Supplementary Material: Optimal Algorithms for Lipschitz Bandits with Heavy-tailed Rewards

A. Proof of Theorem 1

Theorem 1 *Assume (1) and (2) hold. For sufficiently large T such that*

$$\log T \quad \frac{5}{8}(4 \quad)^{-\frac{1}{\epsilon}}$$

the regret of SDTM with parameter $r > 0$ satisfies

$$\mathbb{E}[R(T)] \quad 2rT + (4 \quad T)^{\frac{1}{1+\epsilon}}(16N_c(r)\log T)^{\frac{\epsilon}{1+\epsilon}}$$

where $N_c(r)$ is the r-covering number of the arm set X.

Proof. Let $x_* \in \arg\max_{x\in\mathcal{X}} \quad (x)$ be an optimal arm. By the definition of the oracle, there must exist $k \in [K]$ such that $x_* \in X_k$ and hence $D(x_*, \bar{x}_k) \quad 2r$. Since the expected reward function is Lipschitz, we have

$$(x_*) \quad (\bar{x}_k) \quad D(x_*, \bar{x}_k) \quad 2r. \tag{17}$$

On the other hand, let $\bar{x}_* \in \arg\max_{x_i, i\in[K]} \quad (\bar{x}_i)$ be an optimal skeleton arm. By theoretical guarantees of UCB policies used with the truncated mean estimator (Bubeck et al. 2013, Proposition 1), the expected difference between the cumulative reward of the pulled arms and that of the optimal skeleton arm $\bar{x}_*$ can be upper bounded as follows

$$\mathbb{E}\left[\sum_{t=1} (\bar{x}_*) \quad \sum_{t=1} (x_t)\right] \quad (4 \quad T)^{\frac{1}{1+\epsilon}}(16K\log T)^{\frac{\epsilon}{1+\epsilon}}. \tag{18}$$

Combining (17) and (18) and recalling that $K \quad N_c(r)$, we obtain

$$\mathbb{E}\left[\sum_{t=1} (x_*) \quad \sum_{t=1} (x_t)\right] \quad 2rT + (4 \quad T)^{\frac{1}{1+\epsilon}}(16N_c(r)\log T)^{\frac{\epsilon}{1+\epsilon}}$$

where we use the fact that $(\bar{x}_k) \quad (\bar{x}_*)$.

B. Proof of Corollary 2

Corollary 2 *We have*

$$\sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_z(r)}{r^{1/\epsilon}} \quad O \quad r_0^{-(d_z+1/\epsilon)}$$

and thus

$$R(T) \quad O \quad \inf_{r_0\in(0,1)} \quad r_0 T + \log T \quad r_0^{-(d_z+1/\epsilon)}$$
$$\Theta \quad T^{\frac{d_z\epsilon+1}{d_z\epsilon+\epsilon+1}}$$

where d_z is the zooming dimension of $(X; \quad)$, defined in (5).

Proof. We have

$$\sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_z(r)}{r^{1/\epsilon}} \quad \sum_{i\in\mathbb{N}:2^{-i}\geq r_0} 2^{id_z+i/\epsilon}Z \quad \sum_{i=0}^{\lfloor\log_2\frac{1}{r_0}\rfloor} 2^{id_z+i/\epsilon}Z \quad \int_0^{\log_2\frac{1}{r_0}+1} 2^{id_z+i/\epsilon}Z\,\mathrm{d}i \quad \frac{(2=r_0)^{d_z+1/\epsilon}Z}{\log(2^{d_z+1/\epsilon})}$$

where Z is the zooming constant of $(X; D)$.

C. Proof of Theorem 3

Theorem 3 *Assume (2) and (12) hold. With probability at least* $1 \quad 2$ *, the regret of ADMM satisfies*

$$R(T) \quad \inf_{r_0 \in (0,1)} \quad r_0 T + 68(102^{-})^{\frac{1}{\epsilon}} \log\left(e^{1/8} T^2 = \right) \quad \sum_{r = 2^{-i}: i \in \mathbb{N}, r \geq r_0} \frac{N_z(r)}{r^{1/\epsilon}}$$

where $^{-}$ *is defined in (13) and* $N_z(r)$ *is the* r*-zooming number of* $(X; \quad)$*. Furthermore, by the first inequality in Corollary 2 we have*

$$R(T) \quad \mathcal{O} \quad T^{\frac{d_z \epsilon + 1}{d_z \epsilon + \epsilon + 1}}$$

where d_z *is the zooming dimension of* $(X; \quad)$*, defined in (5).*

Proof. We use the same notations as in the proof of Theorem 2 and propose the following lemmas, which are counterparts of Lemmas 1, 2, 3, and 4 respectively. For brevity, we only prove Lemmas 5 and 6, and the proofs of Lemmas 7 and 8 can be done in the same way as in appendices G and H respectively.

Lemma 5 *Let* R *be the set comprised of all rounds in which Step* 21 *of Algorithm 4 is executed. Then, with probability at least* $1 \quad 2$ *, for all rounds* $t \ 2 \ R$ *and all active arms* $x \ 2 \ A_t$*, we have*

$$j b_t(x) \quad (x) j \quad r_{t+1}(x):$$

Proof. Fix $t \ 2 \ R$ and $x \ 2 \ A_t$. By Lemma 2 in Bubeck et al. (2013), with probability at least $1 \quad \frac{2\delta}{T^2}$ we have

$$j b_t(x) \quad (x) j \quad (12 \quad)^{\frac{1}{1+\epsilon}} \quad \frac{16 \log\left(e^{1/8} T^2 = \right)}{n_t(x)} \quad^{\frac{\epsilon}{1+\epsilon}} \quad (12^{-})^{\frac{1}{1+\epsilon}} \quad \frac{16 \log\left(e^{1/8} T^2 = \right)}{n_t(x)} \quad^{\frac{\epsilon}{1+\epsilon}} = r_{t+1}(x):$$

Taking the union bound over $x \ 2 \ A_t$ and $t \ 2 \ R$ and noticing $j A_t j \quad T; 8t \ 2 \ R$, we conclude the proof.

Lemma 6 *With probability at least* $1 \quad 2$ *, for all rounds* $t \ 2 \ [T]$ *and all active arms* $x \ 2 \ A_t$*, we have*

$$\Delta(x) \quad 3 \overline{p}_{\overline{2}} r_{t+1}(x):$$

Proof. Fix $t \ 2 \ [T]$. For each active arm $x \ 2 \ A_t$, there exist three different scenarios as follows.

(i) x is pulled by Step 4 or Step 10 of Algorithm 4 in round t. In this scenario, on one hand, we have

$$n_t(x) \quad 16 \log\left(e^{1/8} T^2 = \right) + 1$$

and hence

$$r_{t+1}(x) = (12^{-})^{\frac{1}{1+\epsilon}} \quad \frac{16 \log\left(e^{1/8} T^2 = \right)}{n_t(x)} \quad^{\frac{\epsilon}{1+\epsilon}} \quad (12^{-})^{\frac{1}{1+\epsilon}} \quad \frac{16 \log\left(e^{1/8} T^2 = \right)}{16 \log\left(e^{1/8} T^2 = \right) + 1} \quad^{\frac{\epsilon}{1+\epsilon}}$$
$$(3 \overline{p}_{\overline{2}})^{-\frac{1}{1+\epsilon}} \quad \frac{35}{36} \quad^{\frac{\epsilon}{1+\epsilon}} = \frac{35}{36} \quad 3 \overline{p}_{\overline{2}} \quad \frac{35}{36} \quad^{-\frac{1}{1+\epsilon}} \quad \frac{35}{36} \quad 3 \overline{p}_{\overline{2}} \quad \frac{35}{36} \quad^{-1} \quad \frac{1}{3 \overline{p}_{\overline{2}}}$$

where the second inequality follows from the definition of $^{-}$ in (13) and the following fact: $16 \log\left(e^{1/8} T^2 = \right) > 35$ for $T > 1$ and $2 \ (0; 1=2)$. On the other hand, let $x_* \ 2 \ \arg\max_{x \in \mathcal{X}} \ (x)$ be an optimal arm. We have $\Delta(x) = \ (x_*) \quad (x) \quad D(x_*; x) \quad 1$. Thus, we obtain $3 \overline{p}_{\overline{2}} r_{t+1}(x) \quad \Delta(x)$.

(ii) x is pulled by Step 12 of Algorithm 4 in round t. In this case, we have $t \quad 1 \ 2 \ R$ and the arm selection rule implies

$$b_{t-1}(x) + 2 r_t(x) \quad b_{t-1}(x') + 2 r_t(x'); \ 8x' \ 2 \ A_t: \tag{19}$$

Note that $A_t = A_{t-1}$. By Lemma 5, we get

$$(x) \quad b_{t-1}(x) \quad r_t(x);\ b_{t-1}(x') \quad (x') \quad r_t(x');\ 8x' \ 2\ A_t: \tag{20}$$

Combining (19) and (20), we obtain

$$(x) + 3r_t(x) \quad (x') + r_t(x');\ 8x' \ 2\ A_t: \tag{21}$$

Note that the execution of Step 12 implies $X \quad [_{x\in\mathcal{A}_t} B(x; r_t(x))$. Thus, for the optimal arm x_* there must exist an active arm $\bar{x}_* \ 2\ A_t$ such that

$$D(x_*; \bar{x}_*) \quad r_t(\bar{x}_*)$$

which, together with the Lipschitz property of , indicates

$$(x_*) \quad (\bar{x}_*) + r_t(\bar{x}_*):$$

Combining the above inequality and (21) with substitution $x' = \bar{x}_*$, we obtain

$$(x) + 3r_t(x) \quad (x_*):$$

On the other hand, we have

$$\frac{r_{t+1}(x)}{r_t(x)} = \frac{n_{t-1}(x)}{n_t(x)}^{\frac{\epsilon}{1+\epsilon}} = \frac{n_{t-1}(x)}{n_{t-1}(x)+1}^{\frac{\epsilon}{1+\epsilon}} \quad \frac{1}{2}:$$

Therefore, we get $3\ p\overline{2} r_{t+1}(x) \quad 3r_t(x) \quad \Delta(x):$

(iii) x is not played in round t. In this scenario, let s be the last round in which x is pulled. Then, we have $r_{t+1}(x) = r_{s+1}(x)$ and the proof reduces to (i) or (ii).

Lemma 7 *With probability at least* $1 \quad 2$ *, for all* $i = 0; 1; 2; :::,$

$$j\bar{A}_T(i)j \quad N_z(2^{-i}):$$

Lemma 8 *With probability at least* $1 \quad 2$ *, for all* $i = 0; 1; 2; :::,$

$$\sum_{x\in\mathcal{A}_T(i)} n_T(x)\Delta(x) \quad 2^{\frac{i+1}{\epsilon}} \quad (51^-)^{\frac{1}{\epsilon}} \quad 68\log(e^{1/8}T^2 =)N_z(2^{-i}):$$

The remaining proof is the same as that of Theorem 2 and is omitted here.

D. Proof of Theorem 4

Theorem 4 *Fix an arm set* X *with diameter* 1 *and a parameter of moment* $2\ (0; 1]$. *Define* $= \frac{2^{1/\epsilon}\cdot\epsilon}{\log 2}$ *and*

$$R_c(T) = \inf_{r_0\in(0,1)} \quad r_0 T + \log T \sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_c(r)}{r^{1/\epsilon}}$$

where $N_c(r)$ *is the* r*-covering number of* X*. Then, for any* $T > 2$ *and any positive number* $R \quad R_c(T)$*, there exists a set* I *of problem instances on* X *such that*
(i) *for each problem instance* $I \ 2\ I$*, define*

$$R_z(T) = \inf_{r_0\in(0,1)} \quad r_0 T + \log T \sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_z(r)}{r^{1/\epsilon}}$$

in which $N_z(r)$ *is the* r*-zooming number of* I*. We have* $R_z(T) \quad 3R=(8 \ \log T):$
(ii) *for any algorithm* A*, there exists at least one problem instance* $I \ 2\ I$ *on which the expected regret of* A *satisfies* $\mathbb{E}[R(T)] \quad R=(2560 \ \log T):$

Proof. Our proof is inspired by Slivkins (2014) and makes use of the needle-in-the-haystack technique, which is firstly proposed by Auer et al. (2002b) for analyzing multi-armed bandits and then extended to Lipschitz bandits by Kleinberg et al. (2013).

Step 1 (Constructing instance set $\mathcal{I}$) We begin with the following lemma.

Lemma 9 *Define*

$$R_c'(T) = \inf_{r_0 \in (0,1)} \quad r_0 T + \log T \quad \frac{N_c(r_0)}{r_0^{1/\epsilon}} \quad :$$

Then for any $T > 2$, $R_c(T)$ $R_c'(T)$.

Proof. Since $N_c(r)$ is non-increasing with r, we have

$$R_c(T) = \inf_{r_0 \in (0,1)} \quad r_0 T + \log T \sum_{r = 2^{-i}: i \in \mathbb{N}, r \geq r_0} \frac{N_c(r)}{r^{1/\epsilon}} \qquad \inf_{r_0 \in (0,1)} \quad r_0 T + \log T \quad N_c(r_0) \sum_{r = 2^{-i}: i \in \mathbb{N}, r \geq r_0} r^{-\frac{1}{\epsilon}}$$

in which the last term can be upper bounded as follows

$$\sum_{r = 2^{-i}: i \in \mathbb{N}, r \geq r_0} r^{-\frac{1}{\epsilon}} = \sum_{i=0}^{\lfloor \log_2 \frac{1}{r_0} \rfloor} 2^{\frac{i}{\epsilon}} \int_0^{\log_2 \frac{1}{r_0} + 1} 2^{\frac{i}{\epsilon}} \, \mathrm{d}i = \frac{(2 = r_0)^{1/\epsilon} \quad 1}{\log (2^{1/\epsilon})} \quad \frac{2^{1/\epsilon}}{\log 2 \quad r_0^{1/\epsilon}};$$

Recalling $= \frac{2^{1/\epsilon} \cdot \epsilon}{\log 2} > 1$, we conclude the proof.

Fix $T > 2$ and R $R_c(T)$. Let $r = \frac{R}{2\kappa T(1 + \log T)}$ and $N = \max(2; \lfloor T r^{1+1/\epsilon} c \rfloor)$. Based on Lemma 9, we can bound r and N as follows.

Lemma 10 *We say a subset S X is an r-packing of X if the distance between any two points in S is at least r, i.e., $\inf_{u,v \in S} D(u; v)$ r. Let $N_p(r)$ denote the r-packing number of X, defined as the maximal number of points in an r-packing of X:*

$$N_p(r) = \max \{ |S| : S \text{ is an } r\text{-packing of } X \}:$$

We have

$$r < 1 = 2 \quad \text{and} \quad N \quad N_p(r):$$

Proof. Define function $f(r) = \frac{N_c(r)}{r^{1+1/\epsilon}}$. Since $f(1) = 1$, $\lim_{r \to 0} f(r) = +1$, and $f(r)$ is decreasing on $(0; 1)$, there must exist $b \in (0; 1)$ such that $f(b)$ T $f(b = 2)$. From the first inequality, we obtain

$$R \quad R_c(T) \qquad R_c'(T) \qquad bT + \frac{N_c(b)}{b^{1/\epsilon}} \log T \qquad b \ T(1 + \log T)$$

which implies r $\frac{\widehat{r} \kappa T(1 + \log T)}{2\kappa T(1 + \log T)} = \frac{\widehat{r}}{2} < \frac{1}{2}$. From the second inequality, we have

$$T r^{1+1/\epsilon} \quad T \ (b = 2)^{1+1/\epsilon} \quad \frac{N_c(b = 2)}{(b = 2)^{1+1/\epsilon}} \ (b = 2)^{1+1/\epsilon} \quad N_c(r):$$

We conclude the proof by the fact that $N_c(r)$ $N_p(r)$ (Kleinberg et al., 2013) and 2 $N_p(r)$.

The above lemma ensures that we can find a set of arms $U = \{u_1; \ldots; u_N\}$ X such that $\inf_{x,y \in \mathcal{U}} D(x; y)$ r. Based on U, we construct a set of problem instances $\mathcal{I} = \{I_1; \ldots; I_N\}$. Let us fix $i \in [N]$ and describe the construction of I_i: the expected reward function $_i$ is defined as

$$_i(x) = \begin{cases} \frac{7r}{8}; & x = u_i \\ \frac{3r}{4}; & x = u_j; \ j \in [N] \text{ and } j \neq i \\ \max(\frac{r}{2}; \max_{u \in \mathcal{U}} \ _i(u) \quad D(x; u)); & \text{otherwise} \end{cases} \tag{22}$$

and the reward distributions are defined by

$$\Pr(y|x) = p_i(y|x) = \begin{cases} \ _i(x) r^{1/\epsilon}; & y = r^{-1/\epsilon} \\ 1 \quad \ _i(x) r^{1/\epsilon}; & y = 0 \end{cases} : \tag{23}$$

One can show that for $i = 1; 2; \ldots; N$, $_i$ is Lipschitz and the $(1 + \ )$-th moment of p_i is upper bounded by $7 = 8$.

Step 2 (Proving i) Let I be a problem instance in $\mathcal{I}$. Recall the definition of -optimal region: $X_\rho = \{x \in X : =2 < \Delta(x) \quad g$. It is clear that for $\quad 3r$=4, we have $X_\rho = \varnothing$ and thus $N_z(\ ) = 0$. It follows that

$$R_z(T) = \inf_{r_0 \in (0,1)} \quad r_0 T + \log T \sum_{\rho = 2^{-i}: i \in \mathbb{N}, \rho \geq r_0} \frac{N_z(\ )}{1/\epsilon} \quad \frac{3}{4} r T + \log T \sum_{\rho = 2^{-i}: i \in \mathbb{N}, \rho \geq \frac{3}{4} r} \frac{N_z(\ )}{1/\epsilon} \quad \frac{3R}{8 \ \log T}$$

where the last inequality is due to $r = \frac{R}{2\kappa T(1+\log T)}$.

Step 3 (Proving ii) Following the framework of Kleinberg et al. (2013), we first introduce an auxiliary problem instance I_0 in which the expected reward function $\ _0$ is defined as

$$_0(x) = \begin{cases} \frac{3r}{4}; & x = u_j;\ j \in [N] \\ \max(\frac{r}{2}; \max_{u \in \mathcal{U}} \ _0(u) \quad D(x;u)); & \text{otherwise} \end{cases}$$

and the reward distributions are defined by

$$\Pr(y|x) = p_0(y|x) = \begin{cases} \ _0(x) r^{1/\epsilon}; & y = r^{-1/\epsilon} \\ 1 \quad \ _0(x) r^{1/\epsilon}; & y = 0 \end{cases} :$$

The advantage of this construction of I_0 is that the extent to which any other problem instance $I_i \in \mathcal{I}$ deviates from I_0 can be controlled as follows. Let $S_i = B(u_i$

E. Proof of Lemma 1

Lemma 1 *With probability at least* $1 \quad 2$ *, for all rounds* t *2* $[T]$ *and all active arms* x *2* A_t*, we have*

$$j\mathrm{b}_t(x) \quad (x)j \quad r_{t+1}(x):$$

Proof. Fix t *2* $[T]$ and x *2* A_t. By Lemma 1 in Bubeck et al. (2013), with probability at least $1 \quad \frac{2\delta}{T^2}$ the following holds

$$j\mathrm{b}_t(x) \quad (x)j \quad 4 \ ^{\frac{1}{1+\epsilon}} \quad \frac{\log(T^2= \ )}{n_t(x)} \ ^{\frac{\epsilon}{1+\epsilon}} \quad 4^{-\frac{1}{1+\epsilon}} \quad \frac{\log(T^2= \ )}{n_t(x)} \ ^{\frac{\epsilon}{1+\epsilon}} = r_{t+1}(x):$$

Taking the union bound over x *2* A_t and $t = 1;2;:::;T$ and noticing $jA_tj \quad T;8t$ *2* $[T]$, we conclude the proof.

F. Proof of Lemma 2

Lemma 2 *With probability at least* $1 \quad 2$ *, for all rounds* t *2* $[T]$ *and all active arms* x *2* A_t*, we have*

$$\Delta(x) \quad 3 \ ^{p}\overline{2}r_{t+1}(x):$$

Proof. Fix t *2* $[T]$. For each active arm x *2* A_t, there exist three different scenarios as follows.

(i) x is pulled by Step 7 of Algorithm 2 in round t. In this scenario, on one hand, we have $n_t(x) = 1$ and

$$r_{t+1}(x) = 4^{-\frac{1}{1+\epsilon}} \quad \frac{\log(T^2= \ )}{n_t(x)} \ ^{\frac{\epsilon}{1+\epsilon}} \quad 4^{-\frac{1}{1+\epsilon}} \quad \frac{1}{3}\overline{p}\overline{\frac{1}{2}}$$

where we use the fact that $\log(T^2= \ ) \quad \log 4 \quad 1$ for $T > 1$. On the other hand, let x_* *2* $\arg\max_{x\in\mathcal{X}} \ (x)$ be an optimal arm. We have

$$\Delta(x) = \ (x_*) \quad (x) \quad D(x_*;x) \quad 1$$

where the first inequality holds since $\overline{p}$ is Lipschitz, and the second inequality is due to the assumption in (2). Thus, we obtain $3 \ ^{p}\overline{2}r_{t+1}(x) \quad \Delta(x)$.

(ii) x is pulled by Step 9 of Algorithm 2 in round t. In this case, we have $t \quad 2$ and $n_{t-1}(x) \quad 1$ and the arm selection rule implies

$$\mathrm{b}_{t-1}(x) + 2r_t(x) \quad \mathrm{b}_{t-1}(x') + 2r_t(x'); \ 8x' \ 2 \ A_t: \tag{25}$$

Note that $A_t = A_{t-1}$. By Lemma 1, we get

$$(x) \quad \mathrm{b}_{t-1}(x) \quad r_t(x); \ \mathrm{b}_{t-1}(x') \quad (x') \quad r_t(x'); \ 8x' \ 2 \ A_t: \tag{26}$$

Combining (25) and (26), we obtain

$$(x) + 3r_t(x) \quad (x') + r_t(x'); \ 8x' \ 2 \ A_t: \tag{27}$$

Note that the execution of Step 9 implies $X \quad [_{x\in\mathcal{A}_t} B(x; r_t(x))$. Thus, for the optimal arm x_* there must exist an active arm $\bar{x}_*$ *2* A_t such that

$$D(x_*;\bar{x}_*) \quad r_t(\bar{x}_*)$$

which, together with the Lipschitz property of , indicates

$$(x_*) \quad (\bar{x}_*) + r_t(\bar{x}_*):$$

Combining the above inequality and (27) with substitution $x' = \bar{x}_*$, we obtain

$$(x) + 3r_t(x) \quad (x_*):$$

On the other hand, we have

$$\frac{r_{t+1}(x)}{r_t(x)} = \ \frac{n_{t-1}(x)}{n_t(x)} \ ^{\frac{\epsilon}{1+\epsilon}} = \ \frac{n_{t-1}(x)}{n_{t-1}(x)+1} \ ^{\frac{\epsilon}{1+\epsilon}} \quad \overline{p}\overline{\frac{1}{2}}:$$

Therefore, we get $3 \ ^{p}\overline{2}r_{t+1}(x) \quad 3r_t(x) \quad \Delta(x):$

(iii) x is not played in round t. In this scenario, let s be the last round in which x is pulled. Then, we have $r_{t+1}(x) = r_{s+1}(x)$ and the proof reduces to (i) or (ii).

G. Proof of Lemma 3

Lemma 3 *With probability at least* $1 \quad 2$ *, for all* $i \; 2 \; \mathbb{N}$*,*

$$j\bar{A}_T(i)j \quad N_z(2^{-i}):$$

Proof. By the definition of the r-zooming number, for $i = 0; 1; 2; :::$, the set $\bar{A}_T(i)$ can be covered by not more than $N_z(2^{-i})$ balls of radius at most $\frac{1}{18\times 2^i}$. In the following, we show that each of these balls contains at most one arm from $\bar{A}_T(i)$. In fact, suppose that there exist two arms $u; v \; 2 \; \bar{A}_T(i)$ falling into the same ball. On one hand, we have

$$D(u; v) \quad \frac{1}{9 \quad 2^i}: \tag{28}$$

On the other hand, without loss of generality, we assume arm u is added into the active arm set A_T before arm v. Let t be the time when arm v is added into A_T. The execution of Algorithm 2 ensures $t \quad 2$ and

$$D(u; v) > r_t(u): \tag{29}$$

By Lemma 2, we have

$$r_t(u) = r_{t+1}(u) \quad \frac{\Delta(u)}{3}\overline{p}\overline{2} > \frac{1}{6}\overline{p}\overline{2} \quad 2^i > \frac{1}{9 \quad 2^i}$$

which, together with (28) and (29), leads to a contradiction. Thus, $j\bar{A}_T(i)j \quad N_z(2^{-i})$.

H. Proof of Lemma 4

Lemma 4 *With probability at least* $1 \quad 2$ *, for all* $i \; 2 \; \mathbb{N}$*,*

$$\sum_{x\in\mathcal{A}_T(i)} n_T(x)\Delta(x) \quad 2^{\frac{i+1}{\epsilon}} \quad 17^{\frac{\epsilon+1}{\epsilon}-\frac{1}{\epsilon}} \log(T^2= \;)N_z(2^{-i}):$$

Proof. For any arm $x \; 2 \; \bar{A}_T(i)$, by Lemma 2 we have

$$\Delta(x) \quad 3\overline{p}\overline{2}r_{T+1}(x) \quad 17^{-\frac{1}{1+\epsilon}} \quad \frac{\log(T^2= \;)}{n_T(x)} \quad ^{\frac{\epsilon}{1+\epsilon}}:$$

Rearranging the above inequality, we obtain

$$n_T(x)\Delta(x) \quad 17^{\frac{\epsilon+1}{\epsilon}-\frac{1}{\epsilon}} \log(T^2= \;)\Delta(x)^{-\frac{1}{\epsilon}} \quad 2^{\frac{i+1}{\epsilon}} \quad 17^{\frac{\epsilon+1}{\epsilon}-\frac{1}{\epsilon}} \log(T^2= \;)$$

where the second inequality is due to $\Delta(x) > 2^{-(i+1)}$: We finish the proof by applying Lemma 3.

I. Proof of Lemma 11

Lemma 11 *The Kullback–Leibler divergence from* Q_k *to* Q_0 *satisfies*

$$KL(Q_0; Q_k) \quad 39=200:$$

Proof. We first bound the KL divergence from p_0 to p_k:

$$KL(p_0; p_k) = \quad _0(x)r^{1/\epsilon} \log \quad \frac{_0(x)r^{1/\epsilon}}{_k(x)r^{1/\epsilon}} \quad + (1 \quad \quad _0(x)r^{1/\epsilon}) \log \quad \frac{1 \quad \quad _0(x)r^{1/\epsilon}}{1 \quad \quad _k(x)r^{1/\epsilon}} \quad :$$

In the following, we consider two different scenarios, i.e., $x \; 2 \; X \quad S_k$ and $x \; 2 \; S_k$.
(i) $x \; 2 \; X \quad S_k$. By (24), we have $\quad _k(x) = \quad _0(x)$ and thus

$$KL(p_0; p_k) = 0: \tag{30}$$

(ii) $x \in S_k$. By (24), we have $\nu_0(x) < \nu_k(x) \le \nu_0(x) + r/8$ which implies

$$\begin{aligned}
KL(p_0; p_k) &\le \nu_0(x) r^{1/\epsilon} \log \frac{\nu_0(x) r^{1/\epsilon}}{\nu_0(x) r^{1/\epsilon} - r^{1+1/\epsilon}/8} + (1 - \nu_0(x) r^{1/\epsilon}) \log \frac{1 - \nu_0(x) r^{1/\epsilon}}{1 - \nu_0(x) r^{1/\epsilon} - r^{1+1/\epsilon}/8} \\
&\quad (1 - \nu_0(x) r^{1/\epsilon}) \frac{r^{1+1/\epsilon}/8}{1 - \nu_0(x) r^{1/\epsilon} - r^{1+1/\epsilon}/8} \\
&= (1 - \nu_0(x) r^{1/\epsilon} - r^{1+1/\epsilon}/8 + r^{1+1/\epsilon}/8) \frac{r^{1+1/\epsilon}/8}{1 - \nu_0(x) r^{1/\epsilon} - r^{1+1/\epsilon}/8} \\
&= \frac{r^{1+1/\epsilon}}{8} + \frac{r^{1+1/\epsilon}/8}{1 - \nu_0(x) r^{1/\epsilon} - r^{1+1/\epsilon}/8} \\
&\le \frac{r^{1+1/\epsilon}}{8} + \frac{r^{1+1/\epsilon}/8}{1 - 7r^{1+1/\epsilon}/8} \\
&\le \frac{r^{1+1/\epsilon}}{8} + \frac{r^{1+1/\epsilon}/8}{4r^{1+1/\epsilon} - 7r^{1+1/\epsilon}/8} \\
&= \frac{13}{100} r^{1+1/\epsilon}
\end{aligned} \tag{31}$$

where the second inequality follows from the well-known inequality: $\log a \le a - 1, \forall a > 0$, the third inequality is due to $\nu_0(x) \le [r/2, 3r/4]$, and the last inequality holds since $4r^{1+1/\epsilon} \le 4 (1/2)^{1+1/\epsilon} \le 4 (1/2)^2 = 1$.

We continue the proof of Lemma 11 as follows. Denote by $KL(\cdot; \cdot | \cdot)$ the conditional KL divergence also known as conditional relative entropy (Cover & Thomas, 1991; Kleinberg et al., 2013). For $t = 1, \dots, T$, we have

$$\begin{aligned}
KL(Q_0^t; Q_k^t | h^{t-1}) &= \sum_{h^t \in \mathcal{H}^t} Q_0^t(h^t) \log \frac{Q_0^t(h^t | h^{t-1})}{Q_k^t(h^t | h^{t-1})} \\
&= \sum_{h^t \in \mathcal{H}^t} Q_0^t(h^t) \log \frac{Q_0^t(x_t | h^{t-1})}{Q_k^t(x_t | h^{t-1})} \frac{Q_0^t(y_t | x_t; h^{t-1})}{Q_k^t(y_t | x_t; h^{t-1})} \\
&= \sum_{h^t \in \mathcal{H}^t} Q_0^t(h^t) \log \frac{Q_0^t(y_t | x_t; h^{t-1})}{Q_k^t(y_t | x_t; h^{t-1})}
\end{aligned}$$

where the first equality is the definition of conditional KL divergence and the last equality is due to the fact that the distribution of x_t given h^{t-1} depends only on the algorithm A. We proceed as follows

$$\begin{aligned}
KL(Q_0^t; Q_k^t | h^{t-1}) &= \sum_{h^t \in \mathcal{H}^t} Q_0^t(h^t) \log \frac{Q_0^t(y_t | x_t; h^{t-1})}{Q_k^t(y_t | x_t; h^{t-1})} \\
&= \sum_{h^{t-1} \in \mathcal{H}^{t-1}} \int_{x_t \in \mathcal{X}} \sum_{y_t \in \{0, r^{-1/\epsilon}\}} Q_0^t(y_t | x_t; h^{t-1}) \log \frac{Q_0^t(y_t | x_t; h^{t-1})}{Q_k^t(y_t | x_t; h^{t-1})} \, \mathrm{d}\, Q_0^t(x_t; h^{t-1}) \\
&= \sum_{h^{t-1} \in \mathcal{H}^{t-1}} \int_{x_t \in \mathcal{X}} KL(p_0; p_k) \, \mathrm{d}\, Q_0^t(x_t; h^{t-1}) \\
&\overset{(30)}{=} \sum_{h^{t-1} \in \mathcal{H}^{t-1}} \int_{x_t \in S_k} KL(p_0; p_k) \, \mathrm{d}\, Q_0^t(x_t; h^{t-1}) \\
&\overset{(31)}{\le} \sum_{h^{t-1} \in \mathcal{H}^{t-1}} \int_{x_t \in S_k} \frac{13}{100} r^{1+1/\epsilon} \, \mathrm{d}\, Q_0^t(x_t; h^{t-1}) \\
&= \frac{13}{100} r^{1+1/\epsilon} Q_0^t(x_t \in S_k).
\end{aligned}$$

Finally, by the chain rule of KL divergence we have

$$KL(Q_0; Q_k) = KL(Q_0^T; Q_k^T) = \sum_{t=1}^{T} KL(Q_0^t; Q_k^t | h^{t-1}) \le \sum_{t=1}^{T} \frac{13}{100} r^{1+1/\epsilon} Q_0^t(x_t \in S_k) = \frac{13}{100} r^{1+1/\epsilon} \mathbb{E}_{Q_0}[Z_k].$$

where we use the convention that $h^0 = \varnothing$. Recalling $\mathbb{E}_{Q_0}[Z_k]$ $T=N$ and $N = \max(2; bTr^{1+1/\epsilon}c)$, we obtain

$$\mathbb{E}_{Q_0}[Z_k] \quad \frac{3}{2}r^{-(1+1/\epsilon)}$$

which completes the proof.