# Multi-armed Bandits with Cost Subsidy: Supplementary Material

**Outline** The supplementary material of the paper is organized as follows.

- Appendix A contains technical lemmas used in subsequent proofs.
- Appendix B contains a proof of the lower bound.
- Appendix C contains proofs related to the performance of various algorithms presented in the paper.
- Appendix D gives a detailed description of the CS-ETCalgorithm when the costs of the arms are unknown and random.

## A Technical Lemmas

**Lemma 2** (Taylor's Series Approximation)**.** *For $x > 0$, $\ln(1+x) \geq x - \frac{x^2}{1-x^2}$.*

*Proof.* For $x > 0$,

$$\begin{aligned}
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\
&\geq x - \frac{x^2}{2} - \frac{x^4}{4} - \cdots \qquad \text{(because } x > 0) \\
&\geq x - x^2 - x^4 \\
&= x - x^2(1 + x^2 + x^4 + \cdots) \\
&= x - \frac{x^2}{1-x^2}.
\end{aligned}$$

□

**Lemma 3** (Taylor's Series Approximation)**.** *For $x > 0$, $\ln(1-x) \geq -x - \frac{x^2}{1-x}$.*

*Proof.* For $x > 0$,

$$\begin{aligned}
\ln(1-x) &= -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\
&\geq -x - x^2 - x^3 - x^4 - \cdots \qquad \text{(because } x > 0) \\
&= -x - x^2(1 + x + x^2 + \cdots) \\
&= -x - \frac{x^2}{1-x}.
\end{aligned}$$

□

**Lemma 4** (Pinsker's inequality)**.** *Let $Ber(x)$ denote a Bernoulli distribution with mean $x$ where $0 \leq x \leq 1$. Then, $KL(Ber(p); Ber(p+\epsilon)) \leq \frac{4\epsilon^2}{p}$ where $0 < p \leq \frac{1}{2}$, $0 < \epsilon \leq \frac{p}{2}$ and $p + \epsilon < 1$ and the KL divergence between two Bernoulli distributions with mean $x$ and $y$ is given as $KL(Ber(x); Ber(y)) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$.*

*Proof.* $KL(Ber(p); Ber(p+\epsilon)) = p \ln \frac{p}{p+\epsilon} + (1-p) \ln \frac{1-p}{1-p-\epsilon}$

$$
\begin{aligned}
KL(Ber(p); Ber(p+\epsilon)) &= p \ln \frac{p}{p+\epsilon} + (1-p) \ln \frac{1-p}{1-p-\epsilon} \\
&= -p \ln\left(1 + \frac{\epsilon}{p}\right) - (1-p) \ln\left(1 - \frac{\epsilon}{1-p}\right) \\
&\leq -p\left(\frac{\epsilon}{p} - \frac{\frac{\epsilon^2}{p^2}}{1-\frac{\epsilon^2}{p^2}}\right) - (1-p)\left(-\frac{\epsilon}{1-p} - \frac{\frac{\epsilon^2}{(1-p)^2}}{1-\frac{\epsilon}{1-p}}\right) \\
&\qquad \text{(Using Lemmas 2 and 3. )}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
KL(Ber(p); Ber(p+\epsilon)) &\leq -\epsilon + \frac{\epsilon^2}{p\left(1-\frac{\epsilon^2}{p^2}\right)} + \epsilon + \frac{\epsilon^2}{(1-p)\left(1-\frac{\epsilon}{1-p}\right)} \\
&\leq \frac{\epsilon^2}{p\left(1-\frac{1}{4}\right)} + \frac{\epsilon^2}{(1-p)\left(1-\frac{1}{2}\right)} \qquad \text{(because } \frac{\epsilon}{1-p} \leq \frac{\epsilon}{p} \leq \frac{1}{2}\text{)} \\
&= \frac{4\epsilon^2}{3p} + \frac{2\epsilon^2}{1-p} \\
&\leq \frac{2\epsilon^2}{p} + \frac{2\epsilon^2}{1-p} \\
&\leq \frac{2\epsilon^2}{p} + \frac{2\epsilon^2}{p} \qquad \text{(because } p \leq \frac{1}{2}\text{)} \\
&= \frac{4\epsilon^2}{p}.
\end{aligned}
$$

□

## B Proof of Lower Bound

*Proof of Lemma 1.* In the family of instances $\Phi_{\theta,p,\epsilon}$, the costs of the arms are same across instances. Arm 0 is the cheapest arm in all the instances. With this, we define a modified notion of quality regret which penalizes the regret only when this cheap arm is pulled as

$$
\mathsf{Mod\_Quality\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \boldsymbol{c}) = \sum_{t=1}^{T} \max\{\mu_{m*} - \mu_{\pi_t}, 0\} \mathbf{I}(c_{i_t} = 0). \tag{2}
$$

An equivalent notation for denoting the modified regret of policy $\pi$ on an instance $I$ of the problem is $\mathsf{Mod\_Quality\_Reg}_\pi(T, \alpha, I)$. This modified quality regret is at most equal to the quality regret. For proving the lemma, we will show a stronger result that there exists an instance $\phi_{0,p,\epsilon}$ such that $\mathsf{Mod\_Quality\_Reg}(T, 0, \phi_{0,p,\epsilon}) + \mathsf{Cost\_Reg}(T, 0, \phi_{0,p,\epsilon})$ is $\Omega\left(pK^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ which will imply the required result.

Let us first consider any deterministic policy (or algorithm) $\pi$. For a deterministic algorithm, the number of times an arm is pulled is a function of the observed rewards. Let the number of times arm $j$ is played be denoted by $N_j$ and let the total number of times any arm with cost 1 i.e. an expensive arm is played be $N_{exp} = 1 - N_0$. For any $a$ such that $1 \leq a \leq K$, we can use the proof of Lemma A.1 in Auer et al. (2002b), with function $f(\mathbf{r}) = N_{exp}$ to get

$$
\mathbf{E}^a\left[N_{exp}\right] \leq \mathbf{E}^0\left[N_{exp}\right] + 0.5T\sqrt{2\mathbf{E}^0[N_a] KL(Ber(p); Ber(p+\epsilon))}
$$

where $\mathbf{E}^j$ is the expectation operator with respect to the probability distribution defined by the random rewards in instance $\Phi^j_{0,p,\epsilon}$. Thus, using Lemma 4, we get,

$$
\mathbf{E}^a\left[N_{exp}\right] \leq \mathbf{E}^0\left[N_{exp}\right] + 0.5T\sqrt{\mathbf{E}^0[N_a] 8\epsilon^2/p}. \tag{3}
$$

Now, let us look at the regret of the algorithm for each instance in the family $\Phi_{0,p,\epsilon}$. We have

1. $\mathsf{Cost\_Reg}_\pi(T, \alpha, \Phi^0_{0,p,\epsilon}) = \mathbf{E}^0[N_{exp}],\ \mathsf{Mod\_Quality\_Reg}_\pi(T, \alpha, \Phi^0_{0,p,\epsilon}) = 0$
2. $\mathsf{Cost\_Reg}_\pi(T, \alpha, \Phi^a_{0,p,\epsilon}) = 0,\ \mathsf{Mod\_Quality\_Reg}_\pi(T, \alpha, \Phi^a_{0,p,\epsilon}) = \epsilon\,(T - \mathbf{E}^a[N_{exp}])$.

Now, define randomized instance $\phi_{0,p,\epsilon}$ as the instance obtained by randomly choosing from the family of instances $\Phi_{0,p,\epsilon}$ such that $\phi_{0,p,\epsilon} = \Phi^0_{0,p,\epsilon}$ with probability $1/2$ and $\phi_{0,p,\epsilon} = \Phi^a_{0,p,\epsilon}$ with probability $1/2K$ for $1 \le a \le K$. The expected regret of this randomized instance is

$$
\begin{aligned}
&\mathbf{E}\left[\mathsf{Mod\_Quality\_Reg}_\pi(T, 0, \phi_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T, 0, \phi_{0,p,\epsilon})\right] \\
&= \frac{1}{2}\left(\mathsf{Mod\_Quality\_Reg}_\pi(T, \alpha, \Phi^0_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T, \alpha, \Phi^0_{0,p,\epsilon})\right) + \\
&\quad \frac{1}{2K}\sum_{a=1}^{K}\left(\mathsf{Mod\_Quality\_Reg}_\pi(T, \alpha, \Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T, \alpha, \Phi^a_{0,p,\epsilon})\right) \\
&= \frac{1}{2}\mathbf{E}^0[N_{exp}] + \frac{1}{2K}\sum_{a=1}^{K}\epsilon(T - \mathbf{E}^a[N_{exp}]) \\
&\ge \frac{1}{2}\mathbf{E}^0[N_{exp}] + \frac{1}{2K}\sum_{a=1}^{K}\epsilon\left(T - \mathbf{E}^0\left[N_{exp}\right] - \frac{1}{2}T\sqrt{\mathbf{E}^0[N_a]\frac{8\epsilon^2}{p}}\right) \qquad \text{(using (3))} \\
&= \frac{1}{2}\left[\epsilon T + (1-\epsilon)\sum_{a=1}^{K}\mathbf{E}^0[N_a] - \frac{T\epsilon}{2K}\sum_{a=1}^{K}\sqrt{\frac{8\epsilon^2}{p}\mathbf{E}^0[N_a]}\right] \\
&= \frac{1}{2}\sum_{a=1}^{K}\left[\frac{\epsilon T}{K} + (1-\epsilon)(\mathbf{E}^0[N_a])^2 - T\mathbf{E}^0[N_a]\epsilon^2\frac{\sqrt{2}}{K\sqrt{p}}\right] \\
&= \frac{1}{2}\sum_{a=1}^{K}\left[\left(\sqrt{1-\epsilon}\mathbf{E}^0[N_a] - \frac{\epsilon^2 T}{2K}\sqrt{\frac{2}{p(1-\epsilon)}}\right)^2 + \frac{\epsilon T}{K} - \frac{\epsilon^4 T^2}{2pK^2(1-\epsilon)}\right] \\
&\ge \frac{1}{2}\sum_{a=1}^{K}\frac{\epsilon T}{K} - \frac{\epsilon^4 T^2}{2pK^2(1-\epsilon)} \\
&= \frac{\epsilon T}{2} - \frac{\epsilon^4 T^2}{4pK(1-\epsilon)}
\end{aligned}
$$

Taking $\epsilon = \frac{p}{2}(\frac{K}{T})^{\frac{1}{3}}$, we get $\mathbf{E}\left[\mathsf{Mod\_Quality\_Reg}_\pi(T, 0, \phi_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T, 0, \phi_{0,p,\epsilon})\right]$ is $\Omega(pK^{1/3}T^{2/3})$ when $K \le T$.

Using Yao's principle, for any randomized algorithm $\pi$, there exists an instance $\Phi^j_{0,p,\epsilon}$ with $0 \le j \le K$ such that $\mathsf{Mod\_Quality\_Reg}_\pi(T, 0, \Phi^j_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T, 0, \Phi^j_{0,p,\epsilon})$ is $\Omega(pK^{1/3}T^{2/3})$. Also, since $\mathsf{Mod\_Quality\_Reg}_\pi(T, 0, \Phi^j_{0,p,\epsilon}) \le \mathsf{Quality\_Reg}_\pi(T, 0, \Phi^j_{0,p,\epsilon})$, we have $\mathsf{Quality\_Reg}_\pi(T, 0, \Phi^j_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_\pi(T, 0, \Phi^j_{0,p,\epsilon})$ is $\Omega(pK^{1/3}T^{2/3})$. □

*Proof of Theorem 2.* **Notation**: For any instance $\phi$, we define the arms $m^\phi_*$ and $i^\phi_*$ as $m^\phi_* = \arg\max_i \mu^i_\phi$ and $i^\phi_* = \arg\min_i c^i_\phi$ s.t. $q_{i_\phi} \ge (1-\theta)q_{m^*_\phi}$. When the instance is clear, we will use the simplified notation $i_*$ and $m_*$ instead of $i^\phi_*$ and $m^\phi_*$.

**Proof Sketch:** Lemma 1 establishes that when $\alpha = 0$, for any given policy, there exists an instance on which the sum of quality and cost regret are $\Omega(K^{1/3}T^{2/3})$. Now, we generalize the above result for $\alpha = 0$ to any $\alpha$ for $0 \le \alpha \le 1$. The main idea in our reduction is to show that if there exists an algorithm $\pi_\alpha$ for $\alpha > 0$ that achieves $o(K^{1/3}T^{2/3})$ regret on every instance in the family $\Phi_{\alpha,p,\epsilon}$, then we can use $\pi_\alpha$ as a subroutine to construct an algorithm $\pi_0$ for problem (1) that achieves $o(K^{1/3}T^{2/3})$ regret on every instance in the family $\Phi_{0,p,\epsilon}$, thus contradicting the lower bound of Lemma 1. This will prove the theorem by contradiction. In order to construct the aforementioned sub-routine, we leverage techniques from *Bernoulli factory* to generate a sample from a Bernoulli random variable with parameter $\mu/(1-\alpha)$ using samples from a Bernoulli random variable with parameter $\mu$, for any $\mu, \alpha < 1$.

**Aside on Bernoulli Factory:** The key tool we use in constructing the algorithm $\pi_0$ from $\pi_\alpha$ is *Bernoulli factory* for the linear function. The Bernoulli factory for a specified scaling factor $C > 1$ i.e. $\mathit{BernoulliFactory}(C)$ uses a sequence of independent and identically distributed samples from $Ber(r)$ and returns a sample from $Ber(Cr)$.The key aspect of a Bernoulli factory is the number of samples needed from $Ber(r)$ to generate a sample from $Ber(Cr)$. We use the Bernoulli factory described in Huber (2013) which has a guarantee on the expected number of samples $\tau$ from $Ber(r)$ needed to generate a sample from $Ber(Cr)$. In particular, for a specified $\delta > 0$,

$$\sup_{r \in [0, \frac{1-\delta}{C}]} E[\tau] \leq \frac{9.5C}{\delta}. \tag{4}$$

**Detailed proof:** For some value of $p, \epsilon$ (to be specified later in the proof) such that $0 \leq p < 1$ and $0 \leq \epsilon \leq p/2$, consider the family of instances $\Phi_{\alpha,p,\epsilon}$ and $\Phi_{0,p,\epsilon}$. Let $\pi_\alpha$ be any algorithm for the family $\Phi_{\alpha,p,\epsilon}$. Using $\pi_\alpha$, we construct an algorithm $\pi_0$ for the family $\Phi_{0,p,\epsilon}$. This algorithm is described in Algorithm 4. We will use $I_l^\alpha = \pi_\alpha([(I_1^\alpha, r_1), (I_2^\alpha, r_2), \cdots (I_{l-1}^\alpha, r_{l-1})])$ to denote the arm pulled by algorithm $\pi_\alpha$ at time $l$ after having observed rewards $r_l$ $\forall 1 \leq i < l$ through arm pulls $I_l$ $\forall 1 \leq i < l$. The function $\mathit{BernoulliFactory}(C)$ returns two values - a random sample from the distribution $Ber(Cr)$ and the number of samples of $Ber(r)$ needed to generate this random sample.

**Algorithm 4:** Derived Algorithm $\pi_0$

```
Result: Arm $I_t^0$ to be pulled in each round $t$, total number of arm pulls $T$
input : Algorithm $\pi_\alpha$, $L$ - Number of arm pulls for algorithm $\pi_\alpha$
$l = 1, t = 1$ ;
for $l \in [L]$ do
    $I_l^\alpha = \pi_\alpha([(I_1^\alpha, r_1), (I_2^\alpha, r_2), \cdots (I_{l-1}^\alpha, r_{l-1})])$ ;
    if $I_l^\alpha = 0$ then
        Pull arm 0 to obtain outcome $r_l$ ;
        $I_t^0 = I_l^\alpha = 0$ ;
        $U_l = \{t\}$ ;
    else
        Call $r_l, n = \mathit{BernoulliFactory}(\frac{1}{1-\alpha})$ on samples generated from repeated pulls of the arm $I_l^\alpha$ ;
        $U_l = \{t, t+1 \cdots t+n-1\}$ ;
        $I_t^0 = I_{t+1}^0 = \cdots I_{t+n-1}^0 = I_l^\alpha$ ;
    end
    $S_l = |U_l|$ ;
    $l = l + 1$ ;
    $t = t + S_l$ ;
end
$T = t$
```

Now, let us analyze the expected modified regret incurred by algorithm $\pi_0$ on an instance $\Phi_{0,p,\epsilon}^a$ for any $0 \leq a \leq K$ where the expectation is with respect to the random variable $T$, total number of arm pulls.

Similarly, we analyze the cost regret incurred by algorithm $\pi_0$ on an instance $\Phi_{0,p,\epsilon}^a$ for any $0 \leq a \leq K$.

$$\begin{aligned}
&\mathbf{E}\left[\mathsf{Mod\_Quality\_Reg}_{\pi_0}(T, 0, \Phi^a_{0,p,\epsilon})\right] + \mathbf{E}\left[\mathsf{Cost\_Reg}_{\pi_0}(T, 0, \Phi^a_{0,p,\epsilon})\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T}\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_t^0 = 0\}\right] + \mathbf{E}\left[\sum_{t=1}^{T} c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right] \\
&= \mathbf{E}\left[\sum_{l=1}^{L}\sum_{t\in U_l}\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_t^0 = 0\}\right] + \mathbf{E}\left[\sum_{l=1}^{L}\sum_{t\in U_l} c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_t^0}^{\Phi^a_{0,p,\epsilon}}\right] \\
&= \mathbf{E}\left[\sum_{l=1}^{L} S_l\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha = 0\}\right] + \mathbf{E}\left[\sum_{l=1}^{L} S_l\left(c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\right] \\
&= \sum_{l=1}^{L}\mathbf{E}\left[\mathbf{E}\left[S_l|I_l^\alpha\right]\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha = 0\}\right] + \mathbf{E}\left[\sum_{l=1}^{L}\mathbf{E}[S_l|I_l^\alpha]\left(c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\right] \\
&\leq \sum_{l=1}^{L}\mathbf{E}\left[\frac{9.5}{\delta(1-\alpha)}\left(\mu_{m^*}^{\Phi^a_{0,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha = 0\}\right] + \mathbf{E}\left[\sum_{l=1}^{L}\frac{9.5}{\delta(1-\alpha)}\left(c_{i_*}^{\Phi^a_{0,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\right] \qquad \text{(Using (4))} \\
&= \frac{9.5}{\delta(1-\alpha)}\sum_{l=1}^{L}\mathbf{E}\left[\left((1-\alpha)\mu_{m^*}^{\Phi^a_{\alpha,p,\epsilon}} - \mu_{I_l^\alpha}^{\Phi^a_{0,p,\epsilon}}\right)\mathbf{I}\{I_l^\alpha = 0\}\right] + \frac{9.5}{\delta(1-\alpha)}\sum_{l=1}^{L}\mathbf{E}\left[c_{i_*}^{\Phi^a_{\alpha,p,\epsilon}} - c_{I_l^\alpha}^{\Phi^a_{\alpha,p,\epsilon}}\right] \\
&\qquad \text{(Because costs of arms are same in all instances, } i_*^{\Phi^a_{\alpha,p,\epsilon}} = i_*^{\Phi^a_{0,p,\epsilon}} = a \text{ and } \mu_{m^*}^{\Phi^a_{0,p,\epsilon}} = (1-\alpha)\mu_{m^*}^{\Phi^a_{\alpha,p,\epsilon}}\text{ )} \\
&= \frac{9.5}{\delta(1-\alpha)}\mathsf{Quality\_Reg}_{\pi_\alpha}(L, \alpha, \Phi^a_{\alpha,p,\epsilon}) + \frac{9.5}{\delta(1-\alpha)}\mathsf{Cost\_Reg}_{\pi_\alpha}(L, \alpha, \Phi^a_{\alpha,p,\epsilon}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathsf{Quality\_Reg}_{\pi_\alpha}(L, \alpha, \Phi^a_{\alpha,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_\alpha}(L, \alpha, \Phi^a_{\alpha,p,\epsilon}) \\
&\geq \frac{\delta(1-\alpha)}{9.5}\mathbf{E}\left[\mathsf{Mod\_Quality\_Reg}_{\pi_0}(T, 0, \Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_0}(T, 0, \Phi^a_{0,p,\epsilon})\right] \qquad (5) \\
&\geq \frac{\delta(1-\alpha)}{9.5}\mathbf{E}\left[\mathsf{Mod\_Quality\_Reg}_{\pi_0}(L, 0, \Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_0}(L, 0, \Phi^a_{0,p,\epsilon})\right] \qquad \text{(because } L \leq T\text{)} \\
&\geq \frac{\delta(1-\alpha)}{9.5}\left(\mathsf{Mod\_Quality\_Reg}_{\pi_0}(L, 0, \Phi^a_{0,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi_0}(L, 0, \Phi^a_{0,p,\epsilon})\right) \qquad (6)
\end{aligned}$$

Using Lemma 1 and choosing $p = \frac{1-\alpha}{3}, \delta = \frac{1}{2}, \epsilon = \frac{p}{2}(\frac{K}{T})^{1/3}$, we get for any randomized algorithm $\pi_\alpha$, there exists instance $\Phi^b_{\alpha,p,\epsilon}$ (for some $0 \leq b \leq K$) such that $\mathsf{Quality\_Reg}_{\pi}(T, \alpha, \Phi^b_{\alpha,p,\epsilon}) + \mathsf{Cost\_Reg}_{\pi}(T, \alpha, \Phi^b_{\alpha,p,\epsilon})$ is $\Omega\left((1-\alpha)^2 K^{1/3} T^{2/3}\right)$.

□

## C Performance of Algorithms

We use the following fact in the proof of Theorem 1.

**Fact 1.** *(Abramowitz and Stegun, 1948) For a Normal random variable $Z$ with mean $m$ and variance $\sigma^2$, for any $z$,*

$$\Pr\left(|Z - m| > z\sigma\right) > \frac{1}{4\sqrt{\pi}}\exp(-\frac{7z^2}{2}).$$

*Proof of Theorem 1.* This proof is inspired by the lower bound proof in Agrawal and Goyal (2017b). For any given $\alpha, K$ and $T$, we construct an instance on which the CS-TS algorithm (Algorithm 1) gives linear regret in cost.

Consider an instance $\phi$ with $K$ arms where the costs and mean reward of the $j$-th arm are

$$c_j = \begin{cases} 0 & j = 0 \\ 1 & j \neq 0 \end{cases}, \qquad \mu_j = \begin{cases} (1-\alpha)q + \frac{d}{\sqrt{T}} & j = 0 \\ q & j \neq 0 \end{cases}$$

where $q = \frac{d}{(1-\alpha)\sqrt{T}}$ for some $0 < d < \min\{\sqrt{T}/2, (1-\alpha)\sqrt{T}\}$. Moreover, the reward of each arm is deterministic though this fact is not known to the agent. As in the SMS application, we assume that the cost rewards of all arms are known a priori to the agent.

Let the prior distribution that the agent assumes over the mean reward of each arm be $\mathcal{N}(0, \sigma_0^2)$ for some prior variance $\sigma_0^2$ . Further, the agent assumes that the observed qualities to be normally distributed with noise variance $\sigma_n^2$. As such at the start of period $t$, the agent will consider a normal posterior distribution for each arm $i$ with mean

$$\hat{\mu}_i(t) = \frac{T_i(t)}{\frac{\sigma_n^2}{\sigma_0^2} + T_i(t)} \mu_i \tag{7}$$

and variance

$$\sigma_i(t)^2 = \left( \frac{1}{\sigma_0^2} + \frac{T_i(t)}{\sigma_n^2} \right)^{-1}. \tag{8}$$

As $d < q\alpha\sqrt{T}$, the highest quality across all arms is $q$. Thus, note that all arms are *feasible* in terms of quality i.e. have their quality within $(1-\alpha)$ factor of the best quality arm. Hence, quality regret $\mathsf{Quality\_Reg}_{\mathsf{CS-TS}}(t, \alpha, \phi) = 0 \; \forall t > 0$ (for any algorithm) on this instance.

The first arm is the optimal arm ($i_*$). Thus, the cost regret equals the number of times any arm but the first arm is pulled. In particular, let

$$R_c(T) = \sum_{t=1}^{T} \max\{c_{I_t} - c_{i_*}, 0\} = \sum_{t=1}^{T} \mathbf{1}\{I_t \neq 1\},$$

so that $\mathsf{Cost\_Reg}_{\mathsf{CS-TS}}(T, \alpha, I) = \mathbf{E}\left[R_C(T)\right]$.

Define the event $A_{t-1} = \{\sum_{i \neq 1} T_i(t) \leq sT\sqrt{K}\}$ for a fixed constant $s > 0$. For any $t$, if the event $A_{t-1}$ is not true, then $R_c(T) \geq R_c(t) \geq sT\sqrt{K}$. We can assume that $\Pr(A_{t-1}) \geq 0.5 \; \forall t \leq T$. Otherwise

$$\begin{aligned} \mathsf{Cost\_Reg}_{\mathsf{CS-TS}}(T, \alpha, \phi) &= \mathbf{E}\left[R_C(T)\right] \\ &\geq 0.5 E[R_C(T) | A_{t-1}^c] \\ &= \Omega(T\sqrt{K}). \end{aligned}$$

Now, we will show that whenever $A_{t-1}$ is true, probability of playing a sub-optimal arm is at least a constant. For this, we show that the probability that $\mu_1^{score}(t) \leq \mu_1$ and $\mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}$, for some $1 < i \leq K$ is lower bounded by a constant.

Now, given any history of arm pulls $\mathcal{F}_{t-1}$ before time $t$, $\mu_1^{score}(t)$ is a Gaussian random variable with mean $\hat{\mu}_1(t) = \frac{T_i(t)}{\frac{\sigma_n^2}{\sigma_0^2} + T_i(t)} \mu_1$. By symmetry of Gaussian random variables, we have

$$\begin{aligned} \Pr\left( \mu_1^{score}(t) \leq \mu_1 \middle| \mathcal{F}_{t-1} \right) &\geq \Pr\left( \mu_1^{score}(t) \leq \frac{T_i(t)}{\frac{\sigma_n^2}{\sigma_0^2} + T_i(t)} \mu_1 \middle| \mathcal{F}_{t-1} \right) \\ &= \Pr\left( \mu_1^{score}(t) \leq \hat{\mu}_1(t) \middle| \mathcal{F}_{t-1} \right) \\ &= 0.5. \end{aligned}$$

Based on (7) and (8), given any realization $F_{t-1}$ of $\mathcal{F}_{t-1}$, $\mu_i^{score}(t)$ for $i \neq 1$ are independent Gaussian random variables with mean $\hat{\mu}_i(t)$ and variance $\sigma_i(t)^2$. Thus, we have

$$\begin{aligned}
&\Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{1}{1-\alpha}\left(q(1-\alpha) + \frac{d}{\sqrt{T}}\right) - \hat{\mu}_i(t)\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{d}{(1-\alpha)\sqrt{T}} + \frac{1}{1 + T_i(t)\frac{\sigma_0^2}{\sigma_n^2}}q\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&\geq \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{d}{(1-\alpha)\sqrt{T}} + q\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) - \hat{\mu}_i(t) \geq \frac{2d}{(1-\alpha)\sqrt{T}}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ (\mu_i^{score}(t) - \hat{\mu}_i(t))\frac{1}{\sigma_i(t)} \geq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \Pr\left(\exists i \neq 1,\ Z_i(t) \geq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right)
\end{aligned}$$

where $Z_i(t)$ are independent standard normal variables for all $i, t$. Thus,

$$\begin{aligned}
&\Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= 1 - \Pr\left(\forall i \neq 1,\ Z_i(t) \leq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \\
&= 1 - \Pi_{i\neq 1}\left(1 - \Pr\left(\ Z_i(t) \geq \left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)\frac{1}{\sigma_i(t)}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right)\right) \\
&\geq 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-\frac{7}{2}\frac{1}{\sigma_i(t)^2}\left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)^2\right)\right) \qquad \text{(Using Fact 1)} \\
&= 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-\frac{7}{2}\left(\frac{1}{\sigma_0^2} + \frac{T_i(t)}{\sigma_n^2}\right)\left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)^2\right)\right) \\
&\geq 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-\frac{7}{2}\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_n^2}\right)T_i(t)\left(\frac{2d}{(1-\alpha)\sqrt{T}}\right)^2\right)\right),
\end{aligned}$$

The last inequality follows from the fact that $T_i(t) \geq 1$.

Now, when the event $A_{t-1}$ holds, we have $\sum_{i\neq 1} T_i(t) \leq sT\sqrt{K}$. Thus, the right hand side would be minimized when $T_i(t) = \frac{sT}{\sqrt{K}},\ \forall i \neq 1$. Substituting this value of $T_i(t)$, the right hand side reduces to $g(K) = 1 - \Pi_{i\neq 1}\left(1 - \frac{1}{8\sqrt{\pi}}\exp\left(-14\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_n^2}\right)\frac{s\sqrt{K}d^2}{(1-\alpha)^2}\right)\right)$. Thus, $\Pr\left(\exists i \neq 1,\ \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\ \middle|\mathcal{F}_{t-1} = F_{t-1}\right) \geq g(K)$ whenever $F_{t-1}$ is such that $A_{t-1}$ holds.

Probability of playing any sub-optimal arm at time $t$ is,

$$\begin{aligned}
\Pr\left(\exists i \neq 1,\ I_t = i\right) &\geq \Pr\left(\mu_1^{score}(t) \leq \mu_1,\ \exists i \neq 1 \text{ s.t. } \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\right)\\
&= \mathbf{E}\left[\Pr\left(\mu_1^{score}(t) \leq \mu_1,\ \exists i \neq 1 \text{ s.t. } \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\right)\Big|\mathcal{F}_{t-1}\right]\\
&\geq \mathbf{E}\left[\Pr\left(\mu_1^{score}(t) \leq \mu_1,\ \exists i \neq 1 \text{ s.t. } \mu_i^{score}(t) \geq \frac{\mu_1}{1-\alpha}\right)\Big|\mathcal{F}_{t-1}, A_{t-1}\right]\Pr(A_{t-1})\\
&\geq \frac{1}{2}\cdot g(K)\cdot\frac{1}{2}
\end{aligned}$$

Thus, at every time instant $t$, the probability of playing a sub-optimal is lower bounded by $\frac{g(K)}{4}$. This implies that the cost regret $\mathsf{Cost\_Reg}_{\mathsf{CS-TS}}(T,\alpha,\phi) \geq 0.25Tg(K)$.

□

*Proof of Theorem 3.* This algorithm has two phases - pure exploration and UCB. In the first phase, the algorithm pulls each arm a specified number of times ($\tau$). In the second phase, the algorithms maintains upper and lower confidence bounds on the mean reward of each arm. Then, it estimates a *feasible* set of arms and pulls the cheapest arm in this set.

We will define the *clean event* $\mathcal{E}$ in this proof as the event that for every time $t \in [T]$ and arm $i \in [K]$, the difference between the mean reward and the empirical mean reward does not exceed the size of the confidence interval ($\beta_i(t)$) i.e. $\mathcal{E} = \{|\hat{\mu}_i(t) - \mu_i| \leq \beta_i(t),\ \forall i \in [K],\ t \in [T]\}$.

Define $\hat{t} = K\tau + 1$ as the first round in the UCB phase of the algorithm. Further, define instantaneous cost and quality regret as the regret incurred in the $t$-th arm pull:

$$\begin{aligned}
&\mathsf{Quality\_Reg}_\pi^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) = \mathbf{E}\left[\max\{(1-\alpha)\mu_{m_*} - \mu_{\pi_t}, 0\}\right],\\
&\mathsf{Cost\_Reg}_\pi^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) = \mathbf{E}\left[\max\{c_{\pi_t} - c_{i_*}, 0\}\right],
\end{aligned} \tag{9}$$

where the expectation is over the randomness in the policy $\pi$.

Let us first assume that the clean event holds. As both the instantaneous regrets are upper bounded by 1, $\sum_{t=1}^{K\tau}\mathsf{Quality\_Reg}_\pi^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) \leq K\tau$ and $\sum_{t=1}^{K\tau}\mathsf{Cost\_Reg}_\pi^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}) \leq K\tau$.

Now, let us look at the UCB phase of the algorithm. Here, $\forall\ \hat{t} \leq t \leq T$, we have

$$\mu_{i_*}^{\mathsf{UCB}}(t) \geq \mu_{i_*} \geq (1-\alpha)\mu_{m_*} \geq (1-\alpha)\mu_{m_t} \geq (1-\alpha)\mu_{m_t}^{\mathsf{LCB}}(t).$$

Here, the first and fourth inequality are because of the clean event. The second and third inequality are from the definition of $i_*$ and $m_*$ respectively.

Thus from the inequality above, the optimal arm $i_*$ is in the set $Feas(t), \forall \hat{t} \leq t \leq T$. This implies that the arm pulled in each time step in the UCB phase, is either the optimal arm or an arm cheaper than it. Thus, instantaneous cost regret is zero for all time steps in the UCB phase of the algorithm.

Now, let us look at the quality regret in the UCB phase i.e. for any $\hat{t} \leq t \leq T$. We have

$$\mu_{I_t} + 2\beta_{I_t}(t) \geq \mu_{I_t}^{\mathsf{UCB}}(t) \geq (1-\alpha)\mu_{m_t}^{\mathsf{LCB}}(t) \geq (1-\alpha)\mu_{m_*}^{\mathsf{LCB}}(t) \geq (1-\alpha)\left(\mu_{m_*} - 2\beta_{m_*}(t)\right) \geq (1-\alpha)\mu_{m_*} - 2\beta_{m_*}(t)$$

The first and fourth inequality hold because the clean event holds. The second and third inequalities follow from the definition of $I_t$ and $m_t$ respectively. Thus,

$$\mathsf{Quality\_Reg}_\pi^{inst}(t,T,\alpha,\boldsymbol{\mu},\mathbf{c}|\mathcal{E}) = (1-\alpha)\mu_{m_*} - \mu_{I_t} \leq 2\left(\beta_{I_t}(t) + \beta_{m_*}(t)\right) \leq 2\left(\sqrt{\tfrac{2\log T}{\tau}} + \sqrt{\tfrac{2\log T}{\tau}}\right) = 4\sqrt{\tfrac{2\log T}{\tau}}.$$

The total regret incurred by the algorithm is the sum of the instantaneous regrets across all time steps in the exploration and the UCB phase. Thus,

$\mathsf{Quality\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq K\tau + 4(T - K\tau)\sqrt{\frac{2\log T}{\tau}} \leq K\tau + 4T\sqrt{\frac{2\log T}{\tau}}$ and $\mathsf{Cost\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq K\tau$. Substituting $\tau = (T/K)^{2/3}$, we conclude that both cost and quality regret are $O(K^{1/3}T^{2/3}\sqrt{\log T})$.

Now, when the clean event does not hold, the cost and quality regret are at most $T$ each. The probability that the clean event does not hold is at most $2/T^2$ (Lemma 1.6 in Slivkins (2019)). Thus, the expected cost and quality regret obtained by averaging over the clean event holding and not holding is $O(K^{1/3}T^{2/3}\sqrt{\log T})$.

□

*Proof of Theorem 4.* As in the previous proof, we will define the *clean event* $\mathcal{E}$ as the event that for every time $t \in [T]$ and arm $i \in [K]$, the difference between the mean reward and the empirical mean reward does not exceed the size of the confidence interval ($\beta_i(t)$) i.e. $\mathcal{E} = \{|\hat{\mu}_i(t) - \mu_i| \leq \beta_i(t),\ \forall i \in [K],\ t \in [T]\}$. Also, define the quality and cost gap of each arm as $\Delta_{\mu,i} = \max\{(1-\alpha)\mu_{m^*} - \mu_i, 0\}$ and $\Delta_{c,i} = \max\{c_{i_*} - c_i, 0\}$.

When the clean event does not hold, both cost and quality regrets are upper bounded by $T$. Let us look at the case when the clean event holds and analyze the cost and quality regret.

**Quality Regret:** Let $t_i$ be the last time $t$ when $i \in Feas(t)$ i.e. $t_i = \max\{K,\ \max\{t : i \in Feas(t)\}\}$. Thus, $T_i(T) = T_i(t_i)$.

Consider any arm $i$ which would incur a quality regret on being pulled i.e. arm $i$ such that $\mu_i < (1-\alpha)\mu_{m_*}$. We have

$$\mu_i + 2\beta_i(t_i) \geq \mu_i^{\mathsf{UCB}}(t_i) \geq (1-\alpha)\mu_{m_{t_i}}^{\mathsf{UCB}}(t_i) \geq (1-\alpha)\mu_{m_*}^{\mathsf{UCB}}(t_i) \geq (1-\alpha)\mu_{m_*}.$$

The first and fourth inequality hold because of the clean event. The third inequality is from the definition of $m_{t_i}$.

Thus, $(1-\alpha)\mu_{m_*} - \mu_i \leq 2\beta_i(t_i)$. Using the definition of $\beta_i(t_i)$, we get $T_i(T) = T_i(t_i) \leq \frac{8\log T}{\Delta_{\mu,i}^2}$.

Using Jensen's inequality,

$$\begin{aligned}
\left(\frac{\sum_{i=1}^K T_i(T)\Delta_{\mu,i}}{T}\right)^2 &\leq \frac{\sum_{i=1}^K T_i(T)\Delta_{\mu,i}^2}{T} \\
&= \frac{\sum_{i=1:\ \Delta_{\mu,i}>0}^K T_i(T)\Delta_{\mu,i}^2}{T} \\
&\leq \sum_{i=1:\Delta_{\mu,i}>0}^K \frac{8\log T}{\Delta_{\mu,i}^2}\frac{\Delta_{\mu,i}^2}{T} \\
&= \frac{8K\log T}{T}
\end{aligned}$$

Thus, $\mathsf{Quality\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq \sqrt{8KT\log T}$.

**Cost Regret:** Let $i$ be an arm such that $c_i > c_{i_*}$. Let $\tilde{t}_i$ be the last time when arm $i$ is pulled. Thus, $i_* \notin Feas(\tilde{t}_i)$. We have, $\mu_{i_*} \leq \mu_{i_*}^{\mathsf{UCB}}(\tilde{t}_i) < \mu_i^{\mathsf{UCB}}(\tilde{t}_i) \leq \mu_i(\tilde{t}_i) + 2\sqrt{(2\log T)/T_i(\tilde{t}_i)}$. Thus,

$$T_i(T) = T_i(\tilde{t}_i) < \frac{8\log T}{(\mu_{i_*} - \mu_i)^2} \leq \frac{8\delta^2\log T}{(c_{i_*} - c_i)^2} = \frac{8\delta^2\log T}{\Delta_{c,i}^2}.$$

Using Jensen's inequality as for the case of quality regret, we get, $\mathsf{Cost\_Reg}_\pi(T, \alpha, \boldsymbol{\mu}, \mathbf{c}|\mathcal{E}) \leq \sqrt{8\delta^2 KT\log T}$.

Note that the probability of the clean event is at least $1 - 2/T^2$ (Lemma 1.6 in Slivkins (2019)). Thus, the sum of the expected cost and quality regret by averaging over the clean event holding and not holding is $O((1+\delta)\sqrt{KT\log T})$.

□