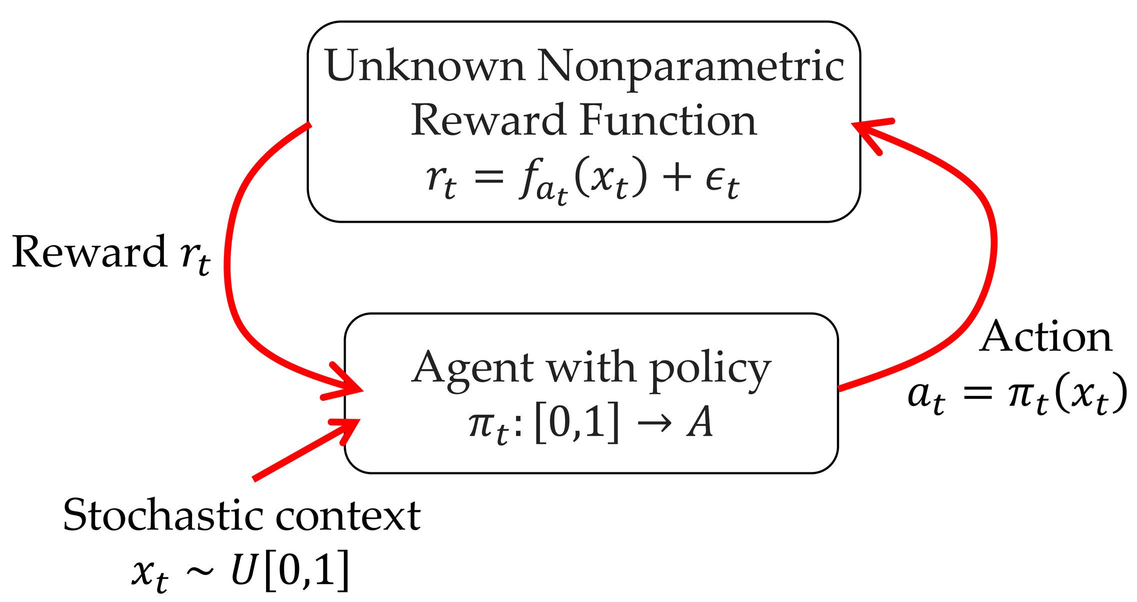
Nonparametric Contextual Bandits in an Unknown Metric Space

Nirandika Wanigasekara — National University of Singapore — nirandiw@comp.nus.edu.sg

Christina Lee Yu — Cornell University — cleeyu@cornell.edu

Problem Setup



- Optimal policy $\pi^*(x) = \arg\max_{a \in \mathcal{A}} f_a(x)$
- Want to minimize expected regret over time horizon T

$$\mathbb{E}\Big[\sum_{t=1}^{T} \big(f_{\pi^*(x_t)}(x_t) - f_{a_t}(x_t)\big)\Big]$$

Suppose $|\mathcal{A}|$ large but finite, with underlying unknown "simple" structure (e.g. drawn from metric space). Can algorithm exploit structure and perform better than treating all arms separately?

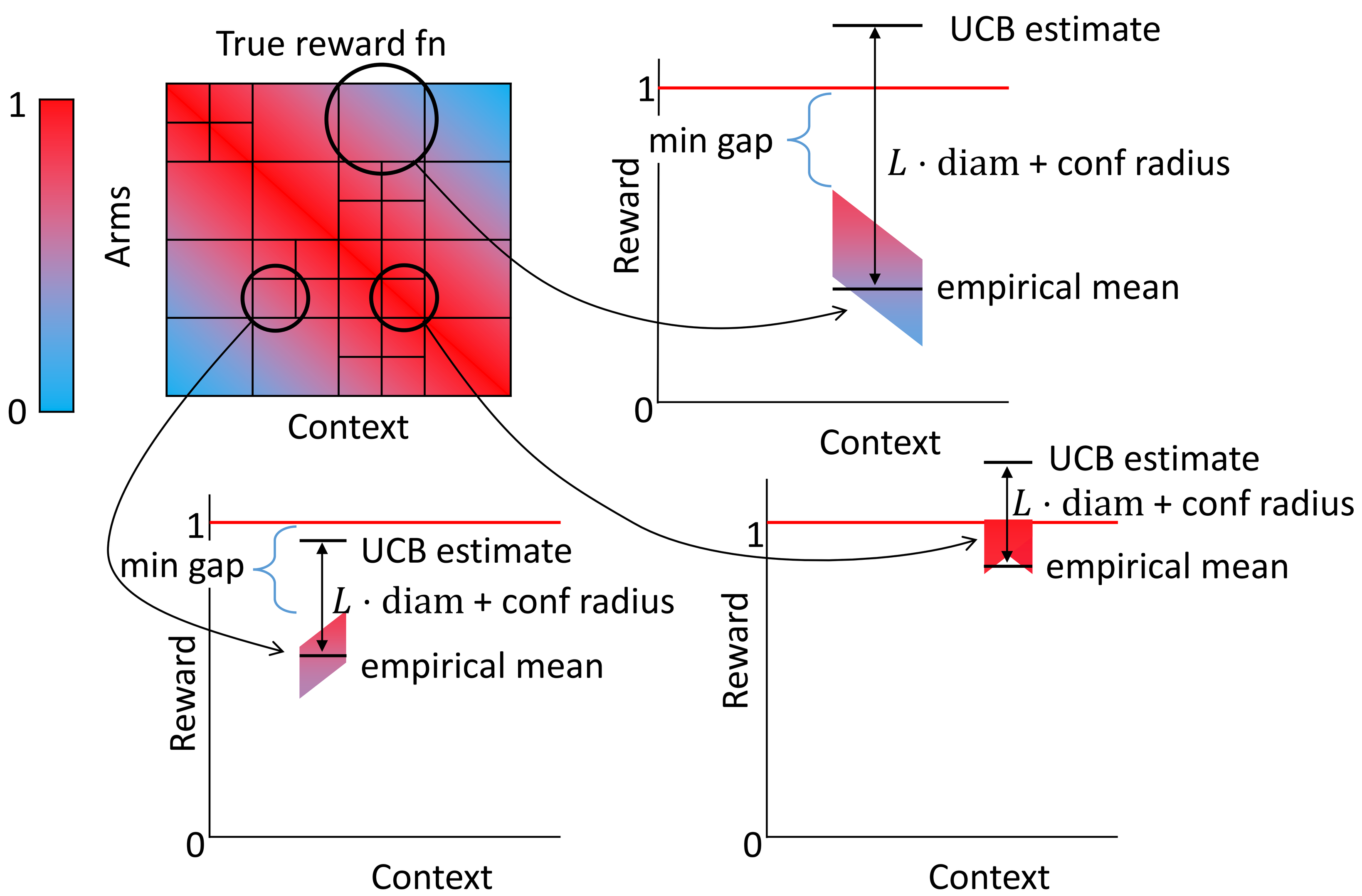
Algorithm for Known Metric [Slivkins 2014]

Zooming algorithm exploits structure via a *known* metric:

- Assume reward function is Lipschitz with respect to metric

$$|f_a(x) - f_{a'}(x')| \leq L\mathcal{D}((x,a),(x',a'))$$

- Key pieces: UCB + Adaptive discretization (Zooming)
- Maintain partition and Upper Confidence Bound estimates
- Select arm in region that maximizes UCB
- Subpartition region if confidence radius $\leq$ bias



Algorithm for Unknown Metric [this paper]

ApproxZooming algorithm estimates clustering between arms

- Maintain UCB estimates for reward in region B

$$UCB_t(B) = \hat{\mu}_t(B) + L\ \mathrm{diam}(B) + \sqrt{\frac{\sigma^2 \ln(T)}{n_t(B)}}$$

- Select "relevant" region that maximizes UCB
  $\to$ exception: flagged regions get priority
- Flag to subpartition when confidence radius $\leq$ bias,

$$n_t(B) \geq \left(\frac{\sigma^2 \ln(T)}{L^2\ \mathrm{diam}(B)}\right)^2$$

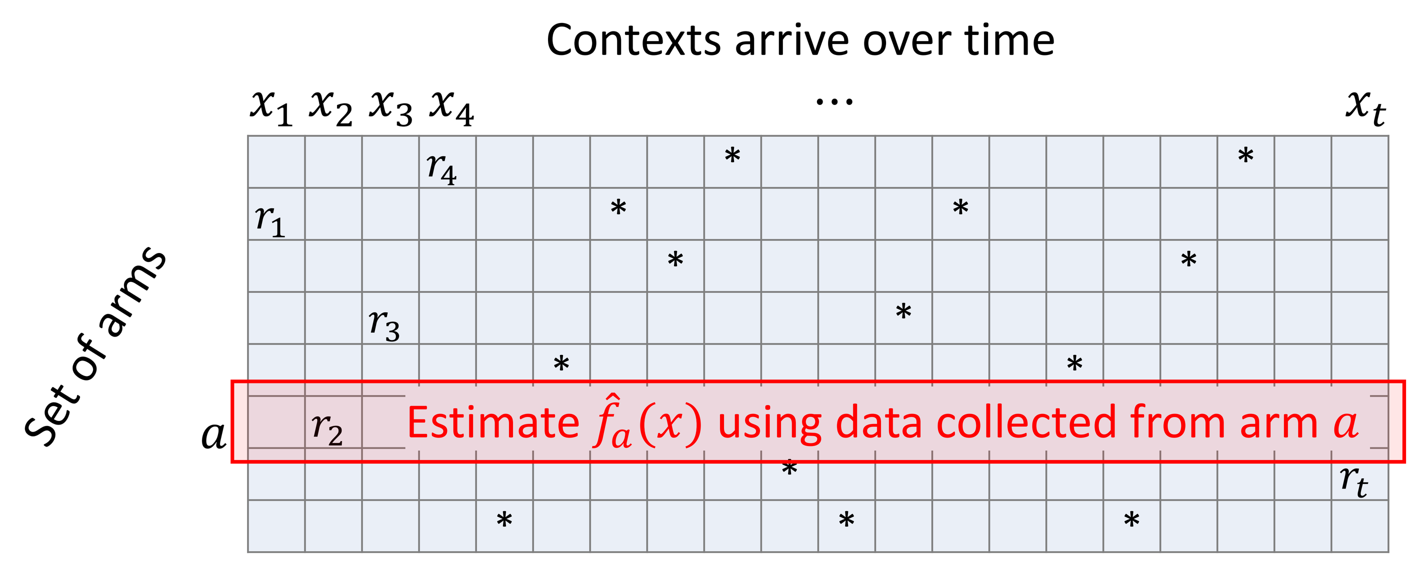
- When flagged region collects sufficient data, subpartition by clustering arms according to estimated L_2 distance

$$D_B(a,b) = \sqrt{\frac{1}{|\mathcal{X}_B|}\int_{x \in \mathcal{X}_B} (\hat{f}_a(x) - \hat{f}_b(x))^2 dx}$$

where $\hat{f}_a(x)$ is estimated via k-nearest neighbor

Inutition

Can one learn distances between arms efficiently?



- Nonparametric minimax rates (for $x_t \sim U([0,1]^d)$)

$$\inf_{\hat{f}_a} \sup_{f_a} \mathbb{E}\big[\sup_x |\hat{f}_a(x) - f_a(x)|^2\big] = \tilde{\Omega}(N^{-2/(d+2)})$$

$$\inf_{\hat{f}_a} \sup_{f_a} \mathbb{E}\big[(\|\hat{f}_a\|_2 - \|f_a\|_2)^2\big] = \tilde{\Omega}(\max(N^{-1/2}, N^{-4/(d+4)}))$$

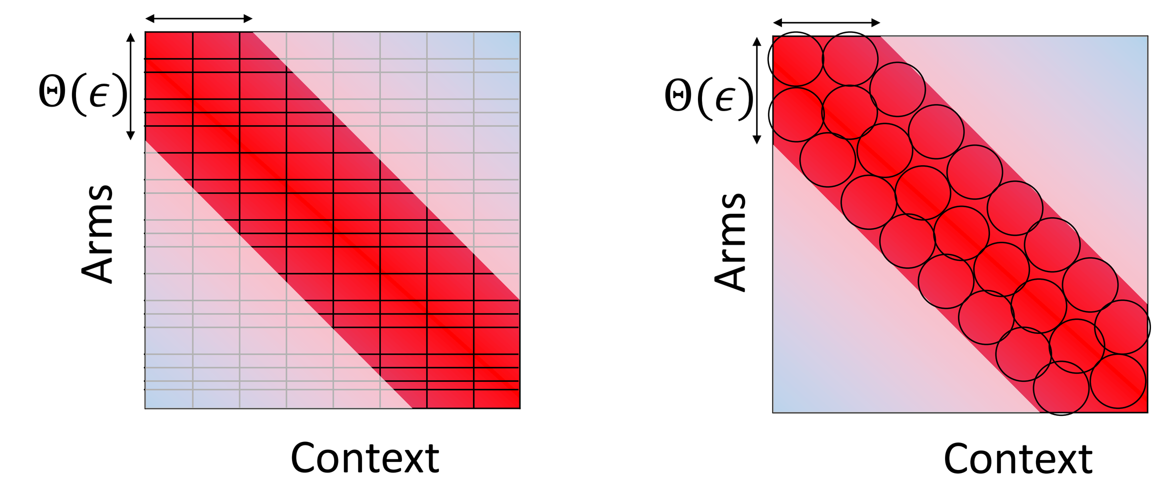
- Maintain partition of $[0,1] \times \mathcal{A}$ s.t. for each region B, $\forall\ (x,a),(x,b) \in B$, $|f_a(x) - f_b(x)| \leq L\ \mathrm{diam}(B)$
- With high prob, $L\ \mathrm{diam}(B) + 2$ conf radius $\leq$ min gap implies region B is never selected again because $UCB_t(B) \leq UCB_t(B^*)$ for B^* containing optimal
- A selected region of diameter ϵ incurs regret at most $O(\epsilon)$, and it is played at most $O(\epsilon^{-2})$ times before flagged
- $O(|\mathcal{A}_B|\epsilon^{-2})$ samples collected to learn clustering

Upper Bound on Regret

Algorithm achieves regret bounded by

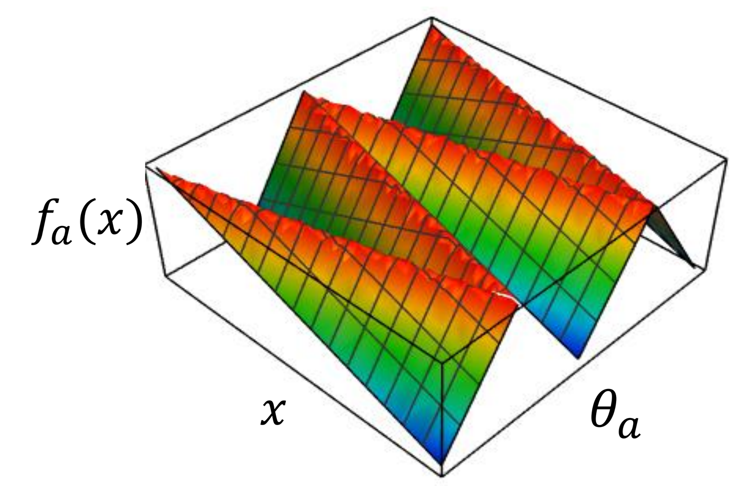
$$R(T) \leq C \inf_{\epsilon_0} \Big(\epsilon_0 T + \sum_{\epsilon \geq \epsilon_0} \frac{M_\epsilon}{\epsilon} \ln(T|\mathcal{A}|)\Big)$$

where M_ϵ denotes number of ϵ-optimal context-arms pairs with context discretization of ϵ. Final regret depends on appropriate "zooming dimension" with respect to discrete metric over arms.

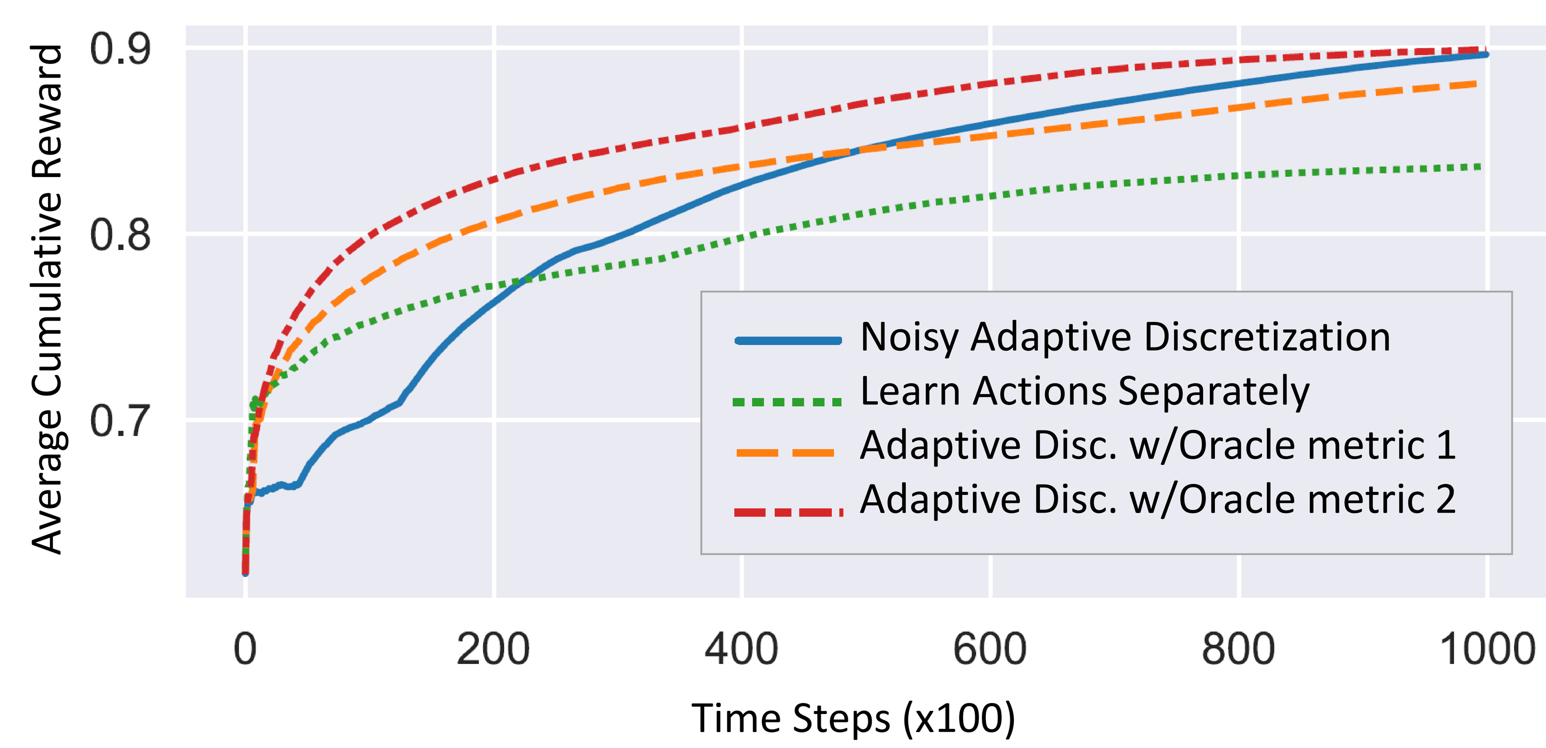


Simulations

- 200 arms, each $a \in \mathcal{A}$ associated to $\theta_a \in [0,1]$
- Reward $f_a(x) = g(x, \theta_a) = 1 - |x - 4\min_{z \in \{0, 0.5, 1\}} |\theta_a - z||$



- Oracle Metric 1: $d_1(a,a') = |\theta_a - \theta_{a'}|$
- Oracle Metric 2: $d_2(a,a') = \big(\int_0^1 (f_a(x) - f_{a'}(x))^2 dx\big)^{1/2}$
- Noisy estimated Metric: $\hat{d}_3(a,a') = \big(\int_0^1 (\hat{f}_a(x) - \hat{f}_{a'}(x))^2 dx\big)^{1/2}$



Our method eventually performs better than naïve oracle metric! Given covariates, could we learn the optimal metric from data?