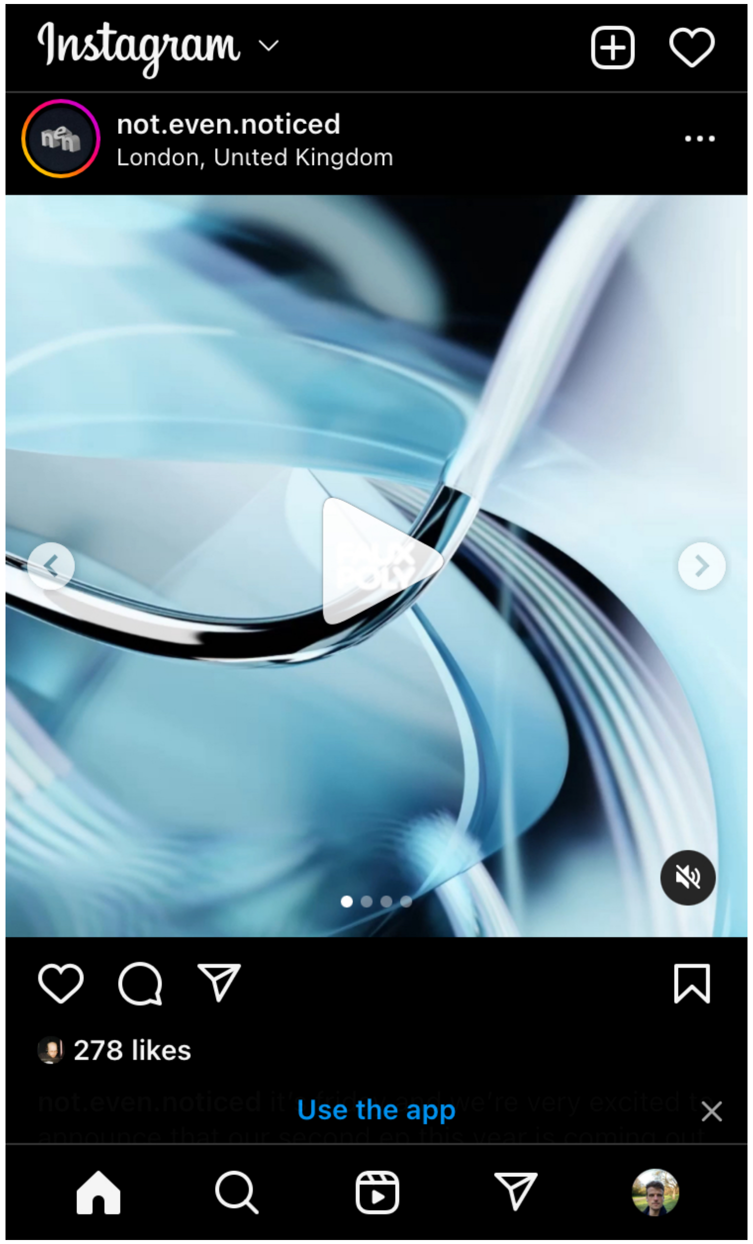
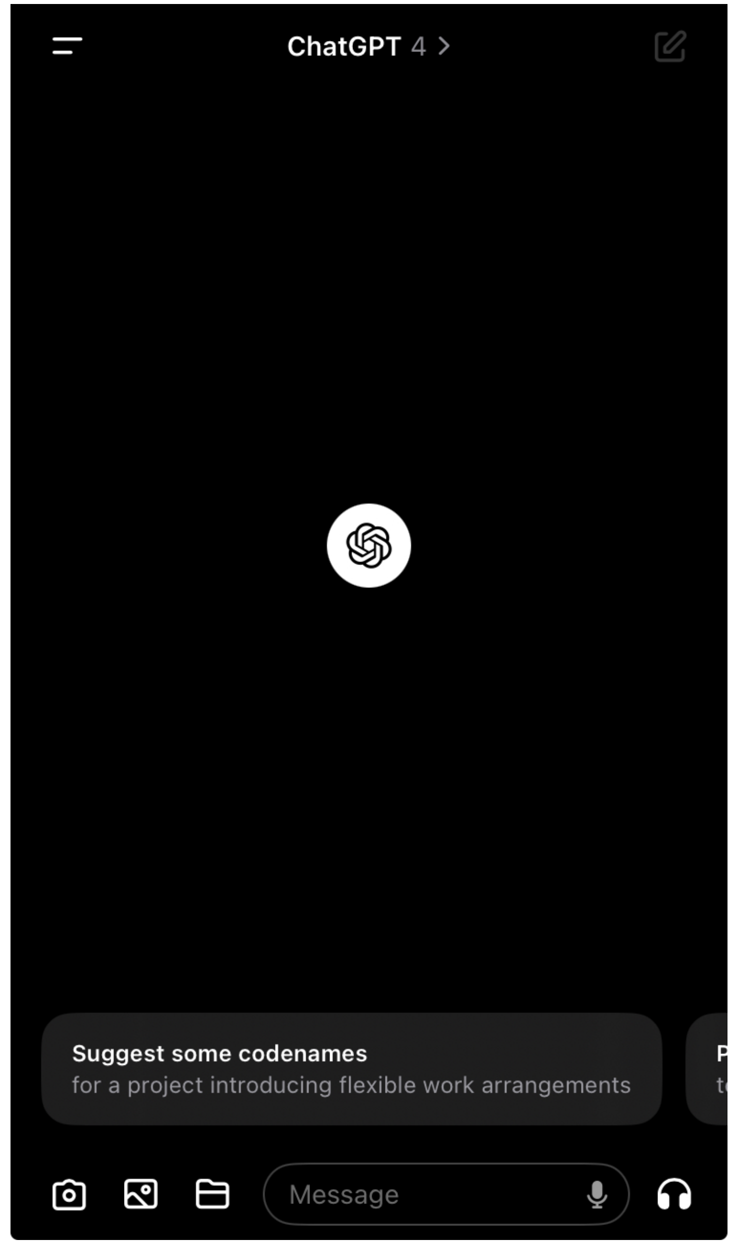
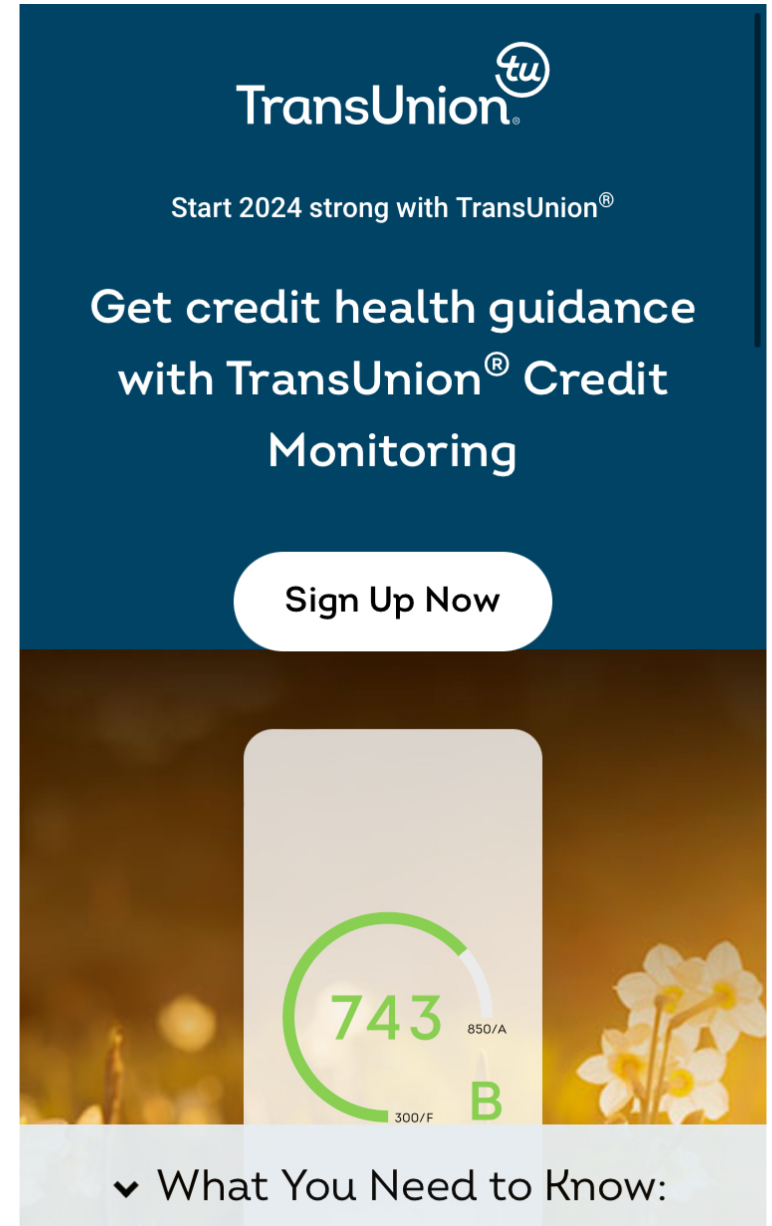
Risk Aversion in Learning Algorithms

Andreas Haupt
Microsoft Research New England
April 16, 2024

Consequential Online Learning



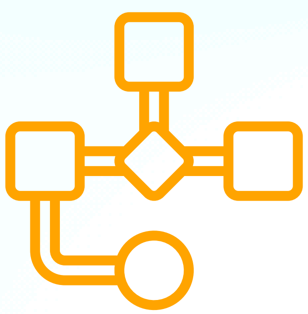




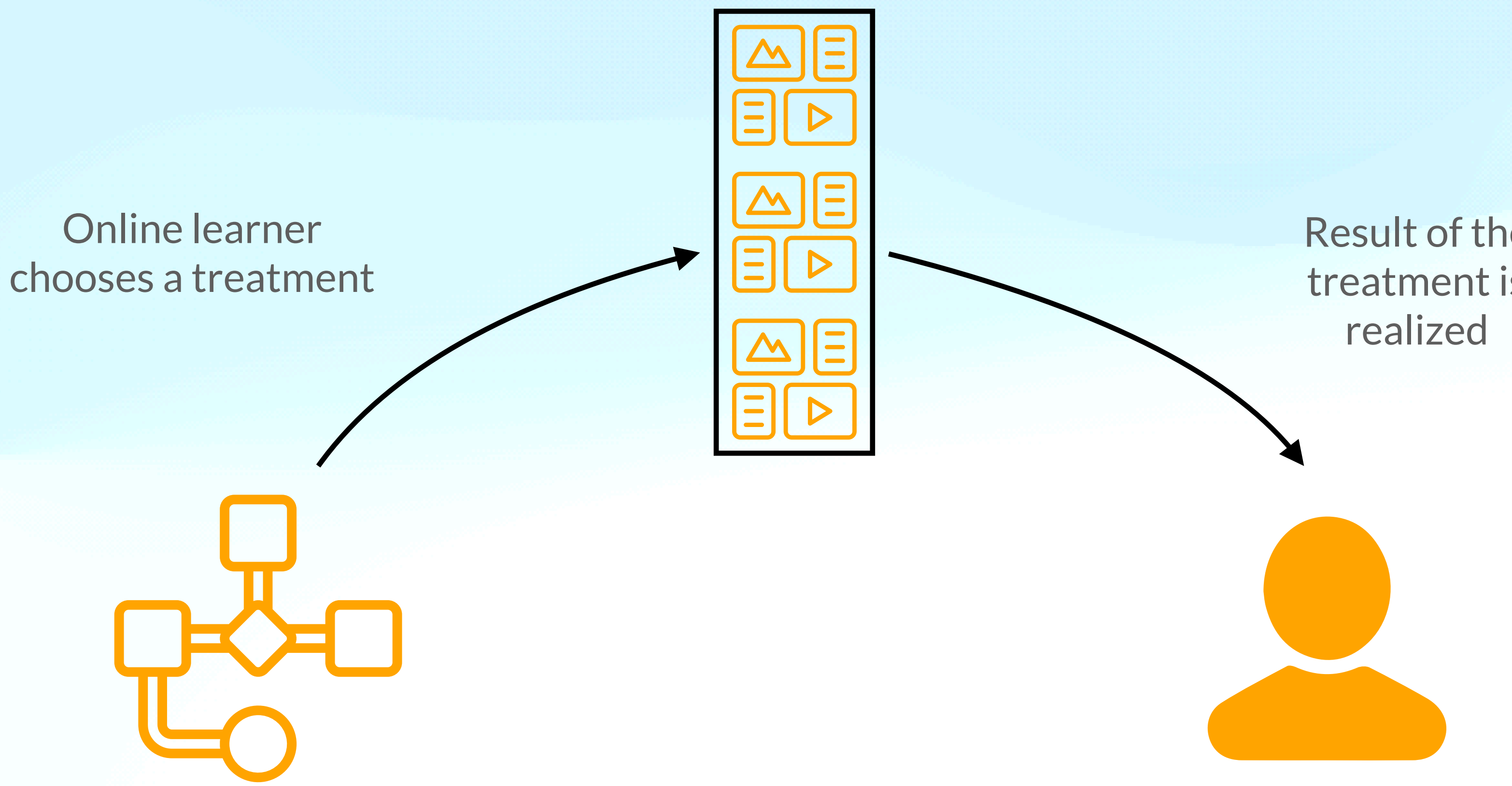
Online Learning Online

User shows up

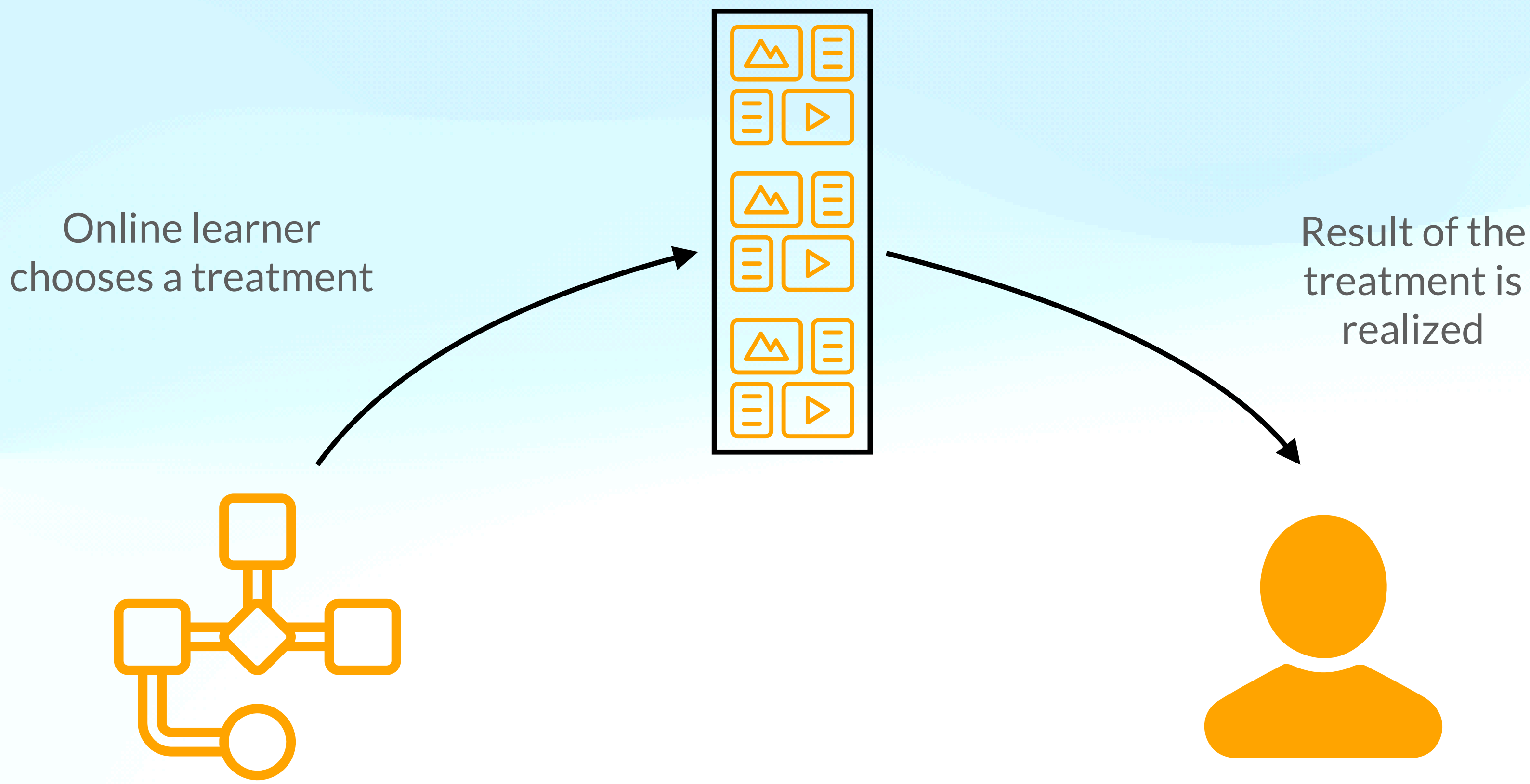
Online learner
chooses a treatment



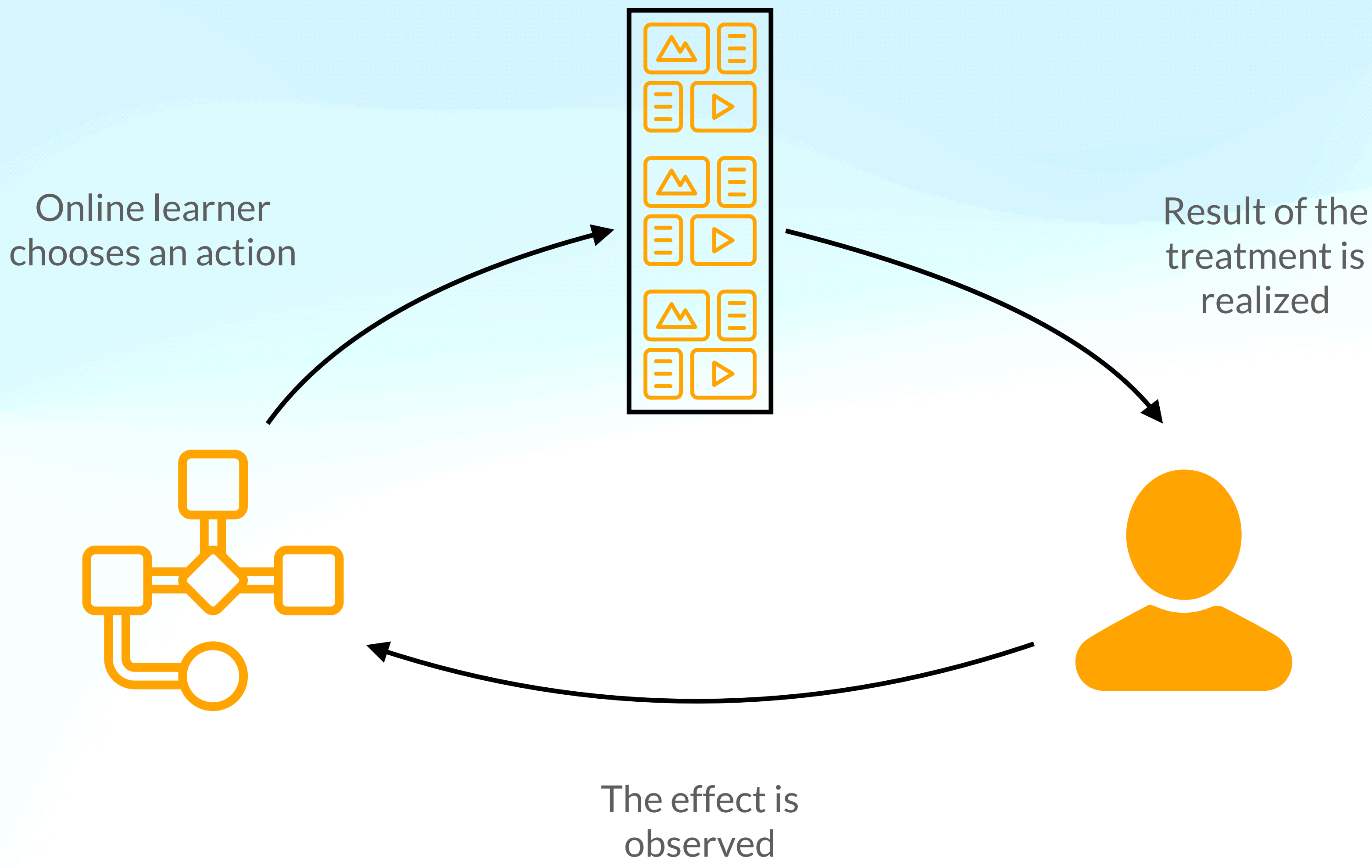
Empfehlungssysteme



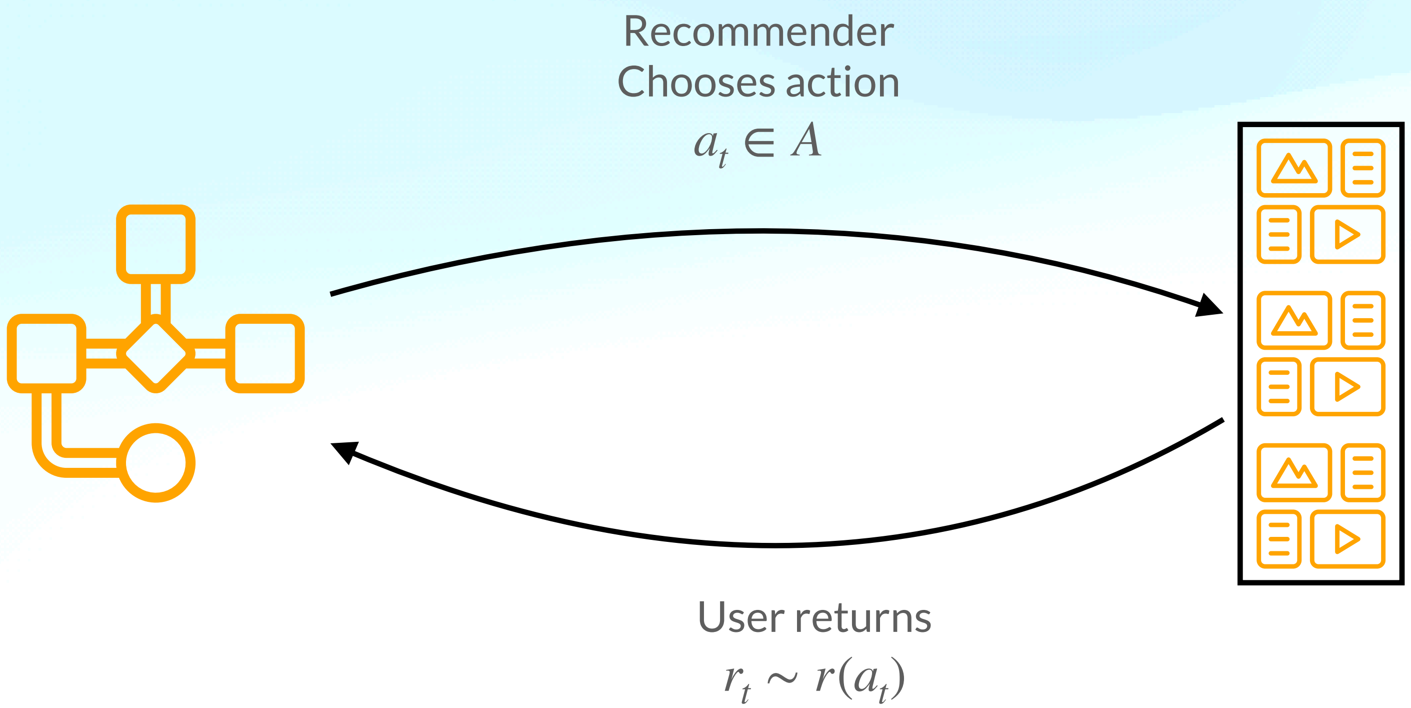
Empfehlungssysteme



Empfehlungssysteme



The Online Learning Problem

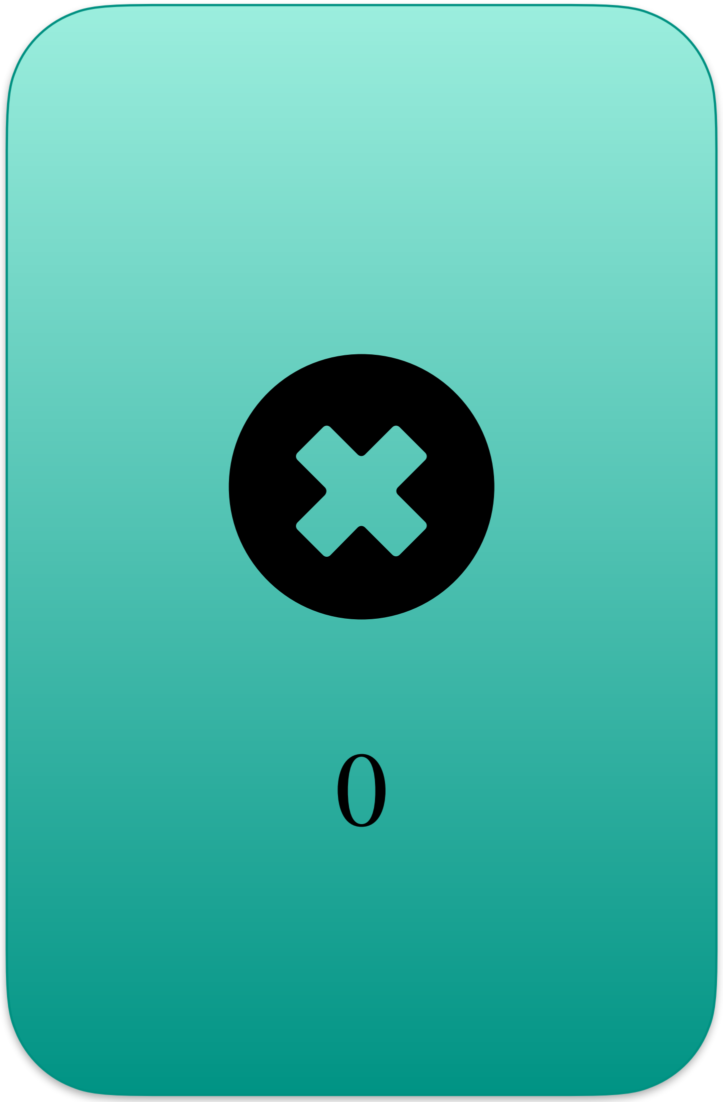


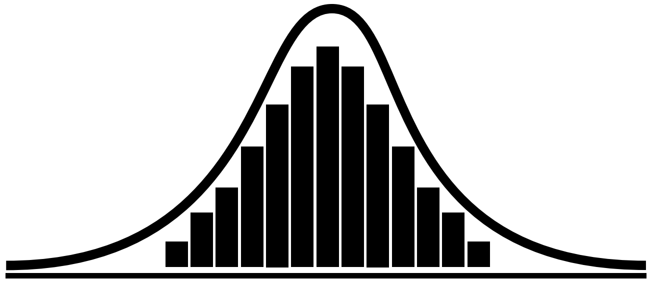
Interactions Online Are Common

- Content online
- Credit Scores
- Hiring

Bad Rep’ is Hard to Get Rid of

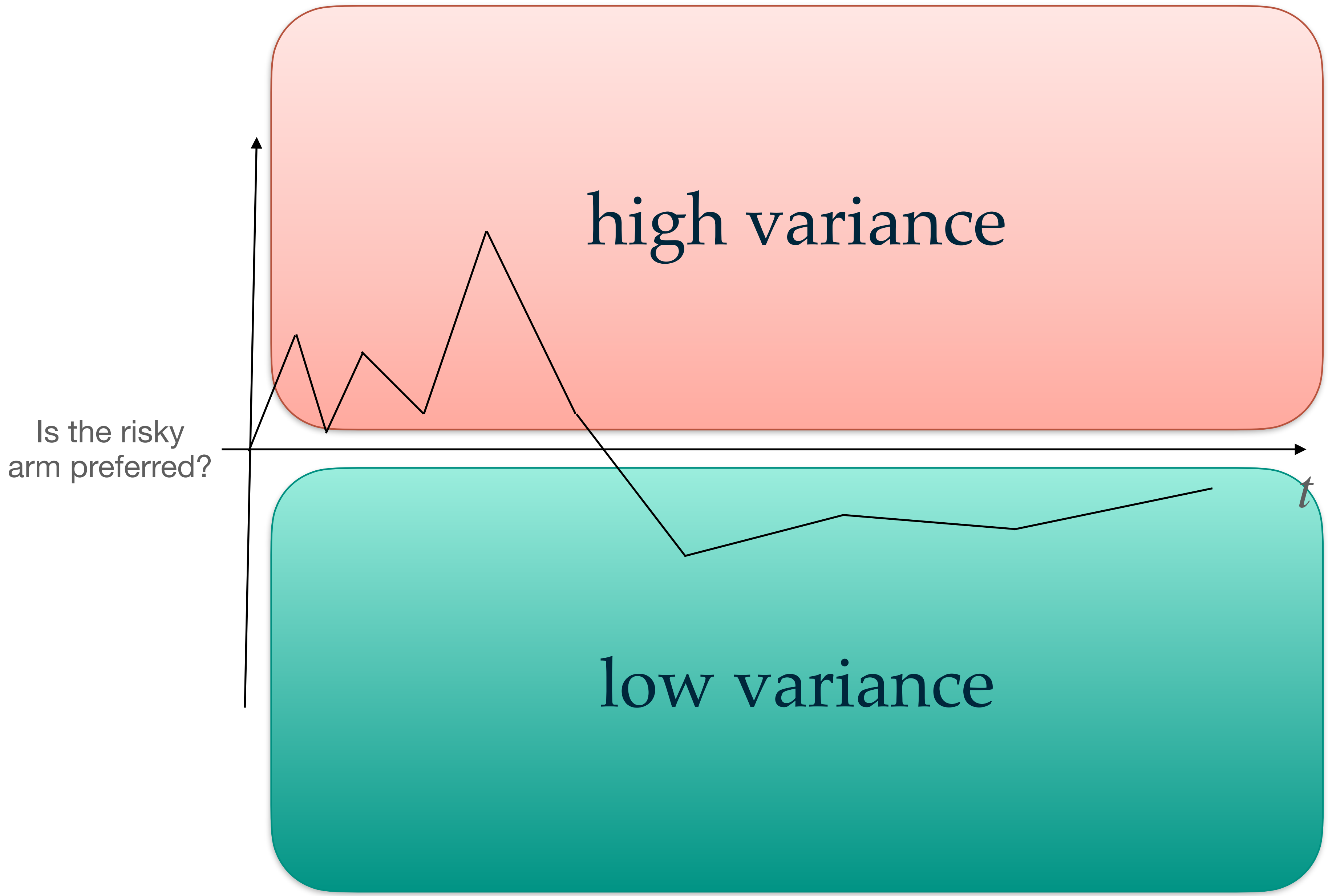
ε- 



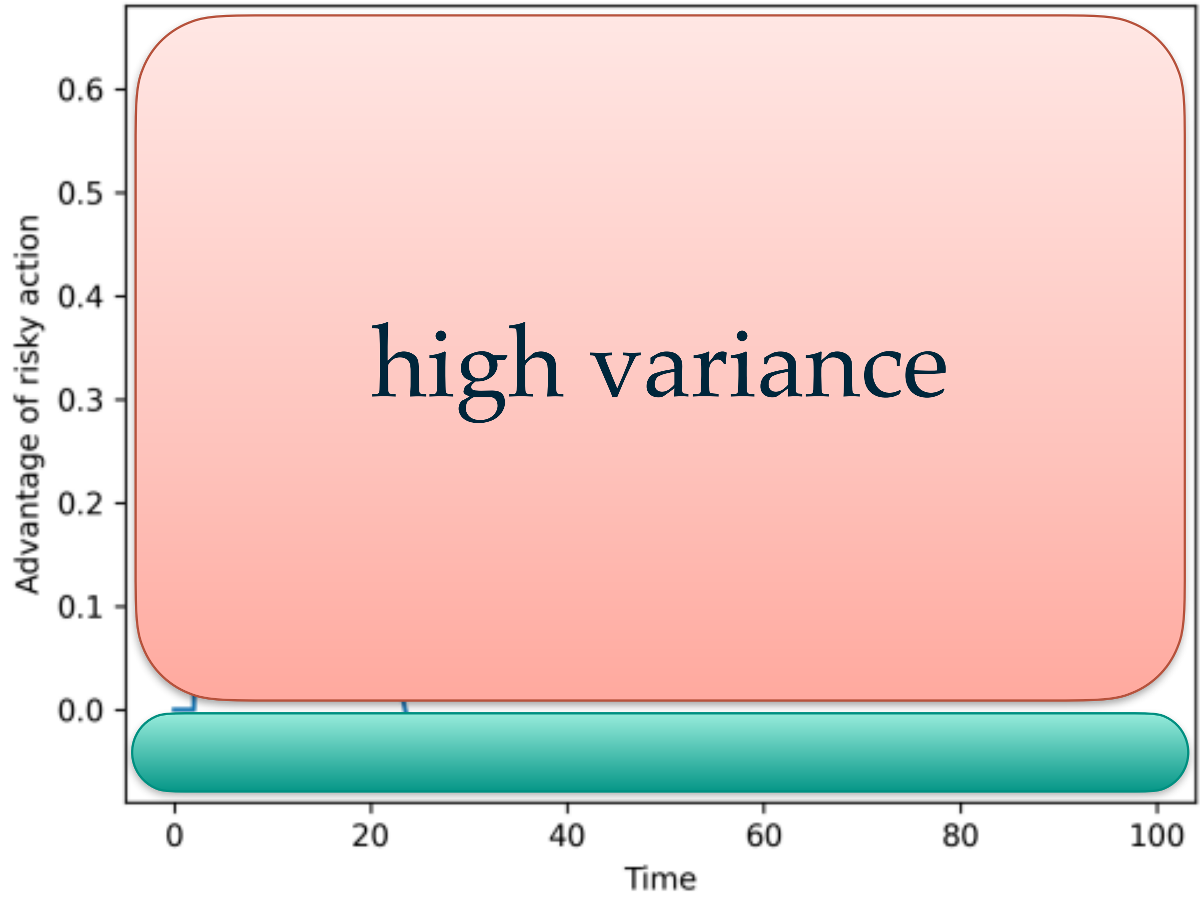


$N(0,1)$

Bad Rep' is hard to get rid of: ε-Greedy



Bad Rep' is hard to get rid of: ε-Greedy



(b) One realization of the advantage walk for ε-Greedy where the safe action has distribution $\mathbb{1}_{\{0\}}$ while the risky action has distribution $U[-1, 1]$

ε-Greedy is risk-averse

Theorem. Let $(\varepsilon_t)_{t\in\mathbb{N}}$ such that $\varepsilon_t \to 0, \Sigma\varepsilon_t = \infty$.
If there is a deterministic action a^* among the optimal actions, and all actions have symmetric reward distributions, $\mathbb{P}[a_t = a^*] \to 1$.

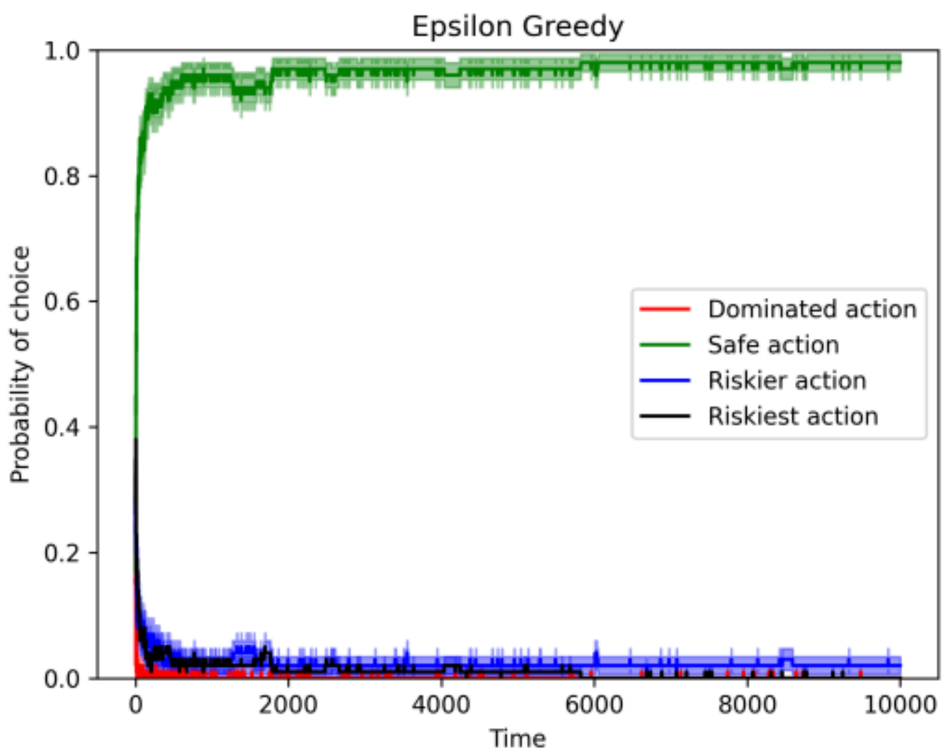
Proof Sketch. Consider the story of aggregate historic rewards $(X_t)_{t\in\mathbb{N}}$.

- Define the last crossing time of zero τ.
- Let E be event that is positive
- $(X_t)_{\tau\leq t'\leq t} \,|\, E$ is a positive symmetric random walk with small variance.

- For constant exploration, get convergence to probability in $(0,1)$.

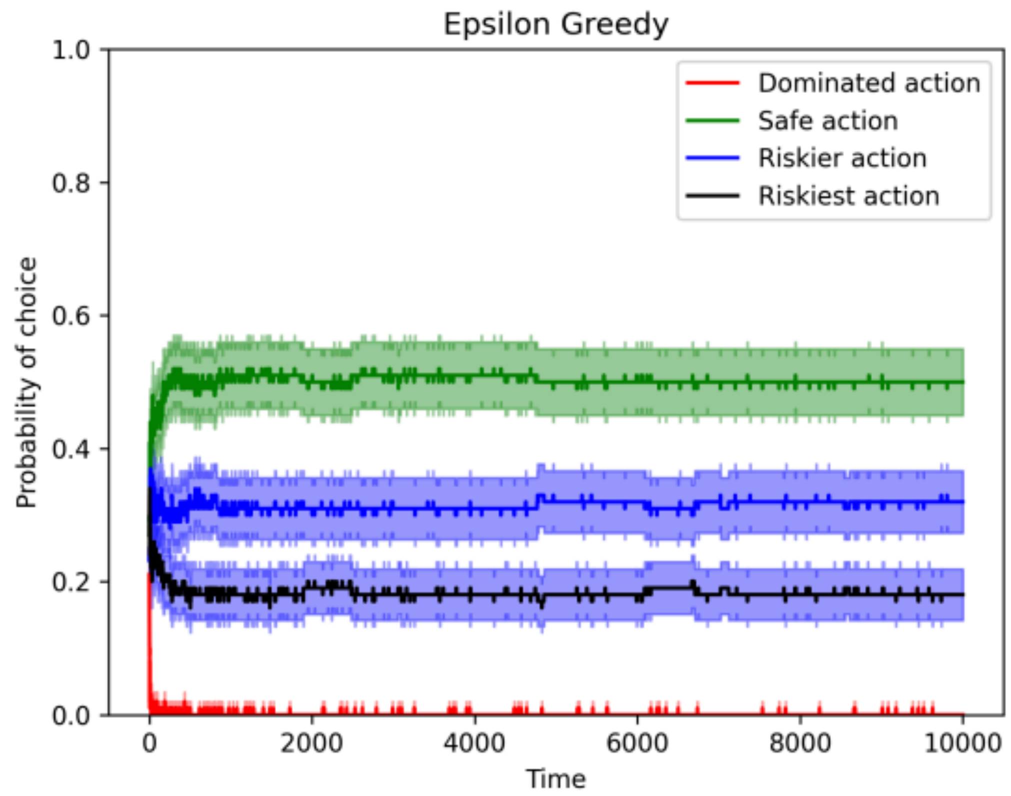
Empirically, the theorem is correct

- Simulation with one deterministic Arm
- Consider prediction policy setting: Known 0 treatment effect



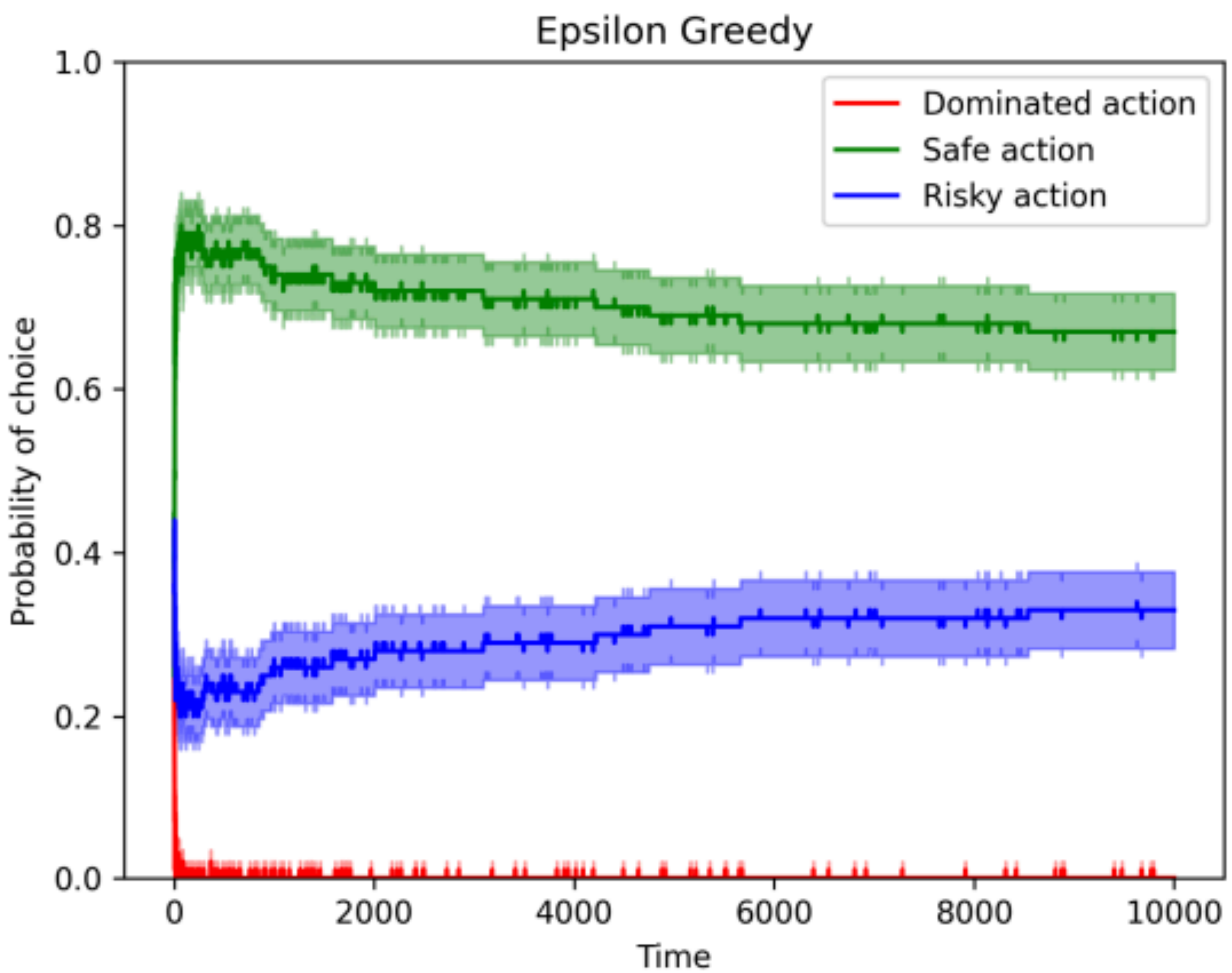
(a) Perfect risk aversion.

Drop Assumptions



(a) ε-Greedy with no optimal safe action.

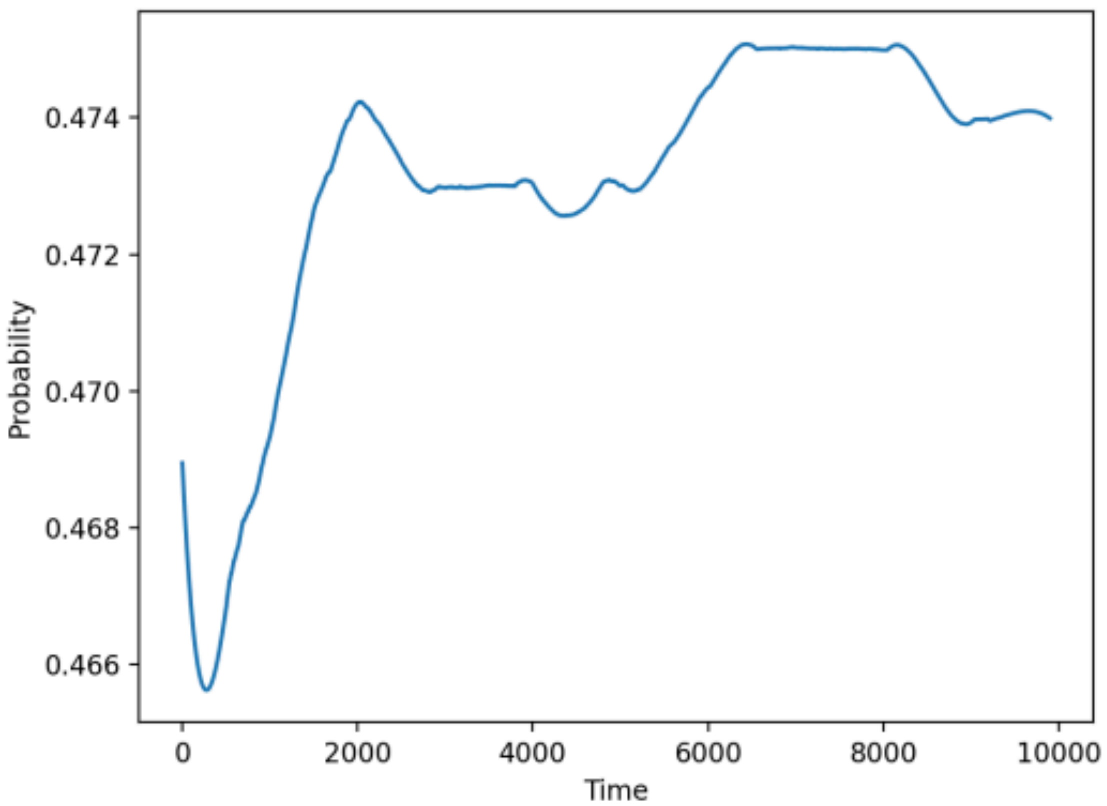
Finite-time effect may be large



(b) ε-Greedy with a strictly better risky action.

Application to a Recommendation System

- Return utility $u_{ij} = x_{ij} + \varepsilon_{ij}$
- Could do fancier simulation with



(a) ϵ-Greedy

Risk Aversion in Reinforcement Learning

Visualization of a Grid World

The Reinforcement Learning Problem

- S: States
- A: Actions
- T: Transitions
- R: Reward Function
- γ: Discount factor

Goal: Maximize $\sum_{t=0}^{T} \gamma^T r(s_t, a_t)$

- Common class of algorithms: Policy Optimization
- $\pi : S \times A \times \Theta \to \Delta(A)$
- Θ: Parameter Space
- Classical algorithm REINFORCE

Visualization of the policy space of a grid world

Bring up high variance low variance point

Bring up high variance low variance point

The Development Process of

- Consider θ_t developing as

$d\theta_t = L(\theta_t)dt$ "gradient flow"

- The real development is finite

$\theta_{t+1} - \theta_t = hL(\theta_t)dt$

- The reality is also noisily observed

$\theta_{t+1} - \theta_t = hL(\theta_t)dt + \sqrt{h}\sigma(\theta_i)dW_t$

- What is the real loss?

How to Correct for Risk Aversion

Recap: What was the issue with risk aversion

- Algorithm affects data distribution
- Noisy data leads to less of such data, not more
- What are ideas for correcting?

- Optimism!
- (If someone of you mentions reweighing: In the paper, some nice maths, we can discuss)

UCB is risk-neutral

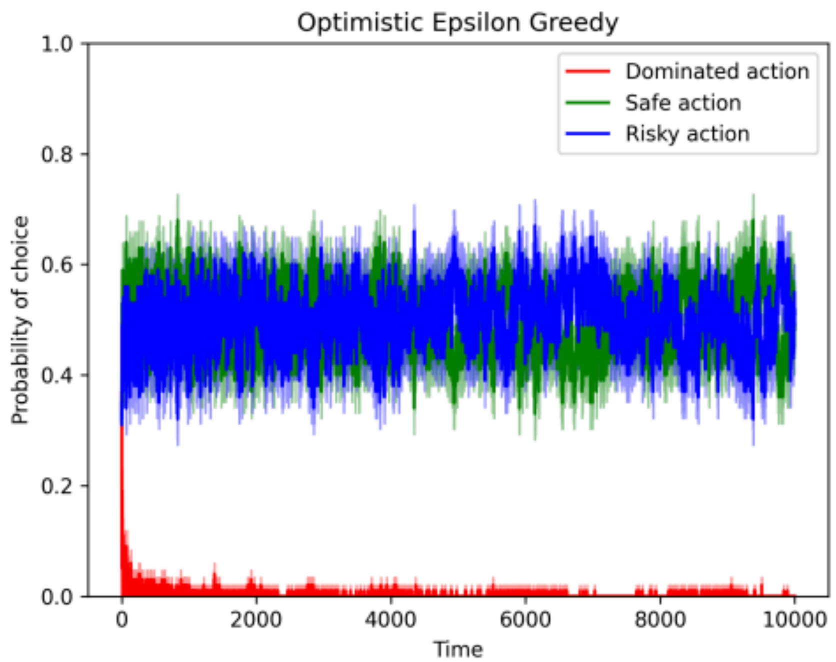
Theorem. There exists $\rho_0 > 1$ such that for any $\rho > \rho_0$ and any $(\varepsilon_t)_{t \in \mathbb{N}}$ with $\varepsilon_t \to 0$, UCB is risk-neutral.

Proof Sketch. Consider the story of aggregate historic rewards $(X_t)_{t \in \mathbb{N}}$.

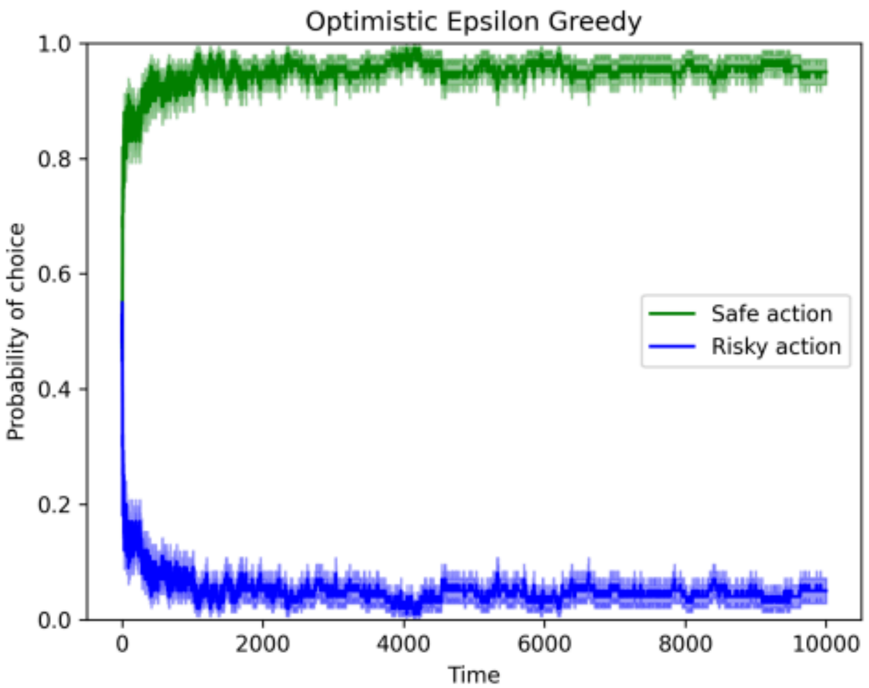
- Prove again that there is
- Let E be event that is positive
- $(X_t)_{\tau \leq t' \leq t} \,|\, E$ is a positive symmetric random walk with small variance.

- For constant exploration, get convergence to probability in $(0,1)$.

Optimism in Bandits

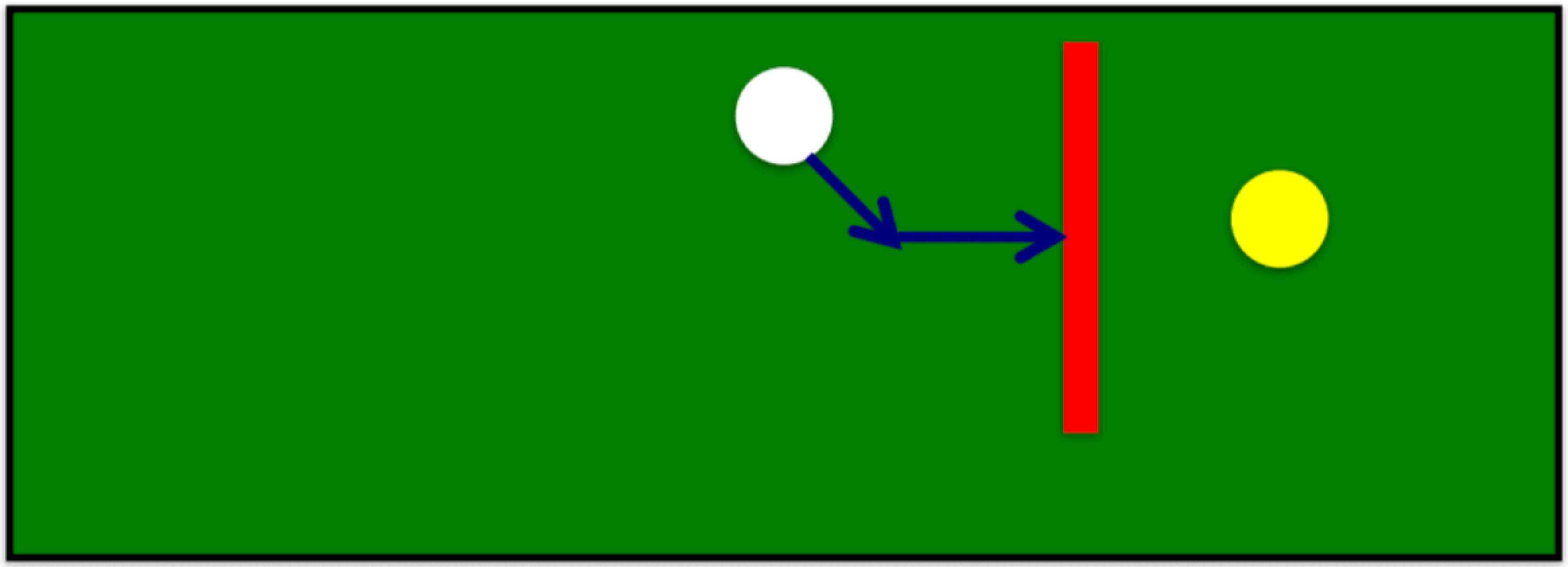


Theorem is correct



Optimism too low

Optimism in Reinforcement Learning



Wisdom from Labor Economics (Li et al. 2020)