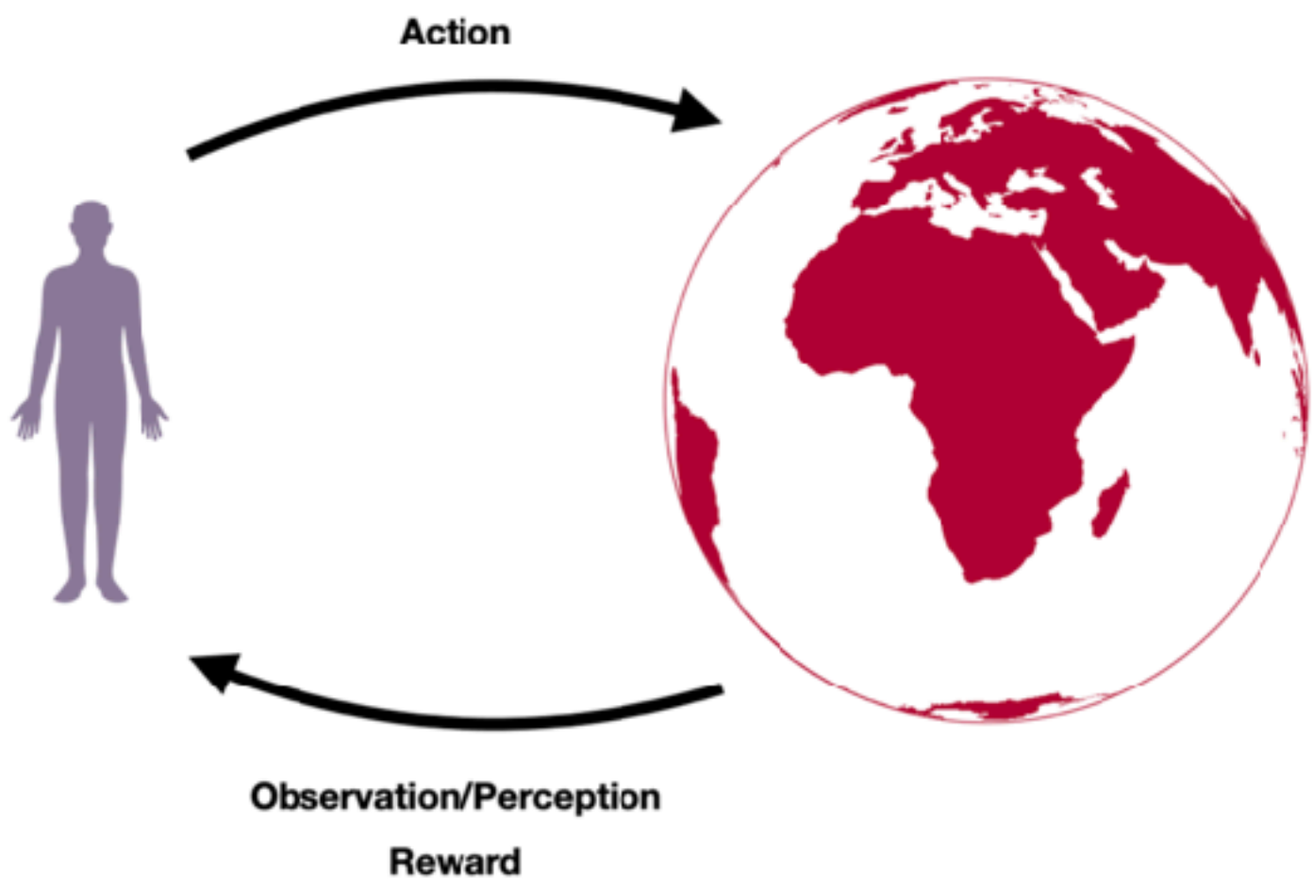
Continuous Bandits

Hédi HADIJI,
L2S, CentraleSupélec/Université Paris-Saclay

**Conference on Control and Inverse Problems,
2023, Monastir**

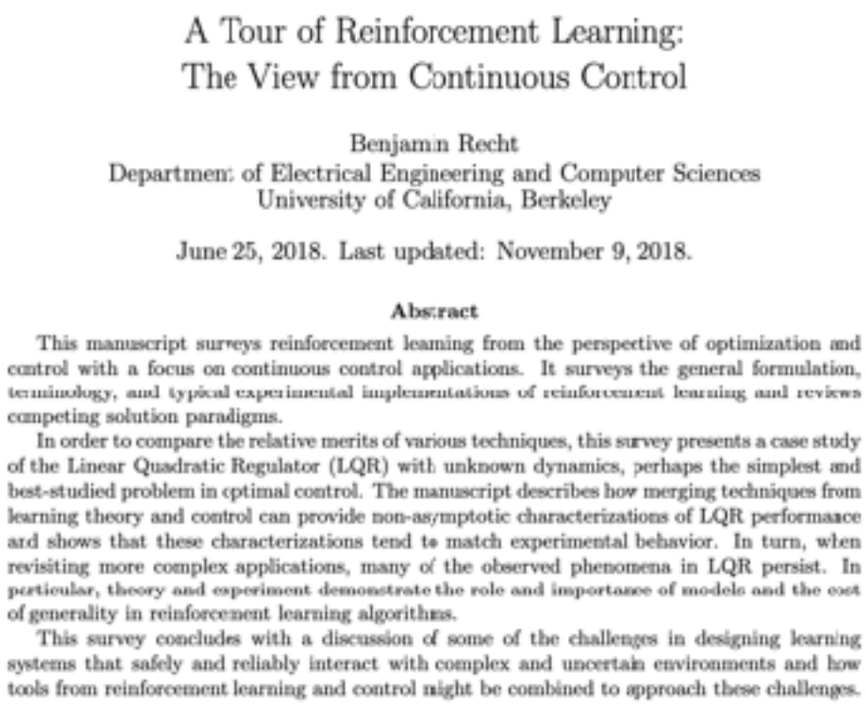
Reinforcement Learning

Agent interacts with
an **unknown environment**
and receives **rewards**



Bertsekas, *RL and Optimal Control*, 2019

Meyn, *Control Systems and RL*, 2022



A Tour of Reinforcement Learning:
The View from Continuous Control

Benjamin Recht
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

June 25, 2018. Last updated: November 9, 2018.

Abstract

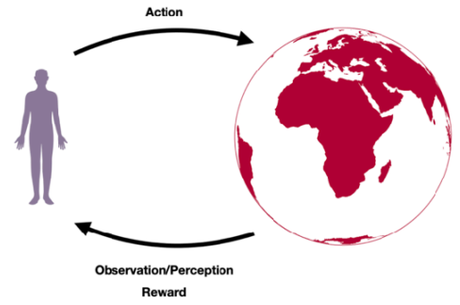
This manuscript surveys reinforcement learning from the perspective of optimization and control with a focus on continuous control applications. It surveys the general formulation, terminology, and typical experimental implementations of reinforcement learning and reviews competing solution paradigms.

In order to compare the relative merits of various techniques, this survey presents a case study of the Linear Quadratic Regulator (LQR) with unknown dynamics, perhaps the simplest and best-studied problem in optimal control. The manuscript describes how merging techniques from learning theory and control can provide non-asymptotic characterizations of LQR performance and shows that these characterizations tend to match experimental behavior. In turn, when revisiting more complex applications, many of the observed phenomena in LQR persist. In particular, theory and experiment demonstrate the role and importance of models and the cost of generality in reinforcement learning algorithms.

This survey concludes with a discussion of some of the challenges in designing learning systems that safely and reliably interact with complex and uncertain environments and how tools from reinforcement learning and control might be combined to approach these challenges.

Recht, *A Tour of RL*, 2018

Formalizing RL



Protocol: Agent and Environment

Input: State set $\mathcal{S}$, Action set $\mathcal{A}$
for $t = 1, \ldots, T, \ldots$ **do**
  Agent selects action $A_t \in \mathcal{A}$
  Environment updates states $S_{t+1} \in \mathcal{S}$, and reward $R_{t+1} \in \mathbb{R}$
  Environment sends observations S_{t+1}, R_{t+1} to agent
end

Assumptions: States and actions form a **Markov Decision Process**, i.e. distribution of next state only depends on current state and action.

Objective: Maximize total rewards $\sum_{t=1}^{T} r_t$

Two aspects of RL

Planning/Control

Actions can have long-term consequences that are hard to assess even when the dynamics are known

Trading-off Exploration and Exploitation

Only want to learn the aspects that are useful for getting large rewards.

Want to learn and get large rewards **in the same process**

Bandits focus exclusively on this aspect, by removing the state

Stochastic Bandits

aka stateless RL

(Multi-armed bandits are in fact way older than RL [Thompson '33])

Action space: $\mathcal{X}$ **Reward distributions:** ν_x **for** $x \in \mathcal{X}$

For $t = 1, \ldots, T, \ldots$:

- Player selects **action** $X_t \in \mathcal{X}$

- Player receives **reward** $Y_t \sim \nu_{X_t}$

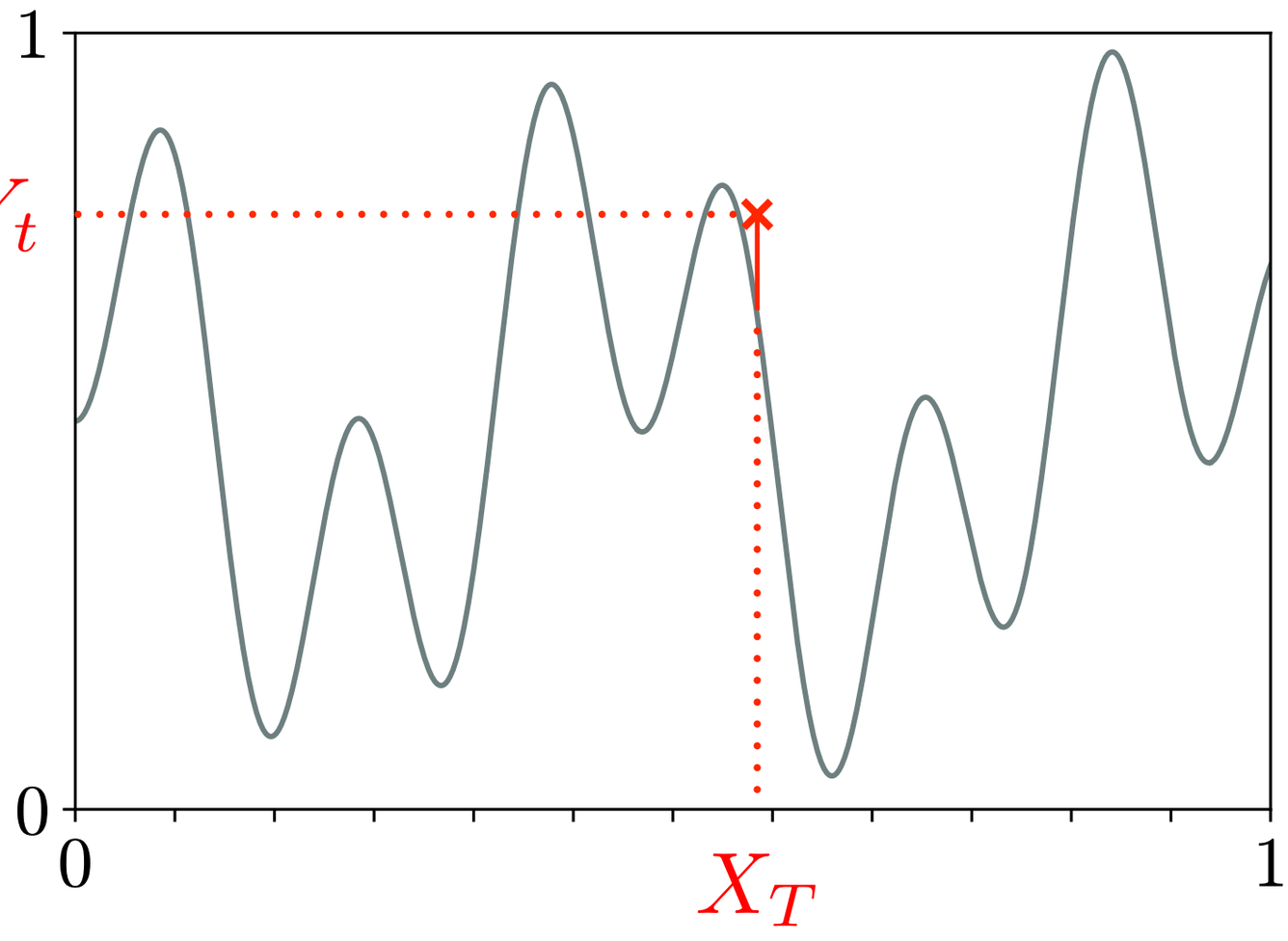
Goal: Maximize cumulative reward, or equivalently minimize **regret**

$$R_T = T \max_{x \in \mathcal{X}} \mathrm{E}[\nu_x] - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

Continuous Bandits

Arm space $\mathcal{X} = [0, 1]$

Unknown mean-payoff function $f \in [0, 1]^{\mathcal{X}}$



Aim : Minimize *regret* $R_T = T \max_{x \in \mathcal{X}} f(x) - \mathbb{E}\left[\sum_{t=1}^{T} f(X_t)\right]$

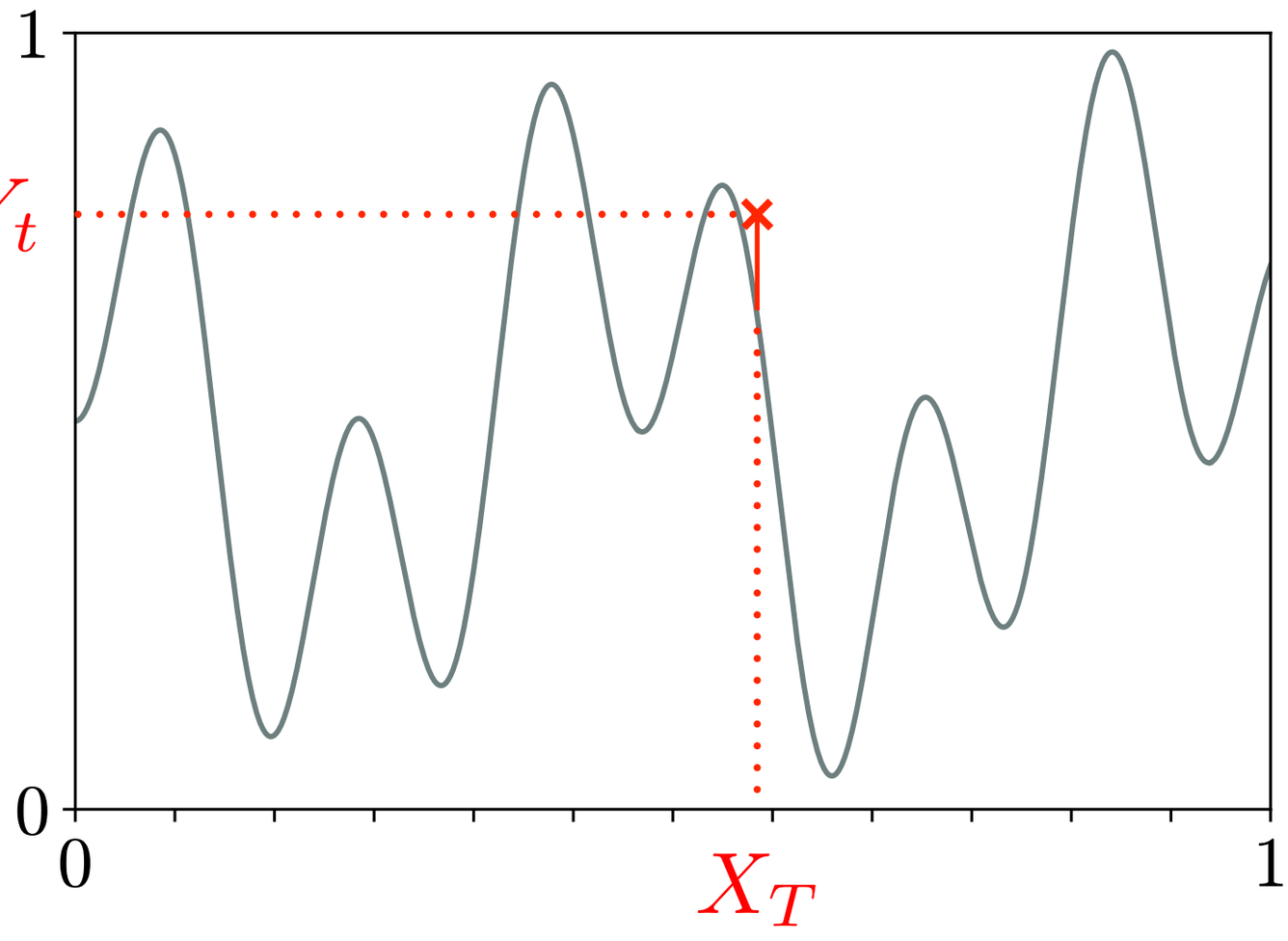
SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1926–1951, November 1995

016

THE CONTINUUM-ARMED BANDIT PROBLEM*

RAJEEV AGRAWAL†

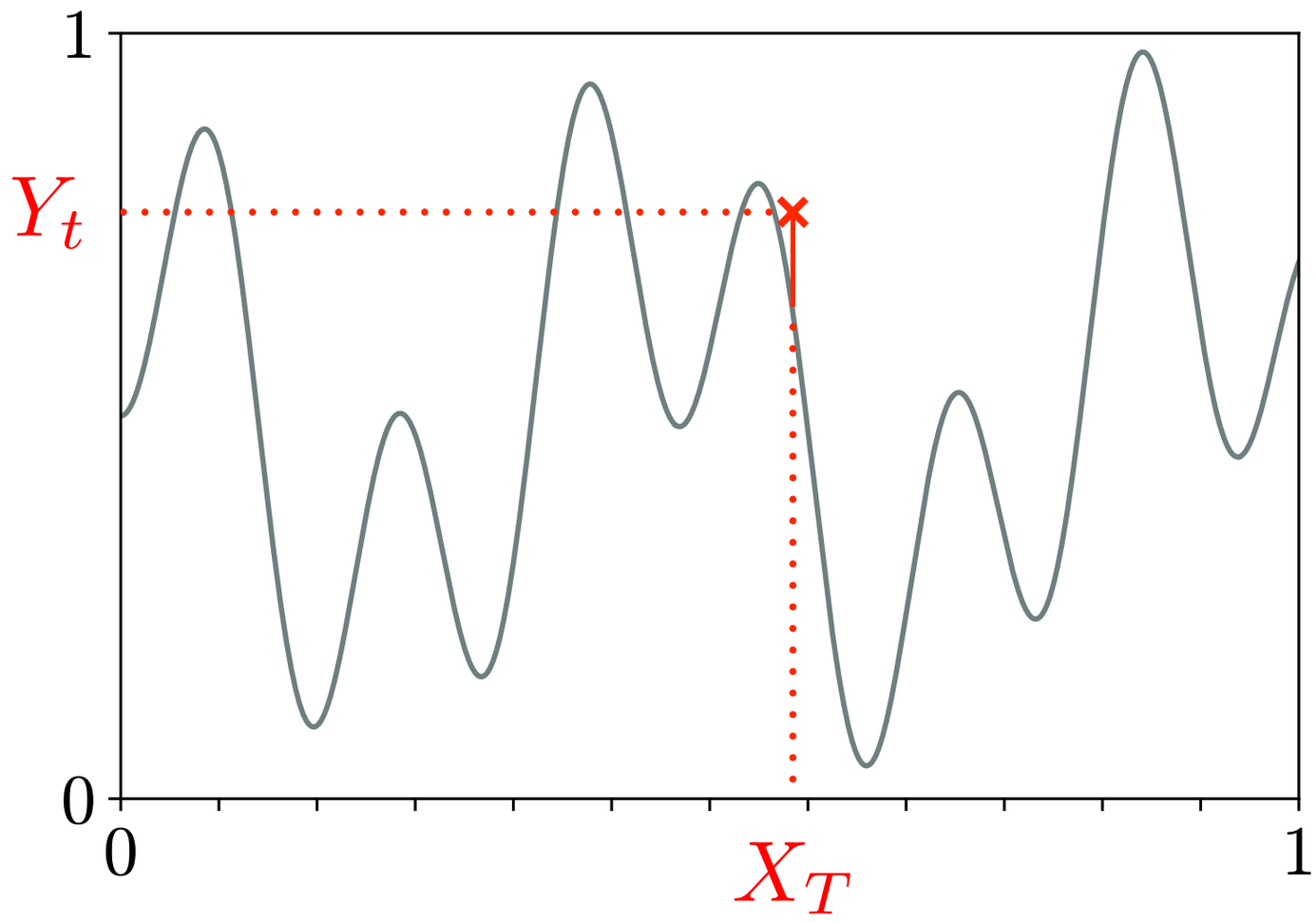
Arm space $\mathcal{X} = [0, 1]$ **Unknown mean-payoff function** $f \in [0, 1]^{\mathcal{X}}$



Aim : Minimize *regret* $R_T = T \max_{x \in \mathcal{X}} f(x) - \mathbb{E}\left[\sum_{t=1}^{T} f(X_t)\right]$

Continuous Bandits: Assumptions

We need some assumptions to make sure that the problem is learnable



Unknown reward function f is Hölder from its maximum:

$$\forall y \in [0,1], \quad |f(x^\star) - f(y)| \leqslant |x^\star - y|^\alpha$$

A Solution to Continuous Bandits

5. The adaptive control scheme. In this section we construct a class of certainty equivalence control with forcing-type adaptive control schemes based on the learning schemes constructed in §3. Let $\{\tau_i\}_{i=1}^{\infty}$ be a positive integer-valued sequence to be specified later. Define the related sequence $\{t_i\}^{\infty}$ as follows:

Prespecify an exploration sequence, scheduled in advance

Build estimate of unknown function and pick the optimum,
as if the estimate was the truth

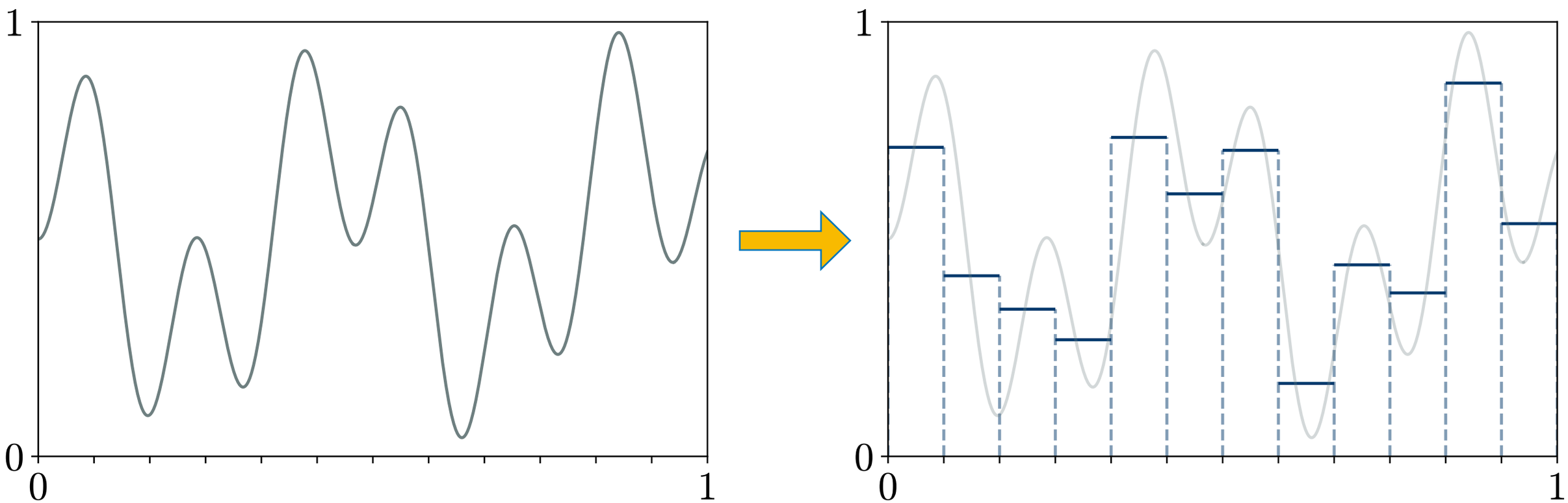
also called `explore-then-exploit' in modern bandit terminology

Issue: Best tuning depends on smoothness of unknown function

Better solution: Discretize

$$\forall y \in [0,1], \quad |f(x^\star) - f(y)| \leqslant |x^\star - y|^\alpha$$

DISCRETIZE $\mathcal{X}$ [Kleinberg '04] into K actions



Yields optimal rates for given Hölder class…
but still not adaptive to the smoothness

$$R_T \leqslant cT^{(\alpha+1)/(2\alpha+1)}$$

Impossibility of adaptation

Theorem : Full adaptation is impossible [Locatelli, Carpentier '18]

Let $\alpha \leqslant \gamma$ be two regularity values

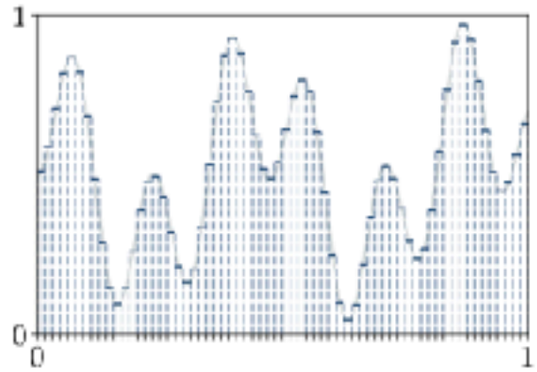
if $\max_{f\,\gamma-\text{Holder}} R_T \leqslant B$ **then** $\max_{f\,\alpha-\text{Holder}} R_T \geqslant cTB^{-\alpha/(1+\alpha)}$

if $B = T^{(\gamma+1)/(2\gamma+1)}$ then $\max_{f\ \alpha\text{-Hölder}} R_T \geqslant T^{\theta}$ with $\theta > \dfrac{\alpha+1}{2\alpha+1}$

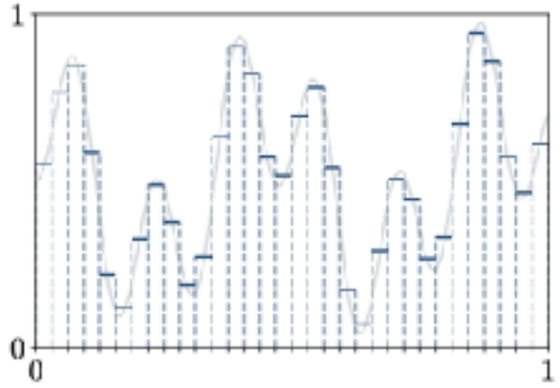
Impossible to guarantee the same results
as if we had known the smoothness in advance

An optimally adaptive algorithm

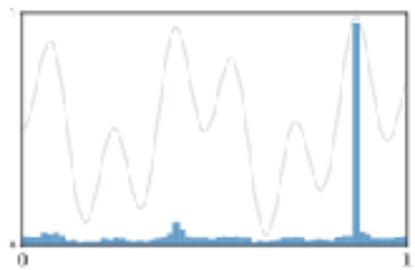
Assume the worst...
start with a very fine discretization
but not for too long



...then ZOOM OUT.
start over with a coarser discretization
but remember what you played before



K actions from the discretization



An extra-action: playing at random among actions selected in the past epoch

Reaches any of the optimal adaptive rates (depending on the initial grid)

Thank you!