universite PARIS-SACLAY

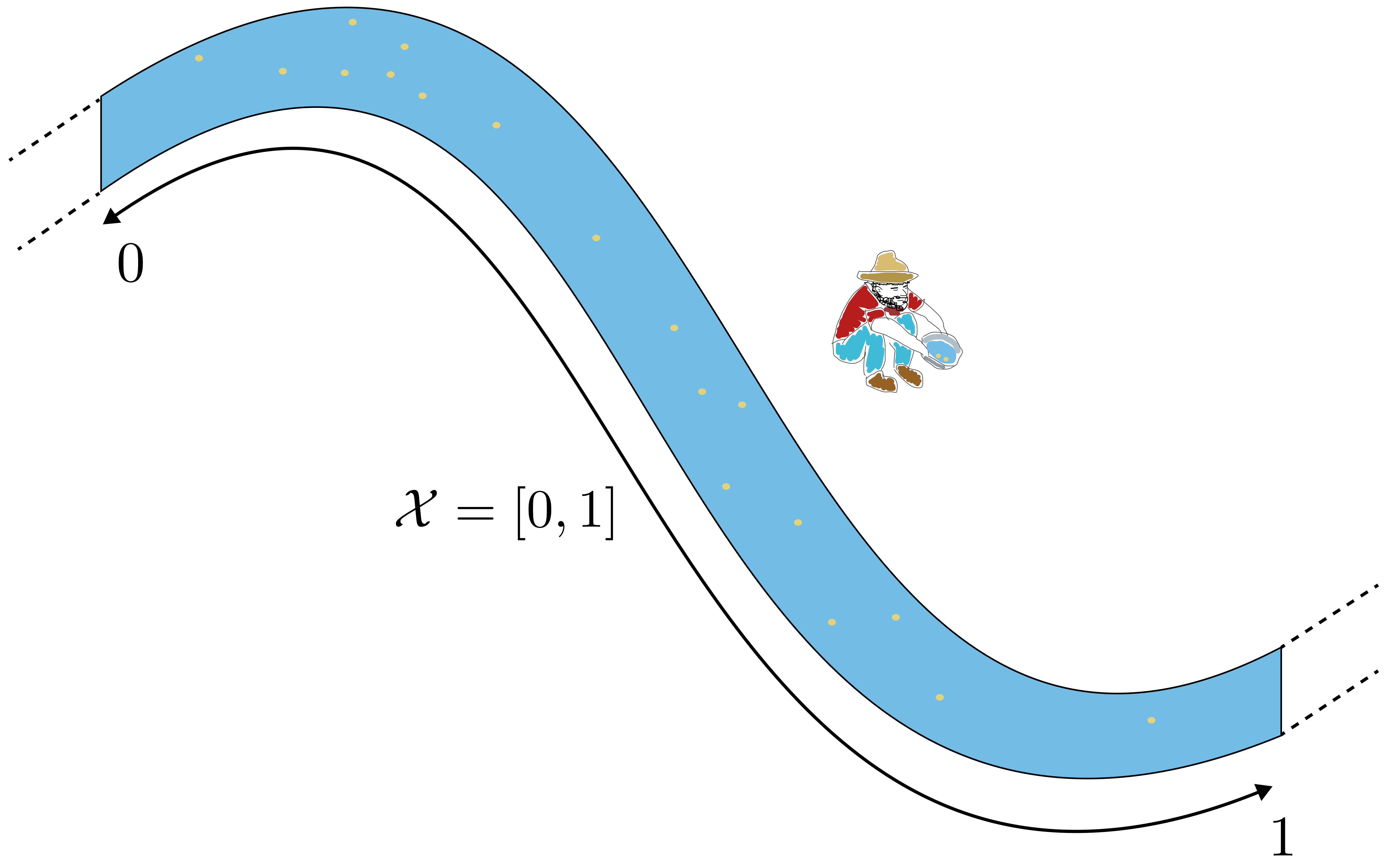
On some adaptivity questions in multi-armed bandits

Thèse de doctorat (virtually) at Orsay

Hédi Hadiji supervised by **Pascal Massart** and **Gilles Stoltz**

4th December 2020

Gold panning in a river



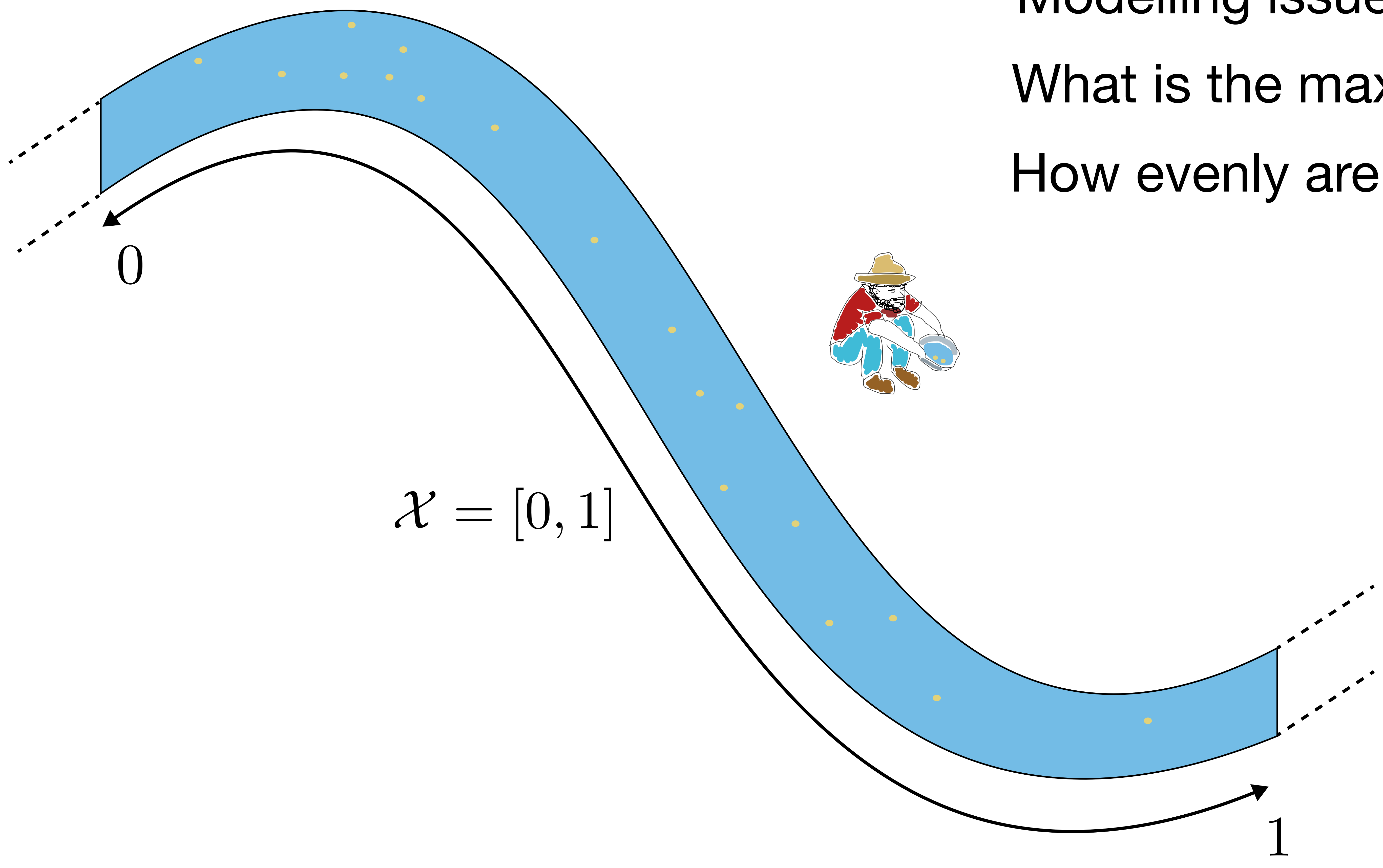
Every day $t = 1, \ldots,$

- Pick a spot $X_t \in [0, 1]$ on the river
- Get Y_t grams of gold

What strategy to
get as much gold as possible?

Making as few assumptions
as necessary

Gold panning in a river



Modelling issues:

What is the maximum reward one can get in a single round?

How evenly are the rewards distributed across space?

Should we just guess?

This will crucially affect our decisions!

what is the maximum reward I can get in a single round?

🤔

🤔

how evenly are the rewards distributed across space?

How does prior information about the rewards affect optimal strategies?

Outline

- I/ Multi-armed bandits
- II/ Adapting to the unknown range of the rewards
- III/ Continuous bandits and smoothness

Multi-armed bandits

[Thompson 1933] [Robbins 1952]

K probability distributions $(\nu_1, \ldots, \nu_K)$ unknown to the player $\mu_i = \mathrm{E}(\nu_i)$

At every time-step t the player:

- chooses an action A_t
- receives and observes reward $Y_t \sim \nu_{A_t}$ given A_t

Minimize **regret**

$$R_T = T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] = T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} \mu_{A_t}\Big]$$

Prior knowledge $=$ assumption on $\nu_1, \ldots, \nu_K$

Upper-Confidence Bounds

[Auer, Cesa-Bianchi, Fischer, '02]

$\nu_1, \ldots, \nu_K$ are supported in $[0,1] \Leftrightarrow$ all rewards are in $[0,1]$

Draw every arm once

Then draw arm A_t maximizing the index

$$U_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{2\log t}{N_a(t)}}$$

Empirical mean of rewards from arm a

Number of times a was picked

Distribution-free

$$R_T \leqslant c\sqrt{KT \ln T} + K$$

Distribution-dependent

$$R_T \leqslant \sum_{a=1}^{K} \frac{8}{\mu^\star - \mu_a} \log T + \sum_{a=1}^{K} (\mu^\star - \mu_a)$$

Improving UCB: two types of optimality

UCB explores a bit too much: we can reduce the confidence bounds and get better guarantees

MOSS [Audibert, Bubeck ’09]
[Degenne, Perchet ’16]

$$R_T \leqslant c\sqrt{KT \cancel{\ln T}} + K$$

KL-UCB [Cappé, Garivier, Maillard, Munos, Stoltz ’13]

$$R_T \leqslant \sum_{a=1}^{K} \frac{\mu^\star - \mu_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} \log T + \text{cst}$$

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) = \inf \left\{ \mathrm{KL}(\nu_a, \nu') \mid \mathrm{E}(\nu') > \mu^\star \text{ and } \nu'([0,1]) = 1 \right\}$$

None are further improvable [Auer, Cesa-Bianchi, Freund, Schapire ’02], [Lai & Robbins 1985]

[Garivier, Ménard ’17] obtain joint optimality in a parametric setting

Theorem: KL-UCB-switch [Garivier, H, Ménard, Stoltz ’18]

KL-UCB-switch is both distribution-dependent and distribution-free optimal

What if the range is unknown?

rewards are in $[m, M]$ instead of $[0, 1]$, and m and M are unknown to the player?

Unknown range: initial remarks

1. $$\frac{Y_t - m}{M - m} \in [0,1]$$

playing with rescaled rewards is equivalent
to standard game with regret scaled by (M-m)
… but we cannot rescale

2. (KL-)UCB, MOSS, etc. are scale-dependent

$$\widehat{\mu}_a(t) + (M - m)\sqrt{\frac{\log t}{N_a(t)}}$$

3. Not knowing the range is harder than knowing the range

$$\sup_{\text{Problems in } [m,M]} R_T \geqslant c\,(M - m)\sqrt{KT}$$

Can we match this lower bound without knowing m and M?

Unknown range: distribution-free adaptation

AdaHedge is from de [Rooij, van Erven, Grünwald, Koolen '13]

[Cesa-Bianchi, Mansour, Stoltz '07]

Theorem: Minimax range adaptation (H, Stoltz 2020)

AdaHedge for bandits with extra-exploration, guarantees

for all $m < M$, $\displaystyle \sup_{\text{Prob in } [m,M]} R_T \leqslant 7(M-m)\sqrt{TK\ln K} + 10(M-m)K\ln K$

Same guarantees as if we had known the range in advance!

But can we get distribution-dependent $\log T$ bounds?

Obstacle: extra-exploration forces at least $R_T \geqslant (M-m)\sqrt{KT}$

Adaptive rates

Distribution-free rate For all T, for all $M > 0$, $\sup_{\text{Problems in } [0,M]} R_T \leqslant M\,\Phi_{\text{free}}(T)$

e.g. Bandit AdaHedge enjoys $\Phi_{\text{free}}(T) = \sqrt{TK \ln K}$

Distribution-dependent rate For all $M > 0$, for all problems in $[0, M]$, $\limsup_{T\to\infty} \frac{R_T}{\Phi_{\text{dep}}(T)} < +\infty$

e.g. $\widetilde{U}_a(t) = \widehat{\mu}_a(t) + (\log t)\sqrt{\frac{\log t}{N_a(t)}}$ enjoys $\Phi_{\text{dep}}(T) = (\log T)^2$ [Lattimore '17]

Can we get $\log T$ and $\sqrt{KT}$ simultaneously?

Lower bound for adaptation to the range

Theorem: Lower bound for range adaptation (H, Stoltz 2020)

For any algorithm enjoying rates $\Phi_{\mathrm{free}}(T)$ and $\Phi_{\mathrm{dep}}(T)$

$$\Phi_{\mathrm{dep}}(T)\,\Phi_{\mathrm{free}}(T) \gtrsim T$$

in particular, $\Phi_{\mathrm{free}}(T) \leqslant \mathcal{O}(\sqrt{T}) \;\Rightarrow\; \Phi_{\mathrm{dep}}(T) \geqslant \Omega(\sqrt{T})$

The cost of distribution-free adaptation is hidden in distribution-dependent rates!

We get T^{α} and $T^{1-\alpha}$ (for $1/2 < \alpha < 1$) by tuning the extra-exploration in Bandit AdaHedge

Adapting to the smoothness

Back to the river: continuous bandits

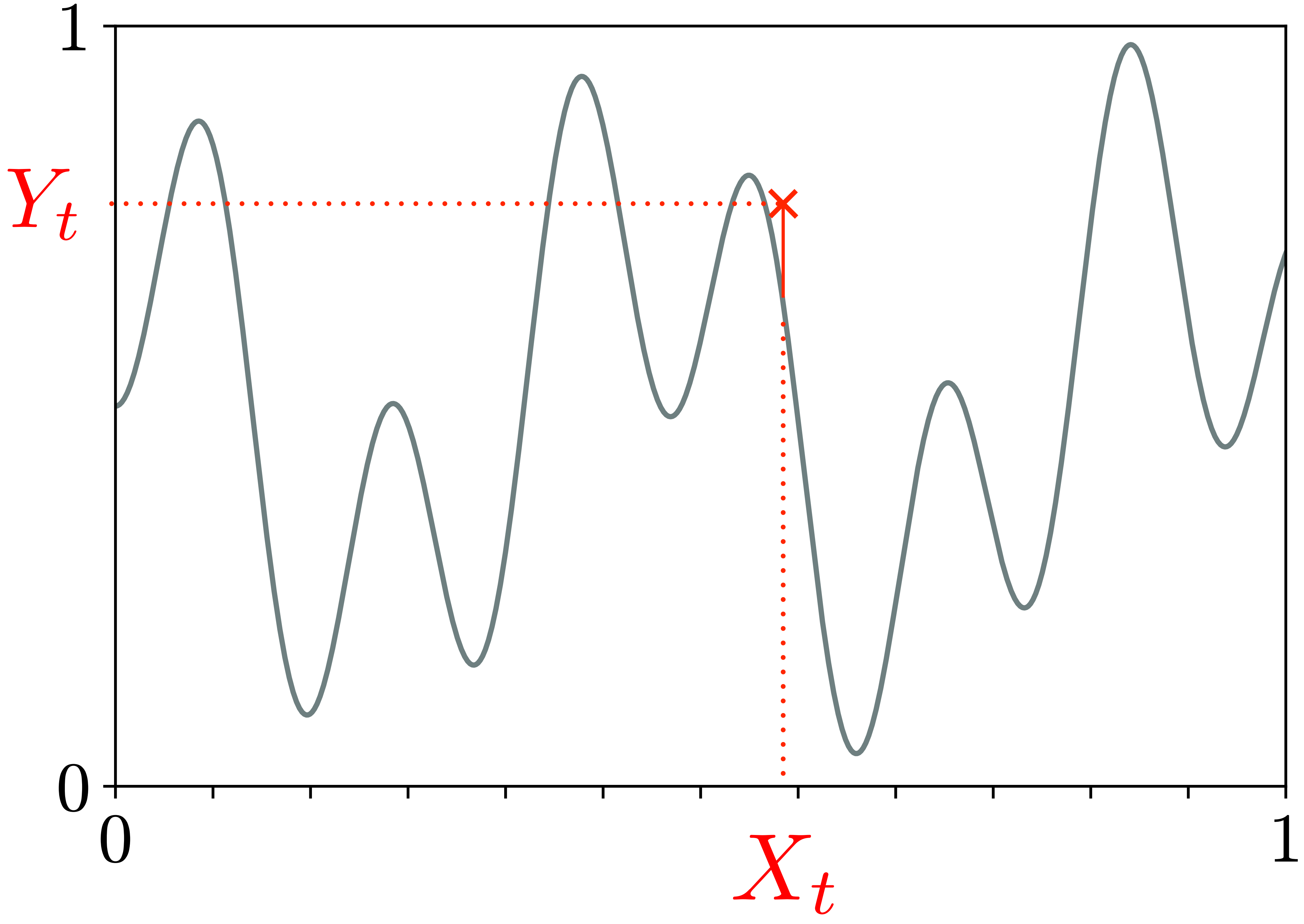
$\mathcal{X} = [0,1]$ Unknown payoff distributions $(\nu_x\,,\, x \in \mathcal{X})$ Mean-payoff function $f : x \mapsto \mathrm{E}(\nu_x)$

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

- pick $X_t \in \mathcal{X}$
- receive and observe $Y_t \sim \nu_{X_t}$ given X_t

Goal

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



We need assumptions

Assumptions

1. Rewards are bounded in $[0,1]$
2. Mean-payoff function is "(L,α)-Hölder around its max" $\quad f(x^\star) - f(x) \leqslant L\,|x^\star - x|^\alpha$

Lots of similar and/or more refined assumptions in the literature

[Agrawal 1995], [Kleinberg '05], [Auer, Ortner, Szepesvári, '07], [Bubeck, Munos, Stoltz, Szepesvári '11]
[Bubeck, Stoltz, Yu '11], [Kleinberg, Slivkins, Upfal, '13], [Bull '15]

Related problems: bandit optimization/simple regret, maximum estimation, maximum location

find $\widehat{X}$
s.t. $f(\widehat{X}) \approx \max f$

[Valko, Carpentier, Munos '13]
[Bartlett, Gabillon, Valko '19] [Shang, Kauffman, Valko '19]

estimate $\max f$

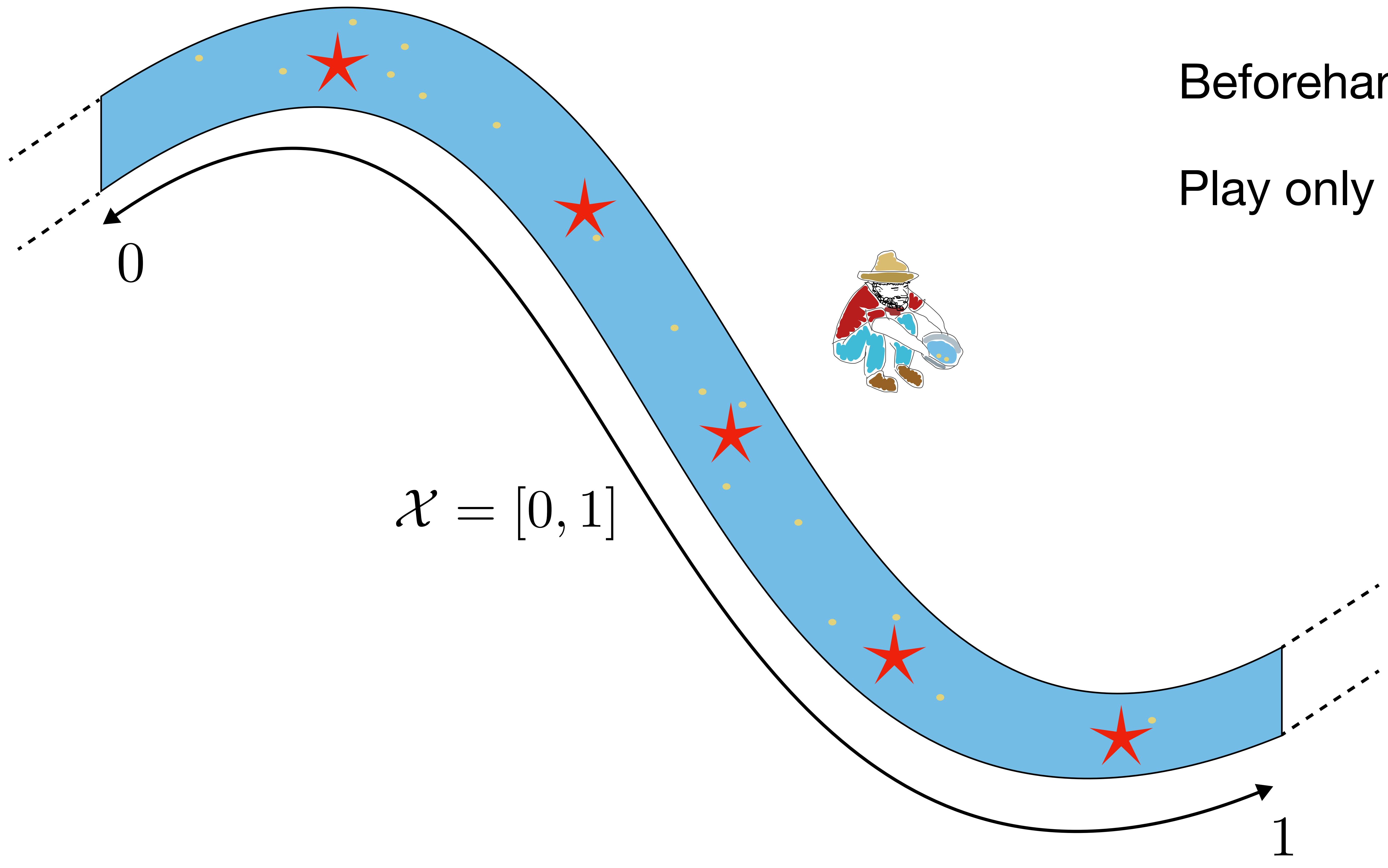
[Lepski 1994]

find $\widehat{X}$
s.t. $\widehat{X} \approx x^\star$

[Muller 1989]

Discretization

[Kleinberg '05]

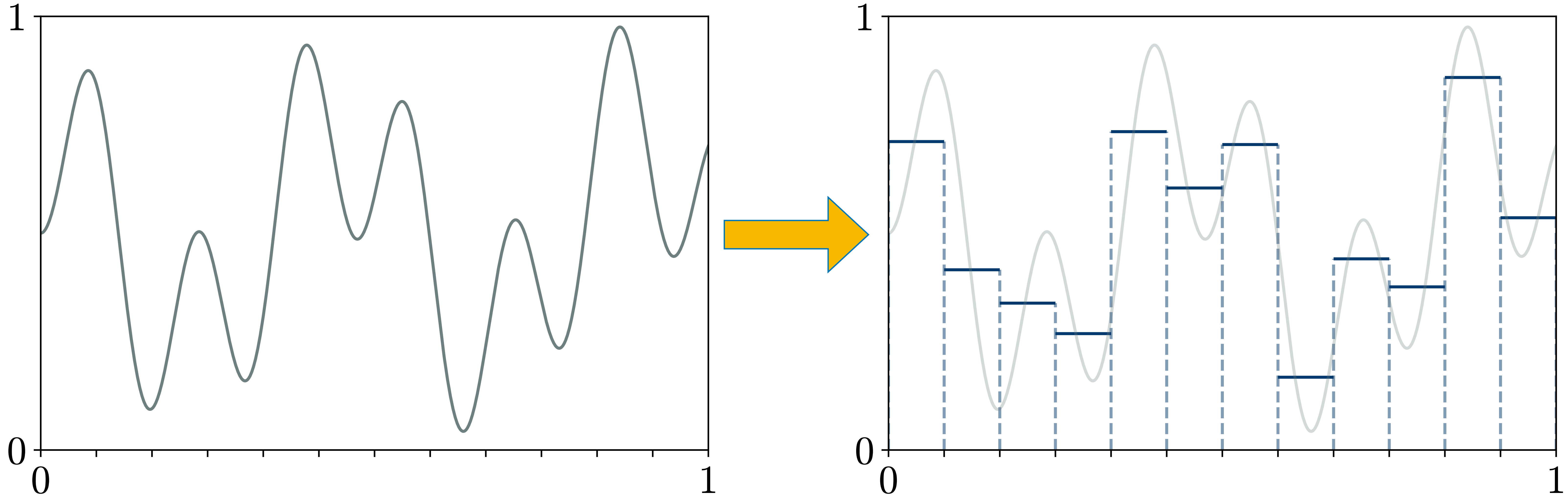


Beforehand, pick a finite number of spots ★

Play only at these spots, using a K-armed bandit algorithm

Back to a finite-armed bandit problem

Bounding the regret of discretization



$$R_T \leqslant c\, L^{1/(2\alpha+1)} T^{(\alpha+1)/(2\alpha+1)} \quad \text{by tuning } K = K^\star(\alpha) \text{ appropriately}$$

Proof:

$$\begin{aligned} R_T &= T\left(\max f - \max_{1\leqslant i\leqslant K} f(x_i)\right) + \max_{1\leqslant i\leqslant K} f(x_i) - \mathbb{E}\left[\sum_{t=1}^{T} f(X_t)\right] \\ &\leqslant \qquad\qquad T\,\frac{L}{K^\alpha} \qquad\qquad + \qquad\qquad c\sqrt{KT} \end{aligned}$$

Impossibility of (full) adaptation

Adaptive rates

An algorithm achieves adaptive rates $\theta : \mathbb{R}_+ \to [1/2,\, 1]$ if

$$\sup_{f\ \alpha\text{-Hölder}} R_T \leqslant c\, T^{\theta(\alpha)} \text{ for all } \alpha$$

Question: how can we obtain adaptive rates $\theta(\alpha) = \dfrac{\alpha+1}{2\alpha+1}$?

Model selection? Cross-validation?

… Exploration is costly

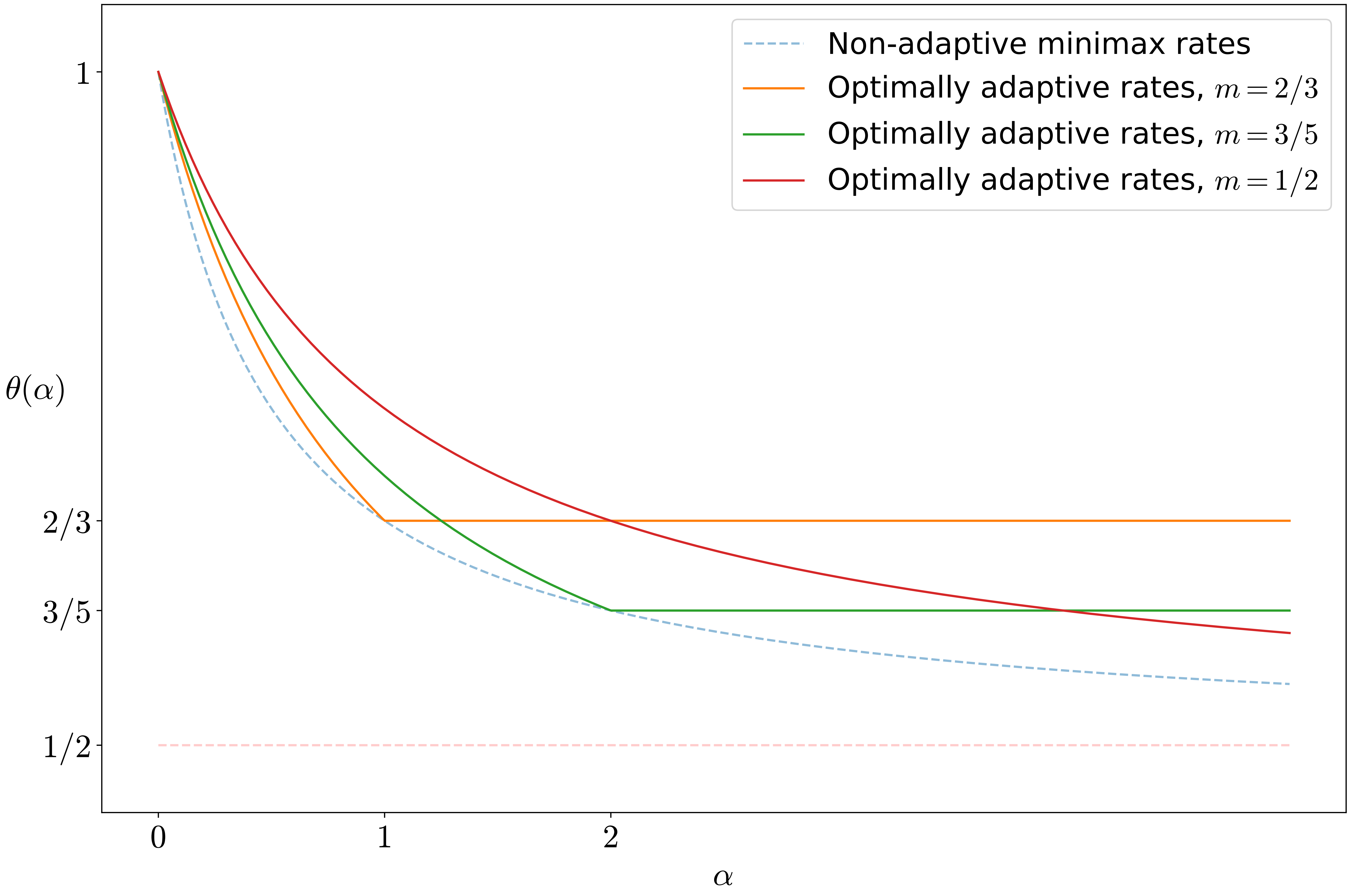
Answer: we cannot

Theorem: Consequence of [Locatelli and Carpentier '18]

If a rate function θ is achieved by some algorithm, then

$$\text{there exists } m \in [1/2,\, 1\,] \text{ s.t} \quad \theta(\alpha) \geqslant \max\left(m\,, 1 - m\frac{\alpha}{\alpha+1}\right)$$

Lower bound(s) on the adaptive rates



if $R_T \leqslant c\, T^{\theta(\alpha)}$
then θ is lower bounded
by one of these rates

Still, can we reach these rates?

A bird’s eye view of past approaches

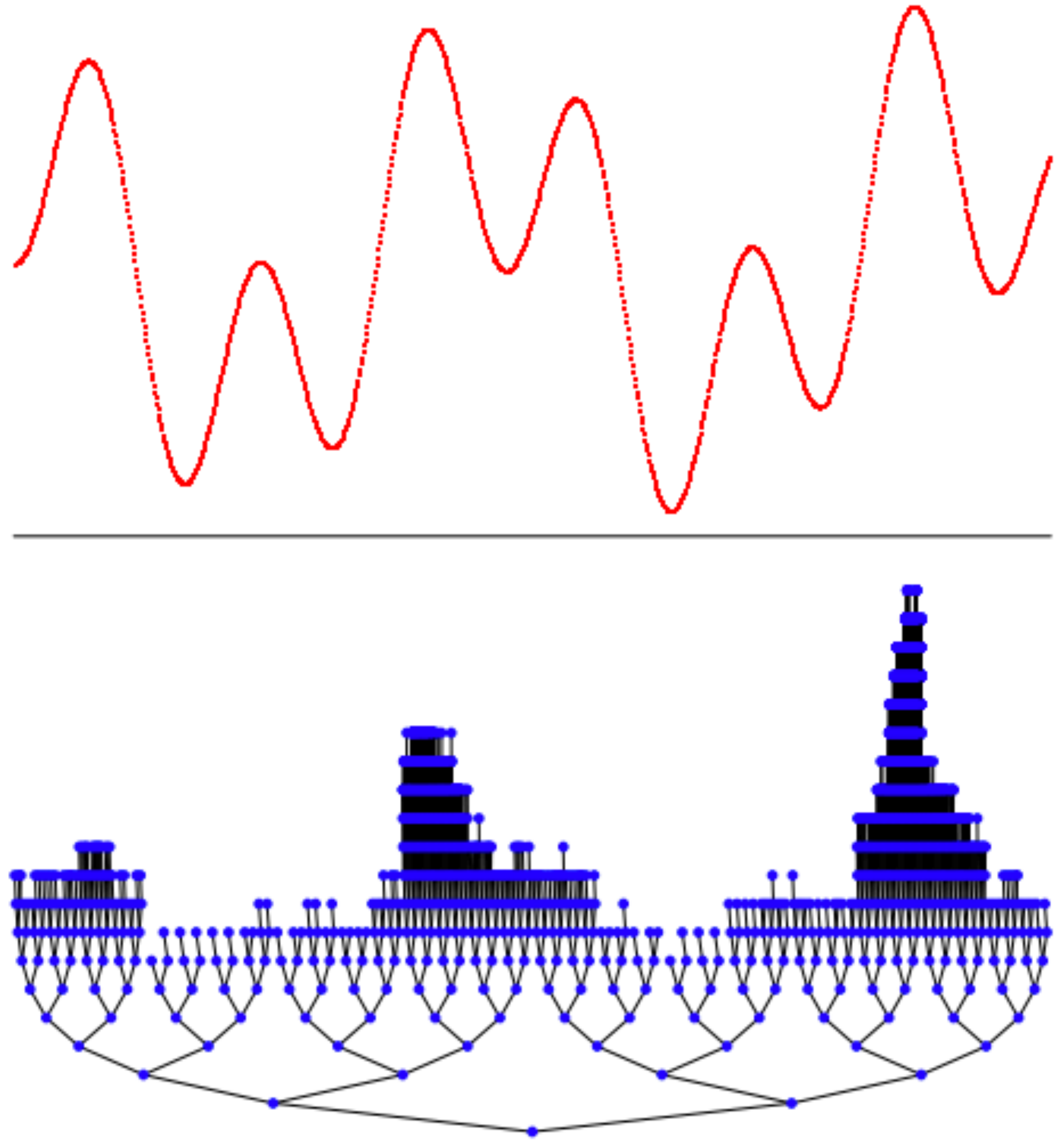
HOO [Bubeck, Munos, Stoltz, Szepesvári ’11]

Zooming algorithm [Kleinberg, Upfal, Slivkins ’13]

Adaptive-treed bandits [Bull ’15]

SR [Locatelli, Carpentier ’18]

e.g.



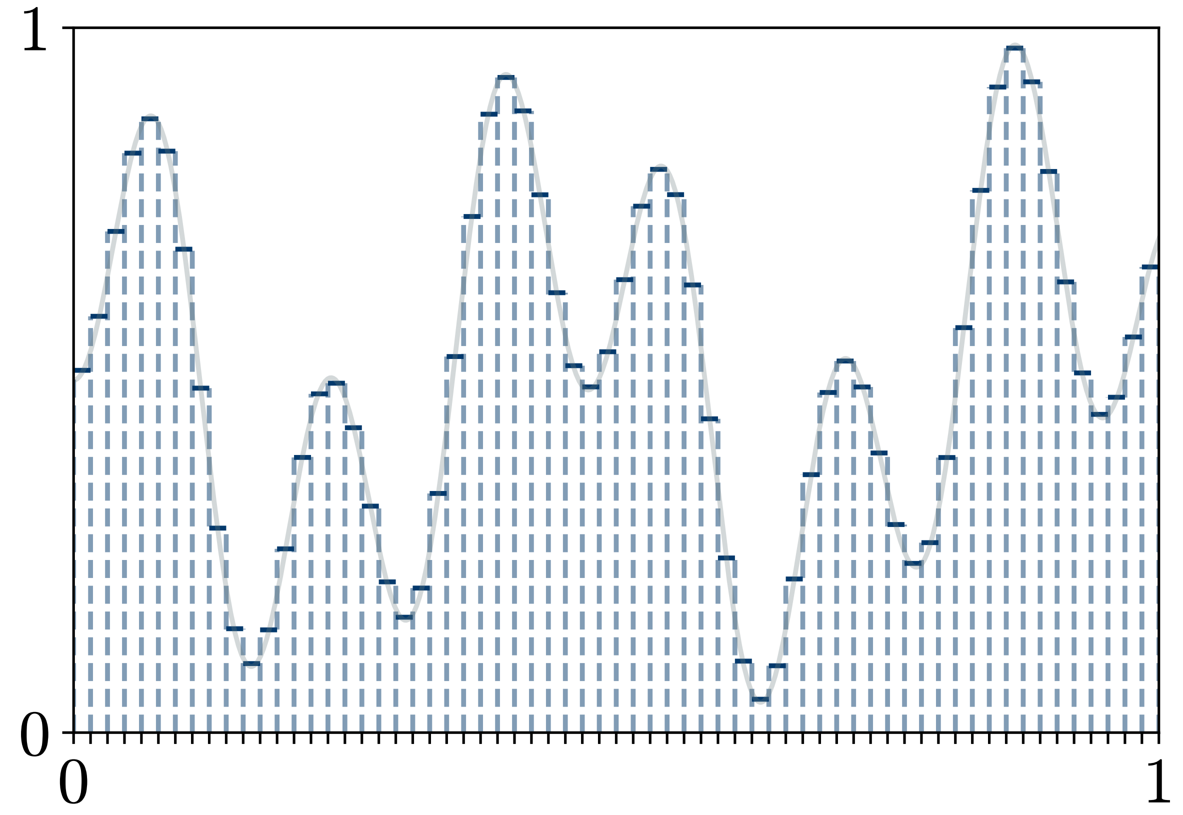
Zoom in on promising regions

Select the promising regions **using the knowledge of the regularity**

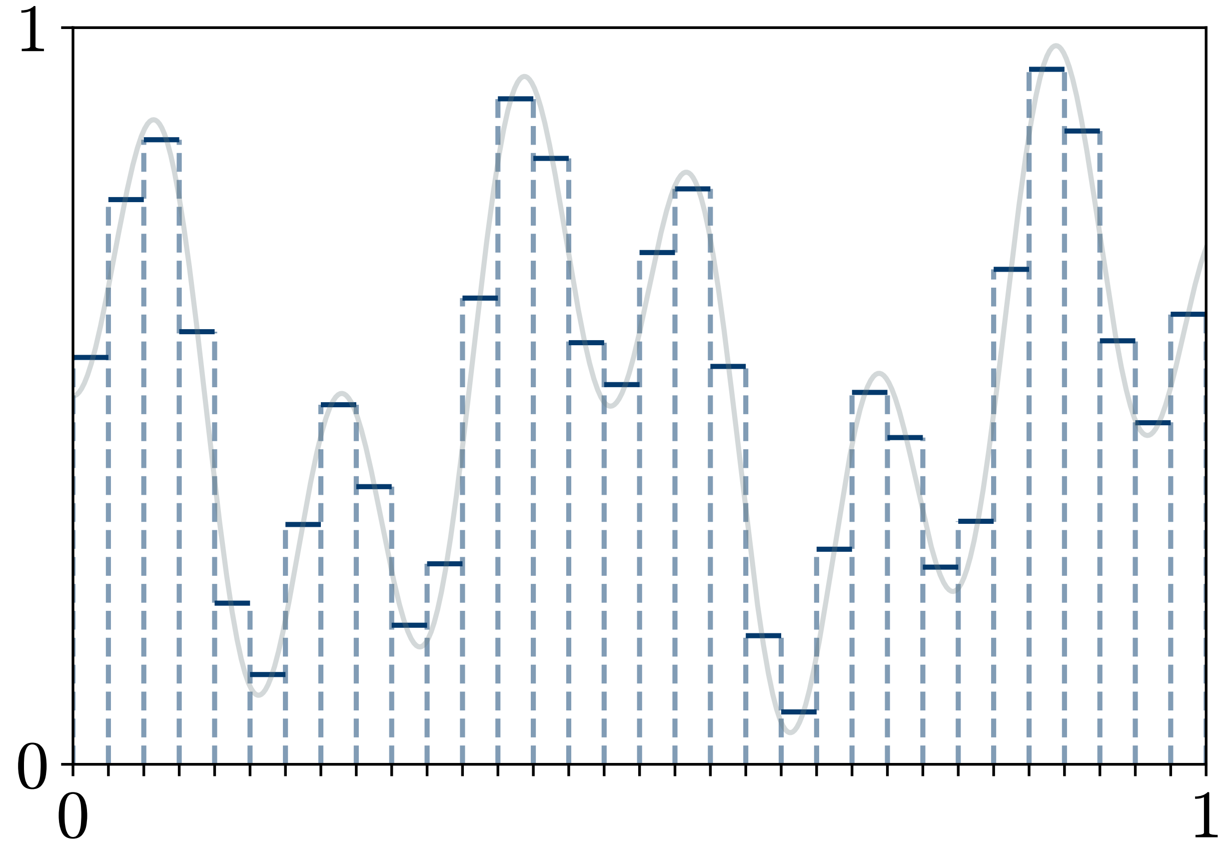
An optimally adaptive algorithm

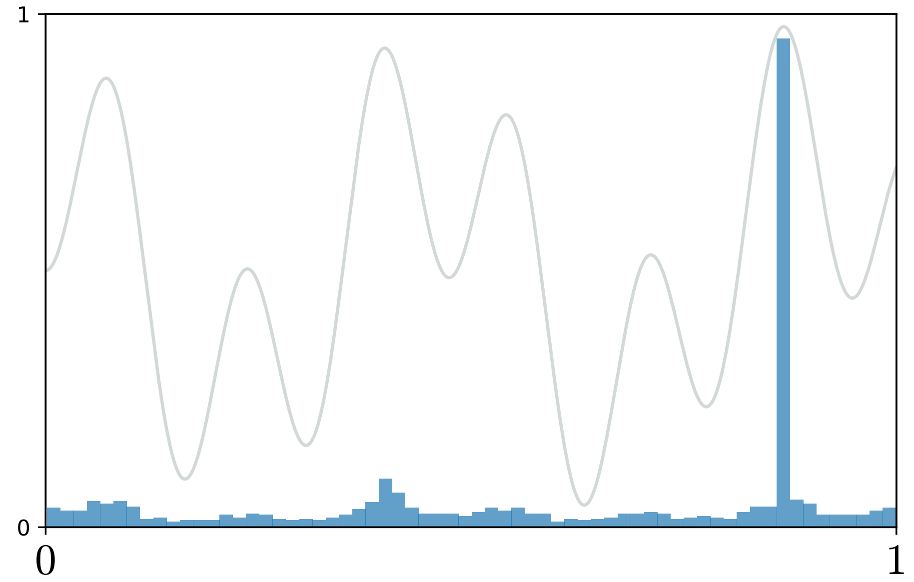
Basic idea

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





Discrete algorithm choosing between: K actions from the discretization

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

Guarantees

Algorithm : Memorize, Discretize, Zoom out

Set $K_i \approx 2^{-i}\sqrt{T}$; $D_i \approx 2^i\sqrt{T}$

For epochs 1 to $\approx \log T$

– For D_i rounds, run a K_i-discretization **with memory of previous plays**

Theorem: Medzo regret bound [H, 2019]

For any (L, α)-Hölder function, without the knowledge of L or α,

$$R_T \leqslant \widetilde{\mathcal{O}}\left(L^{1/(\alpha+1)} T^{(\alpha+2)/(2\alpha+2)}\right)$$

with no assumption on α and L

Conclusion

Adaptation leads to interesting questions with surprising phenomena

Achieving adaptation requires new algorithmic ideas

Ultimately important in practice/getting rid of hyperparameter tuning

Bandits are fun!

Local perspectives

KL-UCB: Unifying notion of optimality in bandits?

Range:
- Minimax and optimal distribution-dependent rates for adapting to the lower end of the range
- Getting rid of the $\sqrt{\log K}$

Continuous bandits:
- Including other types of regularity
- Is there a principled look at Medzo that could be applied elsewhere?
- What do we *need* to know to be able adapt at the usual rates [Locatelli Carpentier ’18]?

General perpectives

- What about non IID payoffs? [Zimmert and Seldin '19]

- Related problem 'Model selection in contextual multi-armed bandits'
[Foster et al. '19], [Chatterji et al '19], [Foster et al. COLT open problem '20]

Observe context $C_t \in \{1, \dots, S\}$ then choose $A_t \in \{1, \dots, K\}$

A policy is mapping from context to action $\pi : \mathcal{C} \to \mathcal{A}$ A model Π is a set of policies

Easy (Exp4): $R_T(\text{Best } \pi \in \Pi) \leqslant c\sqrt{KT \log |\Pi|}$

Difficult: getting the same result with a sequence of nested models

-Dream general approach:
Can we combine a family of bandit algorithms and obtain one that is as good as the best?

[Agrawal et al. '17], [Pacchiano et al. '20], results are very specific

General perpectives

- What about non IID payoffs? [Zimmert and Seldin '19]

- Related problem 'Model selection in contextual multi-armed bandits'
[Foster et al. '19], [Chatterji et al '19], [Foster et al. COLT open problem '20]

Observe context $C_t \in \{1, \dots, S\}$ then choose $A_t \in \{1, \dots, K\}$

A policy is mapping from context to action $\pi : \mathcal{C} \to \mathcal{A}$ A model Π is a set of policies

Easy (Exp4): $R_T(\text{Best } \pi \in \Pi) \leqslant c\sqrt{KT \log |\Pi|}$

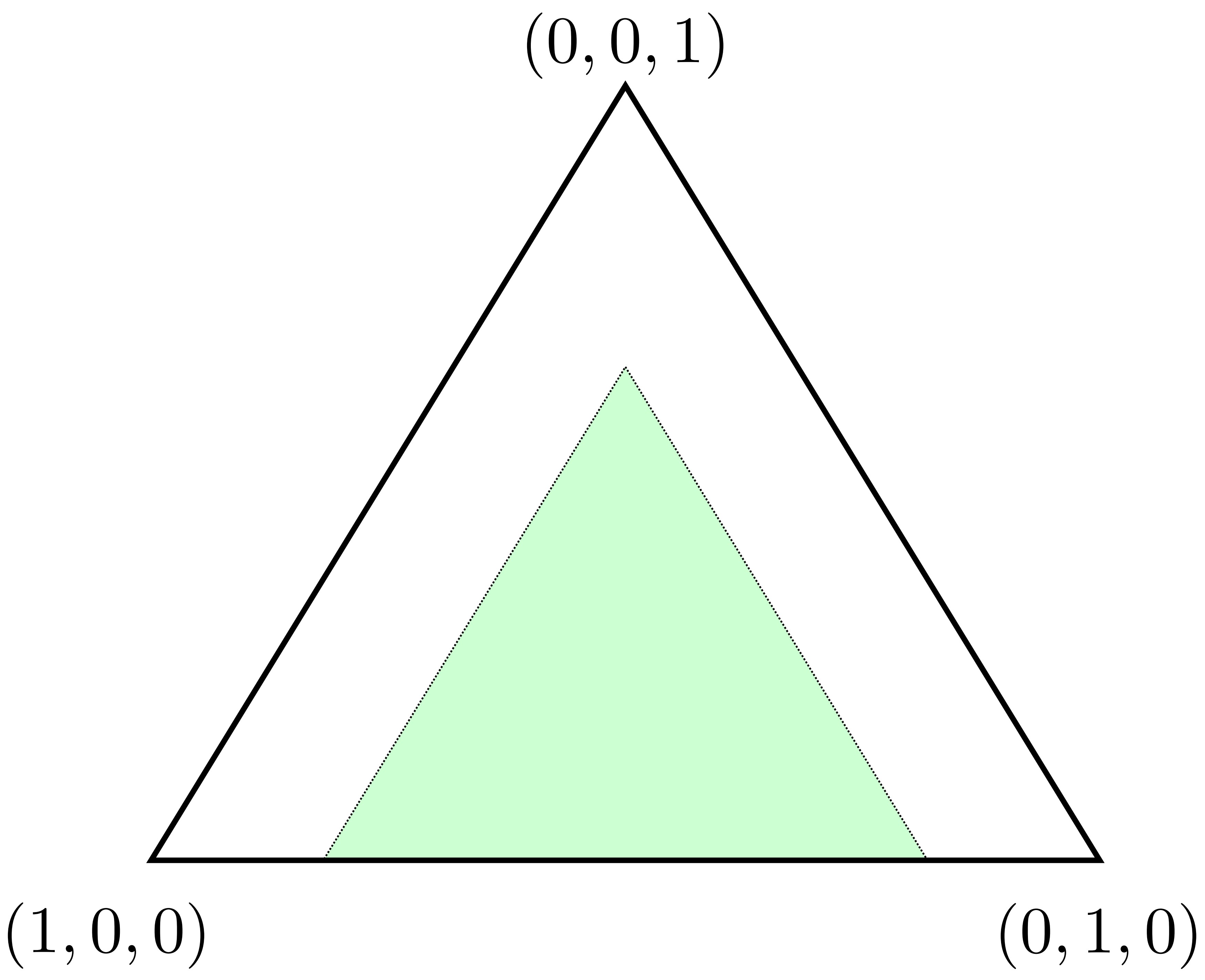
Difficult: getting the same result with a sequence of nested models

-Dream general approach:
Can we combine a family of bandit algorithms and obtain one that is as good as the best?

[Agrawal et al. '17], [Pacchiano et al. '20], results are very specific

Bonus: diversity-preserving bandits

with Sébastien Gerchinovitz, Jean-Michel Loubes and Gilles Stoltz



Play a probability distribution p_t over $\{1,2,3\}$

Observe $A_t \sim p_t$, and $Y_t \sim \nu_{A_t}$

Require that $p_{1,t},\ p_{2,t} \geqslant 0.3$ (for example)

$$R_T = T \max_{p \text{ available}} \sum_{a=1}^{K} p_a \mu_a - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

Bounded regret is possible $\Leftrightarrow$ the best p is in the (relative) interior of the simplex

Thank you Gilles and Pascal!

Thank you everyone!