Smooth Non-stationary Bandits

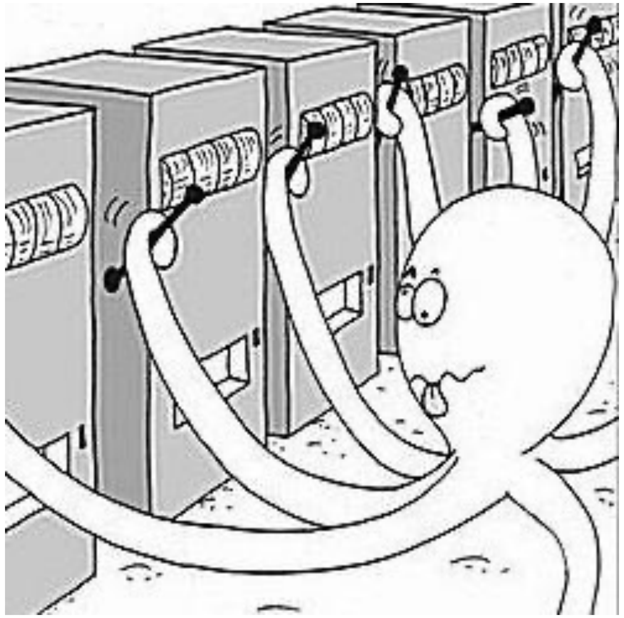
Qian Xie

CS6789 Project Presentation

Based on joint work with Su Jia, Nathan Kallus, and Peter Frazier

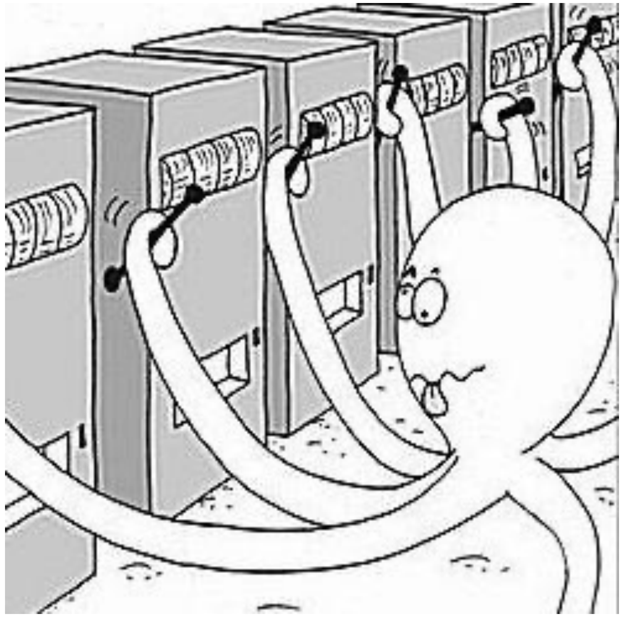
Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Environment changing over time



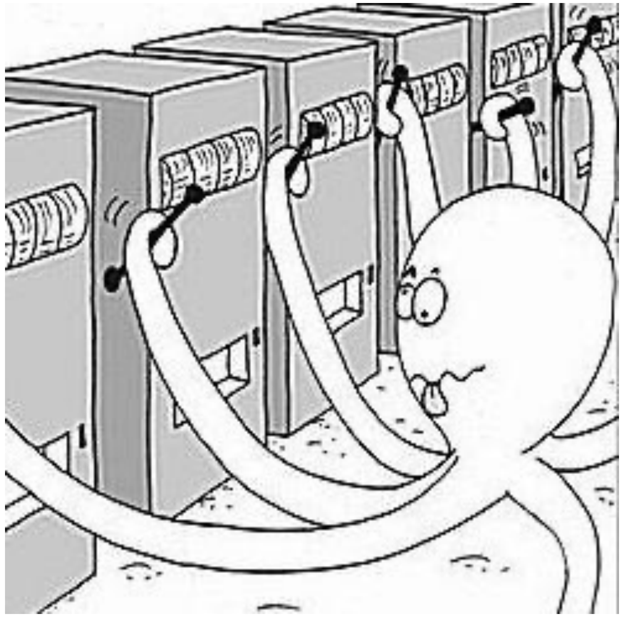
Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Environment changing over time
  - Middle ground between the stochastic bandits and adversarial bandits



Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi’14]
  - Environment changing over time
  - Middle ground between the stochastic bandits and adversarial bandits
  - Adversary chooses mean reward function in advance
  - Rewards are realized stochastically



Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Environment changing over time
  - Middle ground between the stochastic bandits and adversarial bandits
  - Adversary chooses mean reward function $r_a(t)$ in advance
  - Rewards are realized stochastically
  - Mean reward function is Lipschitz and confined by a total variation budget V:

$$\sum_{t=1}^{T} |r_a(t) - r_a(t+1)| \leq V$$

Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Environment changing over time
  - Middle ground between the stochastic bandits and adversarial bandits
  - Adversary chooses mean reward function $r_a(t)$ in advance
  - Rewards are realized stochastically
  - Mean reward function is Lipschitz and confined by a total variation budget V:

$$\sum_{t=1}^{T} |r_a(t) - r_a(t+1)| \leq V$$

  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$

Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Adversary chooses mean reward function $r_a(t) \coloneqq \mathrm{E}[Z_a^t]$ in advance
  - Rewards Z_a^t are realized stochastically
  - Mean reward function is Lipschitz and confined by a total variation budget V
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$

Def. The *regret* of a policy A under instance $r = \{r_a(t)\}$ is defined as

$$\mathrm{Reg}(A, r) = E\left[\sum_{t=1}^{T}\left(r^*(t) - Z_{A_t}^t\right)\right].$$

For a family F of instances, the *worst-case regret* of A is $\max_{r \epsilon F} \mathrm{Reg}(A, r)$.

The *minimax regret* is the minimum achievable worst-case regret among all policies.

Non-stationary Bandits

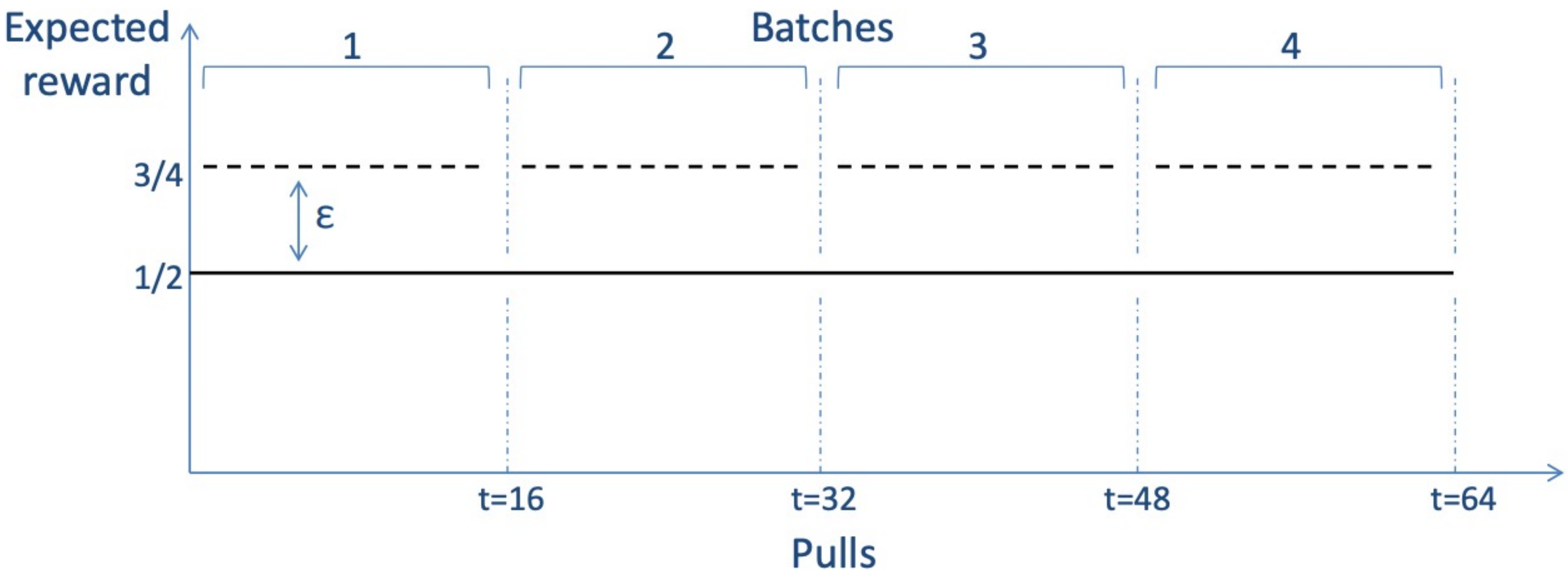
- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Middle ground between the stochastic bandits and adversarial bandits
  - Adversary chooses mean reward function $r_a(t) \coloneqq \mathrm{E}[Z_a^t]$ in advance
  - Rewards Z_a^t are realized stochastically
  - Mean reward function is Lipschitz and confined by a total variation budget V:

$$\sum_{t=1}^{T} |r_a(t) - r_a(t+1)| \leq V$$

  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}} T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope

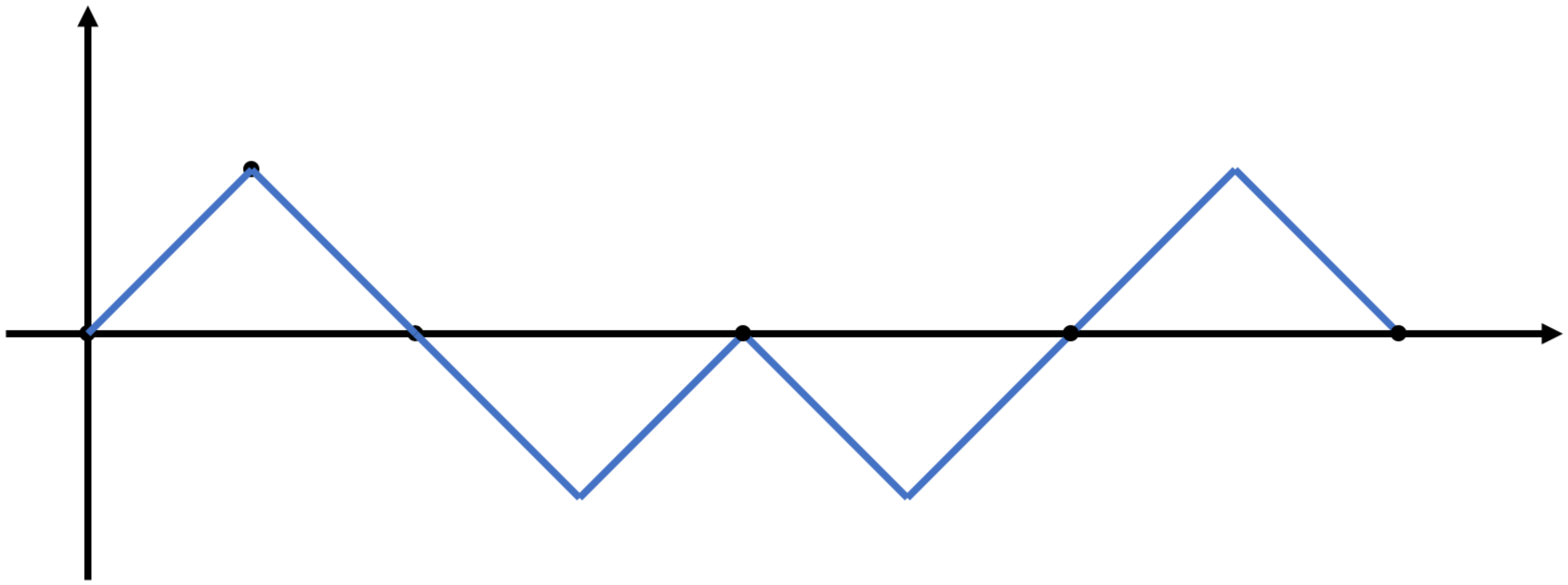
Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Allow the adversary to instantaneously shock the reward function's slope



Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Allow the adversary to instantaneously shock the reward function's slope



Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Middle ground between the stochastic bandits and adversarial bandits
  - Adversary chooses mean reward function $r_a(t)$ in advance
  - Mean reward function is Lipschitz and confined by a total variation budget V:

$$\sum_{t=1}^{T} |r_a(t) - r_a(t+1)| \leq V$$

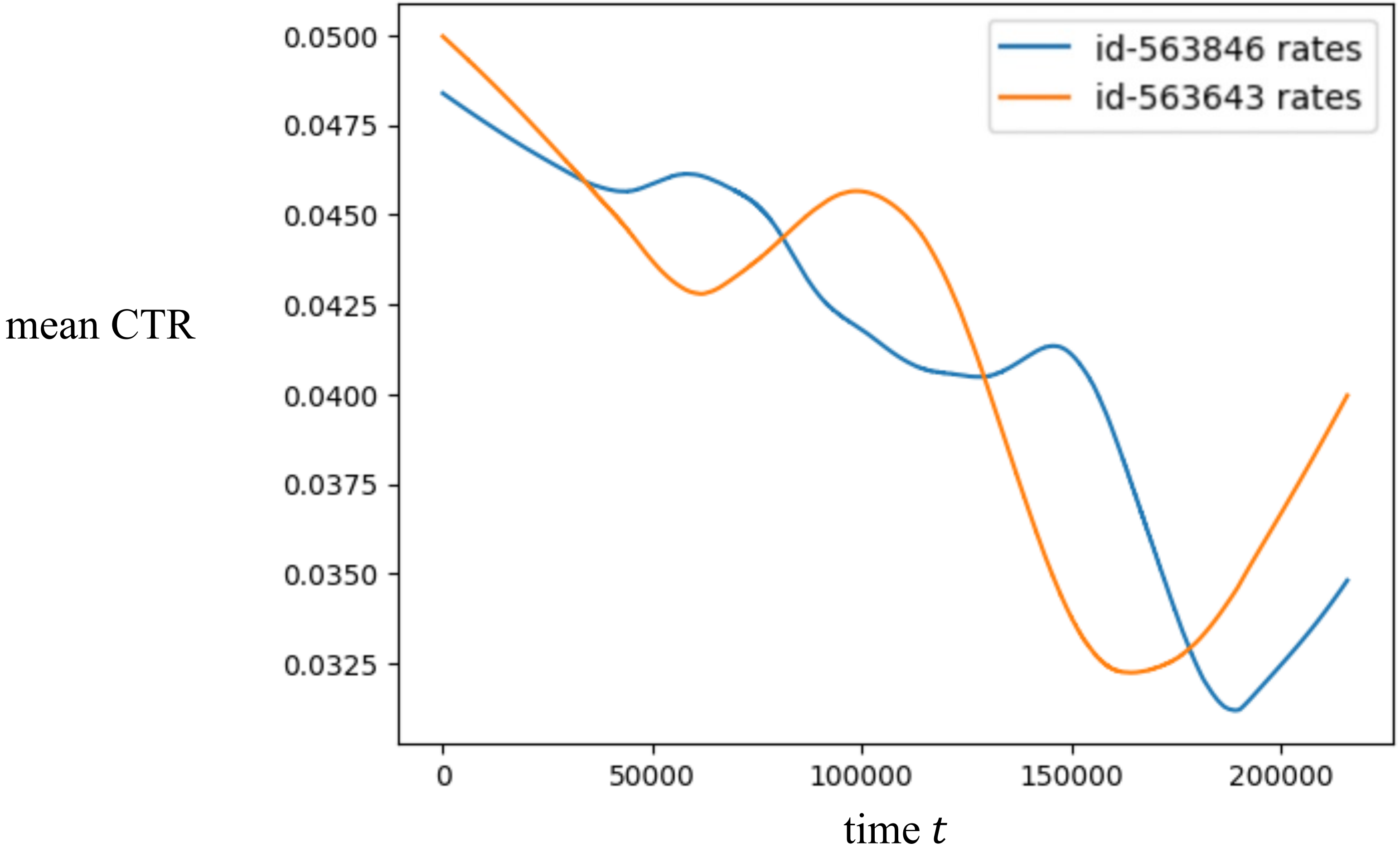
  - Rewards are realized stochastically
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope
  - Overly pessimistic for some applications

Smooth Non-stationary bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope
  - Overly pessimistic for some applications
- Smoothly-changing environment
  - The underlying environment changes in a smooth manner, e.g., temperature, seasonal product demands, economic factors

Smoothly-changing Environment

Yahoo! Front Page Click-Through Rates (CTR)



Smooth Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope
  - Overly pessimistic for some applications
- Smoothly-changing environment
  - The underlying environment changes in a smooth manner, e.g., temperature, seasonal product demands, economic factors
  - Adversaries constrained to choose reward functions that are smooth in time

Smooth Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope
  - Overly pessimistic for some applications
- Smoothly-changing environment
  - The underlying environment changes in a smooth manner, e.g., temperature, seasonal product demands, economic factors
  - Adversaries constrained to choose reward functions that are smooth in time
  - Level of smoothness -- Hölder class!

Hölder Class

Definition 2.2 (Hölder Class). For integers $\beta \geq 1$ and $L > 0$, we say a function $f: [0,1] \to R$ is β-Hölder and write $f \in \Sigma(\beta, L)$ if

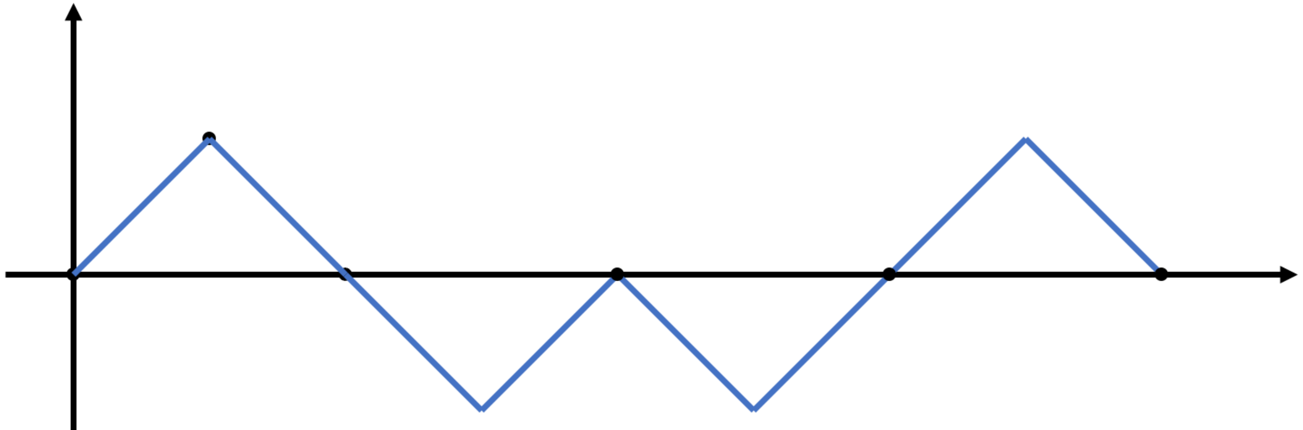
(i) f is $(\beta - 1)$-order differentiable, and

(ii) $f^{(\beta-1)}$ and f are both L-Lipschitz.

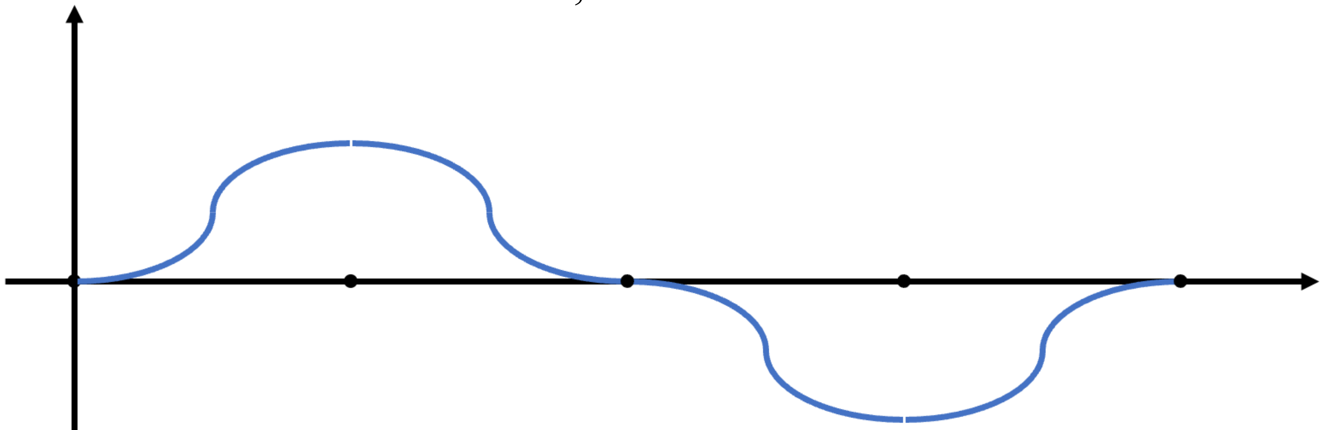
Example.

- $\beta = 1$: $f \in \Sigma(1, L)$ if and only if f is L-Lipschitz
- $\beta = 2$: $f \in \Sigma(2, L)$ if and only if f is differentiable and f' and f are L-Lipschitz

Hölder Class



1-Hölder, but not 2-Hölder



2-Hölder, but not 3-Hölder

Smooth Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope
- Smoothly-changing environment
  - The underlying environment changes in a smooth manner, e.g., temperature, seasonal product demands, economic factors
  - Adversaries constrained to choose reward functions that are smooth in time
  - Level of smoothness -- Hölder class!
  - [Besbes, Gur, Zeevi '14] admits an optimal $T^{2/3}$ regret with $V = O(1)$ for 1-Hölder (Lipschitz) reward function

Smooth Non-stationary Bandits

- Non-stationary bandits [Besbes, Gur, Zeevi'14]
  - Optimal regret bound $\tilde{O}(V^{\frac{1}{3}}T^{\frac{2}{3}})$
  - Allow the adversary to instantaneously shock the reward function's slope
- Smoothly-changing environment
  - The underlying environment changes in a smooth manner, e.g., temperature, seasonal product demands, economic factors
  - Adversaries constrained to choose reward functions that are smooth in time
  - Level of smoothness -- Hölder class!
  - [Besbes, Gur, Zeevi'14] admits an optimal $T^{2/3}$ regret with $V = O(1)$ for 1-Hölder (Lipschitz) reward function
  - Can we break this bound under smooth non-stationarity?

Main Results

- Smoothly-changing environment
  - Level of smoothness -- Hölder class!
  - [Besbes, Gur, Zeevi'14] admits an optimal $T^{2/3}$ regret for 1-Hölder (Lipschitz) reward function
  - Can we break this bound under smooth non-stationarity?
- Main results
  - First separation between the smooth and non-smooth regime
  - A $T^{3/5}$ upper bound for 2-Hölder reward function

Upper Bound

- First separation between the smooth ($\beta \geq 2$) and non-smooth ($\beta = 1$) regime
- Budgeted Exploration algorithm achieves $T^{3/5}$ upper bound for 2-Hölder reward function

$\beta = 1$: $B = T^{1/3}$, $\Delta = T^{-1/3}$
$\beta = 2$: $B = T^{1/5}$, $\Delta = T^{-1/5}$

Algorithm 1 Budgeted Exploration Policy BE(B, Δ)

for $i = 1, \ldots, \Delta^{-1}$ **do**
  Select arm 1 from round $t_i + 1$ until round $t_i + S_i$
  with $S_i = \min\{\tilde{S}_i, \Delta T\}$ where   exploring
  stopping time   epoch size

$$\tilde{S}_i = \min\{s : \sum_{t=t_i}^{t_i+s} Z_1^t \leq -B\}.$$

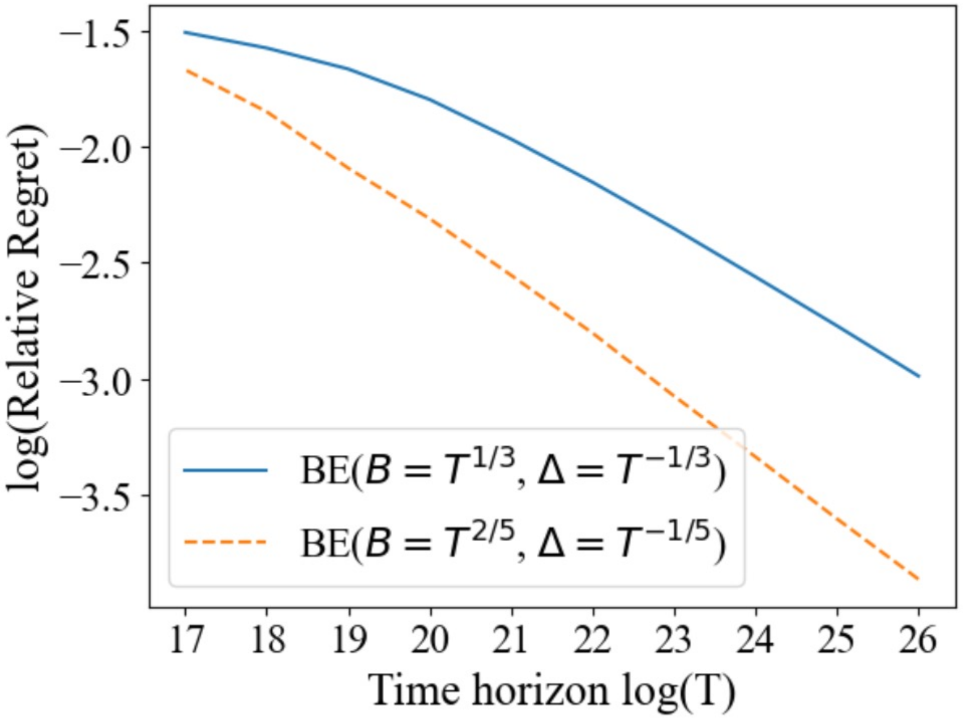
one-arm case

cumulative rewards   budget

  Then select arm 0 from round $t_i + S_i + 1$ till t_{i+1}.
end for   exploiting

Upper Bound

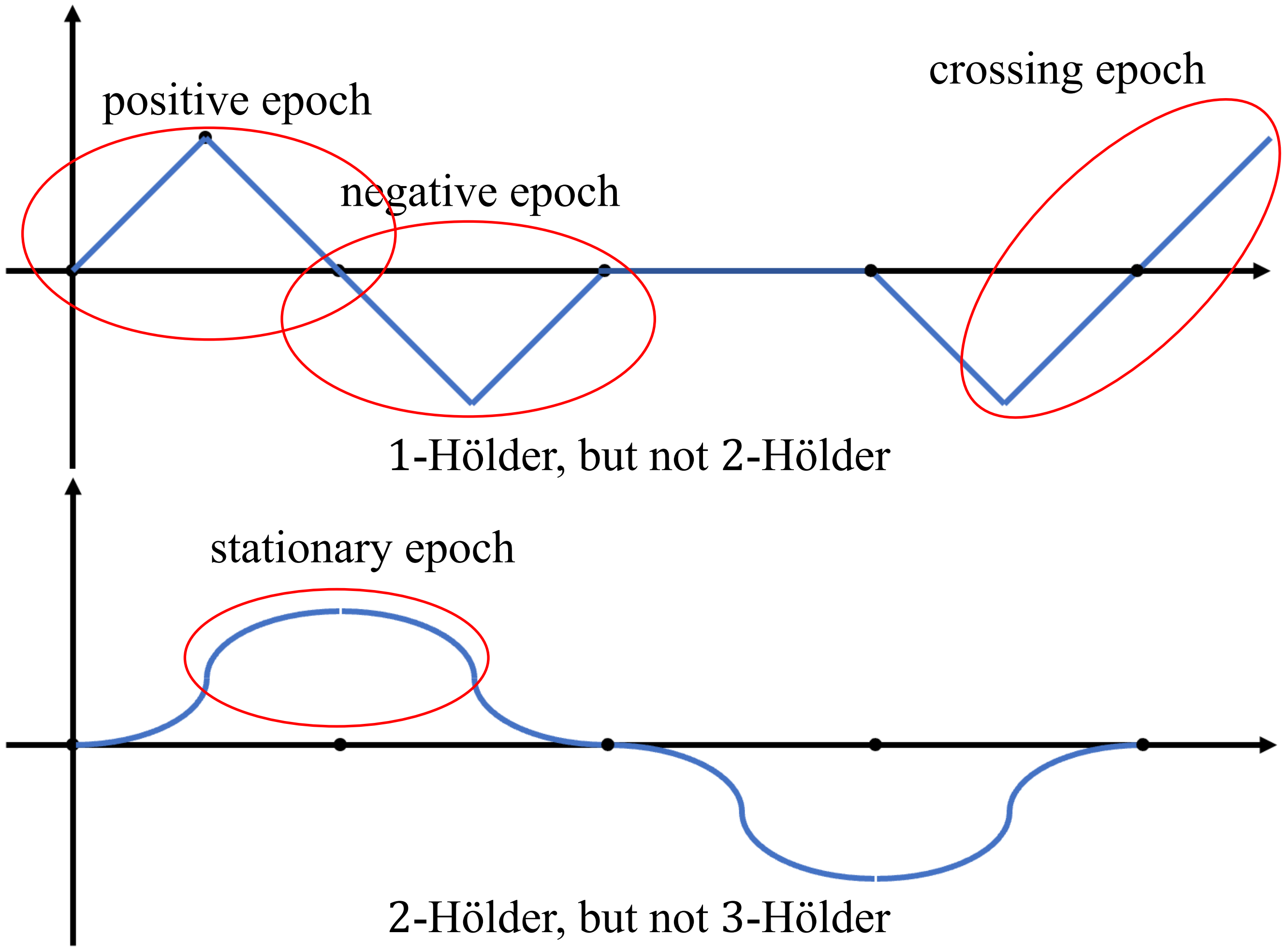
- First separation between the smooth ($\beta \geq 2$) and non-smooth ($\beta = 1$) regime
- Budgeted Exploration algorithm achieves $T^{3/5}$ upper bound for 2-Hölder reward function – power of exploiting smoothness



sinusoidal instances

Upper Bound

proof technique: amortization

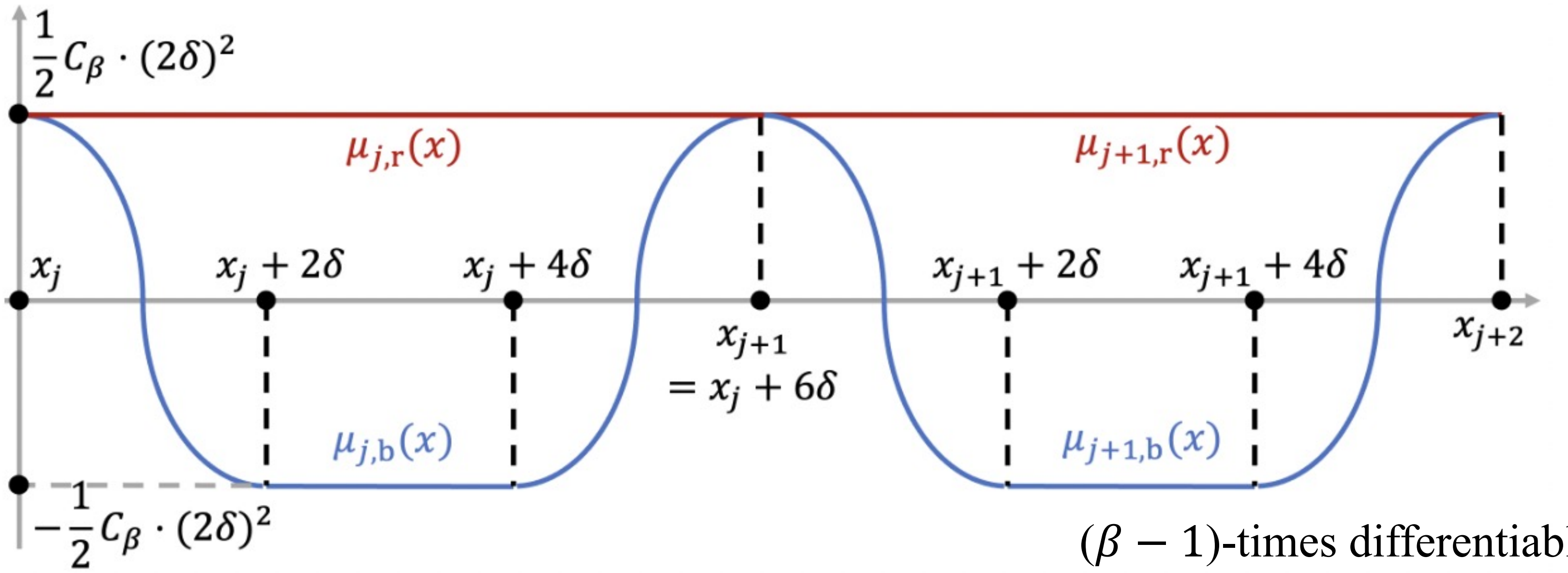


Main Results

- Smoothly-changing environment
  - Level of smoothness -- Hölder class!
  - [Besbes, Gur, Zeevi'14] admits an optimal $T^{2/3}$ regret for 1-Hölder (Lipschitz) reward function
  - Can we break this bound under smooth non-stationarity?
- Main results
  - First separation between the smooth and non-smooth regime
  - A $T^{3/5}$ upper bound for 2-Hölder reward function
  - Matching lower bound: every policy has worst regret $\Omega(T^{\frac{\beta+1}{2\beta+1}})$ for any β-Hölder reward function

Lower Bound

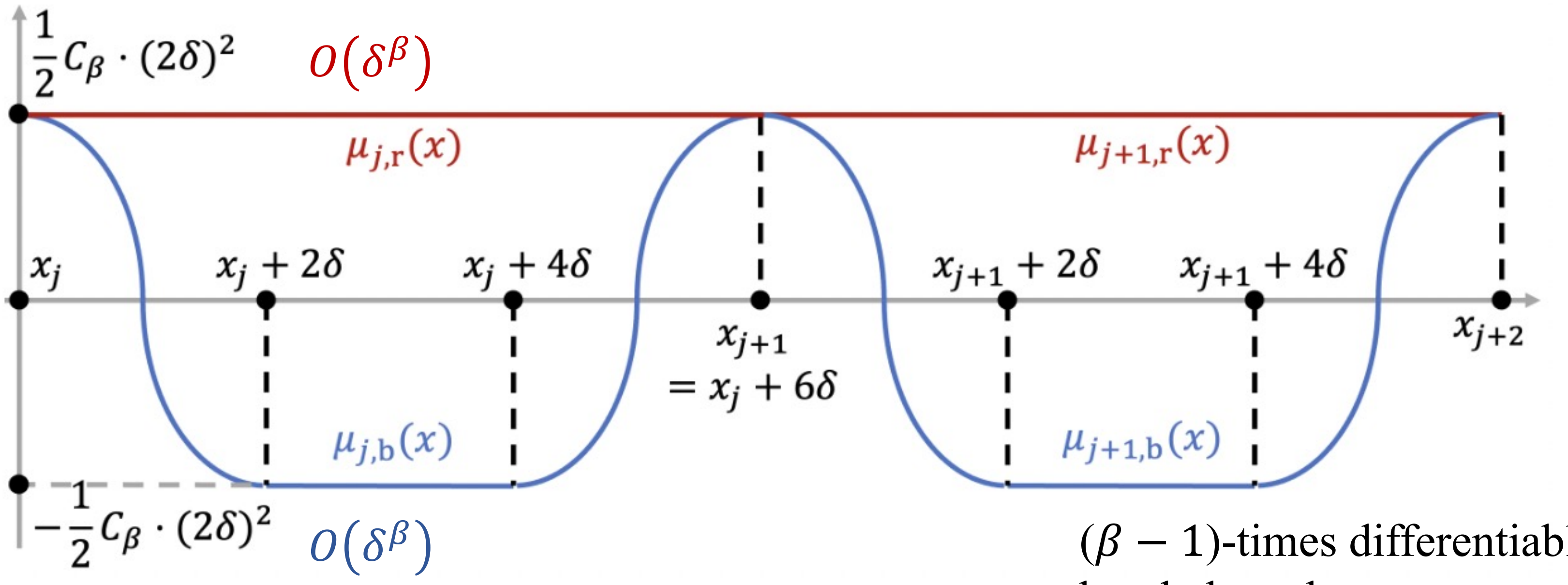
- Every policy has worst regret $\Omega(T^{\frac{\beta+1}{2\beta+1}})$ for β-Hölder reward function
- "Hard" instance for 2-Hölder reward function



$(\beta - 1)$-times differentiable
bowl-shaped curve

Lower Bound

- Every policy has worst regret $\Omega(T^{\frac{\beta+1}{2\beta+1}})$ for β-Hölder reward function
- “Hard” instance for 2-Hölder reward function

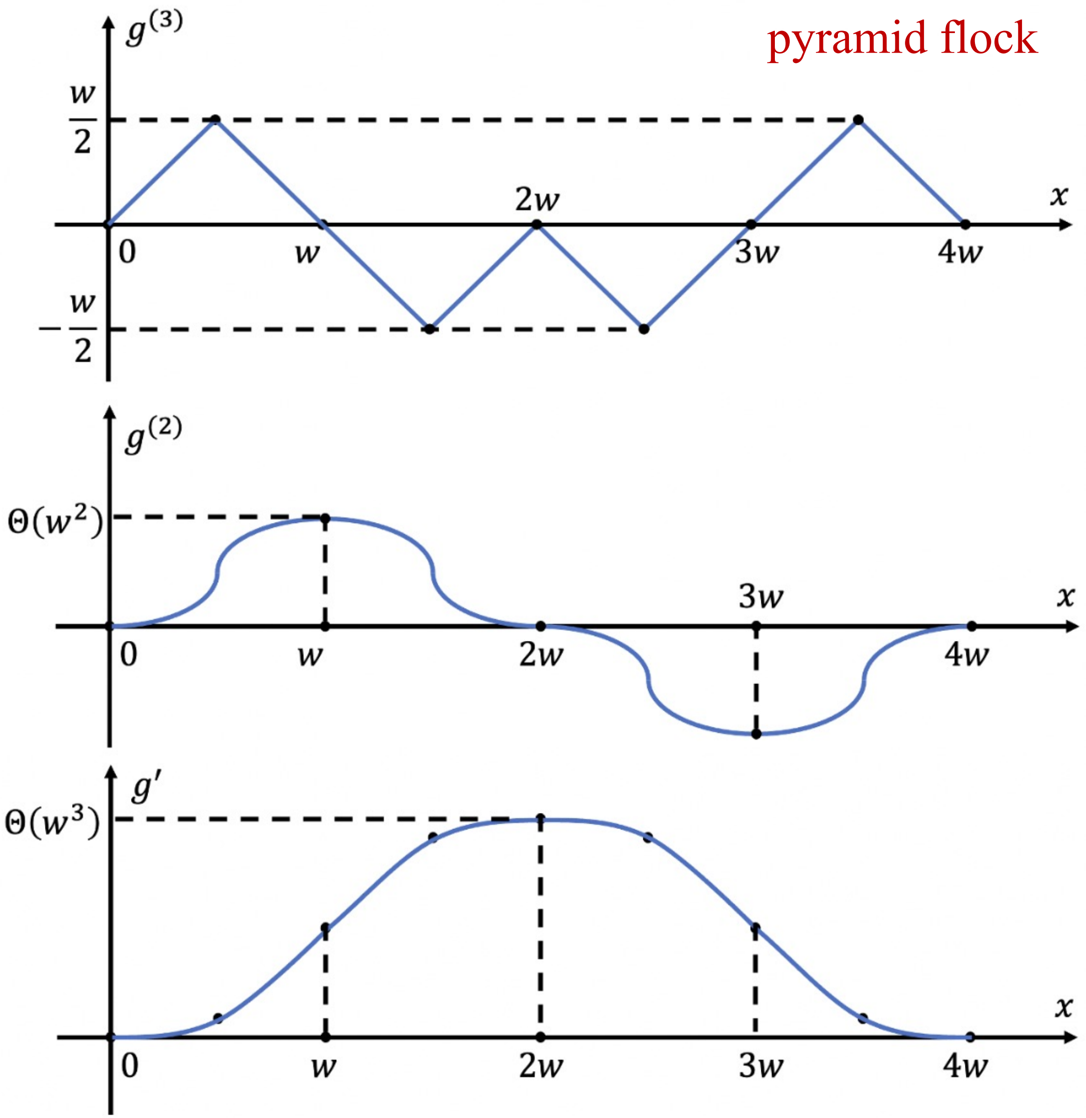


Likelihood to choose a wrong arm $\geq \frac{1}{2}$

$(\beta - 1)$-times differentiable bowl-shaped curve

Lower Bound

- Hard instance construction



Hard instance construction for $\beta = 4$