FPL algorithm at step t and l_t be it's reward. Let the expected number of switches made by the algorithm during the interaction be N_s.

The analysis is similar in spirit to Section 2 of Kalai and Vempala [2005]. We first state the following lemma that bounds the probability of switching the cycle at any step.

Lemma 1.1. $Pr[C_{t+1} \neq c \mid C_t = c] \leq (|S| + 1) \cdot \lambda \cdot \ell_t(s_t(c), a_t(c))$ *for all cycles* c *and times* $t \leq T$.

Proof of Theorem 4.2. We first bound the total loss incurred by the FPL algorithm. Let the expected number of switches made by the algorithm during the interaction be N_s. If the algorithm doesn't switch cycles after time step t, then $\tilde{L}_t^{C_t}$ must be equal to $\tilde{L}_t^{\tilde{c}_t^*}$. Thus, the loss incurred at time step t by C_t is at most $\left(\tilde{L}_t^{\tilde{c}_t^*} - \tilde{L}_{t-1}^{\tilde{c}_{t-1}^*}\right)$. In the steps in which the algorithm switches cycles, the maximum loss incurred is 1. Thus, we have that

$$\begin{aligned}
\mathbb{E}[\text{total loss of FPL}] &\leq \tilde{L}^{\tilde{c}_1^*} + \sum_{i=2}^{T} \left(\tilde{L}_t^{\tilde{c}_t^*} - \tilde{L}_{t-1}^{\tilde{c}_{t-1}^*}\right) + N_s \\
&\leq \tilde{L}^{\tilde{c}_T^*} + N_s \\
&= \tilde{L}^{\tilde{c}^*} + N_s \qquad (1)
\end{aligned}$$

We now bound N_s. From linearity of expectation, we have that

$$N_s = \sum_{t=1}^{T-1} Pr[C_{t+1} \neq C_t].$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

From Lemma 1.1, we have $Pr[C_{t+1} \neq C_t]$ is at most $(|S|+1) \cdot \lambda \cdot \mathbb{E}[l_t]$. This gives us the following bound for N_s.

$$\begin{aligned} N_s &= \sum_{t=1}^{T-1} Pr[C_t + 1 \neq C_t] \\ &\leq \sum_{t=1}^{T-1} (|S|+1) \cdot \lambda \cdot \mathbb{E}[l_t] \\ &\leq (|S|+1) \cdot \lambda \cdot \sum_{t=1}^{T-1} \mathbb{E}[l_t] \\ &\leq (|S|+1) \cdot \lambda \cdot \mathbb{E}[\text{total loss of FPL}] \end{aligned}$$

Combining this with (1) gives us the following.

$$\mathbb{E}[\text{total loss of FPL}] \leq \tilde{L}^{\tilde{c}^*} + (|S|+1) \cdot \lambda \cdot \mathbb{E}[\text{total loss of FPL}] \tag{2}$$

Let $p(c)$ denote the perturbed loss added to cycle c. Since the cycle with lowest perturbed cumulative loss at the end of the interaction is $\tilde{c}^*$, we have

$$\tilde{L}^{\tilde{c}^*} \leq L^{c^*} + p(\tilde{c}^*).$$

Also,

$$\mathbb{E}[p(\tilde{c}^*)] \leq \sum_{i=1}^{|S|} \mathbb{E}\left[\max_{(s,a)} \epsilon_i(s,a)\right] + \mathbb{E}\left[\max_{(s',k)} \delta(s',k)\right] \leq |S| \cdot \frac{(1+\log|S||A|)}{\lambda} + \frac{1+\log|S|^2}{\lambda}.$$

The above inequality comes from the fact that the expectation of the max of k independant exponential random variables with parameter λ is atmost $\frac{1+\log k}{\lambda}$. Plugging this inequality into (2) gives us

$$\mathbb{E}[\text{cost of FPL}] \leq L^* + |S| \cdot \frac{(1+\log|S||A|)}{\lambda} + \frac{1+\log|S|^2}{\lambda} + (|S|+1) \cdot \lambda \cdot \mathbb{E}[\text{cost of FPL}]. \tag{3}$$

Since the maximum cost is T, we have

$$\text{Regret} \leq |S|\frac{(1+\log|S||A|)}{\lambda} + \frac{1+\log|S|^2}{\lambda} + (|S|+1)\lambda T.$$

Setting $\lambda = \frac{\log|S||A|}{\sqrt{T}}$ gives us a bound of $O\left(|S|\sqrt{T\log|S||A|}\right)$ on the regret and expected number of switches. We can also derive first order bounds. From (3), we have

$$\begin{aligned} \mathbb{E}[\text{total loss of FPL}] &\leq L^* + |S| \cdot \frac{(1+\log|S||A|)}{\lambda} + \frac{1+\log|S|^2}{\lambda} + (|S|+1) \cdot \lambda \cdot \mathbb{E}[\text{cost of FPL}] \\ &\leq L^* + 4|S| \cdot \frac{\log|S||A|}{\lambda} + 2|S| \cdot \lambda \cdot \mathbb{E}[\text{total loss of FPL}]. \end{aligned}$$

On rearranging, we get

$$\begin{aligned} \mathbb{E}[\text{total loss of FPL}] &\leq \frac{L^*}{1-2\lambda|S|} + 4|S| \cdot \frac{\log|S||A|}{\lambda(1-2\lambda|S|)} \\ &\leq L^*(1 + (2\lambda|S| + (2\lambda|S|)^2 + \ldots) + 4|S|\frac{\log|S||A|}{\lambda}(1 + 2\lambda|S| + (2\lambda|S|)^2 + \ldots) \\ &\leq L^*(1 + 4\lambda|S|) + 8|S|\frac{\log|S||A|}{\lambda}. \end{aligned}$$

The last two inequalities work when $2\lambda|S| \leq \frac{1}{2}$. Thus,

$$\mathbb{E}[\text{total loss of FPL}] - L^* \leq 4\lambda|S|(L^*) + 8|S|\frac{\log|S||A|}{\lambda}.$$

Set $\lambda = \min\left(\sqrt{\frac{\log|S||A|}{L^*}}, \frac{1}{4|S|}\right)$. This forces $2\lambda|S|$ to be less than $\frac{1}{2}$ and thus the previous inequalities are still valid. On substituting the value of λ, we get that

$$\text{Regret} \leq O\left(|S|\sqrt{L^* \cdot \log|S||A|}\right)$$

when $L^* \geq 16|S|^2 \log|S||A|$. Since the expected number of switches is at most $2\lambda|S| \cdot \mathbb{E}[\text{total loss of FPL}]$, this is also bounded by $O\left(|S|\sqrt{L^* \cdot \log|S||A|}\right)$. □

Proof of Lemma 1.1. Let c be a cycle in the set $\mathcal{C}_{(s,k)}$. Let l_t be shorthand for $\ell_t(s_t(c), a_t(c))$the loss incurred by cycle c at step t. If C_{t+1} is not in $\mathcal{C}_{(s,k)}$, then the algorithm must have switched. Thus, we get the following equation.

$$Pr[C_{t+1} \neq c \mid C_t = c] = Pr[C_{t+1} \notin \mathcal{C}_{(s,k)} \mid C_t = c] + Pr[C_{t+1} \neq c \text{ and } C_{t+1} \in \mathcal{C}_{(s,k)} \mid C_t = c] \tag{4}$$

We now bound both the terms in the right hand side of (4) separately.

First, we study at the first term. We will upper bound this term by proving an appropriate lower bound on the probability of choosing C_{t+1} from $\mathcal{C}_{(s,k)}$. Since $C_t = c$, we know that $\tilde{L}^c_{t-1} \leq \tilde{L}^{c'}_{t-1}$ for all $c' \neq c$. For all $c' \notin \mathcal{C}_{(s,k)}$, the perturbation $\delta(s,k)$ will play a role in the comparison of the perturbed cumulative losses. For $c' \in \mathcal{C}_{(s,k)}$, $\delta(s,k)$ appears on both sides of the comparison and thus gets cancelled out. Thus, we have $\delta(s,k) \geq w$, where w depends only on the perturbations and losses received by c and the cycles not in $\mathcal{C}_{(s,k)}$. Now, if $\delta(s,k)$ was larger than $w + l_t$, then the perturbed cumulative loss of c will be less than that of cycles not in $\mathcal{C}_{(s,k)}$ even after receiving the losses of step t. In this case, C_{t+1} will also be chosen from $\mathcal{C}_{(s,k)}$. This gives us the require probability lower bound.

$$\begin{aligned} Pr[C_{t+1} \in \mathcal{C}_{(s,k)} \mid C_t = c] &\geq Pr[\delta(s,k) \geq w + l_t \mid \delta(s,k) \geq w] \\ &\geq e^{-\lambda \cdot l_t} \\ &\geq 1 - \lambda \cdot l_t \end{aligned}$$

Thus, $Pr[C_{t+1} \notin \mathcal{C}_{(s,k)} \mid C_t = c]$ is at most $\lambda \cdot l_t$.

We now bound the second term. For any two cycles $c' \neq c''$ in $\mathcal{C}_{(s,k)}$, there exists an index $i \leq k$ such that the i^{th} edges of c' and c'' are different and all the smaller indexed edges of the two cycles are the same. We denote this index by $d(c', c'')$. Define $d(c', c'')$ to be zero when c' is from $\mathcal{C}_{(s,k)}$ and $c'' = c'$ or c'' is not from $\mathcal{C}_{(s,k)}$. Now, if C_{t+1} is in $\mathcal{C}_{(s,k)}$ and not equal to c, then $d(C_{t+1}, c)$ is a number between one and k. Thus, we get the following equation.

$$Pr[C_{t+1} \neq c \text{ and } C_{t+1} \in \mathcal{C}_{(s,k)} \mid C_t = c] = \sum_{i=1}^{k} Pr[d(c, C_{t+1}) = i \mid C_t = c] \tag{5}$$

We now bound $Pr[d(c, C_{t+1}) = i \mid C_t = c]$ for any i between 1 and k. Let (s_i, a_i) be the i^{th} edge of c. We prove a lower bound on the probability of choosing C_{t+1} such that $d(c, C_{t+1})$ is not equal to i. Again, since $C_t = c$, we know that $\tilde{L}^c_{t-1} \leq \tilde{L}^{c'}_{t-1}$ for all $c' \neq c$. Consider cycles c' that don't contain the edge (s_i, a_i) in the i^{th} position. The perturbation $\epsilon_i(s_i, a_i)$ will play a role in the comparison of perturbed losses of all such c' with c. Thus, we have $\epsilon_i(s_i, a_i) \geq w$, where w depends only on the perturbations and losses received by c and cycles c' that don't have the (s_i, a_i) edge in the i^{th} position. If $\epsilon_i(s_i, a_i)$ was greater than $w + l_t$, then the perturbed cumulative loss of c will still be less than that of all cycles c' without the (s_i, a_i) edge. In this case, C_{t+1} will be chosen such that it also has the (s_i, a_i) edge. This implies that $d(c, C_{t+1}) \neq i$. Thus, we get the following probability lower bound.

$$\begin{aligned} Pr[d(c, C_{t+1}) \neq i \mid C_t = c] &\geq Pr[\epsilon_i(s_i, a_i) \geq w + l_t \mid \epsilon_i(s_i, a_i) \geq w] \\ &\geq e^{-\lambda \cdot l_t} \\ &\geq 1 - \lambda \cdot l_t \end{aligned}$$

Thus, for all i between 1 and k, $Pr[d(c, C_t + 1) = i \mid C_t = c]$ is at most $\lambda \cdot l_t$. This proves that the term in (5) is at most $k\lambda \cdot l_t$. Since k is at most $|S|$, the second term in the right hand side of (4) is bounded by $|S| \cdot \lambda \cdot l_t$. □

2 REGRET LOWER BOUND

Proof of Theorem 4.5. Let M be an MDP with states labelled $s_0, s_2, \ldots, s_{|S|-1}$. Any action a takes state s_i to s_{i+1}(modulo $|S|$). In other words, the states are arranged in a cycle and every action takes any state to its next state in the cycle. This is the required M.

Consider the problem of *prediction with expert advice* with n experts. We know that for any algorithm $\mathcal{A}$, there is a sequence of losses such that the regret of $\mathcal{A}$ is $\Omega(\sqrt{T \log n})$ over T steps (see **?**). In our case, every policy spends exactly $\frac{T}{|S|}$ steps in each state. Thus, the interaction with M over T steps can be interpreted as a problem of prediction with expert advice at every state where each interaction lasts only $\frac{T}{|S|}$ steps. We have the following decomposition of the regret.

$$R(\mathcal{A}) = \sum_{i=0}^{|S|-1} \sum_{k=0}^{\frac{T}{|S|}-1} \ell_{k|S|+i}\left(s_i, a_{k|S|}\right) - \ell_{k|S|+i}\left(s_i, \pi^*(s_i)\right) \tag{6}$$

In the above equation, a_t is the action taken by $\mathcal{A}$ at step t. The best stationary deterministic policy in hindsight is π^*.

From the regret lower bound for the experts problem, we know that there exists a sequence of losses such that for each i, the inner sum of (6) is atleast $\Omega\left(\sqrt{\frac{T}{|S|}\log|A|}\right)$. By combining these loss sequences, we get a sequence of losses such that

$$R(\mathcal{A}) \geq \sum_{i=0}^{|S|-1} \Omega\left(\sqrt{\frac{T}{|S|}\log|A|}\right) \geq \Omega\left(\sqrt{|S|T\log|A|}\right).$$

This completes the proof. □

3 COMMUNICATING MDPS

3.1 EXISTENCE OF HIGH PROBABILITY CRITICAL LENGTH PATH

We now state an intermediate lemma that will be used to prove Theorem 5.2.

Lemma 3.1. *For any start state s and target $s' \neq s$, we have $\ell_{s,s'} \leq 2D$ and a policy π such that*

$$Pr[T(s' \mid M, \pi, s) = \ell_{s,s'}] \geq \frac{1}{4D}$$

Proof. From the definition of diameter, we are guaranteed a policy $\pi_{s,s'}$ such that

$$\mathbb{E}\left[T(s' \mid M, \pi, s)\right] \leq D$$

From Markov's inequality, we have

$$Pr\left[T(s' \mid M, \pi, s) \leq 2D\right] \geq \frac{1}{2}$$

Since there are only $2D$ discrete values less than $2D$, there exists $\ell_{s,s'} \leq 2D$ such that

$$Pr[T(s' \mid M, \pi, s) = \ell_{s,s'}] \geq \frac{1}{2} \cdot \frac{1}{2D} = \frac{1}{4D}$$

□

We can now prove Theorem 5.2

Proof of Theorem 5.2. From Lemma 3.1, we $\ell_{s'} \leq 4D$ for each s' such that there is a policy $\pi_{s^*,s'}$ that hits the state s' in time ℓ'_s with probability at-least $\frac{1}{4D}$. We take $\ell^* = \max_{s' \neq s^*} \ell_{s'}$. For target state s', the policy $\pi_{s'}$ loops at state s^* for $(\ell^* - \ell_{s'})$ time steps and then starts following policy $\pi_{s,s'}$. Clearly, this policy hits state s' at time ℓ^* with probability at least $\frac{1}{4D}$ □

3.2 CORRECTNESS OF SWITCH_POLICY ROUTINE

We now prove Lemma 5.3

Proof of Lemma 5.3. We want to compute $Pr[S_t = s \mid T_{switch} = t]$.

$$\begin{aligned} Pr[S_t = s \mid T_{switch} = t] &= \frac{Pr[S_t = s, T_{switch} = t]}{Pr[T_{switch} = t]} \\ &= \frac{Pr[S_t = T_t = s, T_{switch} = t]}{Pr[T_{switch} = t]} \\ &= \frac{Pr[T_t = s, S_t = s, T_{switch} = t]}{Pr[T_{switch} = t]} \end{aligned}$$

We now compute the denominator $Pr[T_{switch} = t]$ as follows.

$$\begin{aligned} Pr[T_{switch} = t] &= \sum_{s \in S} Pr[S_t = T_t = s, S_{t-\ell^*} = s^*] \cdot Pr[T_{switch} = t \mid S_t = T_t = s, S_{t-\ell^*} = s^*] \\ &= \sum_{s \in S} Pr[S_t = s \mid T_t = s, S_{t-\ell^*} = s^*] \cdot Pr[T_t = s, S_{t-\ell^*} = s^*] Pr[T_{switch} = t | S_t = T_t = s, S_{t-\ell^*} = s^*] \\ &= \sum_{s \in S} p_s \cdot Pr[T_t = s, S_{t-\ell^*} = s^*] \cdot \frac{p^*}{p_s} \\ &= p^* \sum_{s \in S} Pr[T_t = s, S_{t-\ell^*} = s^*] \\ &= p^* \cdot Pr[S_{t-\ell^*} = s^*] \end{aligned}$$

Now we calculate the numerator.

$$\begin{aligned} Pr[T_t = s, S_t = s, T_{switch} = t] &= Pr[T_t = s, S_t = s, S_{t-\ell^*} = s, T_{switch} = t] \\ &= Pr[S_t = s, T_{switch} = t \mid S_{t-\ell^*} = s^*, T_t = s] \cdot Pr[S_{t-\ell^*} = s^*, T_t = s] \\ &= p^* \cdot Pr[S_{t-\ell^*} = s^*] \cdot Pr[T_t = s \mid S_{t-\ell^*} = s^*] \\ &= p^* \cdot Pr[S_{t-\ell^*} = s^*] \cdot d_\pi^t(s) \end{aligned}$$

Thus, we have

$$Pr[S_t = s \mid T_{switch} = t] = d_\pi^t(s)$$

□

3.3 BOUNDING THE COST OF EACH SWITCH

We now prove Lemma 5.4.

Proof of Lemma 5.4. We bound the expectation using law of total expectations and conditioning on T_{switch}.

$$\mathbb{E}\left[\sum_{t=t_1}^{t_2} \ell_t(s_t, a_t)\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{t=t_1}^{t_2} \ell_t(s_t, a_t) \mid T_{switch}\right]\right]$$

We bound the conditional expectation.

$$\mathbb{E}\left[\sum_{t=t_1}^{t_2} \ell_t(s_t, a_t) \mid T_{switch} = t^*\right] \leq t^* + \mathbb{E}\left[\sum_{t=t^*}^{t_2} \ell_t(s_t, a_t) \mid T_{switch} = t^*\right]$$

From Lemma 5.3, the second term is equal to $\sum_{t=t^*}^{t_2} \hat{\ell}_t(\pi)$ Thus,

$$\mathbb{E}\left[\sum_{t=t_1}^{t_2} \ell_t(s_t, a_t)\right] \leq \mathbb{E}[T_{switch}] + \sum_{t=t_1}^{t_2} \hat{\ell}_t(\pi)$$

Everytime we try to catch the policy from state s^*, we succeed with probability $p^* \geq \frac{1}{4D}$. Thus, the expected number of times we try is $16 \cdot D$ and each attempt takes $\ell^* \leq 2D$ steps. Between each of these attempts, we move at most D steps in expectation to reach s^* again. Thus, in total, we have

$$\mathbb{E}[T_{switch}] \leq 16D^2 + 32D^2 = 48D^2$$

This completes the proof. □

3.4 ANALYSIS OF FPL ALGORITHM FOR COMMUNICATING MDPS WITH UNIFORM START DISTRIBUTION

We now prove Theorem 5.8

Proof of Theorem 5.8. Let L^π denote the total cumulative loss if we followed policy π from the start of the interaction. We use $\tilde{L}^\pi$ to the denote the total perturbed cumulative loss if we followed policy π from the start. Let π^* be the policy with the lowest total cumulative loss. Similarly, let $\tilde{\pi}^*$ be the policy with the lowest perturbed cumulative loss. Let $\tilde{L}_t^\pi$ be the total perturbed cumulative loss till time t. Let π_t be the policy chosen by the FPL algorithm at step t.

Let N_s be the number of times the oracle switches the best policy. As before, we treat each policy as an expert and consider the online learning problem where expert π gets loss $\hat{\ell}_t(\pi) = \mathbb{E}\left[\ell_t(s_t, a_t)\right]$ where $s_1 \sim d_1$ and $a_t = \pi(s_t)$.

Using the arguments from the proof of Theorem 4.3, we get

$$\mathbb{E}\left[\text{total loss of FPL}\right] \leq \tilde{L}^{\tilde{\pi}^*} + N_s.$$

Also, we have $\tilde{L}^\pi = L^\pi + \frac{1}{S}\sum_{i=1}^{s} \epsilon\left(s, \pi(s)\right)$. This comes from the fact that d_1 is the uniform distribution over states.

We know that $N_s = \sum_{t=1}^{T-1} Pr[\pi_{t+1} \neq \pi_t]$. We now bound $Pr[\pi_{t+1} \neq \pi_t]$. Let $\pi_t = \pi$. The algorithm chooses $\pi' \neq \pi$ as π_{t+1} if and only if $\tilde{L}_t^\pi \geq \tilde{L}^{\pi'}$. We now argue that the probability of this happening is low if $\pi_t = \pi$. Since $\pi' \neq \pi$, we have $\pi'(s) \neq \pi(s)$ for some s. Let the smallest state in which π and π' differ be called $d(\pi, \pi')$. Thus,

$$Pr[\pi_{t+1} \neq \pi \mid \pi_t = \pi] = \sum_{s \in S} Pr[d(\pi_{t+1}, \pi) = s \mid \pi_t = \pi].$$

We bound $Pr[d(\pi, \pi_{t+1}) = s \mid \pi_t = \pi]$ for any state s. Consider any policy π' that differs from π in state s. The perturbation $\epsilon(s, \pi(s))$ will play a role in the comparison of perturbed losses of all such π' with π. Since $\pi_t = \pi$, we have $\frac{\epsilon(s,\pi(s))}{|S|} \geq w$ for some w that depends only on the perturbations and losses received by π and policies π' that differ from π in state s. If $\frac{\epsilon(s,\pi(s))}{|S|} \geq w + \hat{\ell}_t(\pi)$, then we would not switch to a policy π' with $\pi(s) \neq \pi'(s)$. Thus,

$$\begin{aligned} Pr[d(\pi, \pi_{t+1} \neq s \mid \pi_t = \pi] &\geq Pr[\epsilon(s, \pi(s)) \geq w|S| + \hat{\ell}_t(\pi)|S| \mid \epsilon(s, \pi(s)) \geq w|S|] \\ &\geq 1 - \lambda\hat{\ell}_t(\pi)|S| \end{aligned}$$

Thus, $Pr[\pi_{t+1} \neq \pi \mid \pi_t = \pi]$ is at-most $|S|^2 \cdot \lambda\hat{\ell}_t(\pi)$. From this, we get $N_s \leq |S|^2 \cdot \lambda \cdot \mathbb{E}[\text{total loss of FPL}]$

Using arguments similar to Section 4.2.2, we get

$$\mathbb{E}[\text{total loss of FPL}] \leq L^{\pi^*} + \frac{(1 + \log|S||A|)}{\lambda} + |S|^2 \cdot \lambda \cdot \mathbb{E}[\text{total loss of FPL}] \tag{7}$$

Let $L^* = L^{\pi^*}$.

On rearranging and simplifying Equation 7 similar to the proof of Theorem 4.2, we have

$$\mathbb{E}[\text{total loss of FPL}] \leq L^*(1 + 2\lambda|S|^2) + 4\frac{\log|S||A|}{\lambda}$$

The above inequality works when $|S|^2\lambda \leq \frac{1}{2}$, Thus, we have

$$\mathbb{E}[\text{Total loss of FPL}] - L^* \leq 2\lambda|S|^2(L^*) + 4\frac{\log|S||A|}{\lambda}.$$

Set $\lambda = \min\left(\frac{1}{|S|}\sqrt{\frac{\log|S||A|}{L^*}}, \frac{1}{2|S|^2}\right)$. On substituting λ into the above equation, we get that

$$\text{Regret} \leq O(|S|\sqrt{L^* \log|S||A|}).$$

Since the expected number of switches is at-most $|S|^2 \cdot \lambda \cdot \mathbb{E}[\text{total loss of FPL}]$, this is also bounded by $O\left(|S|\sqrt{L^* \log|S||A|}\right)$ □