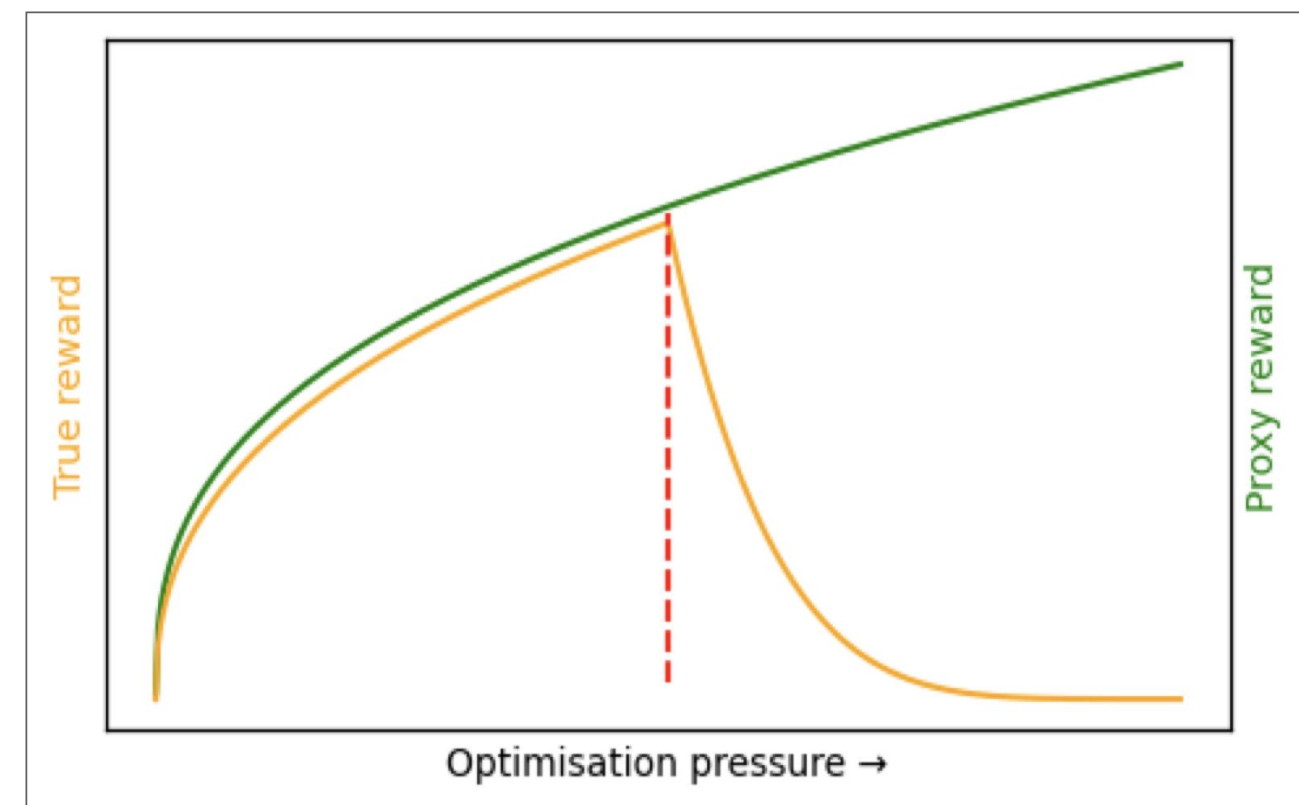


1. What is Goodhart's Law?



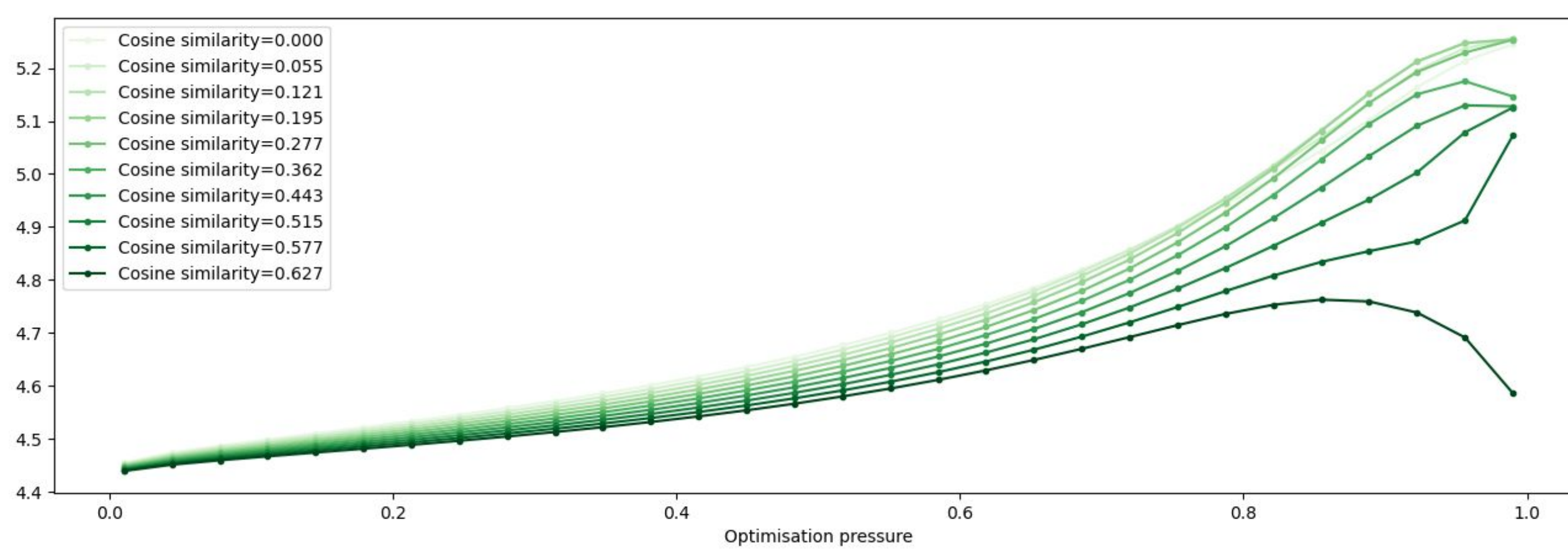
"When a measure becomes a target, it ceases to be a good measure"

2. How often does it occur?

We vary hyperparameters of an MDP:

- Environment type/reward type
- No. states
- No. actions
- True-proxy reward distance
- Discount factor
- Sparsity
- Optimisation pressure operationalisation

In a total: **~30k** environments.



The drop occurs in **19.3%** of them.

Optimisation pressure

- Number of gradient updates
- $e^{-\alpha}$ where $\mathbb{P}[\pi_\alpha(s) = a] \propto \exp(\frac{1}{\alpha} \cdot Q^*(s, a))$ (Boltzmann rationality)
- $e^{-\alpha}$, use training objective $R_\alpha(s, a) := R(s, a) + \alpha H(\pi(s))$ (Maximum causal entropy)

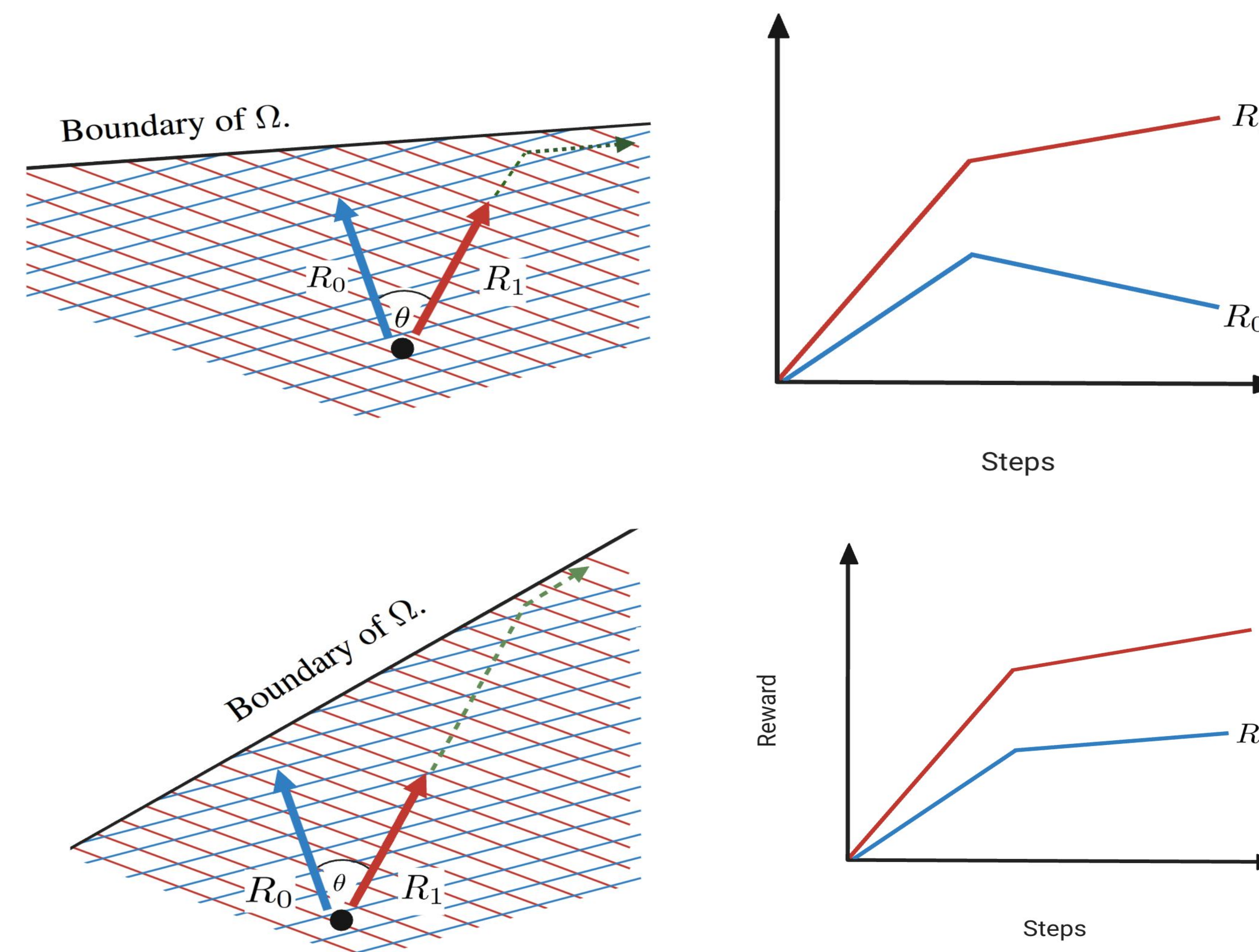
3. Why does Goodharting happen?

We focus on the occupancy measure space

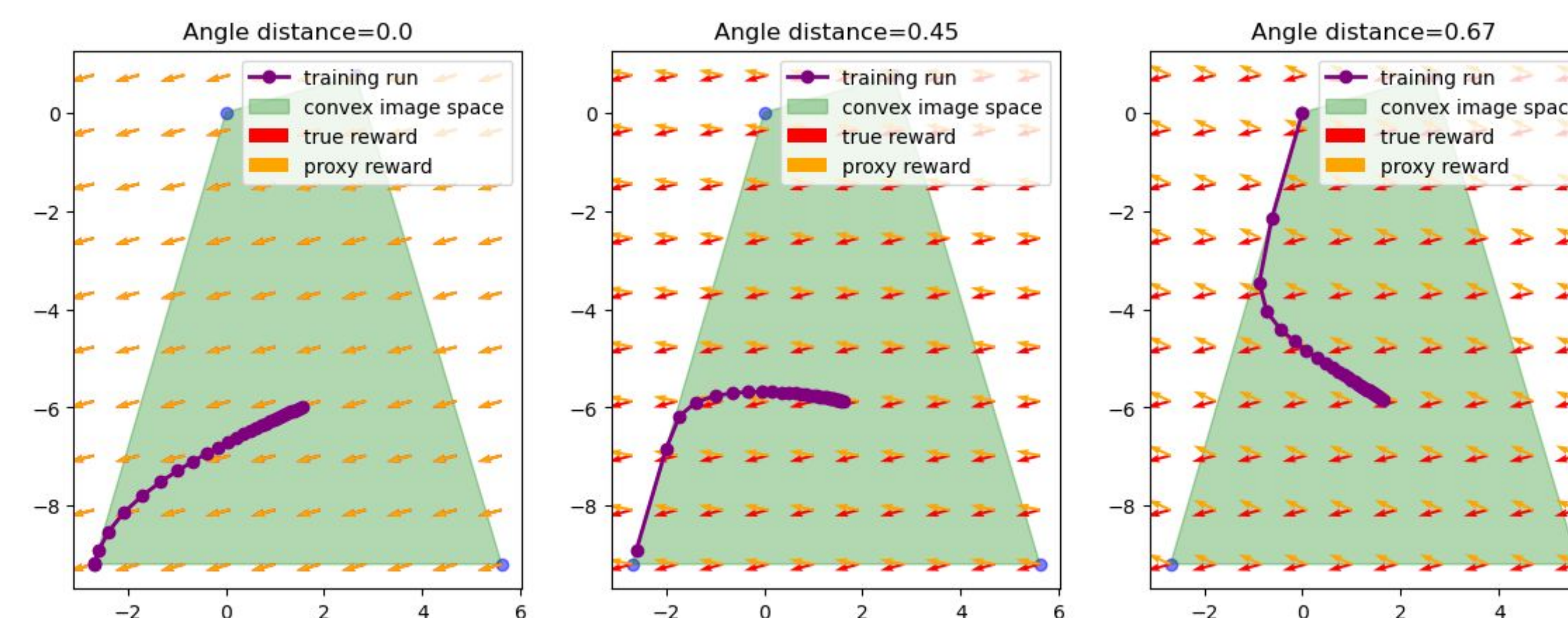
$$\eta^- : \Pi \rightarrow \mathbb{R}^{|S \times A|}$$

$$\eta^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a \mid \xi \sim \pi)$$

Simplifying assumption: perfect gradient ascent



Real dynamics



4. How to prevent it?

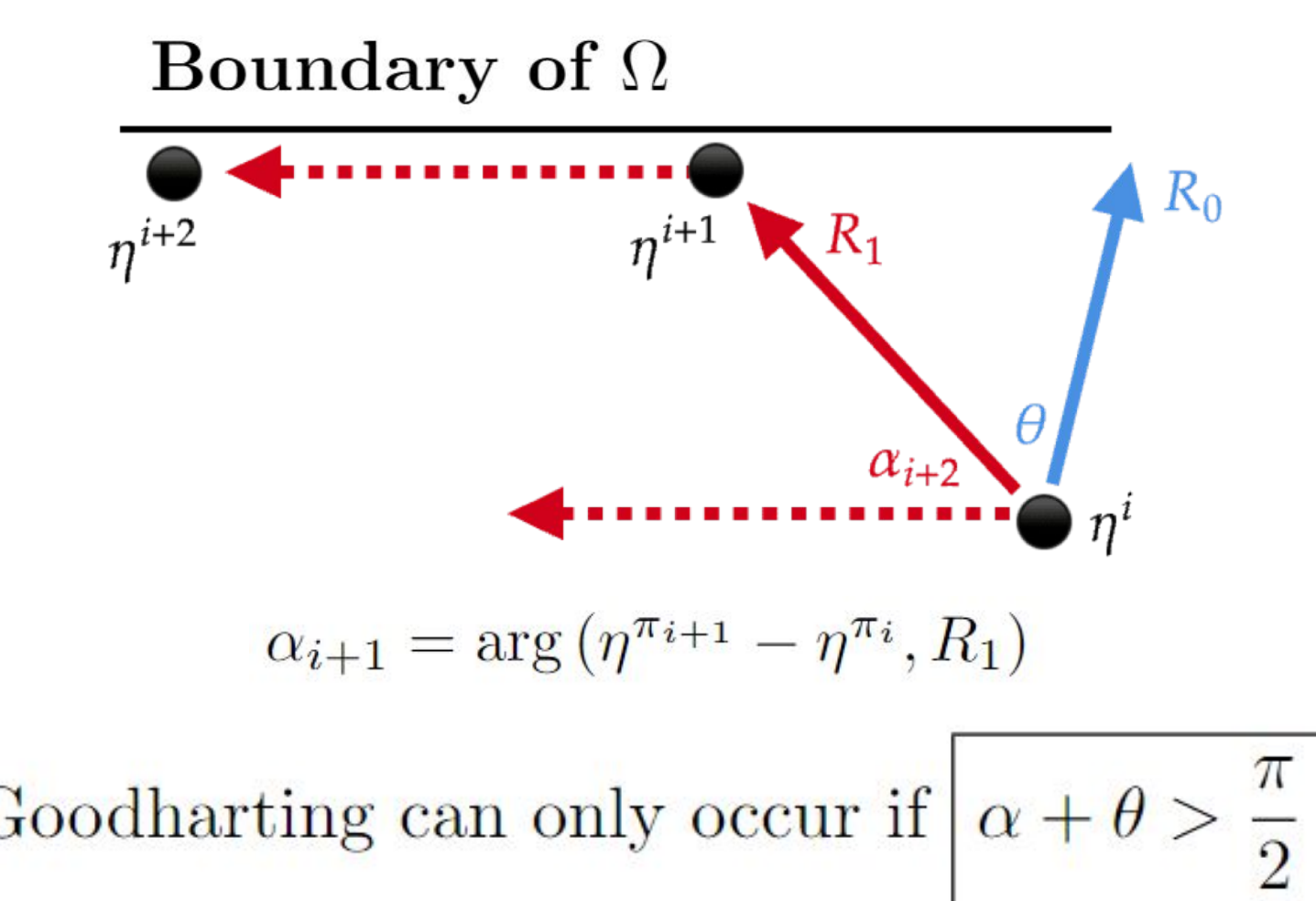
Optimal Stopping: Stopping when

$$\frac{\mathcal{J}_{R_1}(\pi_{t+1}) - \mathcal{J}_{R_1}(\pi_t)}{\|\eta^{\pi_{t+1}} - \eta^{\pi_t}\|} \leq \tau$$

where $\tau = \sin(\theta) \|\text{proj}_P(R_1)\|$ maximizes worst-case returns across training curve

Notation:

- R_1 - proxy reward (known)
- R_0 - true reward (unknown)
- $\mathcal{J}(\pi)$ - returns of policy π
- P - plane of policies in occupancy space
- $\arg(u, v)$ - angle between u, v
- Given θ with $\arg(R_1, R_0) \leq \theta$



Shaped reward:

$$\tilde{\mathcal{J}}_{R_1}(\pi) := \mathcal{J}_{R_1}(\pi) - \kappa \|\eta^\pi\| \arg(\eta^\pi R_1)$$

where $\kappa = \tan(\theta) \|\text{proj}_P(R_1)\|$ has one maximum which globally maximizes worst-case returns.

5. The early-stopping algorithm

```

procedure EARLYSTOPPING( $S, A, \tau, \theta, R$ )
   $\vec{r} \leftarrow M, R$ 
   $\pi \leftarrow \text{Unif}[\mathbb{R}^{S \times A}]$ 
   $\vec{\eta}_0 \leftarrow \eta^\pi$ 
   $\vec{t}_0 \leftarrow \arg\max_{\vec{t} \in T(\vec{\eta}_0)} \vec{t} \cdot R$ 
  while ( $\vec{t}_i \neq \vec{0}$ ) and  $R \cdot \vec{t}_i \leq \sin(\theta) \|R\|$  do
     $\lambda \leftarrow \max\{\lambda : \vec{\eta}_i + \lambda \vec{t}_i \in \Omega\}$ 
     $\vec{\eta}_{i+1} \leftarrow \vec{\eta}_i + \lambda \vec{t}_i$ 
     $\vec{t}_{i+1} \leftarrow \arg\max_{\vec{t} \in T(\vec{\eta}_{i+1})} \vec{t} \cdot R$ 
     $i \leftarrow i + 1$ 
  end while
  return  $(\vec{\eta}_i)^{-1}$ 
end procedure

```

- ✓ Provably prevents Goodharting
- ⚠ Requires knowing the bound on θ
- ⚠ Can lose out on a lot of reward

