Supplemental Material for Multi-objective Contextual Bandit Problem with Similarity Information

Eralp Turğay
Electrical and Electronics Engineering Department
Bilkent University
Ankara, Turkey

Doruk Öner
Electrical and Electronics Engineering Department
Bilkent University
Ankara, Turkey

Cem Tekin
Electrical and Electronics Engineering Department
Bilkent University
Ankara, Turkey

APPENDIX

APPENDIX A - A CONCENTRATION INEQUALITY [1, 2]

Consider a ball B for which the rewards of objective i are generated by a process $\{R_B^i(t)\}_{t=1}^T$ with mean $\mu_B^i = \mathrm{E}[R_B^i(t)]$, where the noise $R_B^i(t) - \mu_B^i$ is 1-sub-Gaussian. Recall that $B(t)$ is the ball selected in round t and $N_B(T)$ is the number of times ball B is selected by the beginning of round T. Let $\hat{\mu}_B^i(T) = \sum_{t=1}^{T-1} \mathrm{I}(B(t) = B) R_B^i(t)/N_B(T)$ for $N_B(T) > 0$ and $\hat{\mu}_B^i(T) = 0$ for $N_B(T) = 0$. Then, for any $0 < \theta < 1$ with probability at least $1 - \theta$ we have

$$\left|\hat{\mu}_B^i(T) - \mu_B^i\right| \leq \sqrt{\frac{2}{N_B(T)}\left(1 + 2\log\left(\frac{(1+N_B(T))^{1/2}}{\theta}\right)\right)} \quad \forall T \in \mathbb{N}.$$

APPENDIX B - PROOF OF LEMMA 1

From the definitions of $\tilde{L}_B^i(t)$, $\tilde{U}_B^i(t)$ and $\tilde{\mathrm{UC}}_B^i$, it can be observed that the event $\tilde{\mathrm{UC}}_B^i$ happens when $\tilde{\mu}_B^i(t)$ remains away from $\mu_{y_B}^i(x_B)$ for some $t \in \{0, \ldots, N_B(T+1)\}$. Using this information, we can use the concentration inequality given in Appendix A. In this formulation expected rewards of the arms must be equal in all time steps, but in our case, $\mu_{\tilde{y}_B(t)}^i(\tilde{x}_B(t))$ changes since the elements of $\{\tilde{x}_B(t), \tilde{y}_B(t)\}_{t=1}^{N_B(T+1)}$ are not identical which makes distributions of $\tilde{R}_B^i(t)$, $t \in \{1, \ldots, N_B(T+1)\}$ different.

In order to overcome this issue, we use the sandwich technique proposed in [3] and later used in [4]. For any ball $B \in \mathcal{B}(T)$, we have $\Pr(\mu_{y_B}^i(x_B) \notin [\tilde{L}_B^i(0) - r(B), \tilde{U}_B^i(0) + r(B)]) = 0$ since $\tilde{\mu}_B^i(0) = 0$, $\tilde{L}_B^i(0) = -\infty$ and $\tilde{U}_B^i(0) = \infty$. Thus, for $B \in \mathcal{B}(T) \setminus \mathcal{B}'(T)$, we have $\Pr(\mathrm{UC}_B | \mathcal{B}(T)) = 0$. Hence, we proceed by bounding the probabilities of the events $\{\mu_{y_B}^i(x_B) \notin [\tilde{L}_B^i(t) - r(B), \tilde{U}_B^i(t) + r(B)]\}$, for $t > 0$ and for the balls in $\mathcal{B}'(T)$. Recall that $\tilde{R}_B^i(t) = \mu_{\tilde{y}_B(t)}^i(\tilde{x}_B(t)) + \tilde{\kappa}_B^i(t)$ and $\tilde{\mu}_B^i(t) = \sum_{l=1}^t \tilde{R}_B^i(l)/t$ (for $t > 0$ and $B \in \mathcal{B}'(T)$). For each $i \in \{1, \ldots, d_r\}$, $B \in \mathcal{B}'(T)$, let

$$\overline{\mu}_B^i = \sup_{(x,y) \in \mathrm{dom}(B)} \mu_y^i(x) \quad \text{and} \quad \underline{\mu}_B^i = \inf_{(x,y) \in \mathrm{dom}(B)} \mu_y^i(x).$$

We define two new sequences of random variables, whose sample mean values will lower and upper bound $\tilde{\mu}_B^i(t)$. The *best sequence* is defined as $\{\bar{R}_B^i(t)\}_{t=1}^{N_B(T+1)}$ where $\overline{R}_B^i(t) := \overline{\mu}_B^i + \tilde{\kappa}_B^i(t)$, and the *worst sequence* is defined as $\{\underline{R}_B^i(t)\}_{t=1}^{N_B(T+1)}$ where $\underline{R}_B^i(t) := \underline{\mu}_B^i + \tilde{\kappa}_B^i(t)$. Let $\overline{\mu}_B^i(t) := \sum_{l=1}^t \overline{R}_B^i(l)/t$ and $\underline{\mu}_B^i(t) := \sum_{l=1}^t \underline{R}_B^i(l)/t$. We have

$$\underline{\mu}_B^i(t) \leq \tilde{\mu}_B^i(t) \leq \overline{\mu}_B^i(t) \quad \forall t \in \{1, \ldots, N_B(T+1)\}.$$

Let

$$\begin{aligned}
\overline{L}_B^i(t) &:= \overline{\mu}_B^i(t) - \tilde{u}_B(t) \\
\overline{U}_B^i(t) &:= \overline{\mu}_B^i(t) + \tilde{u}_B(t) \\
\underline{L}_B^i(t) &:= \underline{\mu}_B^i(t) - \tilde{u}_B(t) \\
\underline{U}_B^i(t) &:= \underline{\mu}_B^i(t) + \tilde{u}_B(t).
\end{aligned}$$

It can be shown that

$$\begin{aligned}
&\{\mu_{y_B}^i(x_B) \notin [\tilde{L}_B^i(t) - r(B), \tilde{U}_B^i(t) + r(B)]\} \qquad (1) \\
&\subset \{\mu_{y_B}^i(x_B) \notin [\overline{L}_B^i(t) - r(B), \overline{U}_B^i(t) + r(B)]\} \\
&\cup \{\mu_{y_B}^i(x_B) \notin [\underline{L}_B^i(t) - r(B), \underline{U}_B^i(t) + r(B)]\}.
\end{aligned}$$

Moreover, the following inequalities can be obtained from Assumption 1:

$$\mu_{y_B}^i(x_B) \leq \overline{\mu}_B^i \leq \mu_{y_B}^i(x_B) + r(B) \qquad (2)$$

$$\mu_{y_B}^i(x_B) - r(B) \leq \underline{\mu}_B^i \leq \mu_{y_B}^i(x_B). \qquad (3)$$

Using (2) and (3) it can be shown that

$$\{\mu_{y_B}^i(x_B) \notin [\overline{L}_B^i(t) - r(B), \overline{U}_B^i(t) + r(B)]\}$$

$$\subset \{\overline{\mu}^i_B \notin [\overline{L}^i_B(t), \overline{U}^i_B(t)]\},$$
$$\{\mu^i_{y_B}(x_B) \notin [\underline{L}^i_B(t) - r(B), \underline{U}^i_B(t) + r(B)]\}$$
$$\subset \{\underline{\mu}^i_B \notin [\underline{L}^i_B(t), \underline{U}^i_B(t)]\}.$$

Plugging this to (1), we get

$$\{\mu^i_{y_B}(x_B) \notin [\tilde{L}^i_B(t) - r(B), \tilde{U}^i_B(t) + r(B)]\}$$
$$\subset \{\overline{\mu}^i_B \notin [\overline{L}^i_B(t), \overline{U}^i_B(t)]\} \bigcup \{\underline{\mu}^i_B \notin [\underline{L}^i_B(t), \underline{U}^i_B(t)]\}.$$

Using the equation above and the union bound we obtain

$$\Pr(\mathrm{UC}^i_B | \mathcal{B}(T)) \leq \Pr\left(\bigcup_{t=1}^{N_B(T+1)} \{\overline{\mu}^i_B \notin [\overline{L}^i_B(t), \overline{U}^i_B(t)]\}\right) + \Pr\left(\bigcup_{t=1}^{N_B(T+1)} \{\underline{\mu}^i_B \notin [\underline{L}^i_B(t), \underline{U}^i_B(t)]\}\right).$$

Both terms on the right-hand side of the inequality above can be bounded using the concentration inequality in Appendix A. Using $\theta = \delta/(2d_rT)$, in Appendix A gives $\Pr(\mathrm{UC}^i_B|\mathcal{B}(T)) \leq \delta/(d_rT)$, since $1 + N_B(t) \leq 1 + N_B(T+1) \leq 2T$. Then, using the union bound over all objectives, we obtain $\Pr(\mathrm{UC}_B|\mathcal{B}(T)) \leq \delta/T$.

APPENDIX C - PROOF OF LEMMA 5

The maximum number of times a radius r ball B can be selected before it becomes a parent ball is upper bounded by $1 + 2r^{-2}(1 + 2\log(2\sqrt{2}d_rT^{\frac{3}{2}}/\delta))$. From the result of Lemma 3, we know that the Pareto regret in each of these rounds is upper bounded by $14r$. Note that after ball B becomes a parent ball, it will create a new radius $r/2$ child ball every time it is selected. From Lemma 4, we know that the Pareto regret in each of these rounds is bounded above by $12(r/2)$. Therefore, we can include the Pareto regret incurred in a round in which a new child ball with radius r is created from a parent ball as a part of the child ball's (total) Pareto regret. Hence, the Pareto regret incurred in a radius r ball is upper bounded by

$$14r\left(1 + 2r^{-2}(1 + 2\log(2\sqrt{2}d_rT^{\frac{3}{2}}/\delta))\right) + 12r$$
$$\leq 14r\left(2 + 2r^{-2}(1 + 2\log(2\sqrt{2}d_rT^{\frac{3}{2}}/\delta))\right)$$
$$\leq 56r^{-1}\log(2\sqrt{2}d_rT^{\frac{3}{2}}e/\delta).$$

Let $r_l := 2^{\lceil \log(r_0)/\log(2)\rceil}$. We have $r_0/2 \leq r_l/2 \leq r_0 \leq r_l \leq 2r_0$. The one-round Pareto regret of the balls whose radii are smaller than r_l is bounded by $14r_l$ on event UC^c according to Lemma 3. Also, we know that $14r_l \leq 28r_0$ by the above inequality. Therefore, the Pareto regret due to all balls with radii smaller than r_l by time T is bounded by $28Tr_0$, and the Pareto regret due to all balls with radii $r = 2^{-i} \geq r_0$ is bounded by $56r^{-1}N_r\log(2\sqrt{2}d_rT^{\frac{3}{2}}e/\delta)$. Thus, summing this up for all possible balls, we obtain the following Pareto regret bound on event UC^c:

$$\mathrm{Reg}(T) \leq 28Tr_0 + \sum_{r=2^{-i}: i\in\mathbb{N}, r_0\leq r\leq 1} 56r^{-1}N_r\log(2\sqrt{2}d_rT^{\frac{3}{2}}e/\delta).$$

APPENDIX D- SIMULATIONS

We evaluate the performance of PCZ on a synthetic dataset. We take $\mathcal{X} = [0,1]$, $\mathcal{Y} = [0,1]$, and generate $\mu^1_y(x)$ and $\mu^2_y(x)$ as shown in Figure 1. To generate $\mu^1_y(x)$, we first define a line by equation $8x + 10y = 8$ and let $y_1(x) = (8 - 8x)/10$. For all context arm pairs (x, y), we set $\mu^1_y(x) = \max\{0, (1 - 5|y - y_1(x)|)\}$. Similarly, to generate $\mu^2_y(x)$, we define the line $8x + 10y = 10$ and let $y_2(x) = (10 - 8x)/10$. Then, we set $\mu^2_y(x) = \max\{0, (1 - 5(y_2(x) - y))\}$ for $y \leq y_2(x)$ and $\mu^2_y(x) = \max\{0, (1 - (y - y_2(x))/4)\}$ for $y > y_2(x)$.

Based on the definitions given above, the Pareto optimal arms given context x lie in the interval $[y_1(x), y_2(x)]$. To evaluate the fairness of PCZ, we define six bins that correspond to context-arm pairs in the Pareto front. Given context x, the 1st bin contains all arms in the interval $[y_1(x), y_1(x) + 1/30]$ and the ith bin $i \in \{2, \ldots, 6\}$ contains all arms in the interval $(y_1(x) + (i-1)/30, y_1(x) + i/30]$. Simply, the first three bins include the Pareto optimal arms whose expected rewards in the first objective are higher than the expected rewards in the second objective and the last three bins include the Pareto optimal arms whose expected rewards in the second objective are higher than the expected rewards in the first objective.

We assume that the reward of arm y in objective i given context x is a Bernoulli random variable with parameter $\mu^i_y(x)$. In addition, at each round t, the context x_t is sampled from the uniform distribution over $\mathcal{X}$.

We compare our algorithm with Contextual Zooming [5] and Random Selection, which chooses in each round an arm uniformly at random from $\mathcal{Y}$. Contextual Zooming only uses the rewards in the first objective to update itself. Both PCZ and Contextual Zooming uses scaled Euclidean distance.[1] We choose $\delta = 1/T$ in PCZ, set $T = 10^5$, run each algorithm 100 times, and report the average of the results in these runs.

[1] We set $D((x,y),(x',y')) = \sqrt{(x-x')^2 + (y-y')^2}/\sqrt{2}$. While this choice does not satisfy Assumption 1, we use this setup to illustrate that learning is still possible when the distance function is not perfectly known by the learner.

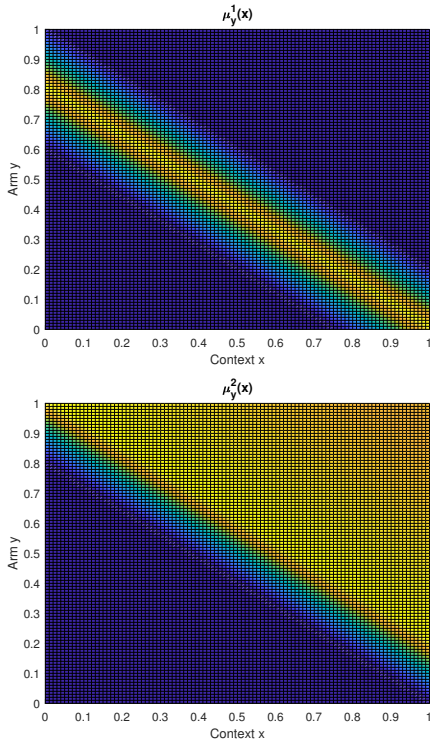

Figure 1: Expected Rewards of Context-Arm Pairs (Yellow Represents 1, Dark Blue Represents 0)

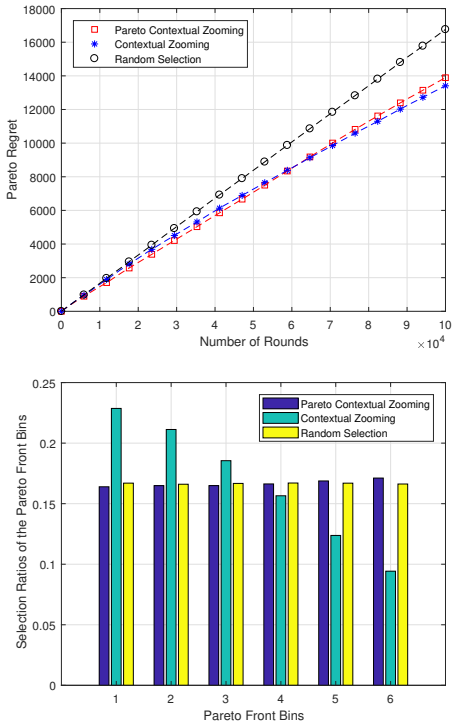

Figure 2: (i) Pareto Regret vs. Number of Rounds (ii) Selection Ratio of the Pareto Front Bins

The Pareto regret is reported in Figure 2(i) as a function of the number of rounds. It is observed that the Pareto regret of PCZ at $T = 10^5$ is 3.61% higher than that of Contextual Zooming and 17.1% smaller than that of Random Selection. We compare the fairness of the algorithms in Figure 2(ii). For this, we report the selection ratio of each Pareto front bin, which is defined for bin i as the number of times a context-arm pair in bin i is selected divided by the number of times a Pareto optimal arm is selected by round T. We observe that the selection ratio of all bins are almost the same for PCZ, while Contextual Zooming selects the context-arm pairs in the 1st bin much more than the other bins.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2312–2320, 2011.

[2] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[3] C. Tekin, J. Yoon, and M. van der Schaar, "Adaptive ensemble learning with confidence bounds," *IEEE Transactions on Signal Processing*, vol. 65, no. 4, pp. 888–903, 2017.

[4] C. Tekin and E. Turgay, "Multi-objective contextual multi-armed bandit problem with a dominant objective," *arXiv preprint arXiv:1708.05655*, 2017.

[5] A. Slivkins, "Contextual bandits with similarity information," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2533–2568, 2014.