Supplementary material for “Gossip-based distributed stochastic bandit algorithms ”

A. Proof of Lemma 4

Proof. The lemma can be proved using induction. First of all, note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{N}\left(w_{j,t+1}^{j',t'}-1\right)^2\right] &= \mathbb{E}\left[\frac{N}{\binom{N}{2}}\sum_{1\le j_1<j_2\le N}\left(\frac{w_{j_1,t}^{j',t'}+w_{j_2,t}^{j',t'}}{2}-1\right)^2\right]\\
&= \mathbb{E}\left[\frac{2}{N-1}\cdot\frac{1}{4}\sum_{1\le j_1<j_2\le N}\left[(w_{j_1,t}^{j',t'}-1)^2+(w_{j_2,t}^{j',t'}-1)^2+2(w_{j_1,t}^{j',t'}-1)(w_{j_2,t}^{j',t'}-1)\right]\right]\\
&= \mathbb{E}\left[\frac{1}{2}\sum_{1\le j\le N}(w_{j,t}^{j',t'}-1)^2+\frac{2}{(N-1)}\sum_{1\le j_1<j_2\le N}(w_{j_1,t}^{j',t'}-1)(w_{j_2,t}^{j',t'}-1)\right]\\
&= \mathbb{E}\Bigg[\frac{N-3}{2(N-1)}\sum_{1\le j\le N}(w_{j,t}^{j',t'}-1)^2\\
&\qquad+\frac{1}{(N-1)}\left(\sum_{1\le j\le N}(w_{j,t}^{j',t'}-1)^2+2\sum_{1\le j_1<j_2\le N}(w_{j_1,t}^{j',t'}-1)(w_{j_2,t}^{j',t'}-1)\right)\Bigg]\\
&\le \frac{1}{2}\,\mathbb{E}\left[\sum_{1\le j\le N}(w_{j,t}^{j',t'}-1)^2\right],
\end{aligned}$$

where the last equality follows because of Lemma 3 since $\sum_{j=1}^{N} w_{j,t}^{j',t'} = N$, and therefore

$$0=\left(N-\sum_{j=1}^{N} w_{j,t}^{j',t'}\right)^2=\sum_{1\le j\le N}(w_{j,t}^{j',t'}-1)^2+2\sum_{1\le j_1<j_2\le N}\left[(w_{j_1,t}^{j',t'}-1)(w_{j_2,t}^{j',t'}-1)\right] .$$

To complete the argument, one can readily check that $\sum_{j=1}^{N}\left(w_{j,t}^{j',t'}-1\right)^2 < N^2$. The last statement follows simply from Lemma 3. □

B. Proof of Lemma 5 (longer version)

Proof. The zero mean is a consequence of Lemma 3. To compute the variance, we have

$$\begin{aligned}
\mathrm{Var}\left[y_{j,t}^{i}\right] &= \mathrm{Var}\left[\sum_{t'=1}^{\mathcal{T}(t/2)-1}\sum_{j'=1}^{N}(1-w_{j,t}^{j',t'})\mathbb{I}_{j',t'}^{i}\right]\\
&= \mathbb{E}\left[\left(\sum_{t'=1}^{\mathcal{T}(t/2)-1}\sum_{j'=1}^{N}(1-w_{j,t}^{j',t'})\mathbb{I}_{j',t'}^{i}\right)^2\right]\\
&= \sum_{t',t''=1}^{\mathcal{T}(t/2)-1}\sum_{j',j''=1}^{N}\mathbb{E}\left[\mathbb{I}_{j',t'}^{i}\mathbb{I}_{j'',t''}^{i}\left(1-w_{j,t}^{j',t'}\right)\left(1-w_{j,t}^{j'',t''}\right)\right]\\
&\le N\sum_{t',t''=1}^{\mathcal{T}(t/2)-1}\sum_{j',j''=1}^{N}\sqrt{\frac{1}{2^{t-t'}}}\sqrt{\frac{1}{2^{t-t''}}} \qquad (11)
\end{aligned}$$

$$
\begin{aligned}
&= N \left(\sum_{t'=1}^{\mathcal{T}(t/2)-1} \sum_{j'}^{N} \sqrt{\frac{1}{2^{t-t'}}} \right) \left(\sum_{t''=1}^{\mathcal{T}(t/2)-1} \sum_{j''=1}^{N} \sqrt{\frac{1}{2^{t-t''}}} \right) && (12)\\
&\leq N \left(N \sum_{t'=t-\mathcal{T}(t/1)-1}^{t} \frac{1}{2^{t'/2}} \right) \left(N \sum_{t''=t-\mathcal{T}(t/1)-1}^{t} \frac{1}{2^{t''/2}} \right)\\
&\leq N^3 \frac{1}{2^{(t-\mathcal{T}(t/2)-1)}} \left(\sum_{t'=1}^{\infty} \frac{1}{2^{t'/2}} \right)^2\\
&\leq N^3 \cdot 2^{-t/2} \left(\frac{1}{1-1/\sqrt{2}} \right)^2\\
&\leq 12 \cdot N^3 \cdot 2^{-t/2} \ ,
\end{aligned}
$$

where (11) follows from Lemma 4 and the Cauchy-Schwarz inequality. □

C. Proof of Theorem 1

Proof. One way to look at the proof is the following[6]. First, in (A) we analyze a version of ϵ-GREEDY, where N independent plays are allowed per iteration. This analysis follows closely the one given by Auer et al. for ϵ-GREEDY (Auer et al., 2002), and just requires only some trivial modifications. Then in (B) we relate this to P2P-ϵ-GREEDY, and we show that the difference is negligible. This requires significantly more effort. In accordance with (A), assume that $t \geq cK/(d^2N)$, use $\epsilon_t = \frac{cK}{d^2tN}$, and put $x_0 = \frac{N}{2K}\sum_{j=1}^{\mathcal{T}(t/2)-1}\epsilon_j$. Now, the probability of choosing some arm i in iteration t at peer p is

$$
\mathbb{P}\left[i_{j,t} = i\right] \leq \frac{\epsilon_t}{K} + (1-\epsilon_t)\,\mathbb{P}\left[\frac{c^i_{j,t}}{d^i_{j,t}} \geq \frac{c^{i^*}_{j,t}}{d^{i^*}_{j,t}}\right], \tag{13}
$$

where $i^* = \arg\max_{1\leq i\leq K}\mu_i$. The second term can be decomposed as

$$
\mathbb{P}\left[\frac{c^i_{j,t}}{d^i_{j,t}} \geq \frac{c^{i^*}_{j,t}}{d^{i^*}_{j,t}}\right] \leq \mathbb{P}\left[\frac{c^i_{j,t}}{d^i_{j,t}} \geq \mu_i + \frac{\Delta_i}{2}\right] + \mathbb{P}\left[\frac{c^{i^*}_{j,t}}{d^{i^*}_{j,t}} \leq \mu_{i^*} - \frac{\Delta_i}{2}\right].
$$

The two terms on the right hand side are bounded the same way. However, at this point it is not possible to continue like that in (Auer et al., 2002). The reason is that when using ϵ-GREEDY, the weights are either 0 or 1, but using P2P-ϵ-GREEDY they can take values with a much wider range (like $N, N/2, N/2^{N-2}, N/8+1$, etc.), thus the bias in the estimation of the expected reward cannot be handled as easily.[7] It is thus time to apply the separation mentioned at the beginning of proof.

Now let $C^i_t = \sum_{t'=1}^{\mathcal{T}(t/2)-1}\sum_{j'}^{N}\xi_{j',t'}\mathbb{I}^i_{j',t'}$ ((B) in (8)) and $D^i_t = \sum_{t'=1}^{\mathcal{T}(t/2)-1}\sum_{j'}^{N}\mathbb{I}^i_{j',t'}$. Using the union bound, we bound the first term of (10) by

$$
\begin{aligned}
\mathbb{P}\left[\frac{c^i_{j,t}}{d^i_{j,t}} \geq \mu_i + \frac{\Delta_i}{2}\right] \leq & \,\mathbb{P}\left[C^i_t - \mu_i D^i_t \geq \frac{\Delta_i}{8}D^i_t\right] + \mathbb{P}\left[c^i_{j,t} - C^i_t \geq \frac{\Delta_i}{8}D^i_t\right]\\
& + \mathbb{P}\left[\mu_i\big(D^i_t - d^i_{j,t}\big) \geq \frac{\Delta_i}{8}D^i_t\right] + \mathbb{P}\left[\frac{\Delta_i}{2}\big(D^i_t - d^i_{j,t}\big) \geq \frac{\Delta_i}{8}D^i_t\right] && (14)\\
= & T_1 + T_2 + T_3 + T_4 && (15)
\end{aligned}
$$

Let us first upper bound T_1, which corresponds to considering (B) in (8). Let η^i_t denote the number of pulling an arm i in the whole network following an exploration step up to iteration $\mathcal{T}(t/2)-1$. In other words η^i_t is

[6] The analogy is not perfect. It is only meant as an aid to intuition.

[7] Note that $\mathbb{I}^i_{j,t} = 1$ depends both on $\mathbf{s}_{j,t}$ and $\mathbf{n}_{j,t}$.

the number of times arm i was selected at random in every peers up to iteration $\mathcal{T}(t/2)-1$. Denoting by ζ_ℓ^i the ℓth reward received on arm i in the whole network, one can show that

$$\begin{align*}
T_1 &= \sum_{g=1}^{N(\mathcal{T}(t/2)-1)} \mathbb{P}\left(\sum_{t=1}^{\mathcal{T}(t/2)-1}\sum_{j=1}^{N}\mathbb{I}_{j,t}^i\xi_{j,t} - \mu_i g \geq \frac{\Delta_i}{8}g \quad \wedge \quad g = \sum_{t=1}^{\mathcal{T}(t/2)-1}\sum_{j=1}^{N}\mathbb{I}_{j,t}^i\right) \\
&= \sum_{g=1}^{N(\mathcal{T}(t/2)-1)} \mathbb{P}\left(\sum_{\ell=1}^{g}\zeta_\ell^i - \mu_i g \geq \frac{\Delta_i}{8}g \quad \wedge \quad g = \sum_{t=1}^{\mathcal{T}(t/2)-1}\sum_{j=1}^{N}\mathbb{I}_{j,t}^i\right) \\
&\leq \sum_{g=1}^{N(\mathcal{T}(t/2)-1)} \mathbb{P}\left(g = \sum_{t=1}^{\mathcal{T}(t/2)-1}\sum_{j=1}^{N}\mathbb{I}_{j,t}^i \middle| \sum_{\ell=1}^{g}\zeta_\ell^i - \mu_i g \geq \frac{\Delta_i}{8}g\right)\mathbb{P}\left(\sum_{\ell=1}^{g}\zeta_\ell^i - \mu_i g \geq \frac{\Delta_i}{8}g\right) \\
&\leq \sum_{g=1}^{N(\mathcal{T}(t/2)-1)} \mathbb{P}\left(g = \sum_{t=1}^{\mathcal{T}(t/2)-1}\sum_{j=1}^{N}\mathbb{I}_{j,t}^i \middle| \sum_{\ell=1}^{g}\zeta_\ell^i - \mu_i g \geq \frac{\Delta_i}{8}g\right)\exp\left(\frac{-\Delta_i^2 g}{2}\right) \tag{16} \\
&\leq \sum_{g=1}^{x_0}\mathbb{P}\left(g = \sum_{t=1}^{\mathcal{T}(t/2)-1}\sum_{j=1}^{N}\mathbb{I}_{j,t}^i \middle| \sum_{\ell=1}^{g}\zeta_\ell^i - \mu_i g \geq \frac{\Delta_i}{8}g\right) + \frac{2}{\Delta_i^2}\exp\left(\frac{-\Delta_i^2\lfloor x_0\rfloor}{2}\right) \tag{17} \\
&\leq \sum_{g=1}^{x_0}\mathbb{P}\left(g \geq \eta_t^i \middle| \sum_{j=1}^{g}\zeta_j^i - \mu_i g \geq \frac{\Delta_i}{8}g\right) + \frac{2}{\Delta_i^2}\exp\left(\frac{-\Delta_i^2\lfloor x_0\rfloor}{2}\right) \\
&\leq x_0\,\mathbb{P}\left(x_0 \geq \eta_t^i\right) + \frac{2}{\Delta_i^2}\exp\left(\frac{-\Delta_i^2\lfloor x_0\rfloor}{2}\right), \tag{18}
\end{align*}$$

where (16) follows from the Hoeffding-bound, (17) can be obtained by using $\sum_{g=x+1}^{\infty}\exp\left(-\kappa g\right) \leq \frac{1}{\kappa}\exp(-\kappa x)$, and (18) follows from the fact that the random selection of an arm is independent from all previous choices of the policy. All that left is to bound $\mathbb{P}\left(x_0 \geq \eta_t^i\right)$. Then

$$\mathbb{E}[\eta_t^i] = \frac{1}{K}\sum_{t'=1}^{\mathcal{T}(t/2)-1} N\epsilon_{t'} = 2\frac{N}{2K}\sum_{t'=1}^{\mathcal{T}(t/2)-1}\epsilon_{t'} = 2x_0$$

and

$$\mathrm{Var}[\eta_t^i] = \sum_{t'=1}^{\mathcal{T}(t/2)-1}\frac{N\epsilon_{t'}}{K}\left(1 - \frac{N\epsilon_{t'}}{K}\right) \leq \frac{N}{K}\sum_{t'=1}^{\mathcal{T}(t/2)-1}\epsilon_{t'} = 2x_0,$$

therefore, applying Bernstein's inequality, one gets

$$\mathbb{P}\left(x_0 \geq \eta_t^i\right) = \mathbb{P}\left(2x_0 - x_0 \geq \eta_t^i\right) \leq \exp\left(-\frac{1/2x_0^2}{2x_0 + (1/2)x_0}\right) = \exp\left(-\frac{x_0}{5}\right).$$

Let us continue with considering (A) in (8), that is, with bounding terms T_2, T_3 and T_4. To upper bound T_2 one can use an argument similar to the one applied in Lemma 6 and obtain for

$$c_{j,t}^i - C_t^i = \sum_{t'=1}^{\mathcal{T}(t/2)-1}\sum_{j'=1}^{N}(1 - w_{j,t}^{j',t'})\mathbb{I}_{j',t'}^i\xi_{j',t'} = z_{j,t}^i$$

to have expected value $\mathbb{E}\left[z_{j,t}^i\right] = 0$ and variance $\mathrm{Var}\left[z_{j,t}^i\right] \leq 12\cdot N^3 2^{-t/2}$. Here the expectation and variance are taken over the peers and the random choices. Now, apply Chebyshev's inequality for $z_{j,t}^i$ to get

$$\mathbb{E}\left[T_2\right] \leq \mathbb{P}\left(\left|z_{j,t}^i\right| \geq \frac{\Delta_i}{8}\right) \leq \frac{768}{\Delta_i^2}N^3 2^{-t/2}$$

$\mathbb{E}[T_3]$ and $\mathbb{E}[T_4]$ can be found in the same way using Lemma 5. Therefore,

$$\mathbb{E}[T_2] + \mathbb{E}[T_3] + \mathbb{E}[T_4] \leq \frac{2304}{\Delta_i^2} N^3 2^{-t/2}. \tag{19}$$

Finally, let us derive a bound on x_0 for $t \geq t' = cK/(d^2N)$. Recall that $\epsilon_t = \frac{cK}{d^2tN}$. Then we have

$$\begin{aligned} x_0 =& \frac{N}{2K} \sum_{\tau=1}^{\mathcal{T}(t/2)-1} \epsilon_\tau = \frac{N}{2K} \sum_{\tau=1}^{t'} \epsilon_t + \frac{N}{2K} \sum_{\tau=t'+1}^{\mathcal{T}(t/2)-1} \epsilon_\tau \geq \frac{Nt'}{2K} + \frac{c}{d^2} \ln \frac{\mathcal{T}(t/2)-1}{t'} \\ \geq& \frac{c}{2d^2} + \frac{c}{d^2} \ln \frac{(\mathcal{T}(t/2)-1)d^2N}{cK} \geq \frac{c}{d^2} \ln \frac{Ntd^2e^{1/2}}{4cK}. \end{aligned}$$

Summing up, for an arm $i \neq i^*$, at any peer j and in iteration t, we have

$$\begin{aligned} \mathbb{P}\left(\mathbb{I}_{j,t}^i = 1\right) \leq& \frac{\epsilon_t}{K} + 2x_0 \exp\left(-\frac{x_0}{5}\right) + \frac{4}{\Delta_i^2} \exp\left(\frac{-\Delta_i^2 \lfloor x_0 \rfloor}{2}\right) + \frac{4608}{\Delta_i^2} N^3 2^{-t/2} \\ =& \frac{c}{d^2tN} + 2\left(\frac{c}{d^2} \ln \frac{Ntd^2e^{1/2}}{4cK}\right) \left(\frac{4cK}{Ntd^2e^{1/2}}\right)^{\frac{c}{5d}} + \frac{4e}{d^2} \left(\frac{4cK}{Ntd^2e^{1/2}}\right)^{\frac{c}{2}} + \frac{4608}{\Delta_i^2} N^3 2^{-t/2}. \end{aligned}$$

□

D. The pseudocode of P2P-ϵ-greedy.slim

Algorithm 2 P2P-ϵ-GREEDY.SLIM at peer j in iteration t

```
Receive $\mathcal{M}_{j_1,t} = (I_{j_1,t}, c_{j_1,t}, d_{j_1,t}, r_{j_1,t}, q_{j_1,t}, a_{j_1,t}, b_{j_1,t})$ and $\mathcal{M}_{j_2,t} = (I_{j_1,t}, c_{j_2,t}, d_{j_2,t}, r_{j_2,t}, q_{j_2,t}, a_{j_2,t}, b_{j_2,t})$
    from the two current neighbours
if $I_{j_1,t} = I_{j_2,t}$ then
    $\mathcal{M}'_{j,t} =$ AGGREGATE($\mathcal{M}_{j_1,t}, \mathcal{M}_{j_1,t}$)
else
    if $\frac{c_{j_1,t}}{d_{j_1,t}} \geq \frac{c_{j_2,t}}{d_{j_2,t}}$ then                                  ▷ Select the better model
        $\mathcal{M}'_{j,t} = \mathcal{M}'_{j_1,t}$
    else
        $\mathcal{M}'_{j,t} = \mathcal{M}'_{j_2,t}$
if $I'_{j,t} = I_{j,t}$ then
    Pull arm $I_{j,t}$ and receive reward $\xi_{j,t}$
    $\mathcal{M}_{j,t+1} =$ UPDATE($\mathcal{M}'_{j,t}, \xi_{j,t}, t$)
    The model to be sent is $\mathcal{M}_{j,t+1}$
else
    Put $\epsilon_t = \min\left(1, \frac{cK}{d^2tN}\right)$
    With probability $1 - \epsilon_t$ let $I = \mathbb{I}\{\frac{c'_{j,t}}{d'_{j,t}} > \frac{c_{j,t}}{d_{j,t}}\}I'_{j,t} + \mathbb{I}\{\frac{c'_{j,t}}{d'_{j,t}} < \frac{c_{j,t}}{d_{j,t}}\}I_{j,t}$ (choose from $\mathcal{M}'_{j,t}$ and $\mathcal{M}_{j,t}$ the
    one with the higher expected reward estimate) and with probability $\epsilon_t$ choose $I = I_{j,t}$
    Pull arm $I$ and receive reward $\xi_{j,t}$
    if $I = I_{j,t}$ then                              ▷ The arm chosen is based on the model received
        $\mathcal{M}_{j,t+1} =$ UPDATE($\mathcal{M}_{j,t}, \xi_{j,t}, t$)
        $\mathcal{M}'_{j,t+1} =$ STEP($\mathcal{M}'_{j,t}, t$)
    else
        $\mathcal{M}_{j,t+1} =$ STEP($\mathcal{M}_{j,t}t$)
        $\mathcal{M}'_{j,t+1} =$ UPDATE($\mathcal{M}'_{j,t}, \xi_{j,t}, t$)
    if $\frac{c_{j,t+1}}{d_{j,t+1}} > \frac{c'_{j,t+1}}{d'_{j,t+1}}$ then
        The model to be sent is $\mathcal{M}_{j,t+1}$
    else
        The model to be sent is $\mathcal{M}'_{j,t+1}$
function AGGREGATE( $\mathcal{M}' = (I', c', d', r', q', a', b')$ , $\mathcal{M}'' = (I'', c'', d'', r'', q'', a'', b'')$)
    $c = (1/2)(c' + c'')$, $d = (1/2)(d' + d'')$
    $r = (1/2)(r' + r'')$, $q = (1/2)(q' + q'')$
    $f = (1/2)(a' + a'')$, $g = (1/2)(b' + b'')$
    $I = I' = I''$
    return $\mathcal{M} = (I, c, d, r, q, a, b)$
function UPDATE(($\mathcal{M} = (I, c, d, r, q, a, b)$ , $\xi, t$))
    if $t$ is power of 2 then
        $c = c + r, \quad d = d + q$
        $r = a, \quad q = b$
        $a = b = 0$
    $a = a + N\xi$
    $b = b + N$
    return $\mathcal{M}$
function STEP(($\mathcal{M} = (I, c, d, r, q, a, b)$ , $t$))
    if $t$ is power of 2 then
        $c = c + r, \quad d = d + q$
        $r = a, \quad q = b$
        $a = b = 0$
    return $\mathcal{M}$
```

E. Lower bounds

E.1. The cost of information spreading

The following example demonstrates that the term $\Omega(\min(t,N)\log N)$ appearing in our regret bound is unavoidable in some settings.

Example 7. *Consider the following scenario. N is large, $K = N^{\epsilon}$ for some constant $\epsilon > 0$, one arm is constant 1, the rest return constant 0. Then, with high probability ($> 1/2$), in the first $(\epsilon/10)\log N$ rounds at least the half of the peers will not be able to find the optimal arm. Indeed, let $\mathcal{N}_t$ denote the number of arms in iteration t that know which is the optimal arm. The expected number of arms that discover the optimal arm via exploration is $N^{1-\epsilon}$ per round, meanwhile the number of arms that know about it because this information was forwarded to them by some other peers is at most $2\mathcal{N}_{t-1}$ in time t. Thus, after $\epsilon \log_3(N/2)$ iterations, the expected number of peers that know the optimal arm is at most $N/2$. Consequently, the cumulated regret is at least $tN/2$ in iterations $t = 1, \dots, \epsilon \log_3(N/2)$.*

The above argument also sheds some light on the huge jump in the regret in Figure 2 for the case $N = 1000$, compared to the two other cases: the regret must grow (roughly) linearly during the first $N \log_3 10$ arm pulls.

E.2. Detereministic algorithms can be suboptimal

The following example demonstrates why randomization is beneficial in the P2P setting. According to it, deterministic algorithms (like UCB, MOSS, etc.) cannot be applied directly, without generating too large regret. Note also that it is not clear what the suitable randomization for these algorithms would be. (Also consult the following subsection about the necessity of using delays. This further suggests the need for randomization that would avoid generating too large regret in an very long epochs where the estimates don't change significantly.)

Example 8. *Consider the folowing setting. $N = K$ large, one arm returns constantly 1, the rest returns constantly $1/(2K)$, $1/(2K-1)$, $\dots$, $1/(2K+1)$ respectively. Then* UCB *will repeat the same thing at each pear. Assume that each arm is pulled once already, but don't count it in the regret.*

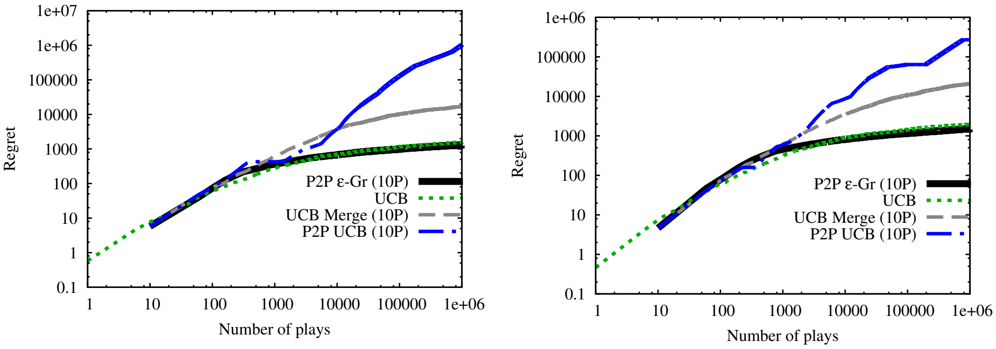
Now, as each peer has exactly the same estimates for each arm, each of them will pull the same arm. However, due to the deterministic nature of the rewards, the new estimates will be again exatly the same at each peer, and even the information sharing will not change this. Therefore, from now on, each peer will do exactly the same as any other peers. Thus bad decisions will have N times the effect they should.

More precisely, UCB *will obviously pull each arm again at least once in the first $O(N)$ round. This leads to a regret of size $\Omega(NK) = \Omega(N^2)$.*

However, the regret of P2P-ϵ-greedy is significantly less in the first $2^{o(N)}$ rounds.

Although the above example is quite extreme, it successfully highlights the risk of using deterministic algorithms in a P2P setting: exploratory, suboptimal actions might be chosen by many peers at the same time in parallel, leading to an unnecessarily large regret.

Besides that, we have also made some experiments to test some direct adaptations of UCB to P2P networks. The results are summarized in Fig. 3. The algorithms we have compared are the P2P-ϵ-Greedy, UCB in a stand alone-version (denoted as UCB), UCB in a P2P setting with the same weight sharing method that was used by P2P-ϵ-Greedy (denoted UCB Merge), and finally UCB in a P2P setting with the same weight sharing and the delayment method that is also used by P2P-ϵ-Greedy (denoted P2P UCB). The results demonstrate nicely the phenomenon mentioned above: due to the parallelization the UCB-based methods indeed have much larger regret. It is even more so when the delaying is applied. Although at first glance it might seem to contradict the argument in the next subsection about the necessity of using delays, in fact it does not: it only shows that randomness and delaying method are really helpful when applied combined. And the reason for P2P UCB having much worse performance than UCB Merge is that, due to the delaying, whenever the estimates at the end of some epoch suggest to pull a suboptimal arm (i.e., it is an exploratory step), P2P UCB pulls that same arm *throughout the whole epoch.*



(a) Regret/PerfectOverlay

(b) Regret/Newscast

Figure 3. Comparison of ϵ-greedy and UCB in terms of regret. We used the PerfectOverlay protocol in 3(a) and the Newscast protocol in 3(b).

E.3. The necessity of using delays

Algorithm P2P-ϵ-Gr-merge(N) (where $N = 10, 100, 1000$ denotes the number of peers in the various experiments) does exactly the same as P2P-ϵ-Greedy, but without the delaying. More precisely, in time t peer j uses $s^i_{j,t}\big(1,t\big)/n^i_{j,t}\big(1,t\big)$ to estimate the expected reward at arm i, and not on $c^i_{j,t}/d^i_{j,t} = s^i_{j,t}\big(1,\mathcal{T}(t/2)-1\big)/n^i_{j,t}\big(1,\mathcal{T}(t/2)-1\big)$. Its performance is shown in Figure 1. From that it is quite apparent that the delay applied in our P2P-ϵ-Greedy is really crucial.

Although it is much harder to argue formally in favor of applying the delay, we try to give some intuition on what causes this really big difference in the performance.

Example 9. *Consider the scenario when $N = K$ large, the first arm returns constantly 0.9, and the rest of the arms have Bernoully distribution with parameter 0.8.*

Consider a suboptimal arm i that is pulled only a few times so far, and assume that the weights of all the rewards for it are well spread. More precisely, assume that the sum of the weights of the rewards for arm i does not exceed $N^{1/2}$ at any peer. Now, if some peer j pulls arm i again in iteration t_0, the resulting reward ξ will have weight 1 at j. Consequently the new reward would clearly dominate the estimate. Assuming furthermore that $\xi = 1$ (which has quite large, 0.8 probability), j will very likely pull arm i again. What is more, the weight for ξ will be higher than $N^{1/2}$ at any peer it reaches (that is, higher then the sum of all the previous rewards at any given peer) even after $\log(N/2)$ iterations.

Furthermore, one can easily show that, with high probability, ξ will not reach the same peer twice during rounds $t_0, t_0+1, \ldots, t_0 + \lfloor(\log N)/2\rfloor - 1$. Indeed, let $\mathcal{N}_t$ denote the set of peers that ξ has reached in rounds $t_0, t_0+1, \ldots, t$. Then $|\mathcal{N}_{t_0}| = 1$ holds obviously, and $|\mathcal{N}_t| = 1 + 2 + \cdots + 2^{t-t_0} = 2^{t-t_0+1} - 1$ if and only if no peer is reached twice during rounds $t_0, t_0+1, \ldots, t$. Now, for $t \le t_0 + (\log N)/2 - 1$

$$
\begin{aligned}
\mathbb{P}\left[|\mathcal{N}_t| = 2^{t-t_0+1} - 1 \,\middle|\, |\mathcal{N}_{t-1}| = 2^{t-t_0} - 1\right] &= \prod_{i=0}^{2^{t-t_0}-1} \left(\frac{N - 2^{t-t_0} + 1 - i}{N}\right) \\
&= \prod_{i=0}^{2^{t-t_0}-1} \left(1 - \frac{2^{t-t_0} - 1 + i}{N}\right) \\
&\ge \left(1 - \frac{2^{t-t_0+1}}{N}\right)^{2^{t-t_0+1}} \\
&= \left(\left(1 - \frac{1}{2^{-t+t_0-1}N}\right)^{2^{-t+t_0-1}N}\right)^{2^{2(t-t_0+1)}/N} \\
&\ge 4^{-2^{2(t-t_0+1)}/N}
\end{aligned}
$$

Thus

$$\begin{aligned}
&\mathbb{P}\left[\left|\mathcal{N}_{t_0+\lfloor(\log N)/2\rfloor-1}\right| \geq \sqrt{N}\right] \\
&\quad\geq \mathbb{P}\left[|\mathcal{N}_t| = 2^{t-t_0+1}-1,\ \ t = t_0, t_0+1, \ldots, t_0+\lfloor(\log N)/2\rfloor-1\right] \\
&\quad= \prod_{t=t_0+1}^{t_0+\lfloor(\log N)/2\rfloor-1} \mathbb{P}\left[|\mathcal{N}_t| = 2^{t-t_0+1}-1 \,\middle|\, |\mathcal{N}_{t'}| = 2^{t'-t_0+1}-1,\ t' = t_0, \ldots, t-1\right] \\
&\quad= \prod_{t=t_0+1}^{t_0+\lfloor(\log N)/2\rfloor-1} \mathbb{P}\left[|\mathcal{N}_t| = 2^{t-t_0+1}-1 \,\middle|\, |\mathcal{N}_{t-1}| = 2^{t-t_0}-1\right] \\
&\quad\geq 4^{-(1/N)\left(1+4+\cdots+4^{\lfloor(\log N)/2\rfloor}\right)} \geq 1/16
\end{aligned}$$

Summing up, there is a high probability that some suboptimal arm i is pulled, and that the resulting reward ξ causes at $\Omega(\sqrt{N})$ peers the estimate for i to be larger then the estimate for the optimal arm at least once during the following $\log_4 N$ rounds. Therefore, unless some other suboptimal arms also get pulled during these rounds, the cumulated regret during these rounds will be $\Omega(\sqrt{N})$, just because of pulling arm i.

If this is repeated for all the suboptimal arms, then it can lead to a cumulated regret of size $N^{3/2}$.

The effect is thus similar to the one described in Example 8.