[32] L. Qin *et al.,* Contextual Combinatorial Bandit and its Application on Diversified Online Recommendation. In *SDM*, pages 461–469, 2014.
[33] X. Zhao *et al.,* Interactive collaborative filtering. In *CIKM*, 2013.
[34] N. Cesa-Bianchi *et al.,* A gang of bandits. In *NIPS*, pages 737–745, 2013.
[35] Y. Yue *et al.,* Hierarchical exploration for accelerating contextual bandits. In *ICML*, 2012.
[36] D. Bouneffouf *et al.,* A contextual-bandit algorithm for mobile context-aware recommender system. In *ICONIP*, pages 324–331, 2012.
[37] K. Amin *et al.,* Graphical Models for Bandit Problems. In *UAI*, 2011.
[38] W. Chu *et al.,* Contextual Bandits with Linear Payoff Functions. In *AISTATS*, 2011.
[39] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.
[40] Y. Abbasi-Yadkori *et al.,* Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.
[41] R. Sutton and A. Barto. Reinforcement learning: An introduction. Cambridge Univ Press, 2011.
[42] L. Li *et al.,* Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, pages 297–306, 2011.
[43] S. Filippi *et al.,* Parametric bandits: The generalized linear case. In *NIPS*, 2010.
[44] W. Li *et al.,* Exploitation and exploration in a performance based contextual advertising system. In *KDD*, pages 27–36, 2010.
[45] L. Li *et al.,* A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670, 2010.
[46] D. Agarwal *et al.,* Explore/exploit schemes for web content optimization. In *ICDM*, pages 1–10, 2009.
[47] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, pages 817–824, 2008.
[48] A. Mahajan and D. Teneketzis. Multi-armed bandit problems. *Foundations and Applications of Sensor Management*, 121–151, 2008.
[49] J. Langford *et al.,* Exploration scavenging. In *ICML*, pages 528–535, 2008.
[50] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, Nov(3):397–422, 2002.
[51] P. Auer *et al.,* Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 2-3(47):235–256, 2002.
[52] P. Auer. Using upper confidence bounds for online learning. In *41st Annual Symposium on Foundations of Computer Science*, pages 270–279, 2000.
[53] P. Auer *et al.,* Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *36th Annual Symposium on Foundations of Computer Science*, 1995.
[54] D. Berry and B. Fristedt. Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability). Springer, 1985.
[55] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 1(6):4–22, Elsevier, 1985.
[56] J. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 148–177, JSTOR, 1979.
[57] R. Luce. Individual choice behavior, a theoretical analysis. *Bull. Amer. Math. Soc*, 66(1960):259-260, 1960.
[58] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 5(58):527–535, 1952.

A REGRET OF SELF-REWARD FOR ISR

Recall that for UCB based algorithm, take (5) and (7) for instance, the choice of item in each round is:

$$i(t) = \arg\max_{j=1,\cdots,K} \left(\hat{r}_j(t) + \hat{c}_j(t) \right), \tag{17}$$

where for each item $j = 1, \cdots, K$, the true mean reward $r_j(t)$ in round t lies in a confidence interval:

$$C_j(t) : \left[\hat{r}_j(t) - \hat{c}_j(t) \quad , \quad \hat{r}_j(t) + \hat{c}_j(t) \right]. \tag{18}$$

To be brief, the estimation of $r_j(t)$ is supposed to be as optimistic as possible and then the item with the best optimistic estimate will be chosen.

As such, we formulate the regret in the vanilla stochastic multi-arm bandit setting as a simpler version of that indicated in (1):

$$R_T = \sum_{t=1}^{T} \left(\mu_* - r_i(t) \right), \tag{19}$$

where u_* denotes the expected reward of the best item. Then [51] shows that after running the UCB based algorithms, with high probability:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} \left(\mu_* - r_i(t) \right) \le \sum_{t=1}^{T} \left(\hat{r}_i(t) + \hat{c}_i(t) - r_i(t) \right) \\ &\le \sum_{t=1}^{T} \left(\hat{r}_i(t) + \hat{c}_i(t) - \left(\hat{r}_i(t) - \hat{c}_i(t) \right) \right) = 2 \sum_{t=1}^{T} \hat{c}_i(t) \,. \end{aligned} \tag{20}$$

Confidence Intervals. It is easy to show that through concatenating all feature vectors into a single "larger" one, the self-reward part of ISR can be treated as a special case of general linear stochastic bandit [40], which in each round chooses the item such that:

$$i(t) = \arg\max_{j=1,\cdots,K} \left(\hat{\mathbf{p}}_t^\intercal \mathbf{q}_{t,j} + c \sqrt{\mathbf{q}_{t,j}^\intercal \Sigma_t^{-1} \mathbf{q}_{t,j}} \right). \tag{21}$$

And the ellipsoid confidence interval for $\mathbf{p}$ is:

$$C_t = \{ \mathbf{p} \mid ||\mathbf{p} - \hat{\mathbf{p}}||_{\Sigma_t^{-1}} \le c \}, \tag{22}$$

where $||x||_\Sigma = \sqrt{\mathbf{x}^\intercal \Sigma \mathbf{x}}$. Given that Σ_t is a symmetric positive definite matrix and:

$$||\mathbf{p} - \hat{\mathbf{p}}_t||_{\Sigma_t^{-1}} = \sqrt{(\mathbf{p} - \hat{\mathbf{p}}_t)^\intercal \Sigma_t^{-1} (\mathbf{p} - \hat{\mathbf{p}}_t)}, \tag{23}$$

if we set Σ_t to be identity matrix, resulting in a norm-2 regularization on $\mathbf{p} - \hat{\mathbf{p}}_t$, then $\hat{\mathbf{p}}_t$ can be estimated through the standard ridge regression:

$$\hat{\mathbf{p}}_t = \arg\min_{\mathbf{p}} \sum_{t'=1}^{t-1} \left(\hat{r}_i(t') - \mathbf{p}^\intercal \mathbf{q}_{t',i} \right) + \lambda ||\mathbf{p}||^2 \,. \tag{24}$$

The corresponding regret is then measured as follows:

$$R_T = \sum_{t=1}^{T} \left(\mathbf{p}_t^\intercal \mathbf{q}_{t,j^*} - \mathbf{p}_t^\intercal \mathbf{q}_{t,j} \right), \tag{25}$$

where $j^* = \arg\max_{j=1,\cdots,K} \mathbf{p}_t^\intercal \mathbf{q}_{t,j}$.

As a common setting, we follow the assumption that everything is Gaussian, e.g., the distribution D described in Section 3 follows a Gaussian distribution with μ and σ as mean and variance respectively. Thus from the solution of ridge regression, we have:

$$\Sigma_t = \lambda_p I + \sum_{t'=1}^{t} \mathbf{q}_{t',i} \mathbf{q}_{t',i}^\intercal \,, \tag{26}$$

making C_t in (22) a valid ellipsoid confidence set containing the true $\mathbf{p}$ with a very high probability controlled by c. Abbasi-Yadkori et al. [40] give a general condition on the use of valid confidence ellipsoid, which says if the linearity of true model and the independence of the rewards with R-sub-Gaussian (with $R \ge 0$) hold, and $\mathbf{p}$ as well as $\mathbf{q}$ are bounded by some constants, i.e., $||\mathbf{p}|| \le S$ and $||\mathbf{q}|| \le L$, then for any $0 \le \delta \le 1$ and all $t \ge 0$, with probability at least $1 - \delta$, the true optimal value $\mathbf{p}_*$ lies in the following ellipsoid confidence set C_t:

$$\mathbf{p} \in \mathbb{R}^d : ||\mathbf{p} - \hat{\mathbf{p}}_t||_{\Sigma_t^{-1}} \le R \sqrt{d \log \left(\frac{1 + tL^2/\lambda}{\delta} \right)} + \lambda^{\frac{1}{2}} S \,. \tag{27}$$

We refer readers to Theorem 2 in [40] for more details.

Therefore, applying (27) with R-sub-Gaussian tails on the noise, $\mathbf{p}$ and $\mathbf{q}$ upper bounded by S and L, C_t in (22) will be at most:

$$O\left(R \sqrt{d |\mathcal{I}| \log \frac{t}{\delta}} + \lambda^{\frac{1}{2}} S \right), \tag{28}$$

where d is the latent feature dimension and $|\mathcal{I}|$ is the number of candidate items.

Regret Bound. Under the assumption that $\lambda \ge \max_{\mathbf{q}} ||\mathbf{q}||^2$ and based on the proof of Theorem 3 in [40], we can further write (20) as follows:

$$R_T \le 2 \sum_{t=1}^{T} c_i(t) = 2 \sum_{t=1}^{T} c_t ||\mathbf{q}_{t,i}||_{\Sigma_t^{-1}} \le 2 \sqrt{\sum_{t=1}^{T} c_t^2 ||\mathbf{q}_{t,i}||^2_{\Sigma_t^{-1}}} \tag{29}$$

$$\le 2 \sqrt{c_T^2 \sum_{t=1}^{T} ||\mathbf{q}_{t,i}||^2_{\Sigma_t^{-1}}} = 2 c_T \sqrt{\sum_{t=1}^{T} ||\mathbf{q}_{t,i}||^2_{\Sigma_t^{-1}}} \,, \tag{30}$$

where (29) is obtained by applying Cauchy-Schwarz inequality [1] and (30) is obtained based on the fact that c_t is monotonically increasing. Again, Abbasi-Yadkori et al. [40] prove that if $\lambda \ge \max_{\mathbf{q}} ||\mathbf{q}||^2$ holds, then:

$$\sum_{t=1}^{T} ||\mathbf{q}_{t,i}||^2_{\Sigma_t^{-1}} \le 2 \log \det(\Sigma_T) \le O\left(d |\mathcal{I}| \log T \right). \tag{31}$$

Last, by putting (29) and (31) together, we have:

$$R_T \le O\left(dRS |\mathcal{I}| \lambda^{\frac{1}{2}} \log \left(\frac{T}{\delta} \right) \sqrt{T} \right), \tag{32}$$

and if we further ignore the logarithmic factors and regards the latent feature dimension parameter d as a constant, then the regret of the self-reward part of ISR is at most $O(\sqrt{T})$.

[1] https://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality