Supplementary material to Median-based Bandits for Unbounded Rewards

A. Supplementary Experiments

We provide more experiments to further analyse our policy. Except mentioned, we will use the simulation environment provided in the main paper as our experiment environment.

A.1. Sub-optimal Draws Confidence Interval

To compare with baseline algorithms, we did not show the confidence interval of the expected sub-optimal draws in Figure 3. In Figure 5, we show the interval between one standard deviation from the estimated expected sub-optimal draws. Settings and environments remain the same as the main paper simulation experiment. We can see the confidence interval shrinks as iteration increases.

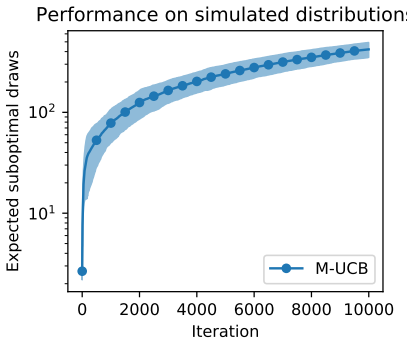

Figure 5. Expected sub-optimal draws with confidence interval for M-UCB ($\alpha = 4, \beta = 1$).

A.2. Estimation of Lower Bound of Hazard Rate

Our policy and bound of expected sub-optimal draws depend on the lower bound of hazard rate. One interesting question is how well we can estimate the lower bound of hazard rate ($\hat{L}_i$), and how sensitivity our policy to the $\hat{L}_i$. Figure 6 shows the estimated value of $\hat{L}_i$ for 10,000 iterations experiment under policy M-UCB. Note that the choice of which arm to be sampled depends on the policy and the number of samples of each arm is not the same. We can see for all arms, the estimated $\hat{L}_i$ converges to the true value (dash line with the same colour) exponentially fast. The estimation remains stable after aiteration 1,000 iterations.

We further show the root mean squared error (RMSE) of lower bound of hazard rate estimation for the first 1,000 iterations in Figure 7. We observe that the error drops exponentially fast for the first 200 iterations and then drops slowly.

To analyse the sensitivity our algorithm to the estimation of the lower bound of hazard rate, we set three groups with fixed value of $\hat{L}_i$ as policy input: 1) L_at_10: take the estimation at 10^{th} iteration, where the $\hat{L}_i$ is poorly estimated; 2) L_at_200: take the estimation at 200^{th} iteration, where the error of estimation has dropped dramatically; 3) True L: the true value of L_i. The comparison is shown in Figure 8.

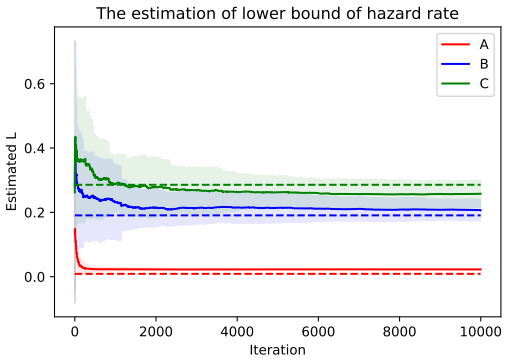

Figure 6. Expected estimation for lower bound of hazard rate. The true value of L_i for each arm is shown in dash-line with corresponding colour.

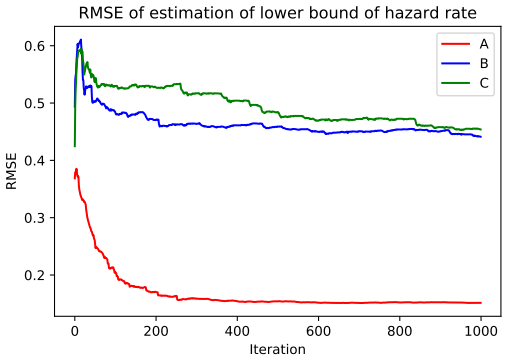

Figure 7. Root mean squared error (RMSE) of lower bound of hazard rate estimation.

As we expected, the "True L" group gives the best performance. The estimated L (one that we used in the main paper) and the fixed estimated $\hat{L}_i$ at 200^{th} iteration give similar performance. Compared with the estimated L, "L_at_200" has a slightly better performance at the beginning, which is because it gets rid of the error caused by very poor estimations at the first 200 iterations. Note the gap between the "Estimated L" and "True L" group is about 500 draws, which is caused by the poor estimation at the beginning state. The increase rates of sub-optimal draws are similar for both groups in the later iterations. The "L_at_10" group has the worst performance, where the gap between it and the "Estimated L" group is also about 500 draws. The increase rate of "L_at_10" group remains a little bit higher than other groups.

From the above comparison, we conclude that being able to estimate the lower bound of hazard rate correctly allows the algorithm to converge faster and therefore to achieve

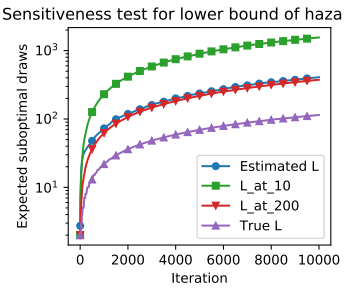

Figure 8. Performance with different lower bound of hazard rate inputs. L_at_10, L_at_200 indicate we use the fixed value of estimation of $\hat{L}_i$ at 10^{th} and 200^{th} iteration respectively.

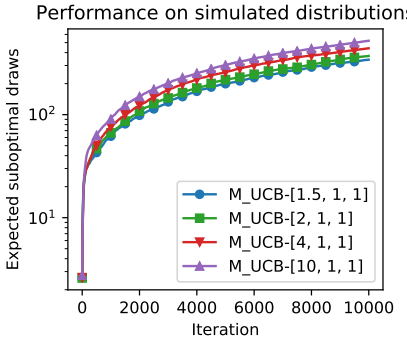

Figure 9. Performance with different hyper-parameters (α). M-UCB with [α, β, dt], where α, β are specified in policy. dt is the interval we estimate the lower bound of hazard rate, see Section 6.1.

a smaller number of sub-optimal draws. However, the algorithm is not highly sensitive to the estimation of lower bound of hazard rate. Even if we use a very poor estimation of $\hat{L}_i$ (L_at_10), we are still able to achieve a reasonable convergence.

A.3. Hyper-parameter Choices

Our policy has two hyper-parameters: α and β. Since our proof is constrained as $\beta = 1$ and $\alpha > 1$, we use $\beta = 1$ for our practical tests. For the choice of α, intuitively, we would like to use a higher value of α for the environments needed more exploration (e.g. small gaps between medians, high variance). We set $\alpha = 1.5, 2, 4, 10$ and show the performance in Figure 9. We observe that the performance is not sensitive to the value of α, where the sub-optimal draws decreases when α becomes smaller.

A.4. More Arms

We analysed the dependence of our algorithm on the number of arms K. In Figure 10, we show the expected sub-optimal draws of our algorithm and U-UCB (Cassel et al., 2018) after 10,000 iterations for K from 3 to 10. The simulation environment of the first three arms remains the same as

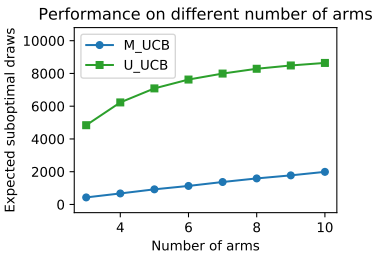

Figure 10. Dependence on number of arms. Note the evaluation is not shown in log scale.

the one we use in the main paper, the rest arms are the absolute Gaussian with mean 1.2 and standard deviation 4 (the same as Arm B). The empirical results show that the expected sub-optimal draws grow linearly with the number of arms for our policy, which is the same as what our bound predicts. Moreover, Our algorithm significantly outperforms U-UCB, although the sub-optimal draws of U-UCB grow in logarithmic rate when the number of arms is smaller than 6.

B. Proof

In this section, we show proofs of the theorems in Section 4.

B.1. Proof for Proposition 1

To prove Proposition 1, we first show R-transform and Rényi's representation of order statistics.

Definition 5 (R-transform)**.** *The R-transform of a distribution F is defined as the non-decreasing function on $[0, \infty)$ by $R(t) = \inf\{x : F(x) \geq 1 - e^{-t}\} = F^{-1}(1 - e^{-t})$.*

Observe that the R-transform defined in Definition 5 is actually the quantile transformation with respect to the CDF of standard exponential distribution, i.e. $F^{-1}\left(F_{exp}\left(t\right)\right)$.

Theorem 5 (*Rényi's representation, Theorem 2.5 in (Boucheron & Thomas, 2012)*)**.** *Let $X_{(1)} \geq \ldots \geq X_{(n)}$ be the order statistics of samples from distribution F, $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(n)}$ be the order statistics of independent samples of the standard exponential distribution, then*

$$\left(Y_{(n)}, \ldots, Y_{(i)}, \ldots, Y_{(1)}\right) \stackrel{d}{=} \left(\frac{E_n}{n}, \ldots, \sum_{k=i}^{n} \frac{E_k}{k}, \ldots, \sum_{k=1}^{n} \frac{E_k}{k}\right), \quad (23)$$

where $E_1, \ldots, E_n$ are independent and identically distributed (i.i.d.) standard exponential random variables, and

$$\left(X_{(n)}, \ldots, X_{(1)}\right) \stackrel{d}{=} \left(R\left(Y_{(n)}\right), \ldots, R\left(Y_{(1)}\right)\right), \quad (24)$$

where $R\left(\cdot\right)$ is the R-transform defined in Definition 5, equality in distribution is denoted by $\stackrel{d}{=}$.

The Rényi's representation shows the order statistics of an Exponential distribution are linear combinations of inde-

pendent Exponentials, which can be extended to the representation for order statistics of a general continuous F by quantile transformation using R-transform.

The following proposition states the connection between the *IHR* and R-transform.

Proposition 2 (Proposition 2.7 (Boucheron & Thomas, 2012))**.** *Let F be an absolutely continuous distribution function with hazard rate h, the derivative of R-transform is* $R' = 1/h(R)$*. Then if the hazard rate h is non-decreasing, then for all* $t > 0$ *and* $x > 0$, $R(t+x) - R(t) \leq x/h(R(t))$.

Based on the above proposition, the expectation of spacing can be bounded as shown in Proposition 1.

Proposition 1. *Let* $S_k = X_{(k)} - X_{(k+1)}$ *be the* k^{th} *spacing, and L be the lower bound of hazard rate of distribution F. For any* $1 \leq k \leq \frac{n}{2}$*, the expectation of spacing* S_k *can be bounded under Assumption 1,*

$$\mathbb{E}[S_k] \leq \frac{1}{kL}. \tag{7}$$

Proof.

$$\begin{aligned}
\mathbb{E}[S_k] &= \mathbb{E}[X_{(k)} - X_{(k+1)}] && (25)\\
&= \mathbb{E}[R\left(Y_{(k+1)} + \frac{E_k}{k}\right) - R\left(Y_{(k+1)}\right)] && (26)\\
&= \int_Y \int_E \left(R\left(y + \frac{z}{k}\right) - R(y)\right) f_Y(y) f_E(z)\,\mathrm{d}z\mathrm{d}y && (27)\\
&\leq \int_Y \int_E \frac{z}{k \times h(R(y))} f_Y(y) f_E(z)\,\mathrm{d}z\mathrm{d}y && (28)\\
&\leq \int_E \frac{z}{kL} f_E(z)\,\mathrm{d}z && (29)\\
&= \frac{1}{kL}. && (30)
\end{aligned}$$

Equation (25) to (26) follows the Rényi's Representation (Theorem 5), where E_k is standard exponentially distributed and independent of $Y_{(k+1)}$. Equation (27) follows the definition of expectation and we denote the value of random variables by lower case letters and the random variables by upper case letters, i.e. z, y is the value of random variable $E_k, Y_{(k+1)}$. Equation (27) to (29) follows the Proposition 2 under the assumption of IHR and with L as the lower bound of hard rate $h(R(y))$. □

B.2. Proof of Lemma 1

Lemma 1. *Let* $S_k = X_{(k)} - X_{(k+1)}$ *be the* k^{th} *spacing of order statistics, and the lower bound of hazard rate as L. Define* $v_n = \frac{n}{k^2L^2}$*, for all* $1 \leq k \leq \frac{n}{2}$*, and all* λ *such that* $0 \leq \lambda < \frac{1}{2}\sqrt{\frac{n}{v_n}}$*, we can bound* $\lambda\mathbb{E}\left[S_k\left(e^{\lambda S_k} - 1\right)\right]$ *under Assumption 1 as following,*

$$\lambda\mathbb{E}\left[S_k\left(e^{\lambda S_k} - 1\right)\right] \leq \frac{2\lambda^2 v_n}{n\left(1 - 2\lambda\sqrt{\frac{v_n}{n}}\right)}. \tag{6}$$

Proof. From Theorem 5, we can represent the spacing as $S_k = X_{(k)} - X_{(k+1)} \overset{d}{=} R\left(Y_{(k+1)+E_k/k}\right) - R\left(Y_{(k+1)}\right)$, where E_k is standard exponentially distributed and independent of $Y_{(k+1)}$. Let $v_n = \frac{n}{k^2L^2}$,

$$\begin{aligned}
&\lambda\mathbb{E}\left[S_k\left(e^{\lambda S_k} - 1\right)\right] && (31)\\
\leq&\lambda \int_E \int_Y \frac{z}{h(R(y))\,k}\left(e^{\frac{\lambda z}{h(R(y))k}} - 1\right)\\
&f_Y(y) f_E(z)\,\mathrm{d}y\mathrm{d}z && (32)\\
\leq& \int_E \lambda\sqrt{\frac{v_n}{n}} z\left(e^{\lambda\sqrt{\frac{v_n}{n}}z} - 1\right) f_E(z)\,\mathrm{d}z && (33)\\
=& \int_0^\infty \lambda\sqrt{\frac{v_n}{n}} z\left(e^{\lambda\sqrt{\frac{v_n}{n}}z} - 1\right) e^{-z}\mathrm{d}z && (34)\\
\leq&\frac{2\lambda^2 v_n}{n\left(1 - 2\lambda\sqrt{\frac{v_n}{n}}\right)}. && (35)
\end{aligned}$$

Similar as the proof of Lemma 1, Equation (31) to (32) follows Proposition 2 with IHR assumption, and from Equation (32) to (33), we assume the lower bound of hazard rate is L. The last step is because for $0 \leq \mu \leq \frac{1}{2}$, $\int_0^\infty \mu z\left(e^{\mu z} - 1\right) e^{-z}\mathrm{d}z = \frac{\mu^2(2-\mu)}{(1-\mu)^2} \leq \frac{2\mu^2}{1-2\mu}$. □

B.3. Proof for Theorem 3

Theorem 3 (Bernstein's inequality for Quantiles)**.** *Let* $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$ *denote the order statistics of* $X_1, \ldots, X_n$*, and* $X_{(qn)}$ *is the empirical q-quantile (assume* qn *is an integer,* $q \in (0, \frac{1}{2}]$*). Define* $v_n = \frac{1}{q^2nL^2}$*, with Assumption 1, for all* λ *such that* $0 \leq \lambda < \frac{1}{2}\sqrt{\frac{n}{v_n}}$*,*

$$\log\mathbb{E}\left[e^{\lambda\left(X_{(qn)} - \mathbb{E}[X_{(qn)}]\right)}\right] \leq \frac{\lambda^2 v_n}{2\left(1 - 2\lambda\sqrt{\frac{v_n}{n}}\right)}. \tag{9}$$

For all $\varepsilon > 0$*, we obtain the concentration inequality*

$$\mathbb{P}\left(X_{(qn)} - \mathbb{E}[X_{(qn)}] \geq \sqrt{2v_n\varepsilon} + 2\varepsilon\sqrt{\frac{v_n}{n}}\right) \leq e^{-\varepsilon}. \tag{10}$$

Proof. Considering the quantile case for Lemma 1, let $S_{qn} = X_{(qn)} - X_{(qn+1)}$ and $v_n = \frac{1}{q^2nL^2}$, then

$$\lambda\mathbb{E}\left[S_{qn}\left(e^{\lambda S_{qn}} - 1\right)\right] \leq \frac{2\lambda^2 v_n}{n\left(1 - 2\lambda\sqrt{\frac{v_n}{n}}\right)}. \tag{36}$$

Then from Theorem 1, for $0 < q \leq \frac{1}{2}$,

$$\begin{aligned}
&\log \mathbb{E} e^{\lambda\left(X_{(qn)}-\mathbb{E}[X_{(qn)}]\right)} \\
\leq &\lambda \frac{qn}{2} \mathbb{E}\left[S_{qn}\left(e^{\lambda S_{qn}}-1\right)\right] && (37)\\
\leq &\frac{q\lambda^2 v_n}{\left(1-2\lambda\sqrt{v_n/n}\right)} && (38)\\
\leq &\frac{\lambda^2 v_n}{2\left(1-2\lambda\sqrt{v_n/n}\right)}. && (39)
\end{aligned}$$

Inequality (9) is thus proved. To equivalently express the logarithmic moment generating function bound into a tail probability bound, we make use of the Cramér-Chernoff method (Boucheron et al., 2013). Markov's inequality implies, for $\lambda > 0$,

$$\mathbb{P}(Z \geq t) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}], \tag{40}$$

where $Z = X_{(qn)} - \mathbb{E}[X_{(qn)}]$. To choose λ to minimise the upper bound, one can introduce $\psi_Z^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda))$, with the $\psi_Z(\lambda)$ being the moment generating function, i.e. $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$. Then we get $\mathbb{P}(Z \geq t) \leq \exp(-\psi_Z^*(t))$. Thus, for $\varepsilon > 0$,

$$\mathbb{P}\left(X_{(qn)} - \mathbb{E}[X_{(qn)}] \geq \sqrt{2v_n\varepsilon} + 2\varepsilon\sqrt{v_n/n}\right) \leq e^{-\varepsilon}. \tag{41}$$

□

B.4. Proof of Theorem 4.

The proof of Theorem 4 is based on three Lemmas shown as following. The proof structure is based on (Bubeck, 2012). We set $\beta = 1$ for the following proof. To simplify the notation, we denote the number of draws of arm i at iteration t as T_i and omit the iteration specification.

Lemma 2. *Suppose we have the following two events.*
A_t) $\hat{m}_{i,T_i} < m_i + D_i(t, T_i)$.
B_t) $\hat{m}_{*,T_*} > m_* + D_*(t, T_*)$.
When A_t and B_t hold, sub-optimal arm i is at most played $\frac{C_i \log N}{(\Delta_i L_i)^2}$ times up to iteration N, where $C_i = 8\alpha\left[(2+\Delta_i L_i) + 2\sqrt{1+\Delta_i L_i}\right]$, $\Delta_i = m_* - m_i$, N is the total iteration, L_i is the lower bound of hazard rate for arm i.

Proof. A_t fails when $\hat{m}_{i,T_i} - m_i \geq D_i(t, T_i)$. By Theorem 3, we have

$$\mathbb{P}(A_t^c) = \mathbb{P}(\hat{m}_{i,T_i} - m_i \geq D_i(t, T_i)) \leq e^{-\varepsilon}.$$

When plug in $\varepsilon = \alpha \log t$, we have $\mathbb{P}(A_t^c) \leq t^{-\alpha}$. Similarly, the probability of event B_t fails is $\mathbb{P}(B_t^c) \leq t^{-\alpha}$.

Based on our policy, a sub-optimal arm i is only played if

$$\hat{m}_{i,T_i} + D_i(t, T_i) > \hat{m}_{*,T_*} + D_*(t, T_*). \tag{42}$$

Suppose both A_t and B_t hold, sub-optimal arm i is pulled due to insufficient sampling up to this point t. From A_t we get,

$$m_i + 2D_i(t, T_i) > \hat{m}_{i,T_i} + D_i(t, T_i). \tag{43}$$

From B_t we get,

$$\hat{m}_{*,T_*} + D_*(t, T_*) > m_*. \tag{44}$$

Chaining (42)(43)(44), and let $\Delta_i = m_* - m_i$, we get,

$$\begin{aligned}
m_i + 2D_i(t, T_i) &> m_*, && (45)\\
D_i(t, T_i) &> \frac{m_* - m_i}{2} = \frac{1}{2}\Delta_i, && (46)\\
T_i &< \frac{C_i \log t}{(\Delta_i L_i)^2} \leq \frac{C_i \log N}{(\Delta_i L_i)^2}, && (47)
\end{aligned}$$

where $C_i = 8\alpha\left[(2+\Delta_i L_i) + 2\sqrt{1+\Delta_i L_i}\right]$, □

Lemma 3. *Let $\hat{m}_{*,s}$ be the empirical median of the reward samples of the optimal arm (i.e. the arm with maximum median) when it has been played s times, Similarly, $\hat{m}_{i,s_i}$ is the empirical median of the reward samples of arm i when it has played s_i times, where $s \geq 1, s_i \geq l$, l is an arbitrary integer, $D_i(t, s_i)$ is the confidence width defined in Section 3.*

$\hat{m}_{*,s} + D_*(t, s) \leq \hat{m}_{i,s_i} + D_i(t, s_i)$ implies that at least one of the following must hold

$$\begin{aligned}
\hat{m}_{*,s} + D_*(t, s) &\leq \mathbb{E}[\hat{m}_{*,s}], && (48)\\
\hat{m}_{i,s_i} - D_i(t, s_i) &\geq \mathbb{E}[\hat{m}_{i,s_i}], && (49)\\
\mathbb{E}[\hat{m}_{*,s}] &< \mathbb{E}[\hat{m}_{i,s_i}] + 2D_i(t, s_i). && (50)
\end{aligned}$$

Proof. Assume all of the three inequalities are not true, then we have

$$\begin{aligned}
\hat{m}_{*,s} + D_*(t, s) &> \mathbb{E}[\hat{m}_{*,s}], && (51)\\
\hat{m}_{i,s_i} - D_i(t, s_i) &< \mathbb{E}[\hat{m}_{i,s_i}], && (52)\\
\mathbb{E}[\hat{m}_{*,s}] &\geq \mathbb{E}[\hat{m}_{i,s_i}] + 2D_i(t, s_i). && (53)
\end{aligned}$$

(51) - (52) we get,

$$\begin{aligned}
&\mathbb{E}[\hat{m}_{*,s}] - \mathbb{E}[\hat{m}_{i,s_i}] \\
<&\hat{m}_{*,s} + D_*(t, s) - (\hat{m}_{i,s_i} - D_i(t, s_i)) && (54)\\
\leq&\hat{m}_{i,s_i} + D_i(t, s_i) - (\hat{m}_{i,s_i} - D_i(t, s_i)) && (55)\\
=&2D_i(t, s_i), && (56)
\end{aligned}$$

which contradicts (53), the assumption that all of the three inequalities are not true doesn't hold. Lemma 3 is proved to be true.

□

Lemma 4. *The expected number of number a sub-optimal arm i is played at iteration N is upper bounded as following,*

$$\mathbb{E}[T_i(N)] \leq \frac{C_i \log N}{(\Delta_i L_i)^2} + \frac{2\alpha}{\alpha - 1},$$

where $C_i = 8\alpha \left[(2 + \Delta_i L_i) + 2\sqrt{1 + \Delta_i L_i}\right]$.

Proof. From Lemma 2 we know, when A_t, B_t hold, sub-optimal arm i is only played at most $\frac{C_i \log N}{(\Delta_i L_i)^2}$ times.

Recall that I_t can only be equal to i for two cases: 1) it has been sampled insufficient times. or 2) event A_t or B_t fails. Thus we have,

$$\mathbb{E}[T_i(N)] = \sum_{t=1}^{N} \mathbb{E}[\mathbb{I}\{I_t = i\}] \tag{57}$$

$$\leq \frac{C_i \log N}{(\Delta_i L_i)^2} + \sum_{t=1}^{N} \mathbb{E}[\mathbb{I}\{A_t^c \cup B_t^c\}] \tag{58}$$

$$\leq \frac{C_i \log N}{(\Delta_i L_i)^2} + \sum_{t=1}^{N} \left(\mathbb{E}[\mathbb{I}\{A_t^c\}] \cup \mathbb{E}[\mathbb{I}\{B_t^c\}]\right) \tag{59}$$

$$\leq \frac{C_i \log N}{(\Delta_i L_i)^2} + \sum_{t=1}^{N} \left(t^{-\alpha} + t^{-\alpha}\right) \tag{60}$$

$$\leq \frac{C_i \log N}{(\Delta_i L_i)^2} + \frac{2\alpha}{\alpha - 1} \tag{61}$$

where the last step is reasoning as bellow,

$$\sum_{t=1}^{N} t^{-\alpha} \leq 1 + \int_1^{\infty} x^{-\alpha} dx = 1 + \frac{-1}{1 - \alpha} = \frac{\alpha}{\alpha - 1}. \tag{62}$$

□

Now, we are ready to prove Theorem 4 based on the above three lemmas. The proof is shown as follows.

Theorem 4 (Sub-optimal draws bound)**.** *Let Δ_i be the difference between median of the optimal arm and arm i, i.e. $\Delta_i = m_* - m_i$, and L_i be the lower bound of hazard rate for arm i. For all $K > 1$, if the proposed policy M-UCB is run on K arms with reward distributions under Assumption 1, then the upper bound of the expected sum of sub-optimal draws $\mathbb{E}[\Psi_N]$ at round N is given by*

$$\mathbb{E}[\Psi_N] \leq \sum_{i \neq i_*} \frac{C_i \log N}{(\Delta_i L_i)^2} + \frac{2\alpha}{\alpha - 1}, \tag{14}$$

where $C_i = 8\alpha[(2 + \Delta_i L_i) + 2\sqrt{1 + \Delta_i L_i}]$.

Proof. We proved the bound of $\mathbb{E}[T_i(N)]$ in Lemma 4, then the sum of the expected draws of each sup-optimal draws is

$$\mathbb{E}[\Psi_N] = \sum_{i:m_i < m_*} \mathbb{E}[T_i(N)] \leq \sum_{i:m_i < m_*} \frac{C_i \log N}{(\Delta_i L_i)^2} + \frac{2\alpha}{\alpha - 1}, \tag{63}$$

where $C_i = 8\alpha \left[(2 + \Delta_i L_i) + 2\sqrt{1 + \Delta_i L_i}\right]$.

□

C. Regret Choices and Median-based Regret

Depending on the design choice of the summary statistic of reward distributions, the regret can have different design choices. In this section, we first show the backgiteration of regret design choice, namely the mean-based regret (Auer et al., 2002; Audibert et al., 2009; Garivier & Cappé, 2011), and the risk-averse regret (using the design choices of U-UCB (Cassel et al., 2018) as an example), then we show the proof of our median-based regret.

C.1. Mean-based Regret

UCB1 and its variants (Auer et al., 2002; Audibert et al., 2009; Garivier & Cappé, 2011) evaluate the reward distributions by the mean and define the optimal arm as the arm with maximum mean.

Definition 6 (Mean-based regret)**.** *The mean-based regret is defined as,*

$$R_N^{\mu} = \sum_{t=1}^{N} X_{*,t} - \sum_{t=1}^{N} \sum_{i=1}^{K} X_{i,T_i(t)} \mathbb{I}\{A_t = i\},$$

where $X_{*,t}$ is the reward sample reward from the optimal arm at iteration t.

Definition 7 (Mean-based Pseudo-regret)**.** *The pseudo-regret is defined in terms of the true mean of arm distributions,*

$$\overline{R}_N^{\mu} = N\mu^* - \sum_{t=1}^{N} \sum_{i=1}^{K} \mu_i \mathbb{I}\{A_t = i\},$$

where μ^*, μ_i are the mean of optimal arm and arm i.

The difference between the regret and the pseudo-regret comes from the randomness of the rewards. We introduce Wald's Identity in Theorem 6 and using it to show the expectations of regret and pseudo-regret are the same (Proposition 3).

Theorem 6 (Wald's Identity (Wald, 1944))**.** *Let $Y_1, Y_2, Y_3, \ldots$ be i.i.d. with finite mean, and N is a stopping time with $\mathbb{E}[N] < \infty$, then $\mathbb{E}[Y_1 + \cdots + Y_N] = \mathbb{E}[Y_1]\,\mathbb{E}[N]$.*

Proposition 3. *The expectations of regret and pseudo-regret defined in Definition 6 and 7 are equivalent and can be decomposed into the weighted sum of sub-optimal draws. i.e.* $\mathbb{E}[R_N^\mu] = \mathbb{E}[\overline{R}_N^\mu] = \sum_{i=1}^K \Delta_i^\mu \mathbb{E}[T_i(N)]$*, where* $\Delta_i^\mu = \mu_* - \mu_i$.

Proof.

$$\mathbb{E}[R_N^\mu] = \mathbb{E}\left[\sum_{t=1}^N X_{*,t} - \sum_{t=1}^N \sum_{i=1}^K X_{i,T_i(t)} \mathbb{I}\{A_t = i\}\right] \tag{64}$$

$$= \mathbb{E}\left[\sum_{t=1}^N X_{*,t}\right] - \mathbb{E}\left[\sum_{t=1}^N \sum_{i=1}^K X_{i,T_i(t)} \mathbb{I}\{A_t = i\}\right] \tag{65}$$

$$= \mu_* N - \sum_{i=1}^K \mu_i \mathbb{E}[T_i(N)]. \tag{66}$$

$$\mathbb{E}[\overline{R}_N^\mu] = \mathbb{E}\left[N\mu^* - \sum_{i=1}^K \mu_i T_i(N)\right] \tag{67}$$

$$= \mu_* N - \mathbb{E}\left[\sum_{i=1}^K \mu_i T_i(N)\right] \tag{68}$$

$$= \mu_* N - \sum_{i=1}^K \mu_i \mathbb{E}\left[T_i(N)\right]. \tag{69}$$

Equation (65) is derived based on the linearity of expectation. By Wald's Identity (Theorem 6), we get Equation (66). The derivation of expected pseudo-regret is based on the linearity of expectation.

The pseudo-regret can be decomposed into the weighted sum of sub-optimal draws, where the weight is the gap $\Delta_i^\mu = \mu_* - \mu_i$,

$$\overline{R}_N^\mu = \sum_{i=1}^K \mu_* T_i(N) - \sum_{i=1}^K \mu_i T_i(N) \tag{70}$$

$$= \sum_{i=1}^K \Delta_i^\mu T_i(N). \tag{71}$$

Then we can also get the decomposed version of expected regret and pseudo-regret,

$$\mathbb{E}[R_N^\mu] = \mathbb{E}[\overline{R}_N^\mu] = \mathbb{E}\left[\sum_{i=1}^K \Delta_i^\mu T_i(N)\right] = \sum_{i=1}^K \Delta_i^\mu \mathbb{E}[T_i(N)]. \tag{72}$$

□

Bounding the expected regret is equal to bounding the expected pseudo-regret and is also equal to bounding the expected number of sub-optimal draws. However, there is still a gap between regret and pseudo-regret. Coquelin & Munos (2007) proved that, with probability at least $1-\beta$, the upper bound of the gap at time N is

$$\left|R_N^\mu - \overline{R}_N^\mu\right| \leq \sqrt{\sum_{i:m_i<m_*} T_i(N) \log(2/\beta)/2.} \tag{73}$$

C.2. Risk-averse Regret

There is no universal definition for risk-averse regret since different summary statistics of reward distributions are chosen. We show the definition based on (Cassel et al., 2018), where they showed a general approach for bandits under risk criteria, under the class of *Empirical Distribution Performance Measures(EDPM)*. An EDPM evaluates performance by means of a function U , which maps $\hat{F}$ to $\mathbb{R}$. (Cassel et al., 2018) analysed their results based on the weighted sum of distribution and empirical distributions,

$$F_N^\pi = \frac{1}{N}\sum_{i=1}^K T_i(N) F^{(i)}, \tag{74}$$

$$\hat{F}_N^\pi = \frac{1}{N}\sum_{t=1}^N \sum_{i=1}^K \hat{F}_{T_i(t)}^{(i)} \mathbb{I}\{A_t = i\}, \tag{75}$$

where $F^{(i)}, \hat{F}_{T_i(t)}^{(i)}$ are the distribution and empirical distribution of arm i. By defining the arm with maximum expected distribution measure U as the optimal arm, the regret and pseudo-regret are defined below.

Definition 8 (Risk-averse Regret)**.** *The risk-averse regret is defined as*

$$R_N^U = U\left(\hat{F}_N^{\pi^*(N)}\right) - U\left(\hat{F}_N^\pi\right) \tag{76}$$

$$= U\left(\frac{1}{N}\sum_{t=1}^N \hat{F}_{T_i(t)}^*\right) - U\left(\frac{1}{N}\sum_{t=1}^N \sum_{i=1}^K \hat{F}_{T_i(t)}^{(i)} \mathbb{I}\{A_t = i\}\right), \tag{77}$$

where $\hat{F}_N^{\pi^*(N)}$ *is the empirical distribution of the optimal arm.*

Definition 9 (Risk-averse Pseudo-regret)**.** *The risk-averse pseudo-regret is defined as*

$$\overline{R}_N^U = U\left(F_{p^*}\right) - U\left(F_N^\pi\right) \tag{78}$$

$$= U\left(F^{(*)}\right) - U\left(\frac{1}{N}\sum_{i=1}^K T_i(N) F^{(i)}\right), \tag{79}$$

where F_{p^*} *are the distribution of the optimal arm.*

Proposition 4 ((Cassel et al., 2018))**.** *When* U *is linear, the expected regret and expected pseudo-regret defined in Definition 8 and 9 are equivalent, and can be decomposed into the weighted sum of sub-optimal draws, i.e.* $\frac{1}{N}\sum_{i\neq i^*} \mathbb{E}\left[T_i(N)\right] \Delta_i$*, where* $\Delta_i = U\left(F_{p^*}\right) - U\left(F^{(i)}\right)$.

Proof.

$$\mathbb{E}[R_N^U] = \mathbb{E}\left[U\left(\hat{F}_N^{\pi^*(N)}\right) - U\left(\hat{F}_N^{\pi}\right)\right] \tag{80}$$

$$= U\left(\mathbb{E}\left[\hat{F}_N^{\pi^*(N)} - \hat{F}_N^{\pi}\right]\right) \tag{81}$$

$$= U\left(\mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{K}\left(\hat{F}_{T_i(t)}^{(i^*)} - \hat{F}_{T_i(t)}^{(i)}\right)\mathbb{I}\{A_t = i\}\right]\right) \tag{82}$$

$$= U\left(\frac{1}{N}\sum_{i=1}^{K}\mathbb{E}\left[\hat{F}_{T_i(t)}^{(i^*)} - \hat{F}_{T_i(t)}^{(i)}\right]\mathbb{E}\left[T_i(t)\right]\right) \tag{83}$$

$$= U\left(\frac{1}{N}\sum_{i=1}^{K}\left(F_{p^*} - F^{(i)}\right)\mathbb{E}\left[T_i(t)\right]\right) \tag{84}$$

$$= \frac{1}{N}\sum_{i=1}^{K} U\left(F_{p^*} - F^{(i)}\right)\mathbb{E}\left[T_i(t)\right]. \tag{85}$$

Step (81) is derived by Jensen's inequality and U is linear. From (82) to (83), we use Wald's Identity. Step (84) is based on the linearity of U.

$$\mathbb{E}[\overline{R}_N^U] = \mathbb{E}\left[U\left(F_{p^*}\right) - U\left(F_N^{\pi}\right)\right] \tag{86}$$

$$= U\left(\mathbb{E}\left[F_{p^*} - F_N^{\pi}\right]\right) \tag{87}$$

$$= U\left(\mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{K}\left(F_{p^*} - F_N^{\pi}\right)\mathbb{I}\{A_t = i\}\right]\right) \tag{88}$$

$$= U\left(\frac{1}{N}\sum_{i=1}^{K}\left(F_{p^*} - F^{(i)}\right)\mathbb{E}\left[T_i(t)\right]\right) \tag{89}$$

$$= \frac{1}{N}\sum_{i=1}^{K} U\left(F_{p^*} - F^{(i)}\right)\mathbb{E}\left[T_i(t)\right]. \tag{90}$$

Similarly, the Step (87) is derived by Jensen's inequality when U is linear. And the last step is based on the linearity of U. □

Proposition 5 ((Cassel et al., 2018)). *When U is non-linear but quasiconvex and is strongly stable EDPM (Definition 3 in (Cassel et al., 2018)), the expected pseudo regret can be bounded by,*

$$\mathbb{E}[\overline{R}_N^U] \leq \frac{L}{N}\sum_{i\neq i^*}\mathbb{E}\left[T_i(N)\right]\left\|F^{(i^*)} - F^{(i)}\right\|, \tag{91}$$

where $L = b\left(1 + \max_{i,j\in\mathcal{K}}\left\|F^{(i)} - F^{(j)}\right\|^{q-1}\right)$.

From Proposition 5 we know, when U is the median of rewards, the expected pseudo-regret can be bounded by the weighted sum of sub-optimal draws, with the weight as $\left\|F^{(i^*)} - F^{(i)}\right\|$.

C.3. Median-based Regret

The risk-averse regret proposed by Cassel et al. (2018) depends on the norm of the difference between CDF and empirical CDF, which could be empirically hard to compute. We propose the median-based regret and pseudo-regret, and show in Proposition 6 that they are in fact equivalent.

Definition 10 (Median-based Regret)**.** *The median-based regret is defined as*

$$R_N = \sum_{t=1}^{N}\hat{m}_{*,t} - \sum_{t=1}^{N}\sum_{i=1}^{K}\hat{m}_{i,T_i(t)}\mathbb{I}\{A_t = i\}. \tag{92}$$

Definition 11 (Median-based Pseudo-regret)**.** *The median-based pseudo-regret is defined based on the medians of arms.*

$$\overline{R}_N = Nm_* - \sum_{i=1}^{K} m_i T_i(N). \tag{93}$$

By taking medians of sample rewards at each iteration, the median-based regret reflects the difference of the cumulative empirical median between the M-UCB policy and the optimal policy. Similar as risk-averse regret (see Appendix C.2 for an example), minimising the median-based regret is not equivalent to maximising the cumulative rewards.

Proposition 6. *The expectations of median-based regret and pseudo-regret defined in Definition 10 and 11 are equivalent and can be decomposed into the weighted sum of sub-optimal draws. i.e.* $\mathbb{E}[R_N] = \mathbb{E}[\overline{R}_N] = \sum_{i=1}^{K}\Delta_i\mathbb{E}[T_i(N)]$, *where* $\Delta_i = m_* - m_i$.

Proof.

$$\mathbb{E}[R_N] = \mathbb{E}\left[\sum_{t=1}^{N}\hat{m}_{*,t}\right] - \mathbb{E}\left[\sum_{t=1}^{N}\sum_{i=1}^{K}\hat{m}_{i,T_i(t)}\mathbb{I}\{A_t = i\}\right] \tag{94}$$

$$= m_*N - \sum_{i=1}^{K} m_i\mathbb{E}[T_i(N)]. \tag{95}$$

$$\mathbb{E}[\overline{R}_N] = m_*N - \mathbb{E}\left[\sum_{i=1}^{K} m_i T_i(N)\right] \tag{96}$$

$$= m_*N - \sum_{i=1}^{K} m_i\mathbb{E}\left[T_i(N)\right]. \tag{97}$$

Equation (94) is derived based on the linearity of expectation. By Wald's Identity (Theorem 6), we get Equation (95). The derivation of expected pseudo-regret is based on the linearity of expectation.

The pseudo-regret for median can be decomposed into the weighted sum of sub-optimal draws, where the weighted is the gap $\Delta_i = m_* - m_i$,

$$\overline{R}_N = \sum_{i=1}^{K} m_* T_i(N) - \sum_{i=1}^{K} m_i T_i(N) \tag{98}$$

$$= \sum_{i=1}^{K} (m_* - m_i) T_i(N) \tag{99}$$

$$= \sum_{i=1}^{K} \Delta_i T_i(N) \tag{100}$$

Then we can also get the decomposed version of expected regret and pseudo-regret,

$$\mathbb{E}[R_N] = \mathbb{E}[\overline{R}_N] = \mathbb{E}\left[\sum_{i=1}^{K} \Delta_i T_i(N)\right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(N)], \tag{101}$$

□

Bounding the expected median-based regret depends on the uppeer bound of the expected number of sub-optimal draws. Recall from Definition 1 that Ψ_N is directly related to the number of draws $T_i(N)$, and hence the expected regret bound can be derived based on Proposition 6 and Theorem 4

Corollary 2 (Regret bound)**.** *Let L_i be the lower bound of hazard rate, and Δ_i be the difference between median of the optimal arm and arm i, i.e. $\Delta_i = m_* - m_i$. For all $K > 1$, if the proposed policy M-UCB is run on K arms with reward distributions under Assumption 1, then the upper bound of the expected regret $\mathbb{E}[R_N]$ is given by*

$$\mathbb{E}[R_N] \leq \sum_{i \neq i_*} \frac{C_i \log N}{\Delta_i L_i^2} + \frac{2\alpha}{\alpha - 1} \left(\sum_{j=1}^{K} \Delta_j\right), \tag{102}$$

where $C_i = 8\alpha \left[(2 + \Delta_i L_i) + 2\sqrt{1 + \Delta_i L_i}\right]$.

Proof. According to the definition of regret shown in Definition 10 and Proposition 6, we derive the upper bound for expected regret based on Theorem 4,

$$\mathbb{E}[R_N] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(N)]$$

$$\leq \sum_{i: m_i < m_*}^{K} \frac{C_i \log N}{\Delta_i L_i^2} + \frac{2\alpha}{\alpha - 1} \left(\sum_{j=1}^{K} \Delta_j\right), \tag{103}$$

where $C_i = 8\alpha \left[(2 + \Delta_i L_i) + 2\sqrt{1 + \Delta_i L_i}\right]$. □

Regret Design Discussion: The definition of regret depends on the choice of summary statistics for reward distributions. When defining the optimal arm as the one with the highest mean, the mean-based regret and pseudo-regret are defined as (Auer et al., 2002),

$$R_N^{\mu} = \sum_{t=1}^{N} X_{*,t} - \sum_{t=1}^{N} \sum_{i=1}^{K} X_{i,T_i(t)} \mathbb{I}\{A_t = i\}, \tag{104}$$

$$\overline{R}_N^{\mu} = N\mu^* - \sum_{t=1}^{N} \sum_{i=1}^{K} \mu_i \mathbb{I}\{A_t = i\}, \tag{105}$$

where $X_{*,t}$ is the sample reward from the optimal arm at iteration t, and μ^*, μ_i are the mean of optimal arm and arm i. Minimising the mean-based regret is equivalent to maximising the cumulative reward. The expected mean-based regret and pseudo-regret are equal and can be decomposed into the weighted sum of sub-optimal draws, i.e. $\mathbb{E}[R_N^{\mu}] = \mathbb{E}[\overline{R}_N^{\mu}] = \sum_{i=1}^{K} \Delta_i^{\mu} \mathbb{E}[T_i(N)]$, where $\Delta_i^{\mu} = \mu_* - \mu_i$.

When evaluating distributions by different summary statistics, regret can be defined accordingly. For example, (Cassel et al., 2018) evaluates performance by means of a function U, which maps the empirical distribution $\hat{F}$ to $\mathbb{R}$. Correspondingly, a risk-averse regret is proposed as the difference between U measure of average empirical distributions with the optimal policy and their policy. Similarly, we define our regret as the difference between the medians of reward distributions.