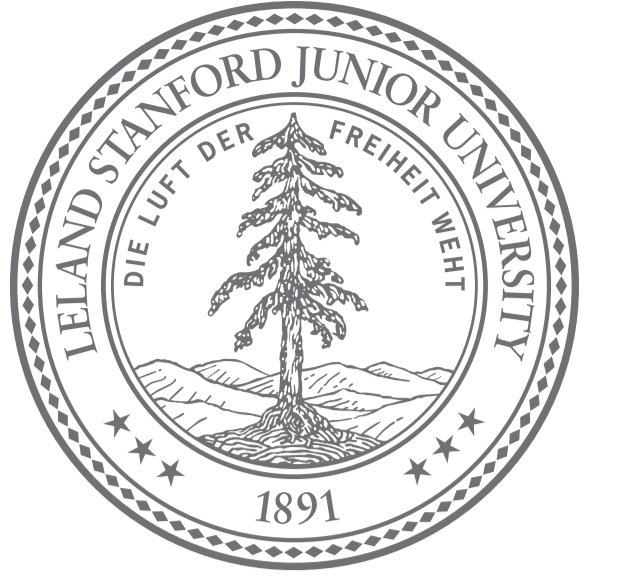


# Learning with Abandonment

Sven Schmit and Ramesh Johari, Stanford University (Presented at International Conference on Machine Learning, ICML 2018)

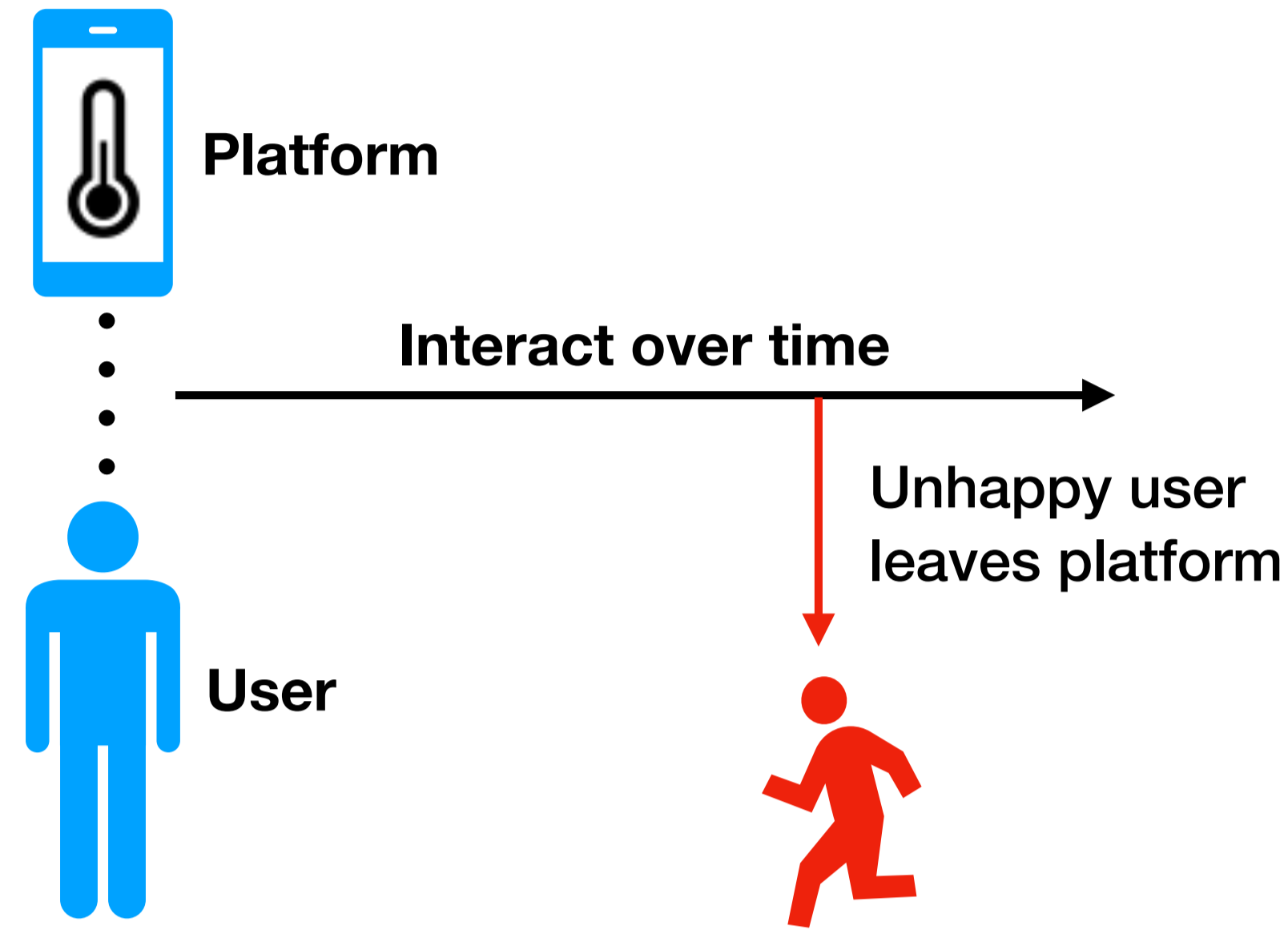
NSF CNS-1544548: CPS: Breakthrough: Collaborative Research: The Interweaving of Humans and Physical Systems: A Perspective From Power Systems; 10/01/2015-09/30/2019

{schmit, rjohari}@stanford.edu



## Personalize at individual level

How does a platform learn a personalized policy for its users?



## Motivation

- Demand response programs have a **high abandonment rate**.
- Once users leave, they are **unlikely to return**.
- Thus the platform is trying to optimize in the face of users that have a **risk of abandonment**.

## Puzzle

- I drew a threshold  $\theta$  uniformly between 0 and 100.
- You guess number  $x$ . If  $x < \theta$ , I pay you  $\$x$  and you can guess another number. If  $x > \theta$ , we stop.
- How can you maximize your discounted sum of rewards?

## Model

- User characterized by thresholds  $\theta_i$  drawn from the population distribution  $F$  (assume known)
- At discrete times  $t = 0, 1, \dots$ , select action  $x_t \in \mathbb{X}$
- If  $x_t < \theta_t$ , obtain (random) reward  $R_t(x_t)$  and continue, otherwise obtain no reward and process stops.

## Objective

Let  $T$  be the first time  $x_t > \theta_t$ :  $T = \min\{t : x_t > \theta_t\}$ . The goal is to **maximize the expected sum of discounted rewards** up to time  $T$ :

$$x^* \in \arg \max_{\{x_t\}_{t=0}^{\infty}} \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t R_t(x_t) \right]$$

## Related work

- Abandonment: Lu et al. [2018]
- Mechanism design and dynamic pricing: Myerson [1981]
- Demand estimation: Kleinberg and Leighton [2003]
- Safe reinforcement learning: Moldovan and Abbeel [2012], Berkenkamp et al. [2017]

## One user

First, we focus on learning the preference of a **single user**. We need to impose additional structure to make the problem tractable.

### Fixed threshold

Suppose  $\theta_t = \theta$  is drawn once from a known threshold distribution  $F$ . Let  $r(x) = \mathbb{E}(R_t(x_t)) > 0$ , also assumed known.

Under minimal assumptions, the optimal policy is a constant policy.

#### Fixed threshold leads to constant policy

Suppose the function  $f : x \rightarrow r(x)(1 - F(x))$  has a global optimum at  $x^*$ . Then, the optimal policy is  $x_t = x^*$  for all  $t$ .

### Intuition

Suppose optimal policy is increasing:  $x_t = y$ ,  $x_{t+1} = z > y$ . Compare to  $x_t = x_{t+1} = z$ .

- $\theta < y$ : identical outcome
- $\theta \geq y$ :  $z$  is optimal, so  $x_t = z$  is better than  $x_t = y$

### Corollary for simple model

For  $\theta \sim U[0, 1]$  and  $R_t(x) = x$  the optimal policy is  $x_t = 1/2$  for all  $t$ .

Now consider  $\theta \sim U[c, 1]$  for any  $c \in [0, 1/2]$ , then the optimal policy remains  $x_t = 1/2$ .

## Independent thresholds

Consider the other extreme:  $\theta_t$  drawn iid from  $F$ . This prohibits any learning across time, so again a **constant policy is optimal**.

## Robustness

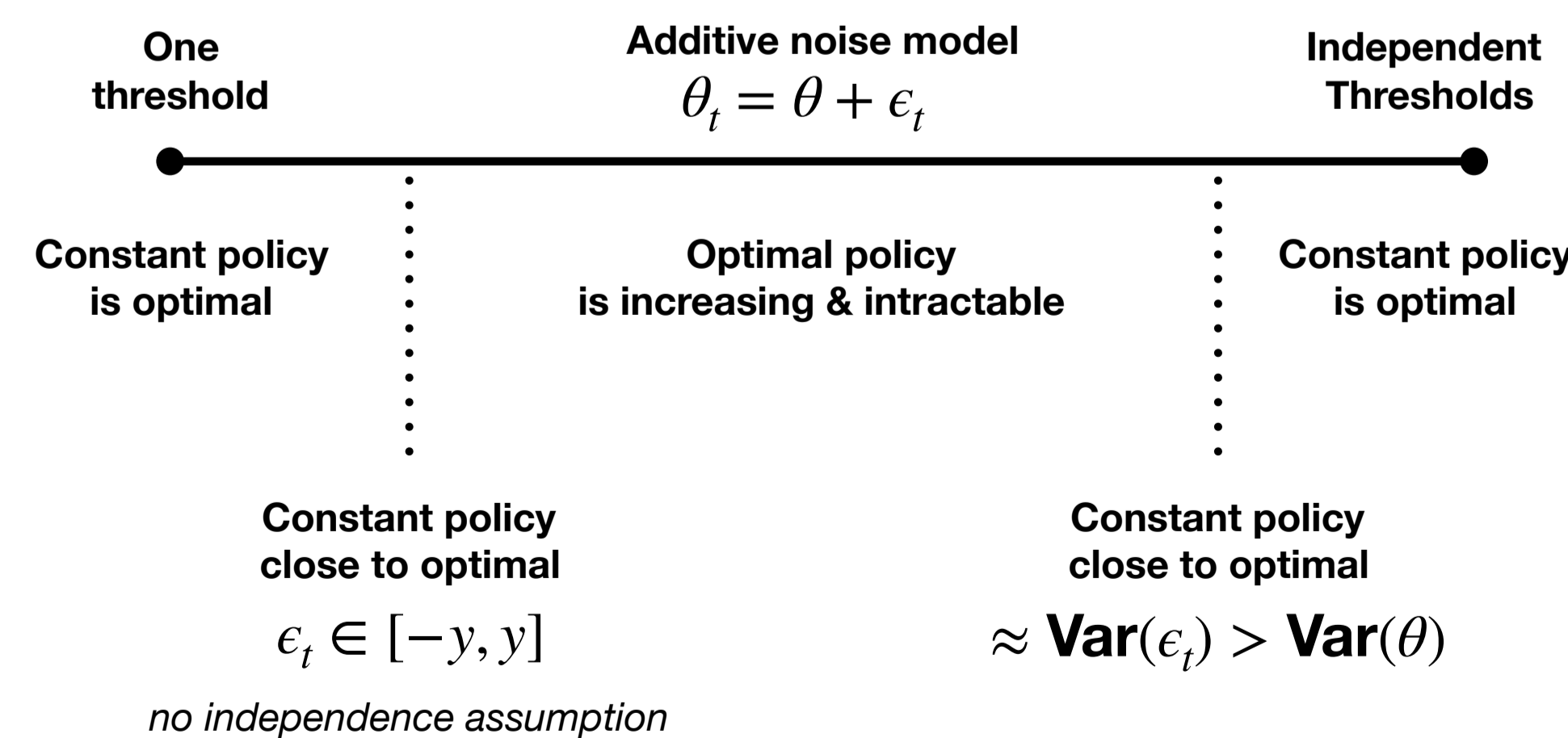
For more general models, **optimal policy is intractable**, but generally increasing.

**Additive noise**  $\theta_t = \theta + \epsilon_t$ .

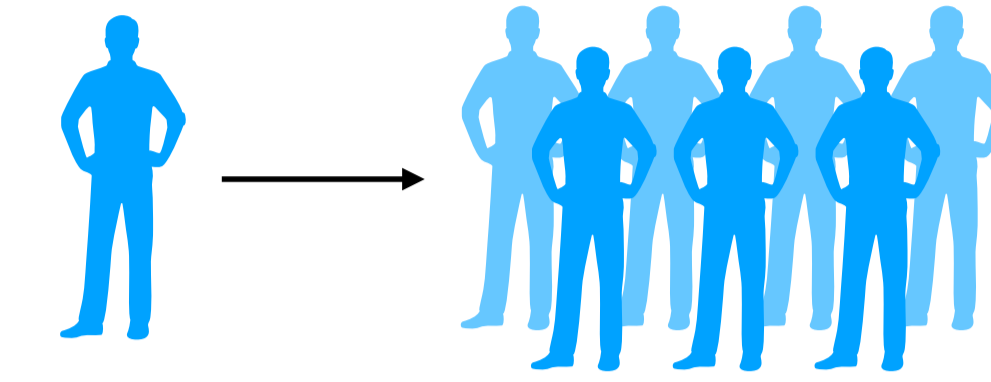
- Small noise** If  $\epsilon_t \in [-y, y]$ , then constant policy is approximately optimal (informal; see paper for details).
- Large noise** If  $\epsilon_t$  are independent, and variance is large, then constant policy is approximately optimal (informal; see paper for details).

## Summary

Our results for the single user model can be summarized as follows.



## Learning across a population



So far, we have assumed that threshold distribution and reward function are **known**. What if we do not know these, but have a population of users to learn from?

## Setting

- Users arrive sequentially
- User  $u$  has fixed threshold  $\theta_u$  drawn from unknown distribution  $F$
- Assumptions:
  - $F$  has support  $[0, 1]$
  - Rewards bounded in  $[0, 1]$ ,  $r(x)$  unknown
  - Profit function  $p(x) = r(x)(1 - F(x))$  is concave\*

### Regret

- Optimal action:  $x^* \in \arg \max_x p(x)$
- Consider constant policies per user, one pull corresponds to lifetime rewards for one user
- $\text{regret}(n) = (1 - \gamma)np(x^*) - (1 - \gamma)\sum_{u=1}^n p(x_u)$

## Learning strategy

We follow the approach of Kleinberg and Leighton [2003].

- Discretize  $[0, 1]$  into  $K = O((n/\log n)^{1/4})$  points
- Run UCB [Auer et al., 2002] / KL-UCB [Garivier and Cappé, 2012] algorithm on discretized actions

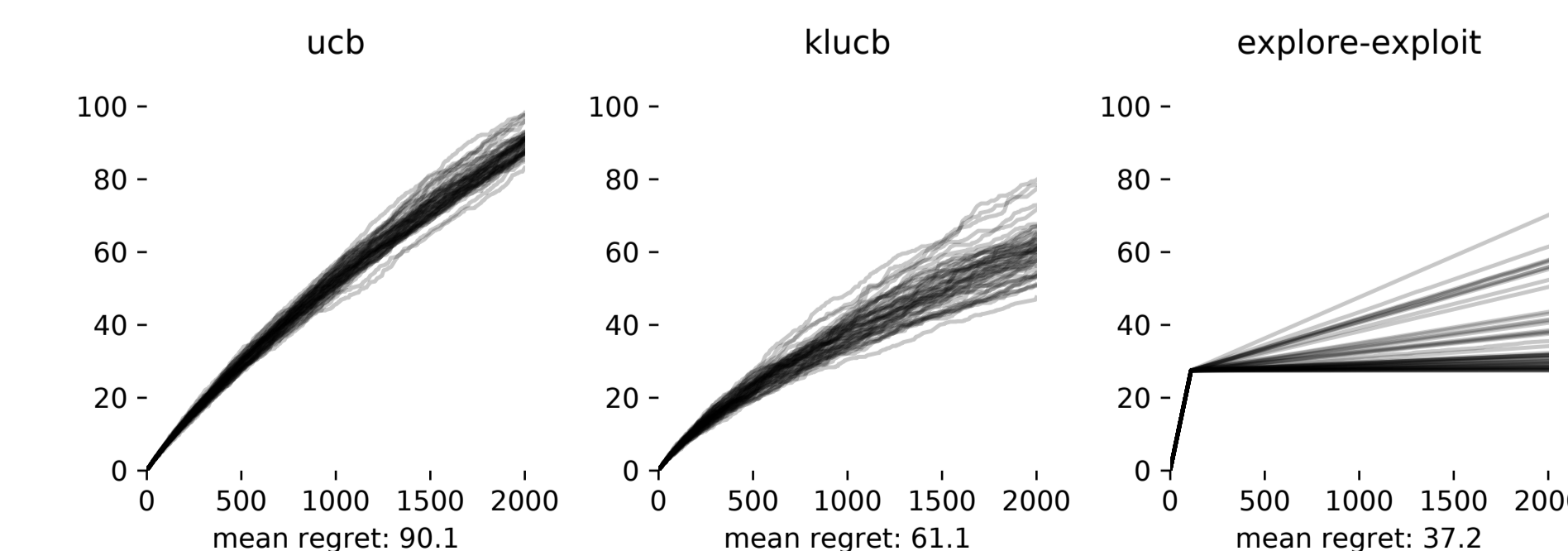
#### Regret of UCB and KL-UCB policies

If  $p(x) = r(x)(1 - F(x))$  satisfies a concavity condition, then UCB and KL-UCB algorithms using a discretized grid with  $K = O((n/\log n)^{1/4})$  achieve  $O(\sqrt{n \log n})$  regret.

Note: Kleinberg and Leighton [2003] provide  $O(\sqrt{n})$  lower bound for constant policies that applies here as well, but dynamic policies could perform better.

## Simulations

We include an optimistic benchmark: **explore-exploit** observes first  $m$  thresholds  $\theta_u$  and thereafter selects the optimal action based on the empirical CDF and known reward function.



Take-away: possible gains possible with dynamic policies that learn more about individual thresholds  $\theta_u$ .

## Feedback

**Key idea:** user does not always abandon immediately.

In a demand response program:

- Suppose a user is unhappy with **platform's thermostat adjustment**.
- The first few times this happens, **the user might just override the settings, rather than abandoning**.
- But eventually, if the experience is negative too frequently, **the user will abandon**.

## Augmenting the general model

User abandons after  $m \sim \text{Geometric}(p)$  violations of threshold.

That is, if  $x_t > \theta$

- with probability  $p$  user abandons, process stops, and
- with probability  $1 - p$ , platform receives no reward, but process continues.

#### Feedback leads to partial learning

For any abandonment risk  $0 \leq p < 1$ , the optimal policy partially learns about the user. That is, at a certain point the optimal policy becomes constant. (Informal)

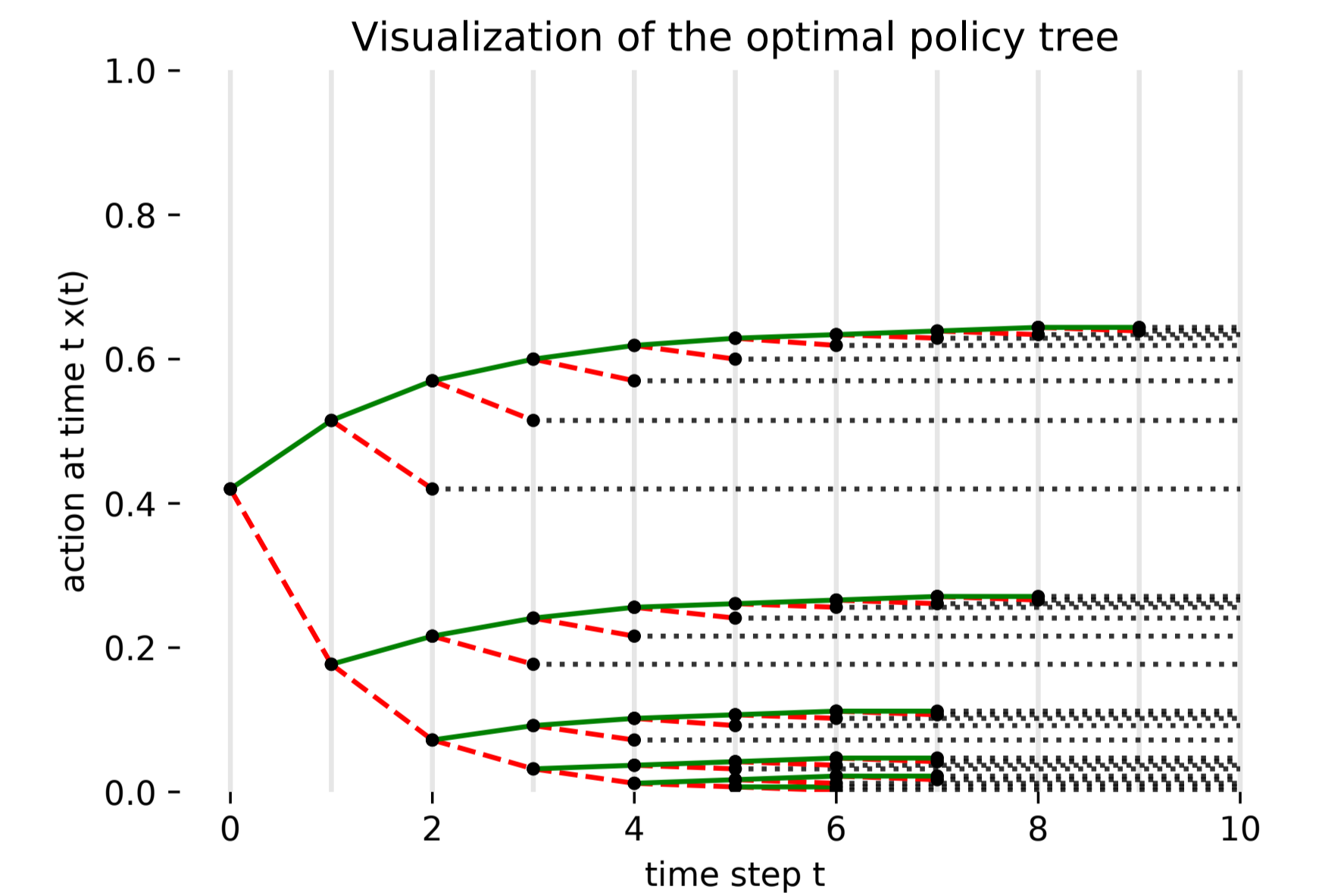


Figure 1: Optimal policy for  $\theta \sim U[0, 1]$ ,  $r(x) = x$ ,  $p = 1/2$  and  $\gamma = 0.9$ .

Note, this is also true for  $p = 0$ ; when there is no risk of abandonment.

## Aggressive and conservative policies

We define aggressive and conservative policies as follows:

- Aggressive policy:**  $x_0 > x_c^*$
- Conservative policy:**  $x_0 < x_c^*$

If  $p \approx 1$  (high risk of abandonment), then optimal policy is **conservative**, provided  $\gamma$  is sufficiently large.

If  $p \approx 0$  (low risk of abandonment), then optimal policy is **aggressive**.