Orientation

What we have seen:

- RLMPC, why does it work & different flavors

Orientation

What we have seen:

- RLMPC, why does it work & different flavors

What we will do now:

- The theory is not about MPC, it is about MDPs (MPC is a special case)
- It has broader implications for AI and model-based decision making
- Provide (tentative) practical ideas for in Sim2Real

AI for Operational Decisions
Some perspectives from recent research

Prof. Sebastien Gros

Department of Cybernetic
Faculty of Information Technology
NTNU

Freiburg PhD School

Outline

Model-Free Pathway - "Pure" Reinforcement Learning

Actions

State

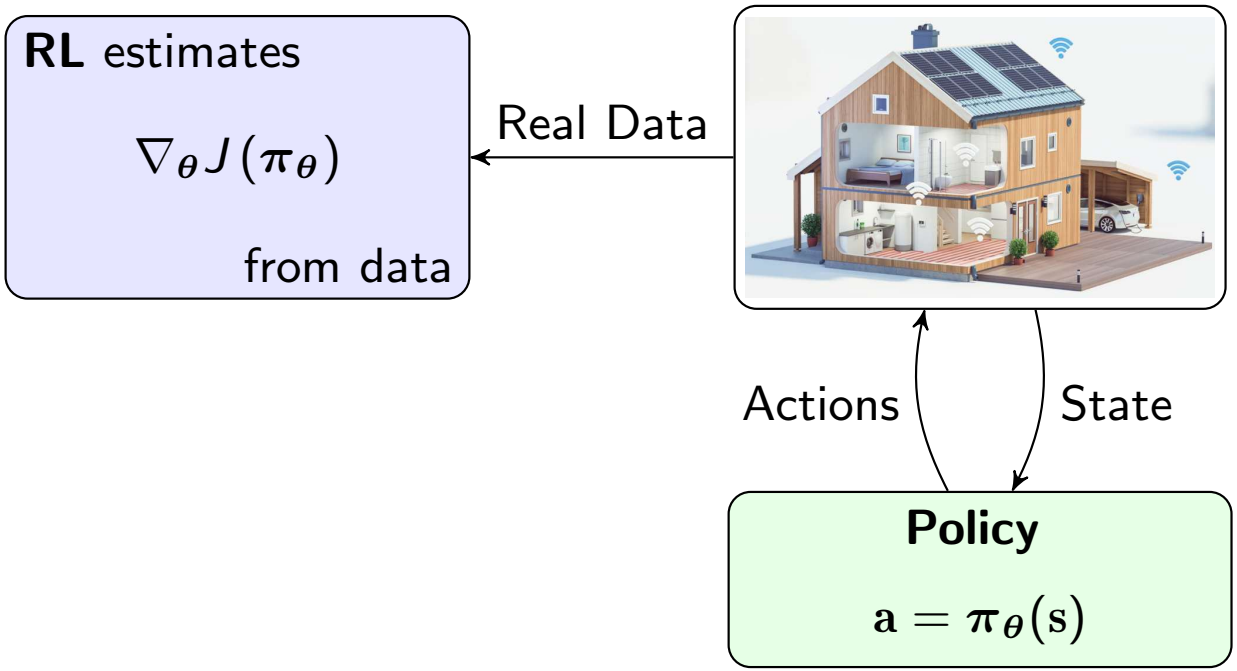
Policy

$$\mathbf{a} = \pi_\theta(\mathbf{s})$$

- Define policy π_θ as a parametrized function
- Optimize policy as:

$$\min_\theta \; J(\pi_\theta)$$

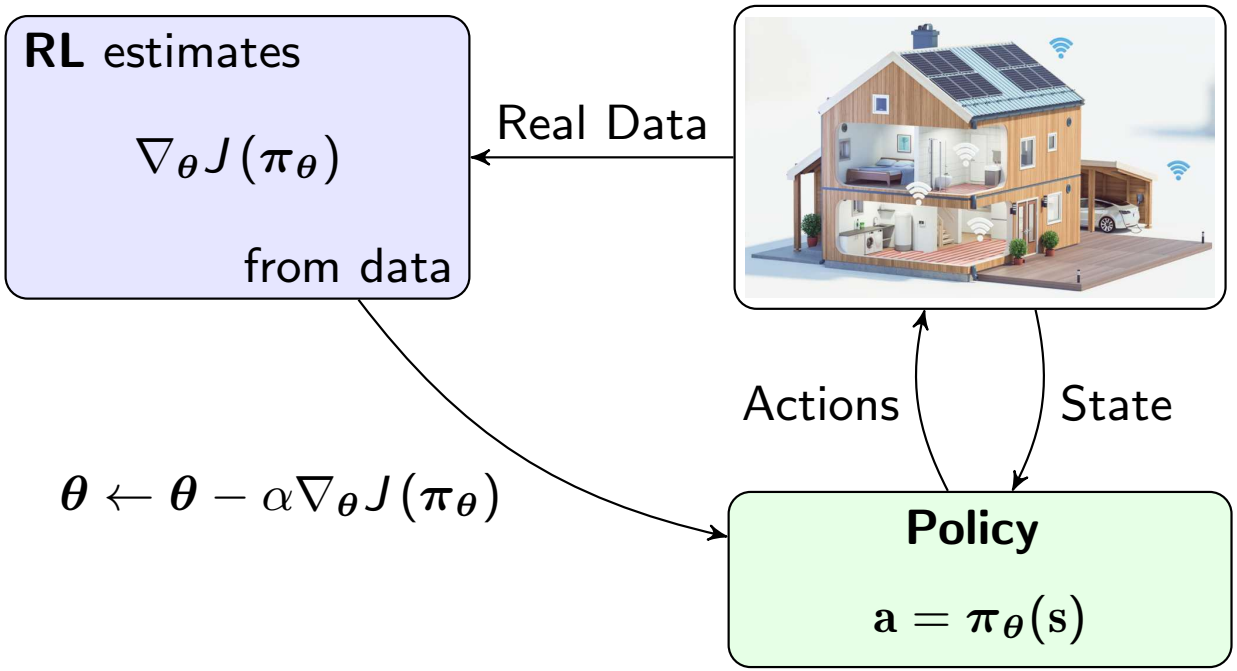
Model-Free Pathway - "Pure" Reinforcement Learning



- Define policy π_θ as a parametrized function
- Optimize policy as:

$$\min_\theta \; J(\pi_\theta)$$

Model-Free Pathway - "Pure" Reinforcement Learning

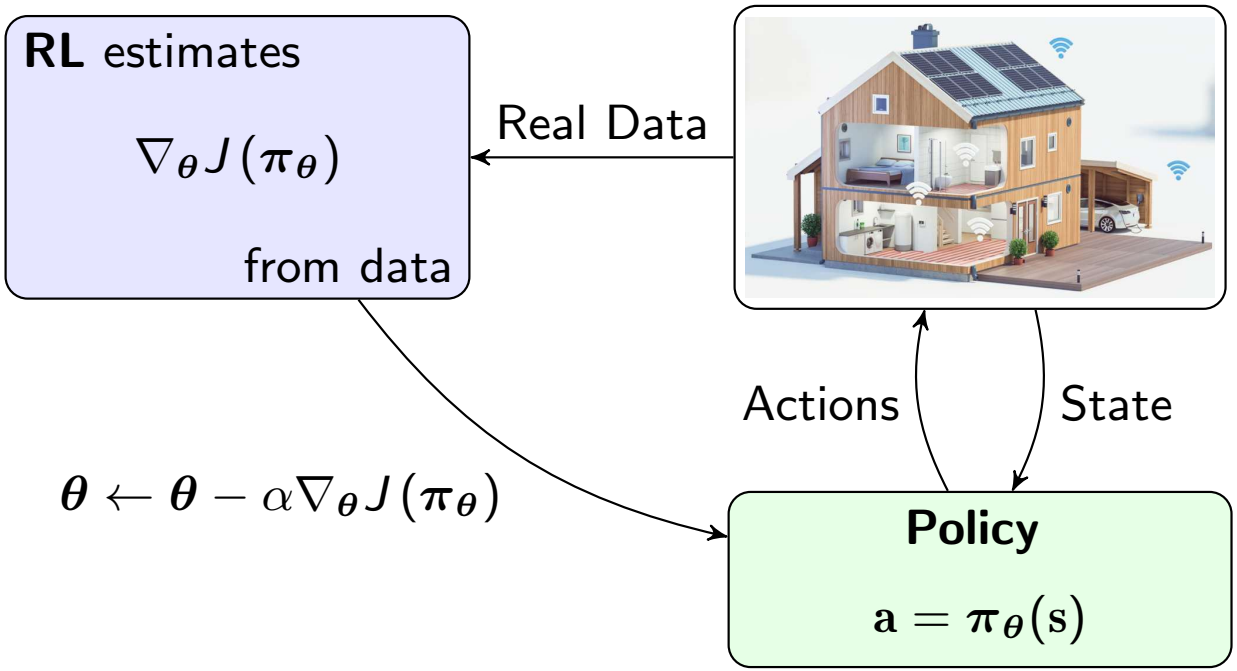


- Define policy π_θ as a parametrized function
- Optimize policy as:

$$\min_\theta \; J(\pi_\theta)$$

- Gradient descent can be used to optimize policy

Model-Free Pathway - "Pure" Reinforcement Learning



- Define policy π_θ as a parametrized function
- Optimize policy as:

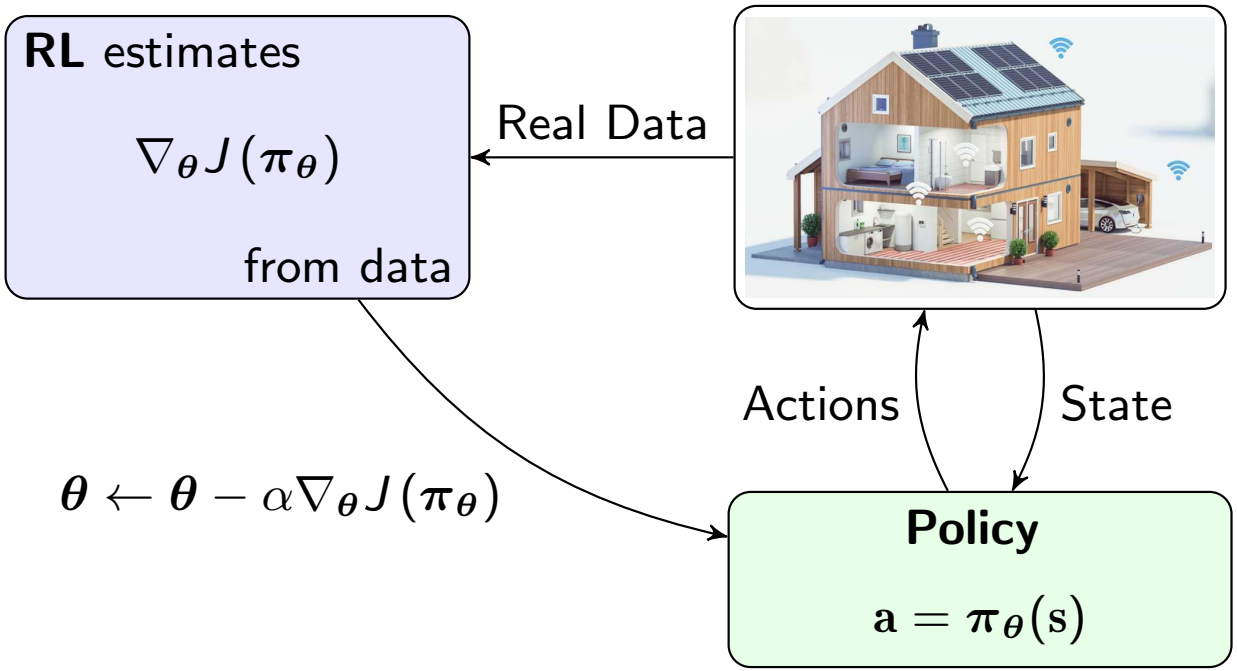
$$\min_\theta \; J(\pi_\theta)$$

- Gradient descent can be used to optimize policy

Remarks

- $\pi_\theta(\mathbf{s})$ often from DNN
- End-to-end training requires a lot of data and effort
- Difficult to provide explainability and guarantees on the policy we obtain
- Poor track record of industrial adoption (exception chatbots)

Model-Free Pathway - "Pure" Reinforcement Learning



- Define policy π_θ as a parametrized function
- Optimize policy as:

$$\min_\theta \; J(\pi_\theta)$$

- Gradient descent can be used to optimize policy

Remarks

- $\pi_\theta(\mathbf{s})$ often from DNN
- End-to-end training requires a lot of data and effort
- Difficult to provide explainability and guarantees on the policy we obtain
- Poor track record of industrial adoption (exception chatbots)

$\rightarrow$policy is often trained on **high-fidelity simulations**: embed knowledge, reduce costs, manage safety, promote explainability, control the training

Model-based Pathway - Predictive AI Models (data-driven, ML, etc)

Real world: $\mathbf{s}_{k+1} \sim \rho(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$

- Data with enough "richness"

Model-based Pathway - Predictive AI Models (data-driven, ML, etc)

One-step model

AI model e.g.

$$\hat{\mathbf{s}}_{k+1} \sim \rho_{\boldsymbol{\theta}}(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$$

fitting data

Real world: $\mathbf{s}_{k+1} \sim \rho(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$

- Data with enough "richness"

Here we no longer need the model $\rho_{\boldsymbol{\theta}}$ to be "optimization-friendly"

Model-based Pathway - Predictive AI Models (data-driven, ML, etc)

One-step model

AI model e.g.

$$\hat{\mathbf{s}}_{k+1} \sim \rho_{\boldsymbol{\theta}}(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$$

fitting data

Real world: $\mathbf{s}_{k+1} \sim \rho(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$

- Data with enough "richness"

Here we no longer need the model $\rho_{\boldsymbol{\theta}}$ to be "optimization-friendly"

Model parameters $\boldsymbol{\theta}$

- Selected such that model $\rho_{\boldsymbol{\theta}}$ "ressembles" reality ρ, using data (+ physics)
- Deterministic models are a special case of $\rho_{\boldsymbol{\theta}}$
- Many methods, from Least Squares and MLE to Bayesian and Adversarial Learning
- Still, in most cases $\rho_{\boldsymbol{\theta}}$ "simplifies" ρ because
  - need for huge amounts of data to decide $\boldsymbol{\theta}$ if model is very rich
  - computationally demanding simulations if $\rho_{\boldsymbol{\theta}}$ is expensive to sample from
- Then
  - biases in case of insufficiently rich model structure
  - validity is limited to some parts of the state-action space
  - only few first moments are correct (typ. mean + variance)

Model-based Pathway - Predictive AI Models (data-driven, ML, etc)

One-step model

AI model e.g.

$$\hat{\mathbf{s}}_{k+1} \sim \rho_{\boldsymbol{\theta}}(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$$

fitting data

Virtual data from "simulating"
$\hat{\mathbf{s}}_{k+1} \sim \rho_{\boldsymbol{\theta}}(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$ forward (Monte Carlo)

Here we no longer need the model $\rho_{\boldsymbol{\theta}}$ to be "optimization-friendly"

Model parameters $\boldsymbol{\theta}$

- Selected such that model $\rho_{\boldsymbol{\theta}}$ "ressembles" reality ρ, using data (+ physics)
- Deterministic models are a special case of $\rho_{\boldsymbol{\theta}}$
- Many methods, from Least Squares and MLE to Bayesian and Adversarial Learning
- Still, in most cases $\rho_{\boldsymbol{\theta}}$ "simplifies" ρ because
  - need for huge amounts of data to decide $\boldsymbol{\theta}$ if model is very rich
  - computationally demanding simulations if $\rho_{\boldsymbol{\theta}}$ is expensive to sample from
- Then
  - biases in case of insufficiently rich model structure
  - validity is limited to some parts of the state-action space
  - only few first moments are correct (typ. mean + variance)

Model-based Pathway - Predictive AI Models (data-driven, ML, etc)

One-step model

AI model e.g.

$$\hat{\mathbf{s}}_{k+1} \sim \rho_{\boldsymbol{\theta}}(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$$

fitting data

Virtual data from "simulating"
$\hat{\mathbf{s}}_{k+1} \sim \rho_{\boldsymbol{\theta}}(\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k)$ forward (Monte Carlo)

In most cases, $\rho_{\boldsymbol{\theta}}$ is a fairly simple representation of reality ρ, easy to sample from

Here we no longer need the model $\rho_{\boldsymbol{\theta}}$ to be "optimization-friendly"

Model parameters $\boldsymbol{\theta}$

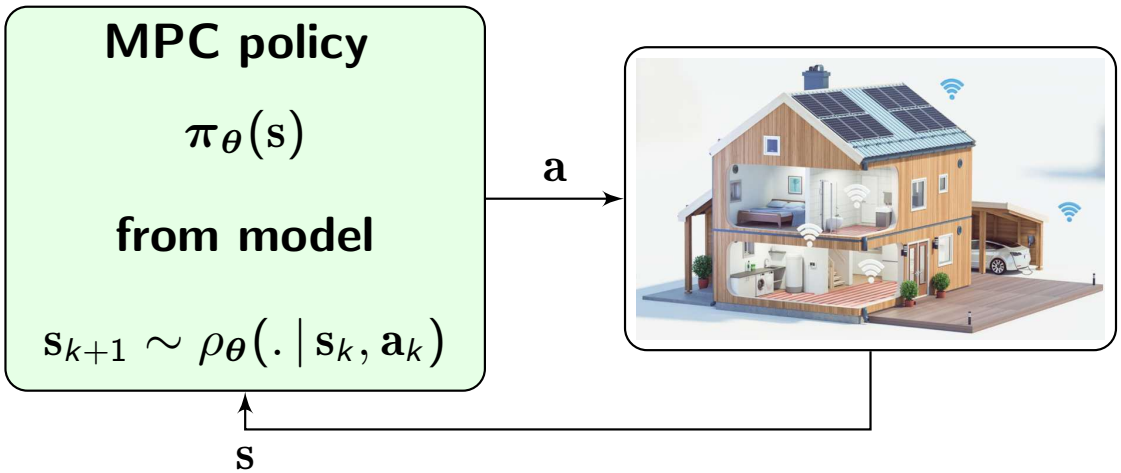
- Selected such that model $\rho_{\boldsymbol{\theta}}$ "ressembles" reality ρ, using data (+ physics)
- Deterministic models are a special case of $\rho_{\boldsymbol{\theta}}$
- Many methods, from Least Squares and MLE to Bayesian and Adversarial Learning
- Still, in most cases $\rho_{\boldsymbol{\theta}}$ "simplifies" ρ because
  - need for huge amounts of data to decide $\boldsymbol{\theta}$ if model is very rich
  - computationally demanding simulations if $\rho_{\boldsymbol{\theta}}$ is expensive to sample from
- Then
  - biases in case of insufficiently rich model structure
  - validity is limited to some parts of the state-action space
  - only few first moments are correct (typ. mean + variance)

Back to MPC - A "Classical" View on Decision Making

Policy $\pi_\theta(\mathrm{s})$ given implicitly by

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbb{E}\left[T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1} L\left(\mathbf{x}_k,\mathbf{u}_k\right)\right] \\ \text{s.t.} \quad & \mathbf{x}_{k+1} \sim \rho_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

use $\pi_\theta(\mathrm{s}) = \mathbf{u}_0^\star$ on the system

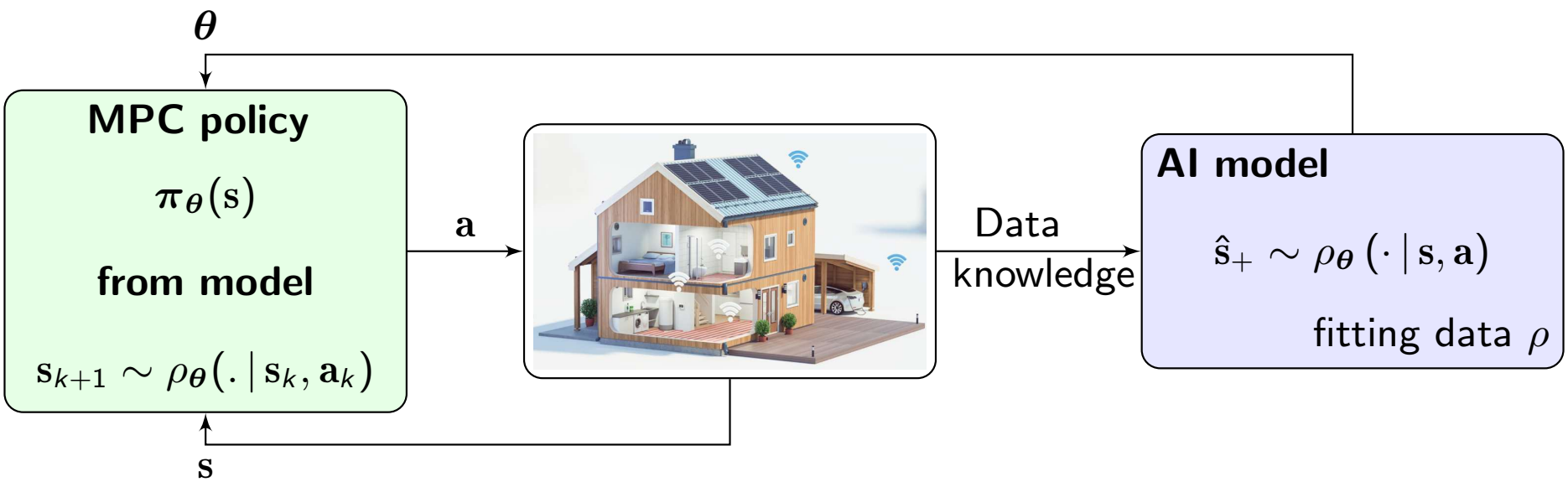


Back to MPC - A "Classical" View on Decision Making

Policy $\pi_\theta(\mathrm{s})$ given implicitly by

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbb{E}\left[T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1} L\left(\mathbf{x}_k,\mathbf{u}_k\right)\right] \\ \text{s.t.} \quad & \mathbf{x}_{k+1} \sim \rho_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

use $\pi_\theta(\mathrm{s}) = \mathbf{u}_0^\star$ on the system



Back to MPC - A "Classical" View on Decision Making

Policy $\pi_\theta(\mathrm{s})$ given implicitly by

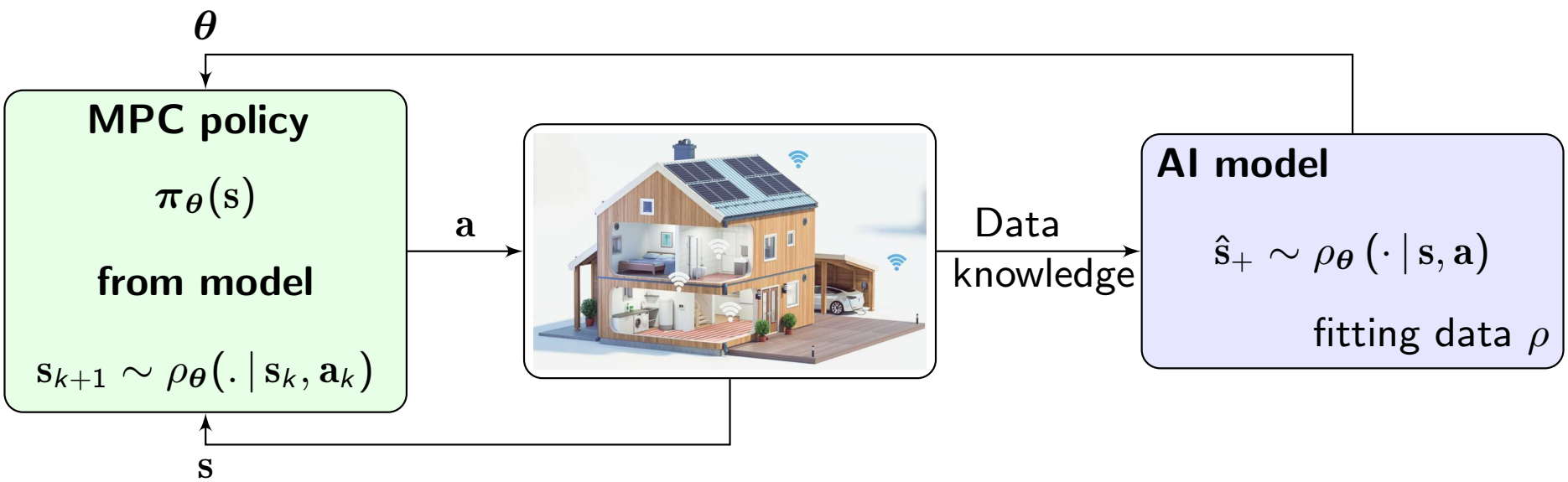
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbb{E}\left[T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \right] \\ \text{s.t.} \quad & \mathbf{x}_{k+1} \sim \rho_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

use $\pi_{\boldsymbol{\theta}}(\mathrm{s}) = \mathbf{u}_0^\star$ on the system

Defines a paradigm...

- Performance from model accuracy

  i.e. $\rho_{\boldsymbol{\theta}}$ "close enough" to ρ
- "Ignore" that we replan all the time



Back to MPC - A "Classical" View on Decision Making

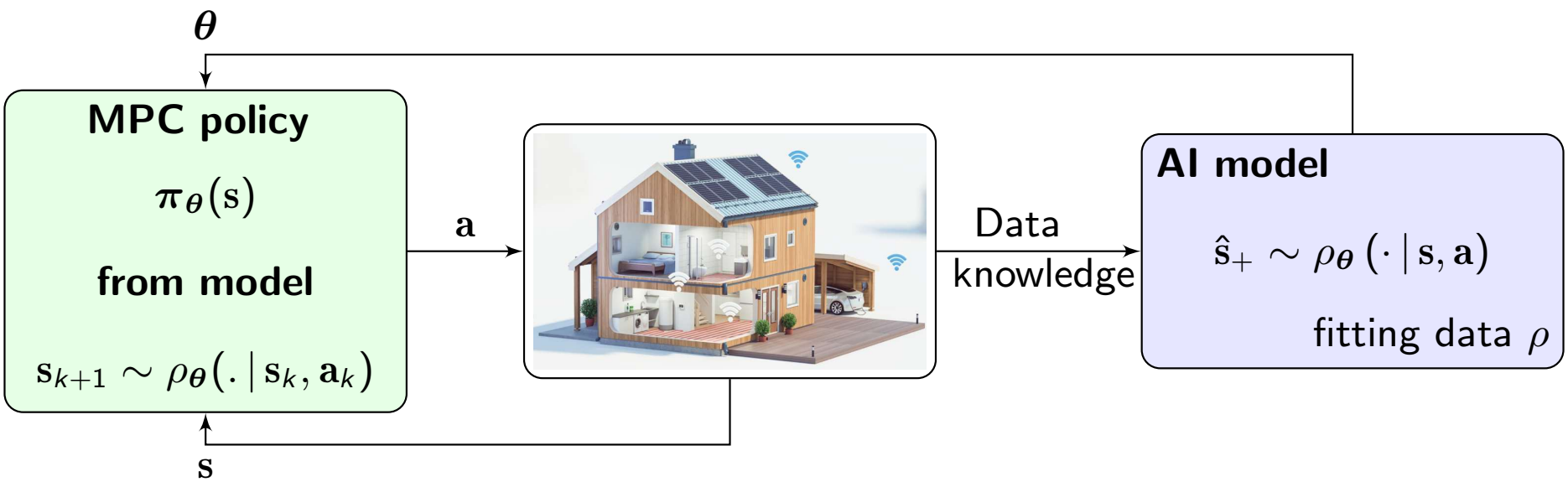
Policy $\pi_\theta(\mathrm{s})$ given implicitly by

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbb{E}\left[T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1} L\left(\mathbf{x}_k,\mathbf{u}_k\right)\right] \\ \text{s.t.} \quad & \mathbf{x}_{k+1} \sim \rho_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

use $\pi_{\boldsymbol{\theta}}(\mathrm{s}) = \mathbf{u}_0^\star$ on the system

Relationship π_θ to $\pi^\star$??

- They match if $\rho_{\boldsymbol{\theta}}$ is exact and deterministic
- But it usually cannot be...



Back to MPC - A "Classical" View on Decision Making

Policy $\pi_\theta(\mathrm{s})$ given implicitly by

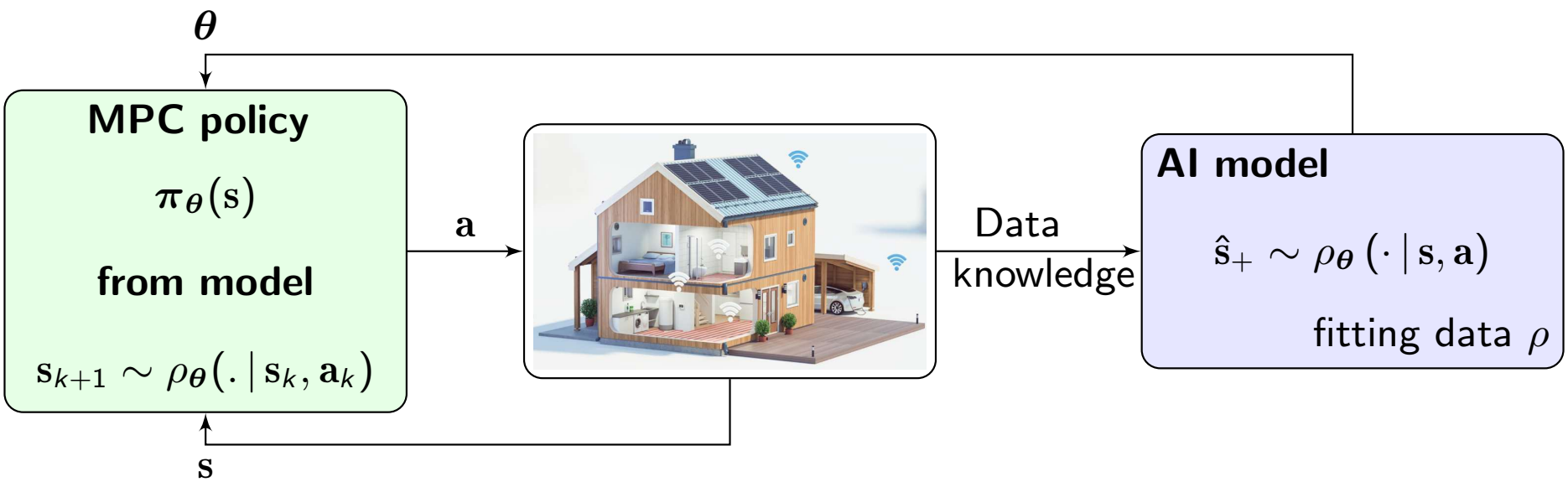
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbb{E}\left[T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right)\right] \\ \text{s.t.} \quad & \mathbf{x}_{k+1} \sim \rho_\theta\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

use $\pi_\theta(\mathrm{s}) = \mathbf{u}_0^\star$ on the system

Relationship π_θ to $\pi^\star$??

- They match if ρ_θ is exact and deterministic
- But it usually cannot be...

"Standard methods" for choosing θ in general do not yield the best π_θ.
RL over MPC can fix that, L becomes part of the model



Back to MPC - A "Classical" View on Decision Making

Policy $\pi_\theta(\mathrm{s})$ given implicitly by

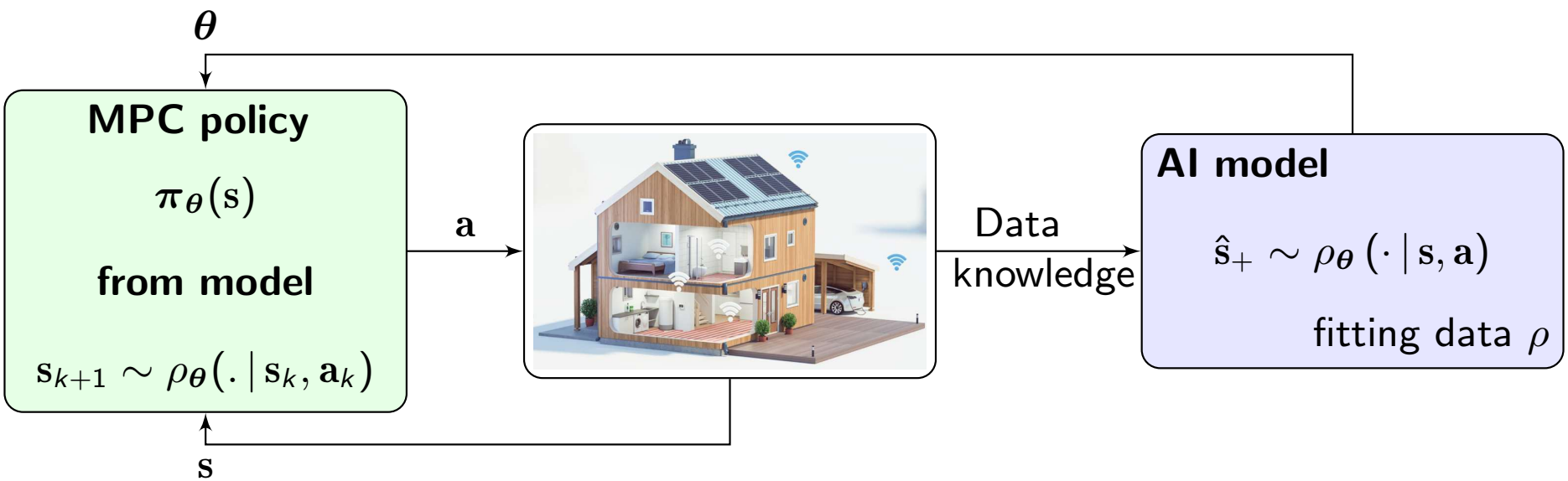
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbb{E}\left[T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1} L\left(\mathbf{x}_k,\mathbf{u}_k\right)\right] \\ \text{s.t.} \quad & \mathbf{x}_{k+1} \sim \rho_\theta\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

use $\pi_\theta(\mathrm{s}) = \mathbf{u}_0^\star$ on the system

Relationship π_θ to $\pi^\star$??

- They match if ρ_θ is exact and deterministic
- But it usually cannot be...

"Standard methods" for choosing θ in general do not yield the best π_θ.
RL over MPC can fix that, L becomes part of the model



This is not the only way of taking decisions from models

Decision policy from AI models (Sim2Real)

AI model

$$\hat{\mathbf{s}}_+ \sim \rho_{\boldsymbol{\theta}}\left(\cdot \,|\, \mathbf{s}, \mathbf{a}\right)$$

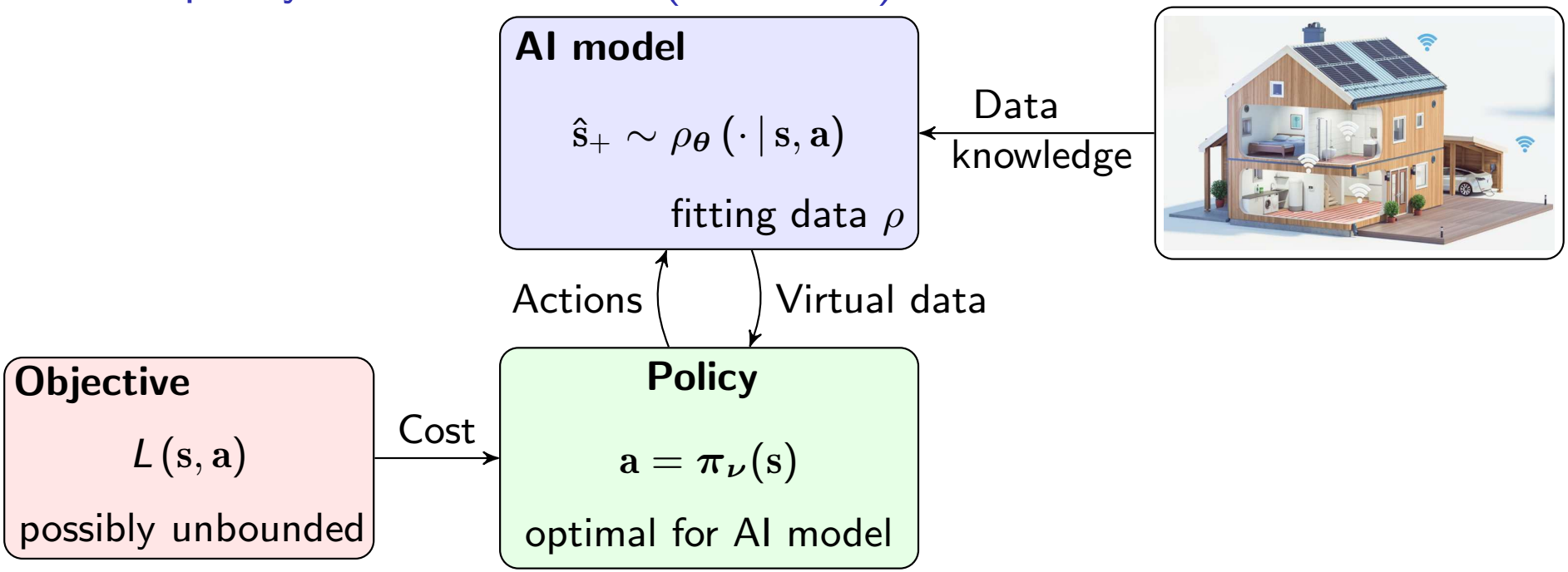
fitting data ρ

Data knowledge

Classical Process:

- Fit AI model $\rho_{\boldsymbol{\theta}}$ to real data

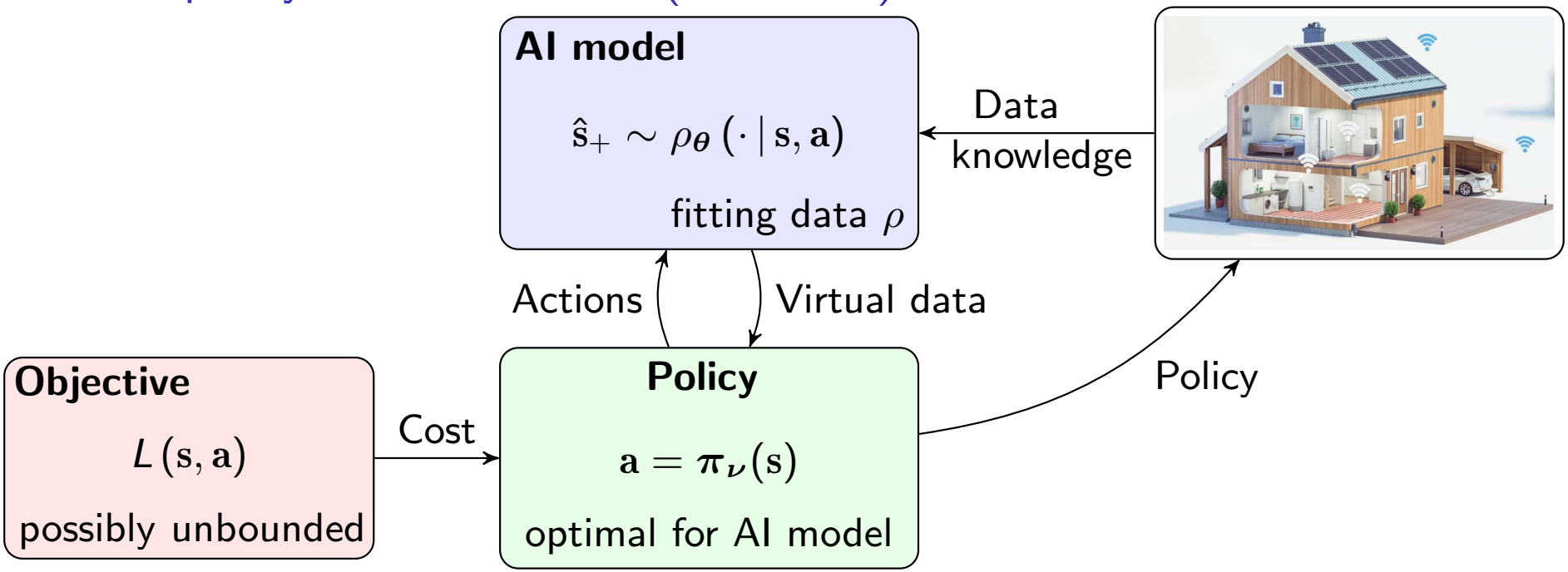
Decision policy from AI models (Sim2Real)



Classical Process:

- Fit AI model $\rho_{\boldsymbol{\theta}}$ to real data
- Develop optimal policy for L and $\rho_{\boldsymbol{\theta}}$ from AI model, e.g. using In-Sim RL
  - Define parametrized policy $\boldsymbol{\pi}_{\boldsymbol{\nu}}$
  - Optimize policy parameters $\boldsymbol{\nu}$ for performance w.r.t. AI model: $\rho_{\boldsymbol{\theta}}, L \rightarrow \boldsymbol{\nu}^\star$

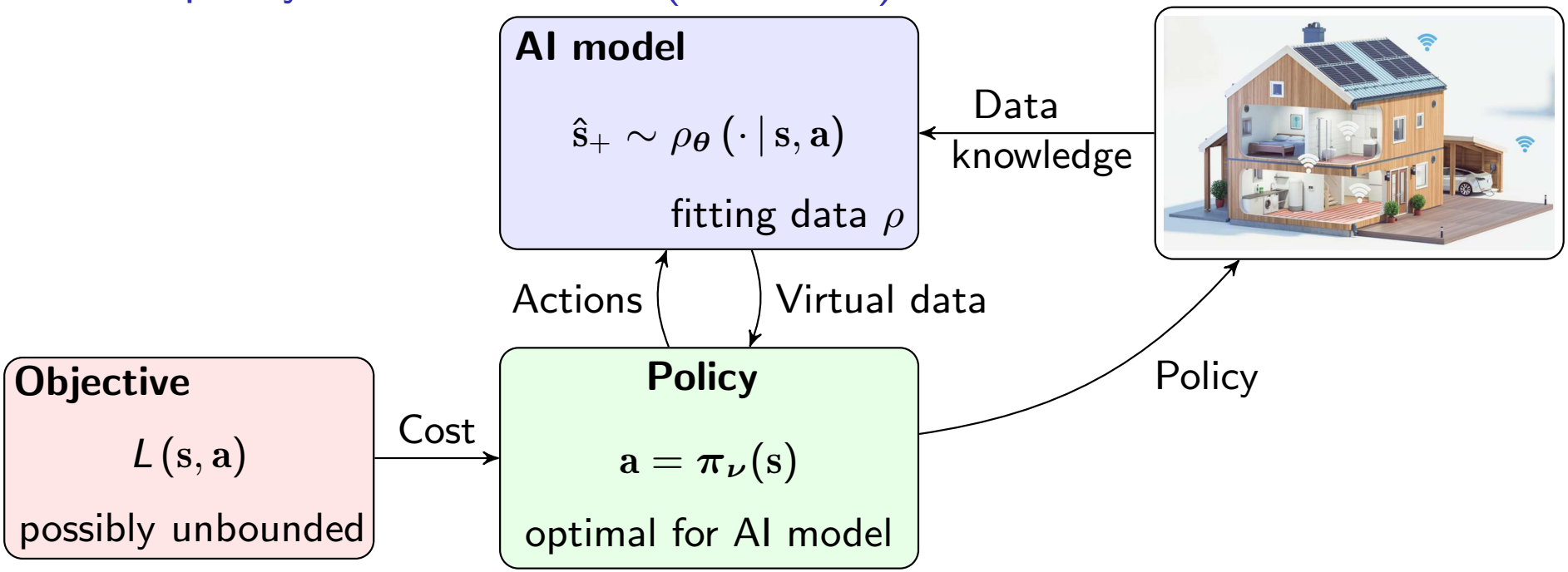
Decision policy from AI models (Sim2Real)



Classical Process:

- Fit AI model $\rho_{\boldsymbol{\theta}}$ to real data
- Develop optimal policy for L and $\rho_{\boldsymbol{\theta}}$ from AI model, e.g. using In-Sim RL
  - Define parametrized policy $\boldsymbol{\pi}_{\boldsymbol{\nu}}$
  - Optimize policy parameters $\boldsymbol{\nu}$ for performance w.r.t. AI model: $\rho_{\boldsymbol{\theta}}, L \rightarrow \boldsymbol{\nu}^\star$
- Transfer policy $\boldsymbol{\pi}_{\boldsymbol{\nu}^\star}$ into the real world

Decision policy from AI models (Sim2Real)

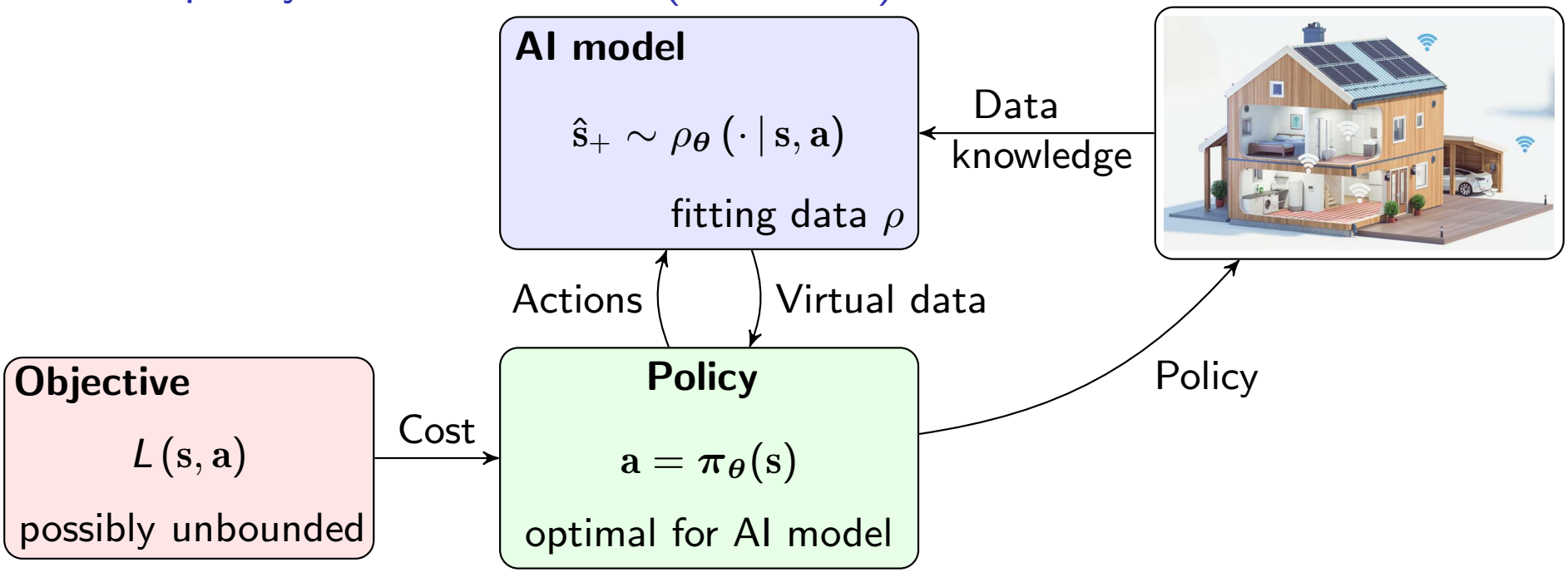


Classical Process:

- Fit AI model $\rho_{\boldsymbol{\theta}}$ to real data
- Develop optimal policy for L and $\rho_{\boldsymbol{\theta}}$ from AI model, e.g. using In-Sim RL
  - Define parametrized policy $\boldsymbol{\pi}_{\boldsymbol{\nu}}$
  - Optimize policy parameters $\boldsymbol{\nu}$ for performance w.r.t. AI model: $\rho_{\boldsymbol{\theta}}, L \rightarrow \boldsymbol{\nu}^\star$
- Transfer policy $\boldsymbol{\pi}_{\boldsymbol{\nu}^\star}$ into the real world

Note: policy parameters $\nu^\star$ optimal for AI model become function of θ
Let's then label $\pi_\theta = \pi_{\nu^\star}$ for simplicity

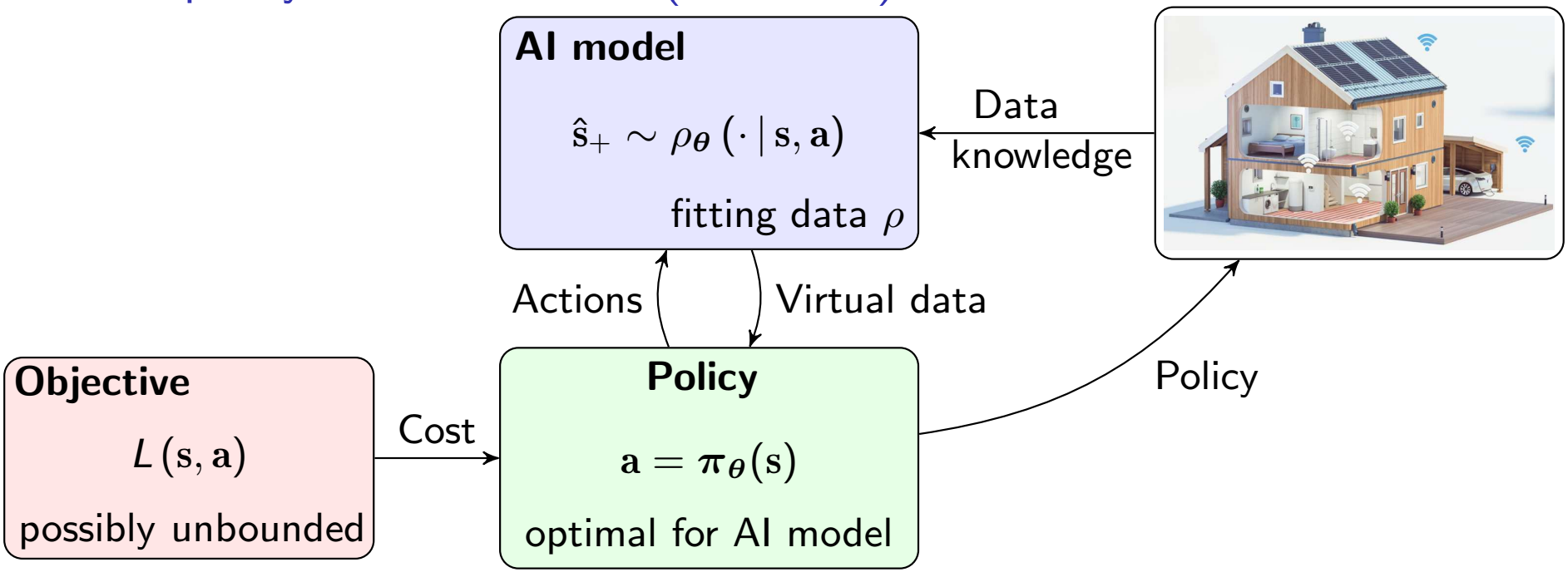
Decision policy from AI models (Sim2Real)



Remarks

- Relationship $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ to $\boldsymbol{\pi}^\star$??
  - They match if $\rho_{\boldsymbol{\theta}} = \rho$, i.e. if model is exact
  - But $\rho_{\boldsymbol{\theta}}$ is (almost always) an approximation

Decision policy from AI models (Sim2Real)



Remarks

- Relationship $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ to $\boldsymbol{\pi}^\star$??
  - They match if $\rho_{\boldsymbol{\theta}} = \rho$, i.e. if model is exact
  - But $\rho_{\boldsymbol{\theta}}$ is (almost always) an approximation

Are "standard methods" for choosing θ resulting in a good policy π_θ??
Empirically the answer is "no". Then how to choose θ to get a good policy?

MPC or Sim2Real - It's the same question...

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

$$Q^\star(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \rho}[V^\star(\mathbf{s}_+)|\mathbf{s}, \mathbf{a}]$$

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

$$Q^\star(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \rho}[V^\star(\mathbf{s}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\boldsymbol{\pi}^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

$$Q^\star(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \rho}[V^\star(\mathbf{s}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\pi^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

Model MDP (simulation environment)

Model MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$\hat{V}^\star(\mathbf{s}) = \min_{\mathbf{a}} \hat{Q}^\star(\mathbf{s}, \mathbf{a})$$

$$\hat{Q}^\star(\mathbf{s}, \mathbf{a}) = \hat{L}(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \hat{\rho}}[V^\star(\hat{\mathbf{s}}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\hat{\pi}^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} \hat{Q}^\star(\mathbf{s}, \mathbf{a})$$

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

$$Q^\star(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \rho}[V^\star(\mathbf{s}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\pi^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

Model MDP (simulation environment)

Model MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$\hat{V}^\star(\mathbf{s}) = \min_{\mathbf{a}} \hat{Q}^\star(\mathbf{s}, \mathbf{a})$$

$$\hat{Q}^\star(\mathbf{s}, \mathbf{a}) = \hat{L}(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \hat{\rho}}[V^\star(\hat{\mathbf{s}}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\hat{\pi}^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} \hat{Q}^\star(\mathbf{s}, \mathbf{a})$$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

MPC or Sim2Real - It's the same question...

The theory is about model-based decisions, and equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

$$Q^\star(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \rho}[V^\star(\mathbf{s}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\pi^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

Model MDP (simulation environment)

Model MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$\hat{V}^\star(\mathbf{s}) = \min_{\mathbf{a}} \hat{Q}^\star(\mathbf{s}, \mathbf{a})$$

$$\hat{Q}^\star(\mathbf{s}, \mathbf{a}) = \hat{L}(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \hat{\rho}}[V^\star(\hat{\mathbf{s}}_+)|\mathbf{s}, \mathbf{a}]$$

Optimal policy

$$\hat{\pi}^\star(\mathbf{s}) = \arg\min_{\mathbf{a}} \hat{Q}^\star(\mathbf{s}, \mathbf{a})$$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Proof: telescopic sum, some non-trivial assumptions to prevent $\infty - \infty$ cancellations

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

In-Sim policy training uses a Model MDP

- Practitioners mostly work on model $\hat{\rho}$, but "unsure" on how to tune it
- In the ML communities, people talk about "value alignment", we are trying to make this "MDP equivalence" understood

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it to the real world (classical fitting)

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it to the real world (classical fitting)
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then we can have support $\hat{\rho} \subset$ support of ρ

More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

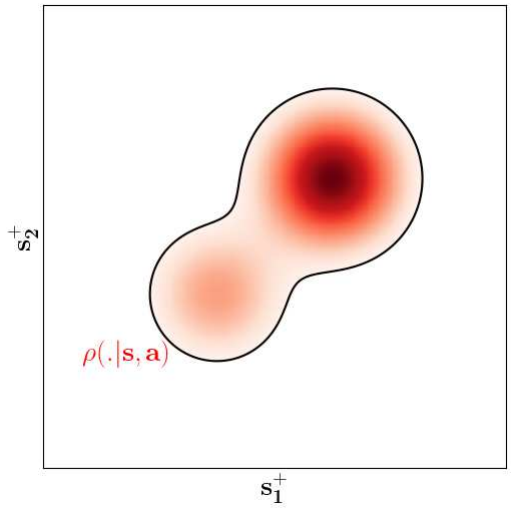
- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it to the real world (classical fitting)
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then we can have support $\hat{\rho} \subset$ support of ρ



More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

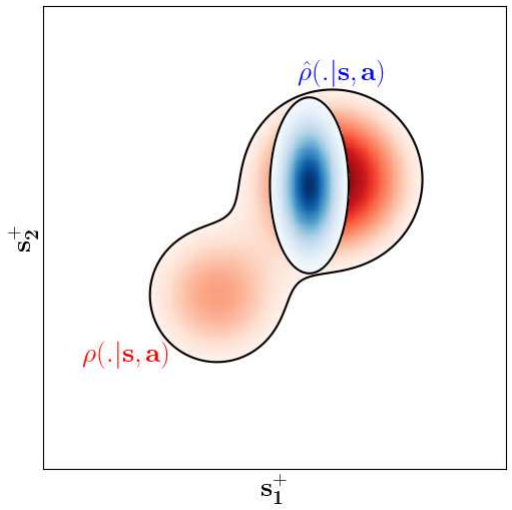
- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it to the real world (classical fitting)
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then we can have support $\hat{\rho} \subset$ support of ρ



More on MDP Equivalence

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

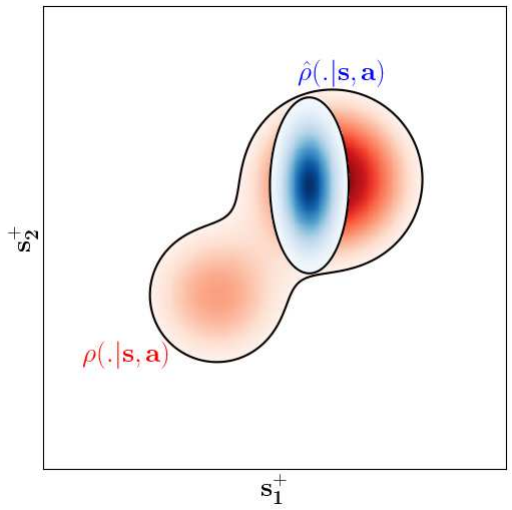
- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a (non-unique) "optimal" model $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

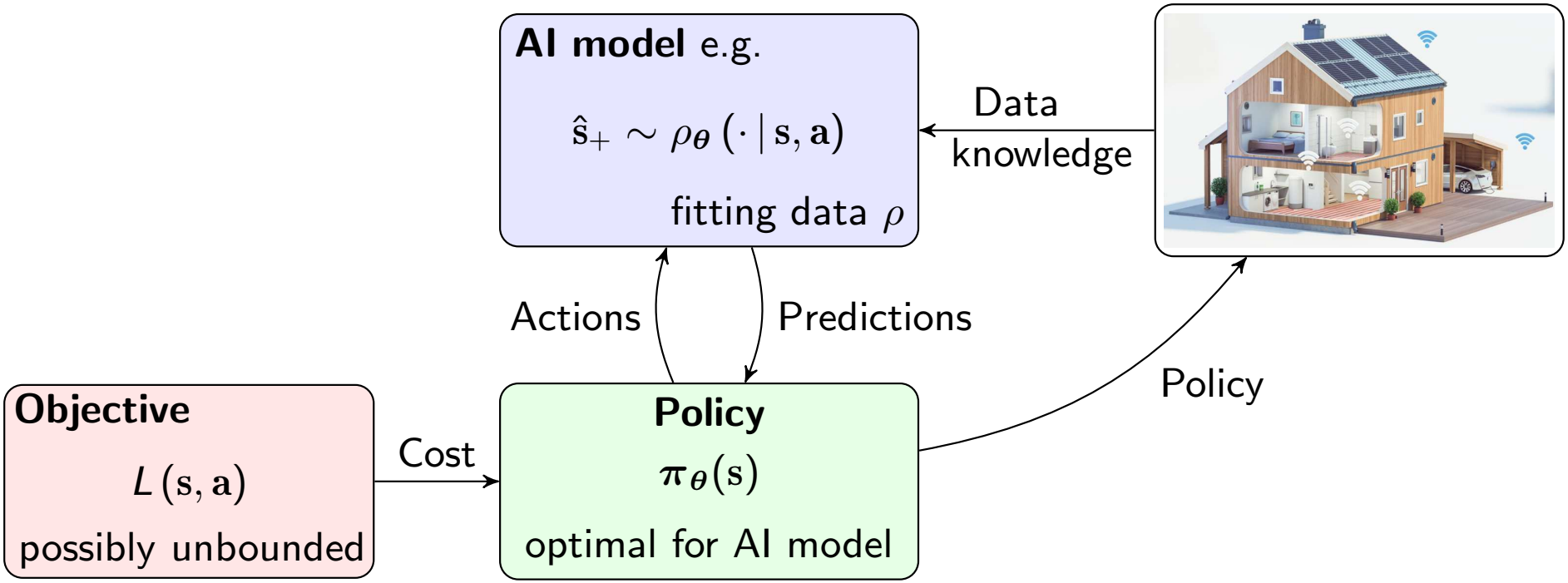
Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it to the real world (classical fitting)
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then we can have support $\hat{\rho} \subset$ support of ρ
- *More simply said*: we can build optimal models $\hat{\rho}$ that make "plausible" predictions about the real world.



Outline

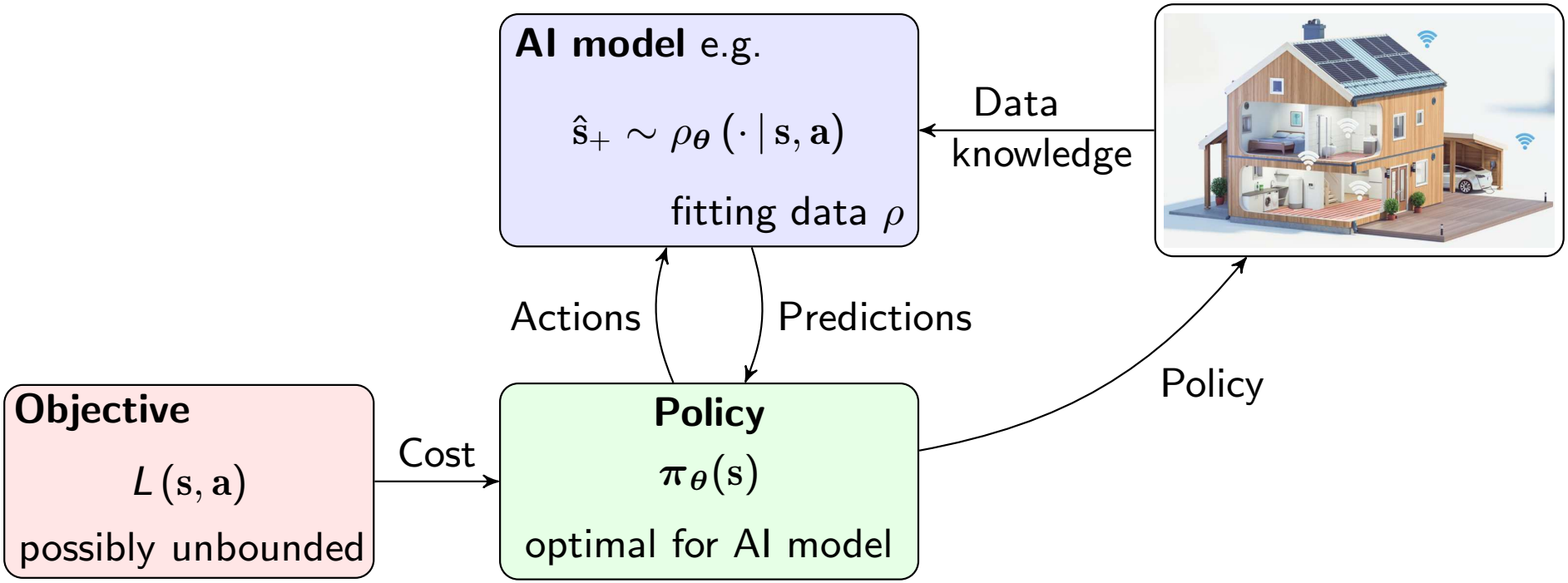
AI models for Decisions?



Learning model ρ_θ should aim at identifying

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) \tag{1}$$

AI models for Decisions?

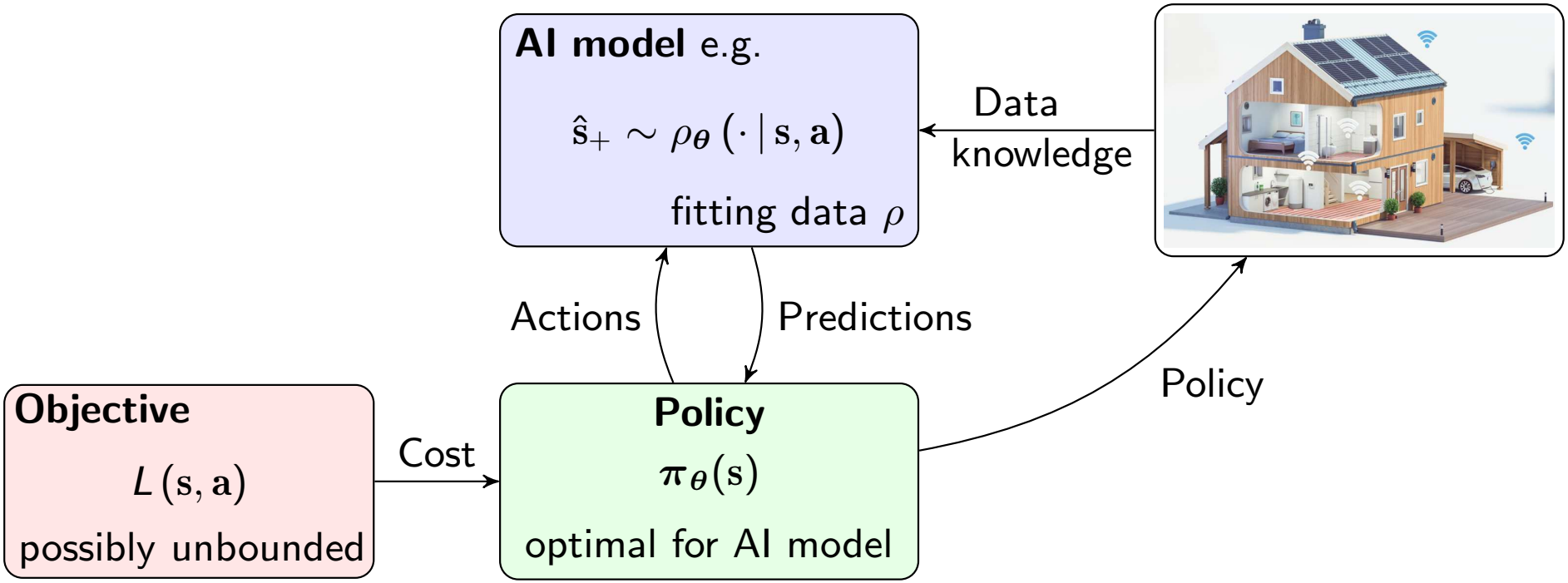


Learning model ρ_θ should aim at identifying

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) \tag{1}$$

- Closed-loop performance J is intricate ($\boldsymbol{\theta} \rightarrow$ simulations $\rightarrow$ policy $\rightarrow$ real system)

AI models for Decisions?

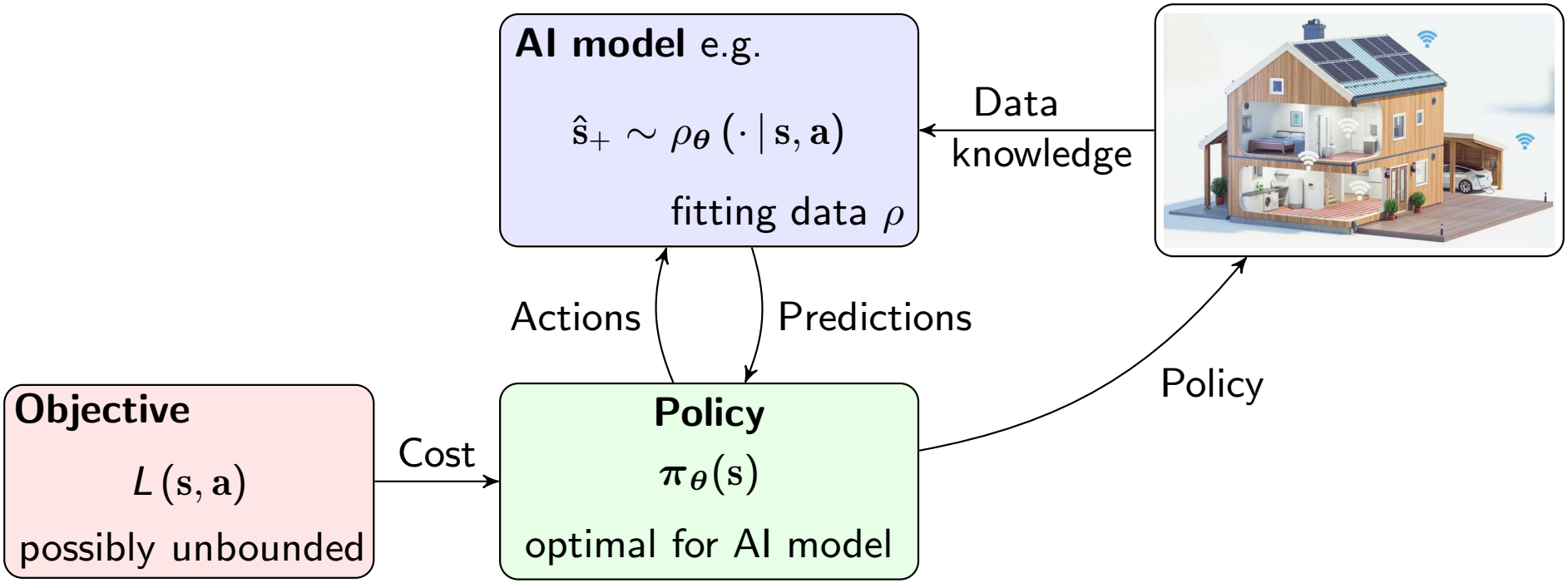


Learning model ρ_θ should aim at identifying

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) \tag{1}$$

- Closed-loop performance J is intricate ($\boldsymbol{\theta} \rightarrow$ simulations $\rightarrow$ policy $\rightarrow$ real system)
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$ is to be estimated from data, i.e. data hungry, noisy, possible biases

AI models for Decisions?

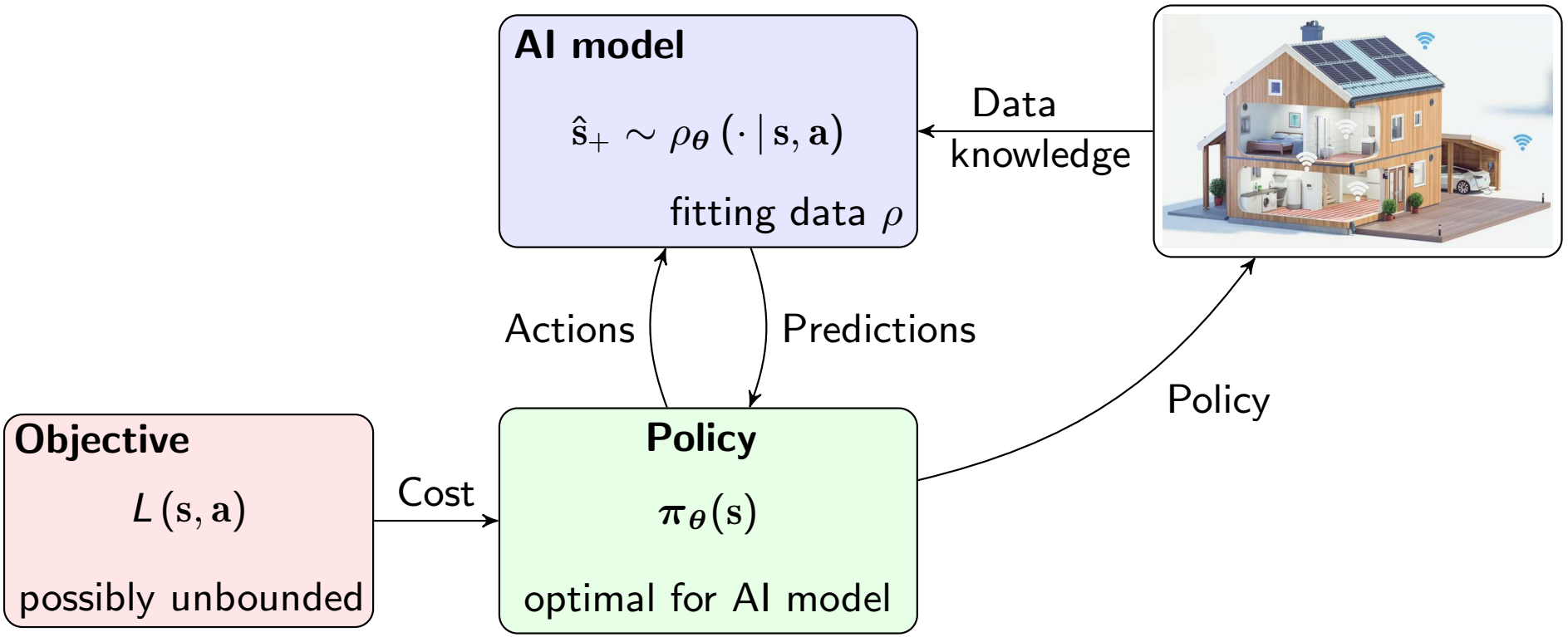


Learning model ρ_θ should aim at identifying

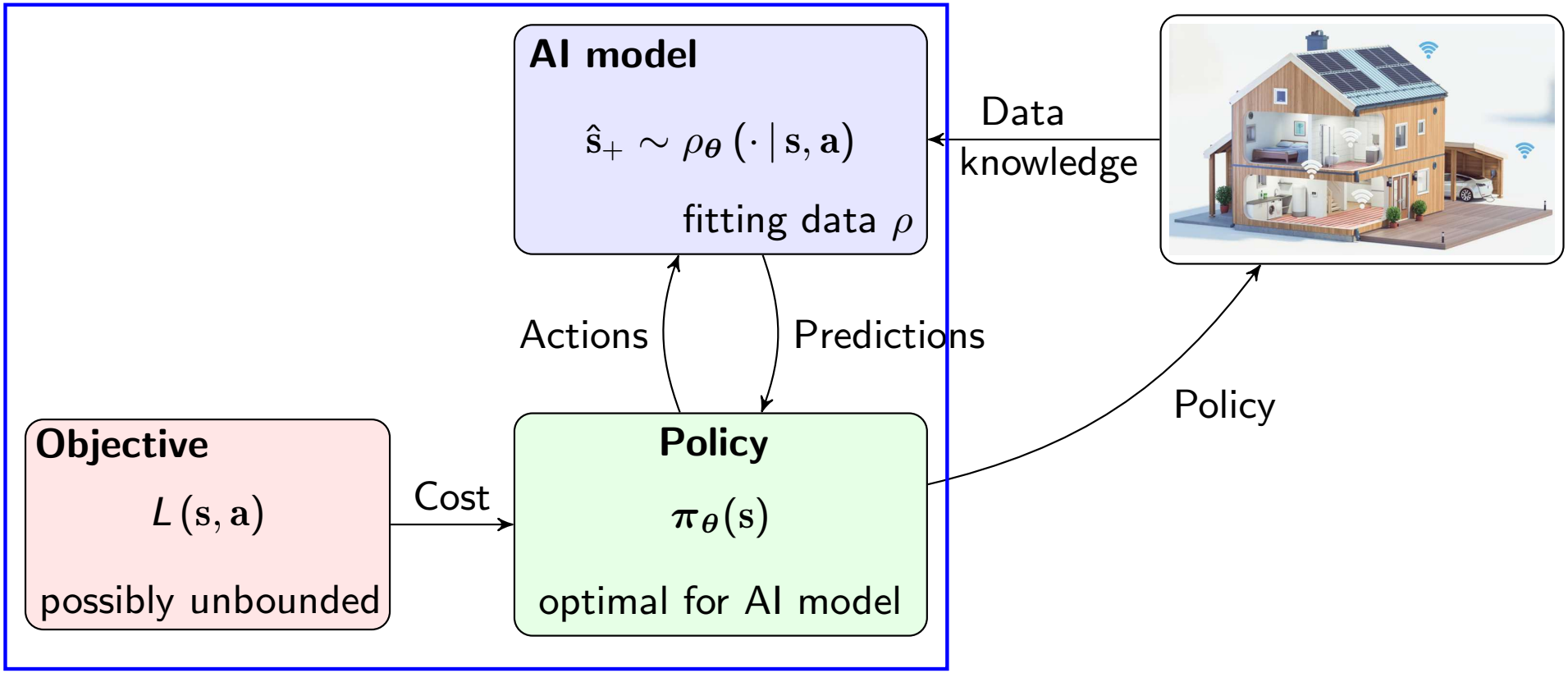
$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) \tag{1}$$

- Closed-loop performance J is intricate ($\boldsymbol{\theta} \rightarrow$ simulations $\rightarrow$ policy $\rightarrow$ real system)
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$ is to be estimated from data, i.e. data hungry, noisy, possible biases
- (1) is in general different than model fitting, i.e. no loss function does (1)

Optimal AI models for Decision - A Paradigm Shift

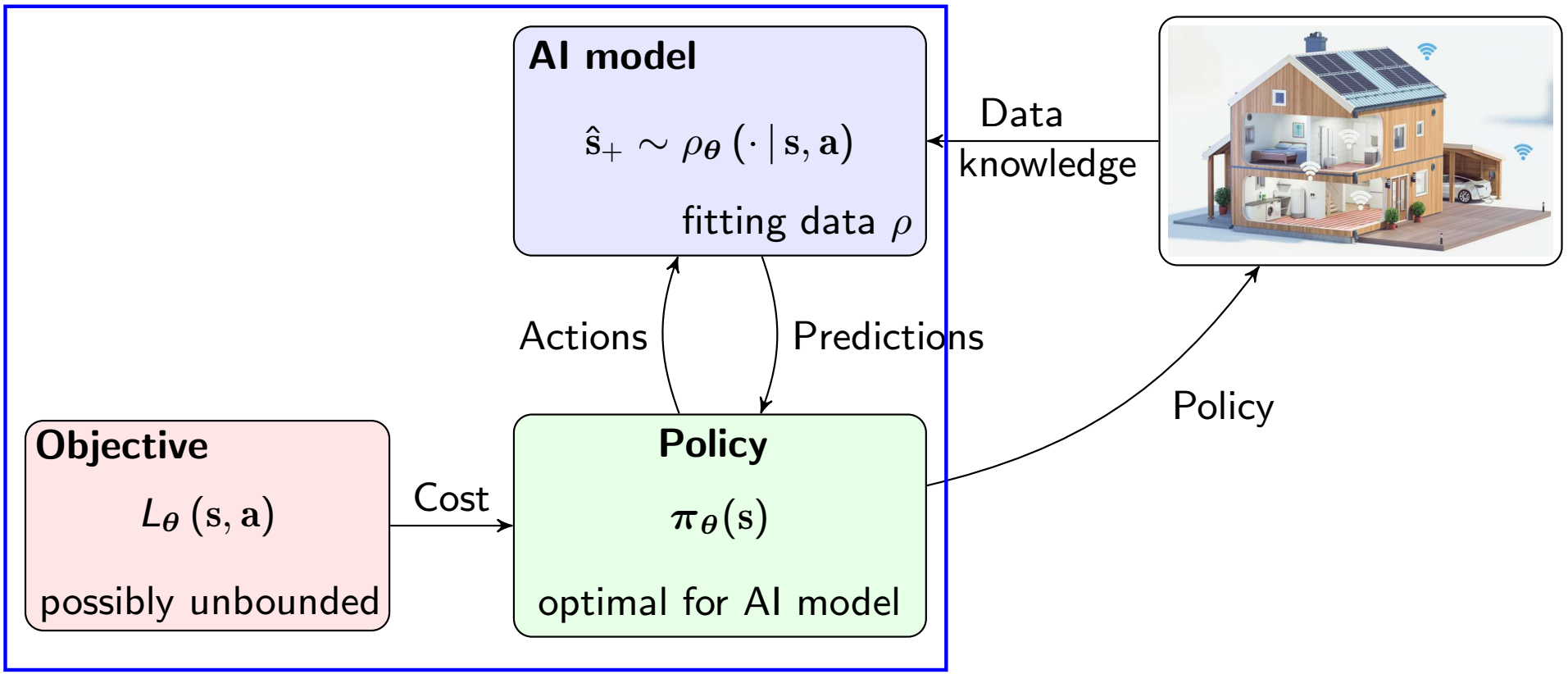


Optimal AI models for Decision - A Paradigm Shift



- The model is not $\rho_{\boldsymbol{\theta}}$, it is the entire "decision-making box"

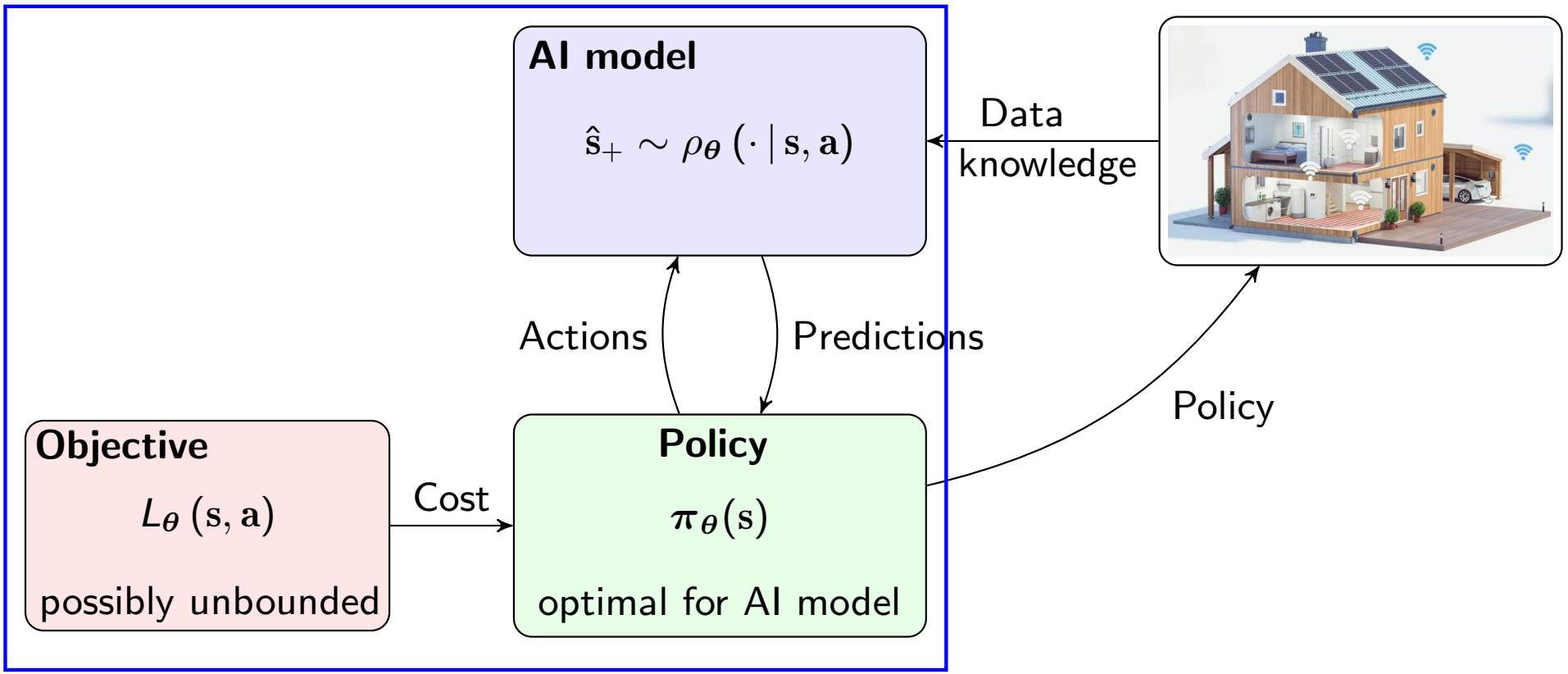
Optimal AI models for Decision - A Paradigm Shift



- The model is not $\rho_{\boldsymbol{\theta}}$, it is the entire "decision-making box"
- The model includes objective $L_{\boldsymbol{\theta}}$ used to build the policy*

* policy performance on real system still assessed via L

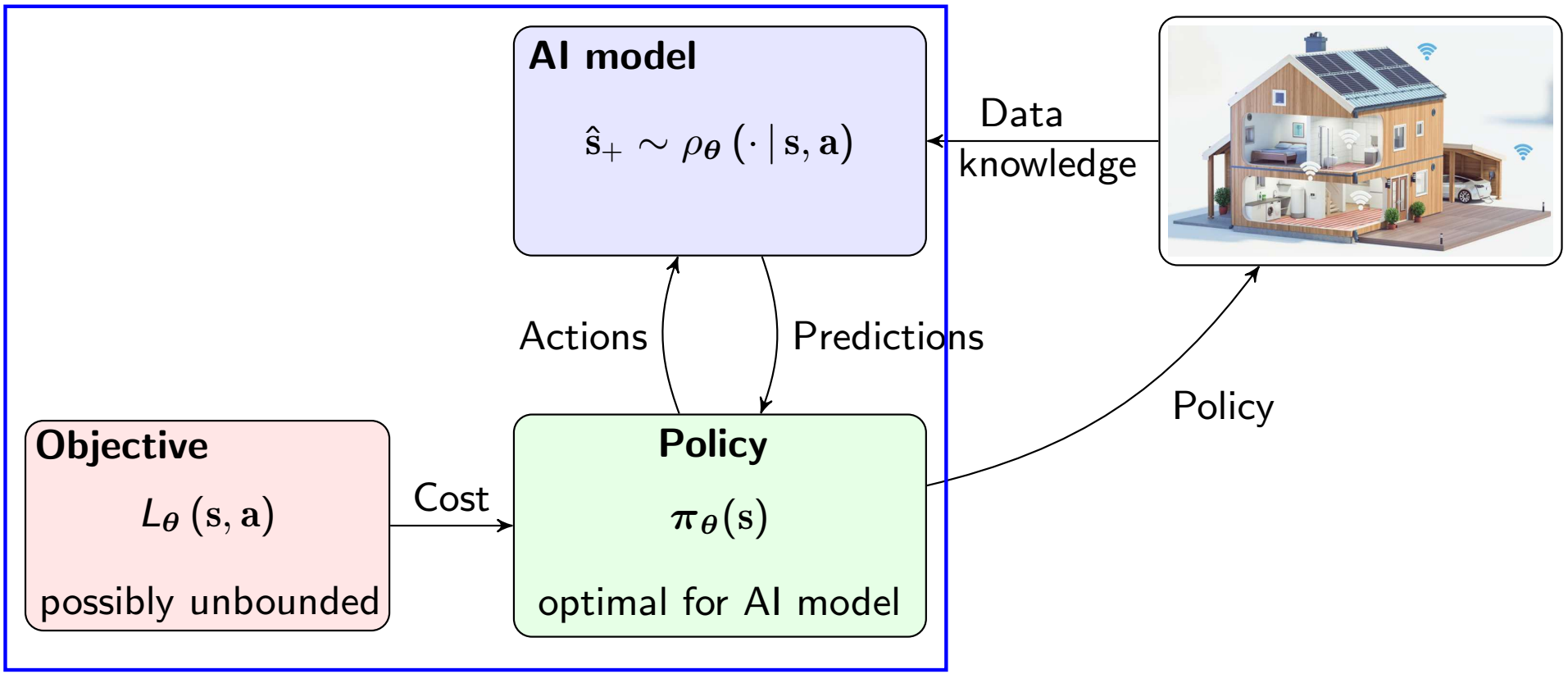
Optimal AI models for Decision - A Paradigm Shift



- The model is not $\rho_{\boldsymbol{\theta}}$, it is the entire "decision-making box"
- The model includes objective $L_{\boldsymbol{\theta}}$ used to build the policy*
- Best model $\rho_{\boldsymbol{\theta}}$ should *not necessarily* represent the data in a "classical sense"

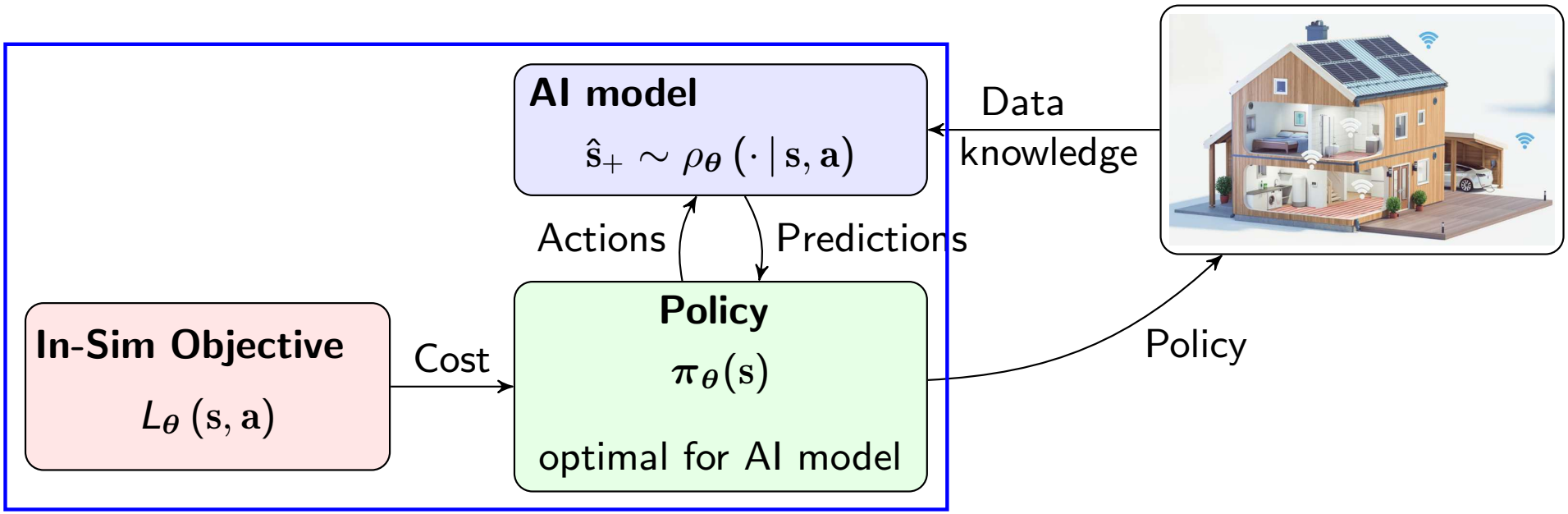
* policy performance on real system still assessed via L

Optimal AI models for Decision - A Paradigm Shift

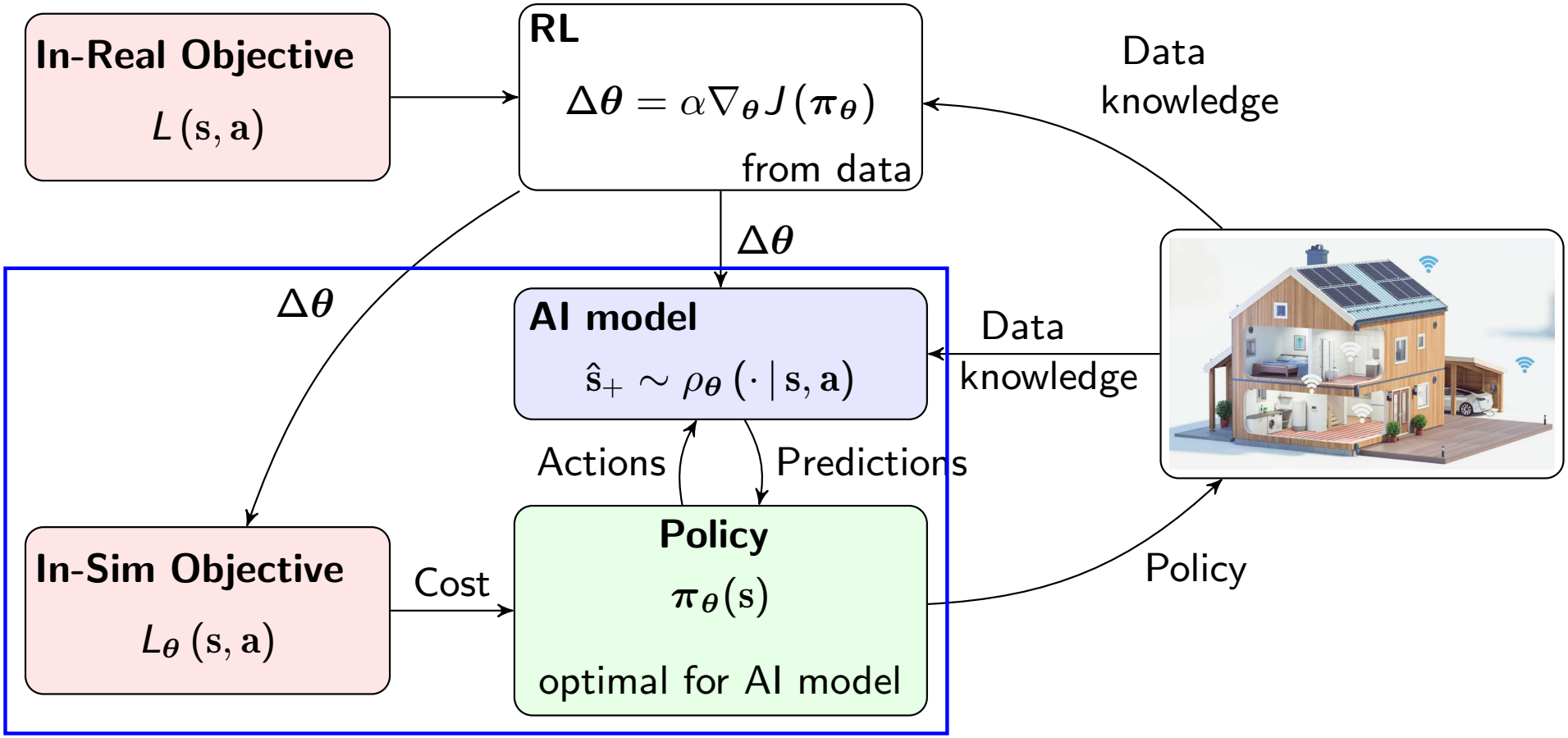


Theorem: under some (technical) assumptions, there is a $L_{\boldsymbol{\theta}}$ such that $\boldsymbol{\pi}_{\boldsymbol{\theta}} = \boldsymbol{\pi}_\star$, even if $\rho_{\boldsymbol{\theta}}$ does not represent real world ρ correctly

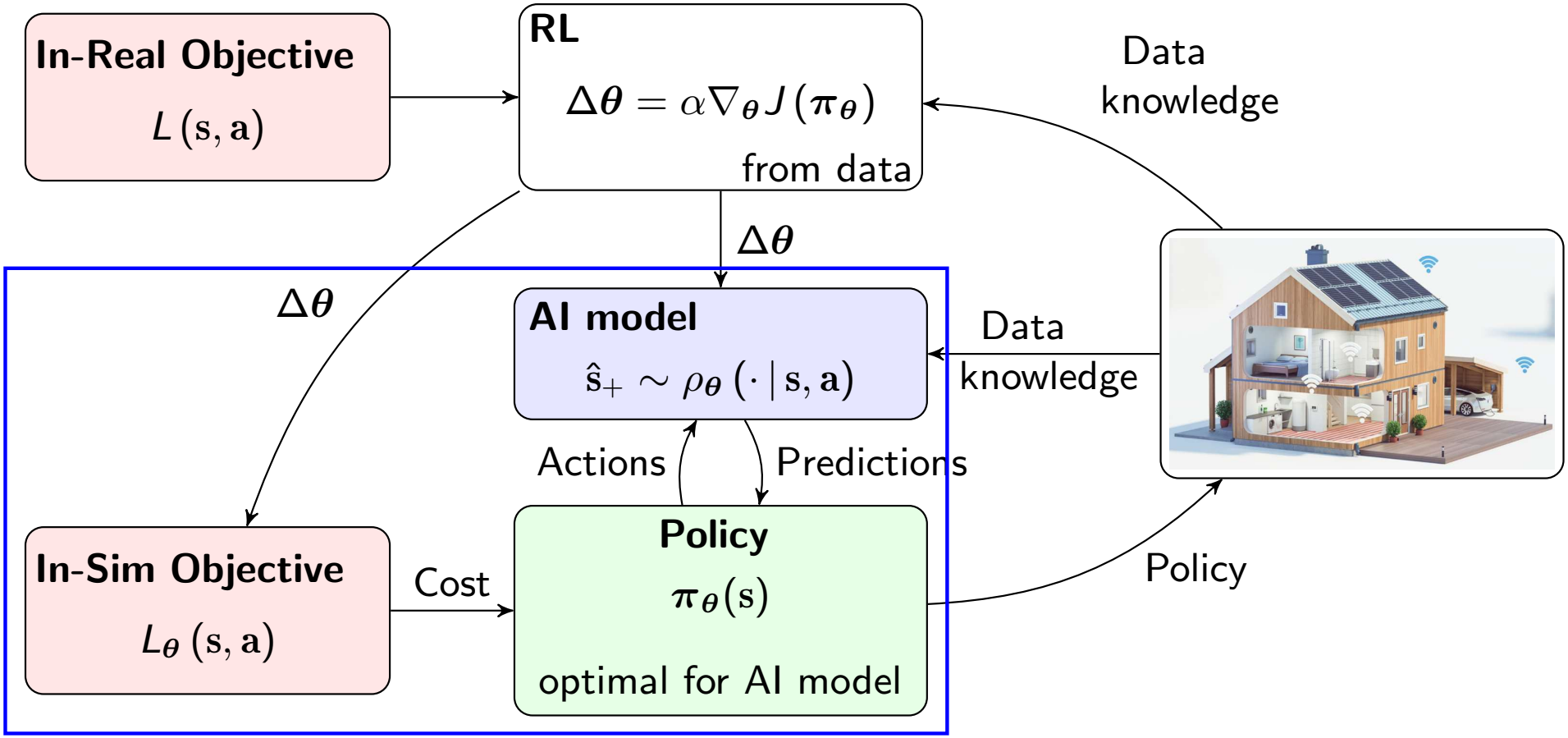
How to use this more generally? RL over Policy Training



How to use this more generally? RL over Policy Training



How to use this more generally? RL over Policy Training

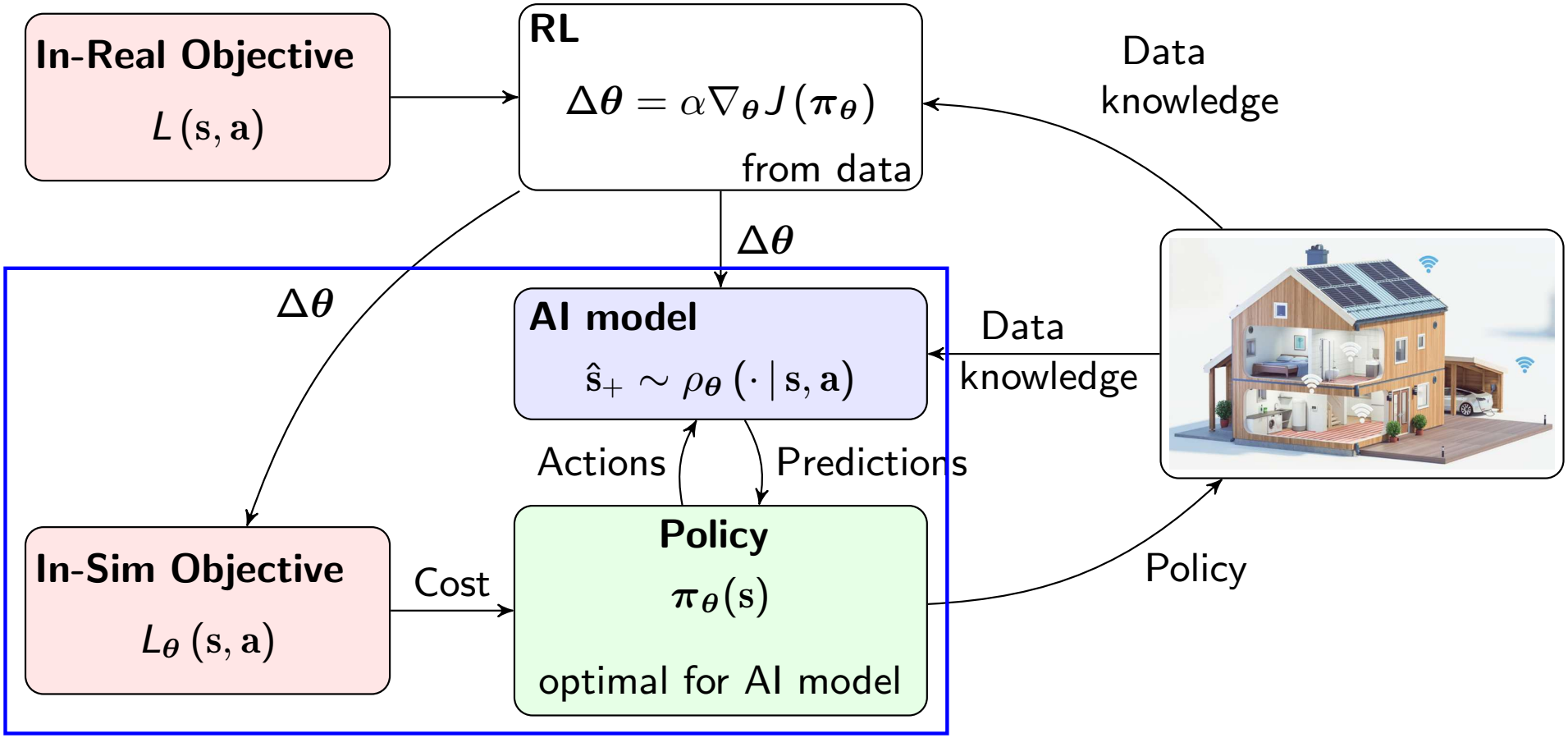


Policy gradient

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{a}} Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right]$$

- $Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is the critic, well-established RL tool
- $\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}$ requires differentiating the closed-loop simulation...

How to use this more generally? RL over Policy Training



Difficulty: computing $\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}$ requires total differentiation through the simulation-based policy optimization process:

$\boldsymbol{\theta} \rightarrow$ simulations $\rightarrow$ optimal policy for $\rho_{\boldsymbol{\theta}}$

Differentiating through policy optimization can be computationally heavy

Policy Differentiation

Performance for a stochastic policy $\pi_{\boldsymbol{\nu}}$, model MDP defined by $L_{\boldsymbol{\theta}}$, $\hat{\rho}_{\boldsymbol{\theta}}$, γ

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\hat{\rho}_{\boldsymbol{\theta}}}\left[\sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k \sim \pi_{\boldsymbol{\nu}}\left[\cdot|\mathbf{s}_k\right]\right]$$

Policy Differentiation

Performance for a stochastic policy $\pi_{\boldsymbol{\nu}}$, model MDP defined by $L_{\boldsymbol{\theta}}$, $\hat{\rho}_{\boldsymbol{\theta}}$, γ

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\hat{\rho}_{\boldsymbol{\theta}}}\left[\sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k \sim \pi_{\boldsymbol{\nu}}\left[\cdot|\mathbf{s}_k\right]\right]$$

Policy given by

$$\boldsymbol{\nu}^\star = \arg\min_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \qquad \text{or equivalently} \qquad \frac{\partial}{\partial \boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = 0$$

Policy Differentiation

Performance for a stochastic policy $\pi_{\boldsymbol{\nu}}$, model MDP defined by $L_{\boldsymbol{\theta}}$, $\hat{\rho}_{\boldsymbol{\theta}}$, γ

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\hat{\rho}_{\boldsymbol{\theta}}}\left[\sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k \sim \pi_{\boldsymbol{\nu}}\left[\cdot|\mathbf{s}_k\right]\right]$$

Policy given by

$$\boldsymbol{\nu}^\star = \arg\min_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \qquad \text{or equivalently} \qquad \frac{\partial}{\partial \boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = 0$$

Policy sensitivity: if we change $\boldsymbol{\theta}$ (cost $L_{\boldsymbol{\theta}}$ and simulation $\hat{\rho}_{\boldsymbol{\theta}}$) how does $\pi_{\boldsymbol{\nu}}$ change?

$$\frac{\partial^2}{\partial \boldsymbol{\nu}^2} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \underbrace{\frac{\mathrm{d}\pi_{\boldsymbol{\nu}}}{\mathrm{d}\boldsymbol{\theta}}}_{=\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}} + \frac{\partial^2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\theta}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = 0$$

Policy Differentiation

Performance for a stochastic policy $\pi_{\boldsymbol{\nu}}$, model MDP defined by $L_{\boldsymbol{\theta}}$, $\hat{\rho}_{\boldsymbol{\theta}}$, γ

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\hat{\rho}_{\boldsymbol{\theta}}}\left[\sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k \sim \pi_{\boldsymbol{\nu}}\left[\cdot|\mathbf{s}_k\right] \right]$$

Policy given by

$$\boldsymbol{\nu}^\star = \arg\min_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \qquad \text{or equivalently} \qquad \frac{\partial}{\partial \boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = 0$$

Policy sensitivity: if we change $\boldsymbol{\theta}$ (cost $L_{\boldsymbol{\theta}}$ and simulation $\hat{\rho}_{\boldsymbol{\theta}}$) how does $\pi_{\boldsymbol{\nu}}$ change?

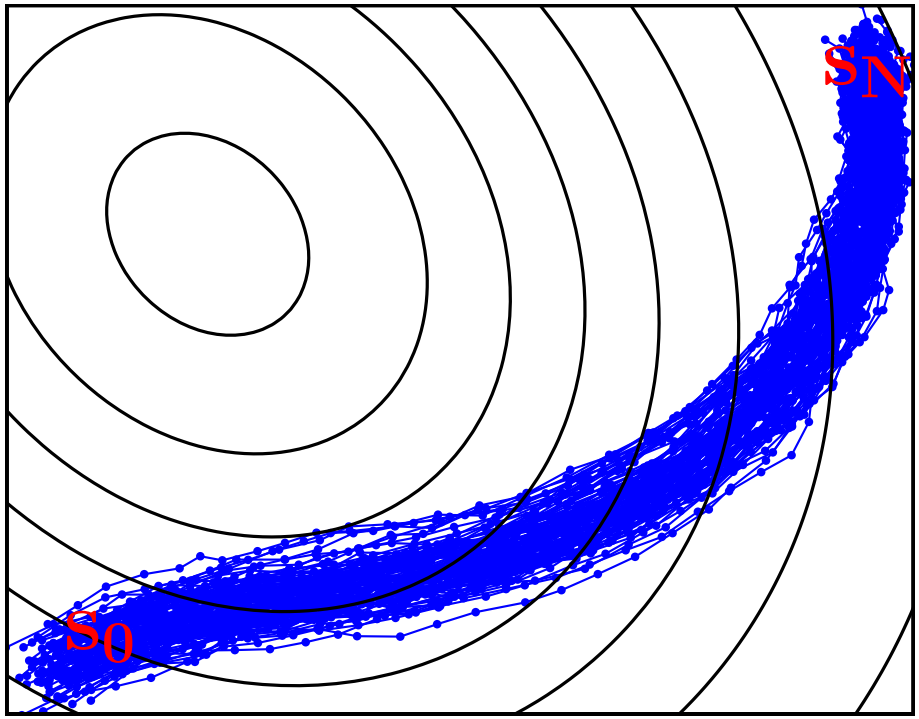
$$\frac{\partial^2}{\partial \boldsymbol{\nu}^2} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \underbrace{\frac{\mathrm{d}\pi_{\boldsymbol{\nu}}}{\mathrm{d}\boldsymbol{\theta}}}_{=\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}} + \frac{\partial^2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\theta}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = 0$$

Simulation: (with $\mathbf{s} \equiv \mathbf{s}_{0,\ldots,N}$ and $\mathbf{a} \equiv \mathbf{a}_{0,\ldots,N}$)

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \int \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \varphi\left(\mathbf{s}, \mathbf{a}\right) \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{a}$$

$$\varphi\left(\mathbf{s}, \mathbf{a}\right) = \rho_0\left[\mathbf{s}_0\right] \prod_{k=0}^{N-1} \hat{\rho}_{\boldsymbol{\theta}}\left[\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k\right] \prod_{k=0}^{N} \pi_{\boldsymbol{\nu}}\left[\mathbf{a}_k|\mathbf{s}_k\right] \qquad \text{(density of the simulation)}$$

Policy Differentiation



- Blue trajectories are realization of φ:

$$\mathbf{s}^i, \mathbf{a}^i \sim \varphi(\cdot, \cdot)$$

- Sample-based evaluation (n samples)

$$\hat{J}_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\nu}}) \approx \frac{1}{n} \sum_{k=i}^{n} \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right)$$

Simulation: (with $\mathbf{s} \equiv \mathbf{s}_{0,\ldots,N}$ and $\mathbf{a} \equiv \mathbf{a}_{0,\ldots,N}$)

$$\hat{J}_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\nu}}) = \int \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}(\mathbf{s}_k, \mathbf{a}_k) \varphi(\mathbf{s}, \mathbf{a}) \, \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{a}$$

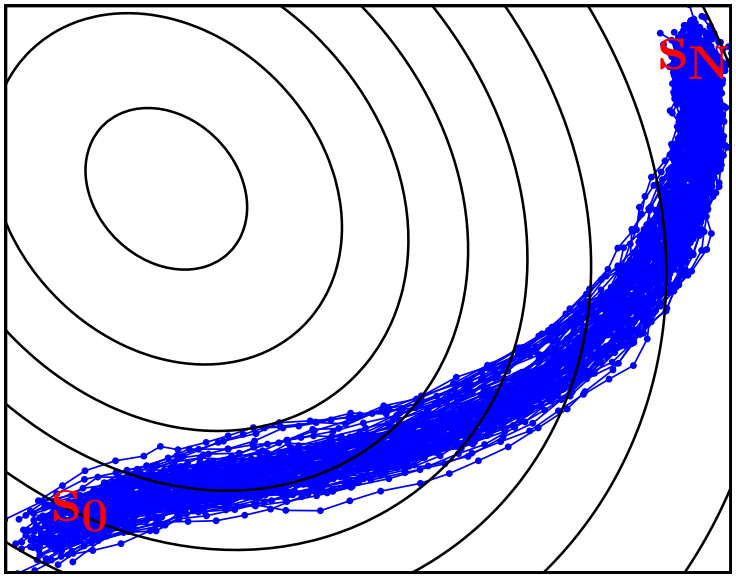
$$\varphi(\mathbf{s}, \mathbf{a}) = \rho_0[\mathbf{s}_0] \prod_{k=0}^{N-1} \hat{\rho}_{\boldsymbol{\theta}}[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k] \prod_{k=0}^{N} \pi_{\boldsymbol{\nu}}[\mathbf{a}_k | \mathbf{s}_k] \qquad \text{(density of the simulation)}$$

Direct Simulation Differentiation

Simulation

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \int \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \varphi\left(\mathbf{s}, \mathbf{a}\right) \mathrm{dsda}$$

$$\varphi\left(\mathbf{s}, \mathbf{a}\right) = \rho_0\left[\mathbf{s}_0\right] \prod_{k=0}^{N-1} \hat{\rho}_{\boldsymbol{\theta}}\left[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k\right] \prod_{k=0}^{N} \pi_{\boldsymbol{\nu}}\left[\mathbf{a}_k | \mathbf{s}_k\right]$$

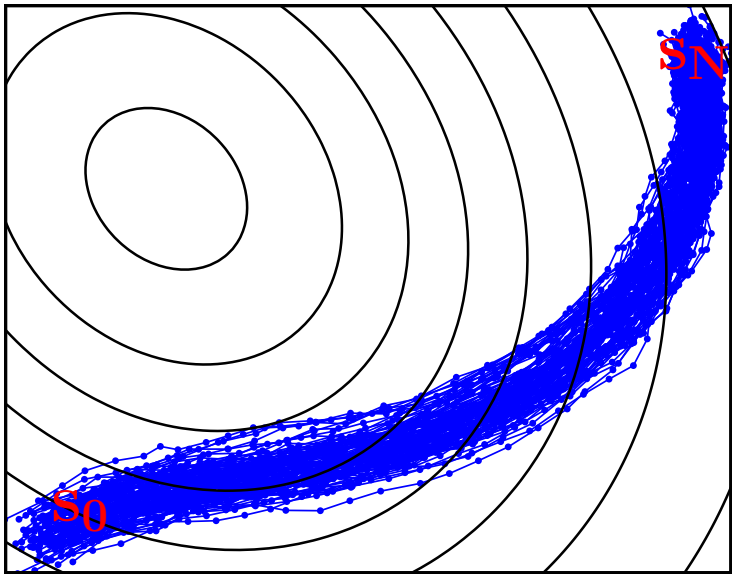


Direct Simulation Differentiation

Simulation

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \int \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \varphi\left(\mathbf{s}, \mathbf{a}\right) \mathrm{dsda}$$

$$\varphi\left(\mathbf{s}, \mathbf{a}\right) = \rho_0\left[\mathbf{s}_0\right] \prod_{k=0}^{N-1} \hat{\rho}_{\boldsymbol{\theta}}\left[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k\right] \prod_{k=0}^{N} \pi_{\boldsymbol{\nu}}\left[\mathbf{a}_k | \mathbf{s}_k\right]$$



Direct Differentiation:

$$\partial \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\hat{\rho}_{\boldsymbol{\theta}}}\left[\sum_{k=0}^{N} \gamma^k \left(\partial L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) + L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \partial \log \varphi\left(\mathbf{s}, \mathbf{a}\right) \right) \middle| \mathbf{a}_k \sim \pi_{\boldsymbol{\nu}}\left[\cdot | \mathbf{s}_k\right] \right]$$

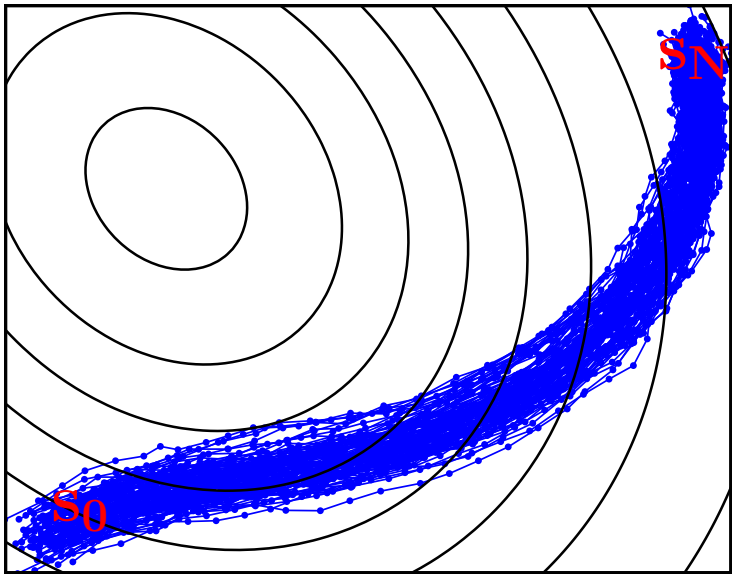
can be evaluated from data

Direct Simulation Differentiation

Simulation

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \int \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \varphi\left(\mathbf{s}, \mathbf{a}\right) \mathrm{dsda}$$

$$\varphi\left(\mathbf{s}, \mathbf{a}\right) = \rho_0\left[\mathbf{s}_0\right] \prod_{k=0}^{N-1} \hat{\rho}_{\boldsymbol{\theta}}\left[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k\right] \prod_{k=0}^{N} \pi_{\boldsymbol{\nu}}\left[\mathbf{a}_k | \mathbf{s}_k\right]$$



Direct Differentiation:

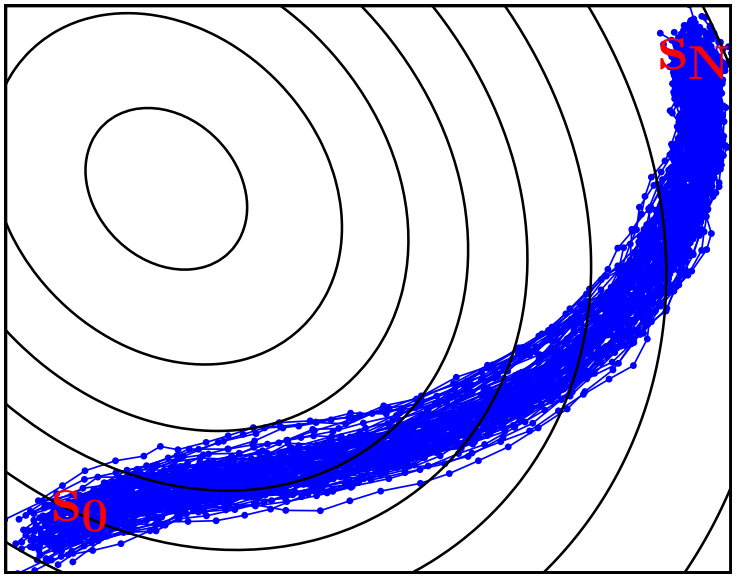
$$\partial \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{k=0}^{N} \gamma^k \left(\partial L_{\boldsymbol{\theta}}\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right) + L_{\boldsymbol{\theta}}\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right) \partial \log \varphi\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right) \right)$$

Direct Simulation Differentiation

Simulation

$$\hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \int \sum_{k=0}^{N} \gamma^k L_{\boldsymbol{\theta}}\left(\mathbf{s}_k, \mathbf{a}_k\right) \varphi\left(\mathbf{s}, \mathbf{a}\right) \mathrm{dsda}$$

$$\varphi\left(\mathbf{s}, \mathbf{a}\right) = \rho_0\left[\mathbf{s}_0\right] \prod_{k=0}^{N-1} \hat{\rho}_{\boldsymbol{\theta}}\left[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k\right] \prod_{k=0}^{N} \pi_{\boldsymbol{\nu}}\left[\mathbf{a}_k | \mathbf{s}_k\right]$$



Direct Differentiation:

$$\partial \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{k=0}^{N} \gamma^k \left(\partial L_{\boldsymbol{\theta}}\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right) + L_{\boldsymbol{\theta}}\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right) \partial \log \varphi\left(\mathbf{s}_k^i, \mathbf{a}_k^i\right) \right)$$

Remarks

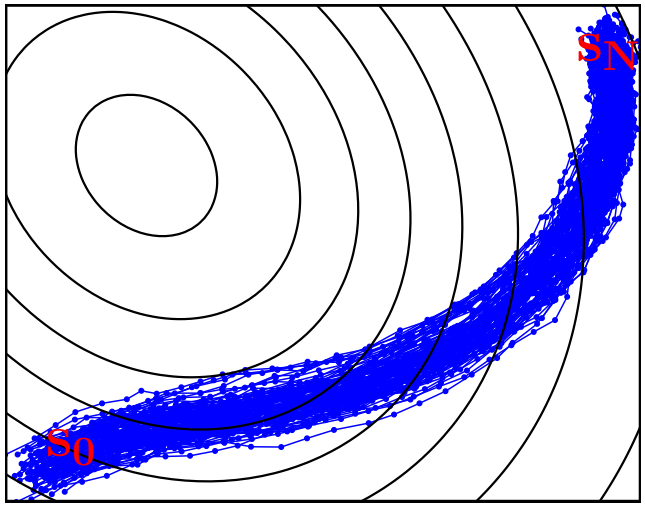
- ✓ Second-order sensitivities are similar
- ✓ Simple and computationally very efficient
- ✓ Can differentiate through discrete state and action sets
- ✘ Sample-based estimations are very noisy

Critic-Based Differentiation

In-Sim Actor-critic does

$$\nabla_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\substack{\mathbf{s} \sim \rho_{\boldsymbol{\theta}} \\ \mathbf{a} \sim \pi_{\boldsymbol{\nu}}}} \left[\nabla_{\boldsymbol{\nu}} \log \pi_{\boldsymbol{\nu}}\left(\mathbf{a}|\mathbf{s}\right) A^{\boldsymbol{\pi}_{\boldsymbol{\nu}}}\left(\mathbf{s}, \mathbf{a}\right)\right] = 0$$

defines relationship $\pi_{\boldsymbol{\nu}^\star}$ with $\boldsymbol{\nu}^\star$ function of $\boldsymbol{\theta}$



Critic-Based Differentiation

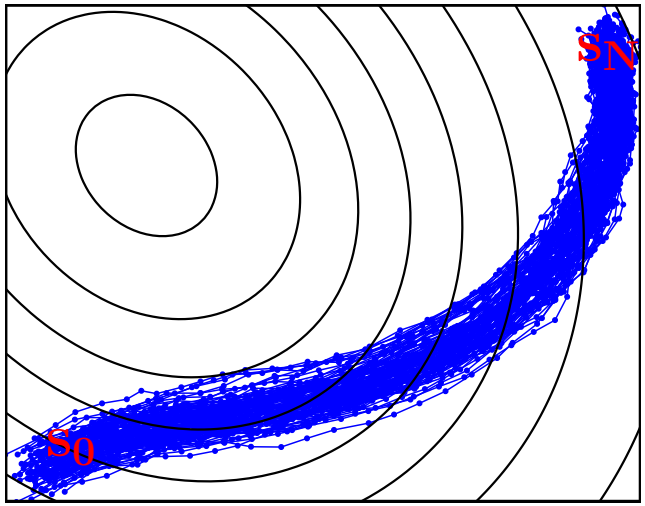
In-Sim Actor-critic does

$$\nabla_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\substack{\mathbf{s} \sim \rho_{\boldsymbol{\theta}} \\ \mathbf{a} \sim \pi_{\boldsymbol{\nu}}}} \left[\nabla_{\boldsymbol{\nu}} \log \pi_{\boldsymbol{\nu}}\left(\mathbf{a}|\mathbf{s}\right) A^{\pi_{\boldsymbol{\nu}}}\left(\mathbf{s}, \mathbf{a}\right)\right] = 0$$

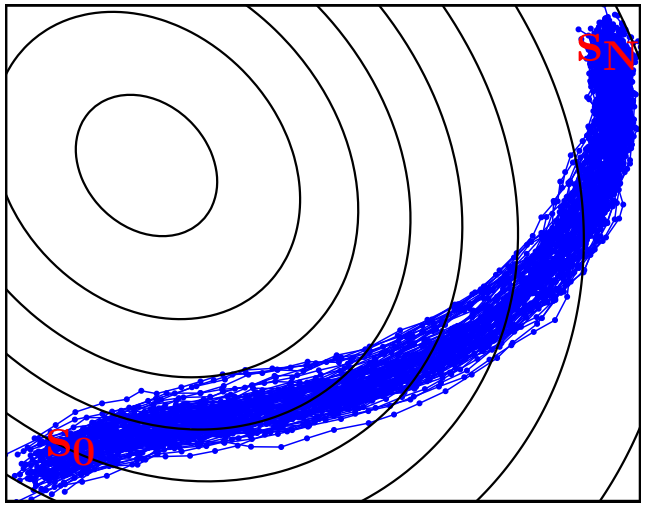
defines relationship $\pi_{\boldsymbol{\nu}^\star}$ with $\boldsymbol{\nu}^\star$ function of $\boldsymbol{\theta}$

In-Real Actor-critic evaluates

$$\nabla_{\boldsymbol{\theta}} J\left(\pi_{\boldsymbol{\nu}^\star}\right) = \mathbb{E}_{\text{Real World}}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}\left(\mathbf{a}|\mathbf{s}\right) A^{\boldsymbol{\nu}^\star}\left(\mathbf{s}, \mathbf{a}\right)\right]$$



Critic-Based Differentiation



In-Sim Actor-critic does

$$\nabla_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right)=\mathbb{E}_{\substack{\mathbf{s} \sim \rho_{\boldsymbol{\theta}} \\ \mathbf{a} \sim \pi_{\boldsymbol{\nu}}}}\left[\nabla_{\boldsymbol{\nu}} \log \pi_{\boldsymbol{\nu}}(\mathbf{a}|\mathbf{s}) A^{\pi_{\boldsymbol{\nu}}}(\mathbf{s}, \mathbf{a})\right]=0$$

defines relationship $\pi_{\boldsymbol{\nu}^\star}$ with $\boldsymbol{\nu}^\star$ function of $\boldsymbol{\theta}$

In-Real Actor-critic evaluates

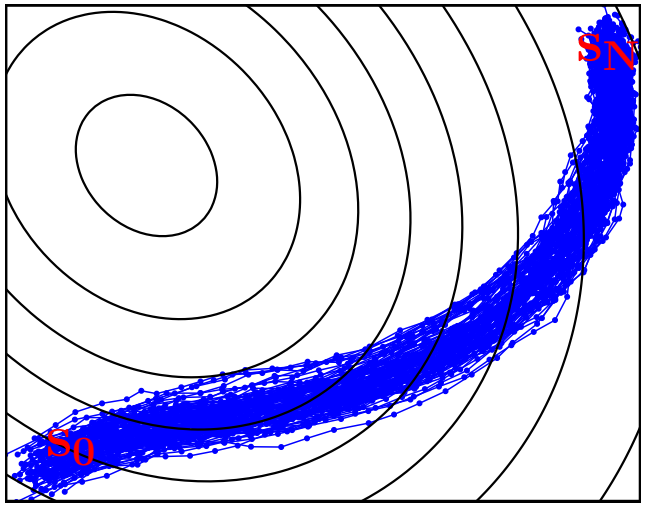
$$\nabla_{\boldsymbol{\theta}} J\left(\pi_{\boldsymbol{\nu}^\star}\right)=\mathbb{E}_{\text{Real World}}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}(\mathbf{a}|\mathbf{s}) A^{\boldsymbol{\nu}^\star}(\mathbf{s}, \mathbf{a})\right]$$

where $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}(\mathbf{a}|\mathbf{s})$ requires $\frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}}$

Sensitivity

$$\frac{\partial^2}{\partial \boldsymbol{\nu}^2} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}^\star}\right) \frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}}+\frac{\partial^2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\theta}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}^\star}\right)=0$$

Critic-Based Differentiation



In-Sim Actor-critic does

$$\nabla_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\substack{\mathbf{s} \sim \rho_{\boldsymbol{\theta}} \\ \mathbf{a} \sim \pi_{\boldsymbol{\nu}}}} \left[\nabla_{\boldsymbol{\nu}} \log \pi_{\boldsymbol{\nu}}\left(\mathbf{a}|\mathbf{s}\right) A^{\pi_{\boldsymbol{\nu}}}\left(\mathbf{s}, \mathbf{a}\right)\right] = 0$$

defines relationship $\pi_{\boldsymbol{\nu}^\star}$ with $\boldsymbol{\nu}^\star$ function of $\boldsymbol{\theta}$

In-Real Actor-critic evaluates

$$\nabla_{\boldsymbol{\theta}} J\left(\pi_{\boldsymbol{\nu}^\star}\right) = \mathbb{E}_{\text{Real World}}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}\left(\mathbf{a}|\mathbf{s}\right) A^{\boldsymbol{\nu}^\star}\left(\mathbf{s}, \mathbf{a}\right)\right]$$

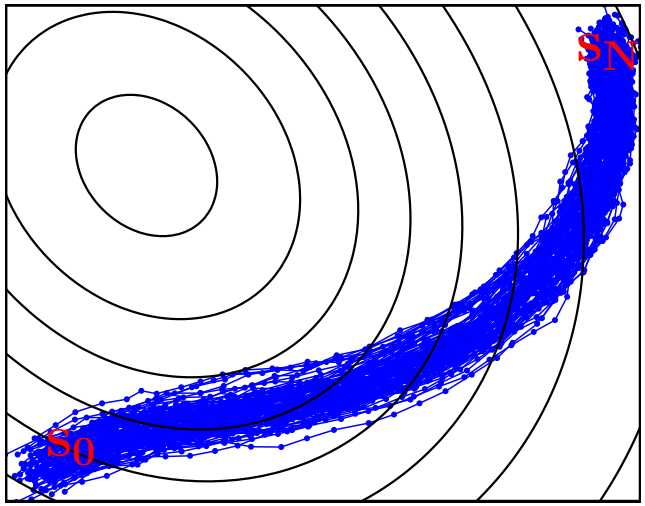
where $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}\left(\mathbf{a}|\mathbf{s}\right)$ requires $\frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}}$

Sensitivity

$$\frac{\partial^2}{\partial \boldsymbol{\nu}^2} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}^\star}\right) \frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}} + \frac{\partial^2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\theta}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}^\star}\right) = 0$$

Idea: compute $\frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}}$ from differentiating the in-sim actor-critic w.r.t. $\frac{\partial}{\partial \boldsymbol{\nu}}$ and $\frac{\partial}{\partial \boldsymbol{\theta}}$

Critic-Based Differentiation



In-Sim Actor-critic does

$$\nabla_{\boldsymbol{\nu}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}}\right) = \mathbb{E}_{\substack{\mathbf{s} \sim \rho_{\boldsymbol{\theta}} \\ \mathbf{a} \sim \pi_{\boldsymbol{\nu}}}} \left[\nabla_{\boldsymbol{\nu}} \log \pi_{\boldsymbol{\nu}}\left(\mathbf{a}|\mathbf{s}\right) A^{\pi_{\boldsymbol{\nu}}}\left(\mathbf{s}, \mathbf{a}\right)\right] = 0$$

defines relationship $\pi_{\boldsymbol{\nu}^\star}$ with $\boldsymbol{\nu}^\star$ function of $\boldsymbol{\theta}$

In-Real Actor-critic evaluates

$$\nabla_{\boldsymbol{\theta}} J\left(\pi_{\boldsymbol{\nu}^\star}\right) = \mathbb{E}_{\text{Real World}}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}\left(\mathbf{a}|\mathbf{s}\right) A^{\boldsymbol{\nu}^\star}\left(\mathbf{s}, \mathbf{a}\right)\right]$$

where $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\nu}^\star}\left(\mathbf{a}|\mathbf{s}\right)$ requires $\frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}}$

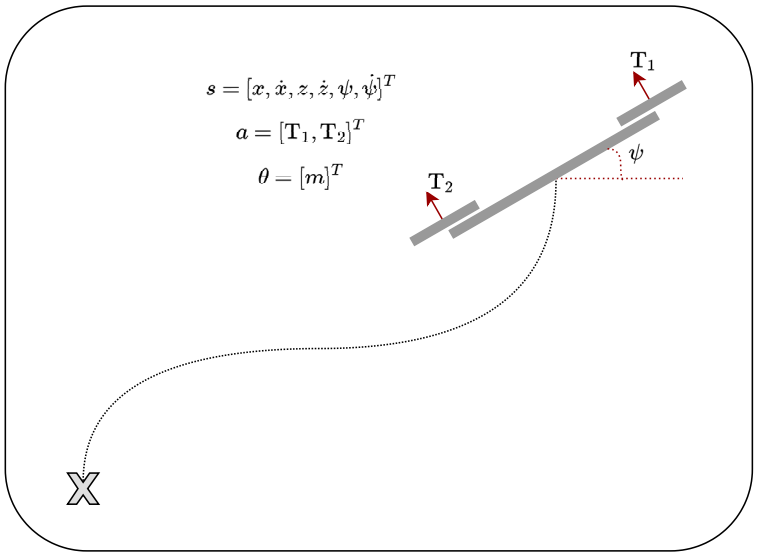
Sensitivity

$$\frac{\partial^2}{\partial \boldsymbol{\nu}^2} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}^\star}\right) \frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}} + \frac{\partial^2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\theta}} \hat{J}_{\boldsymbol{\theta}}\left(\pi_{\boldsymbol{\nu}^\star}\right) = 0$$

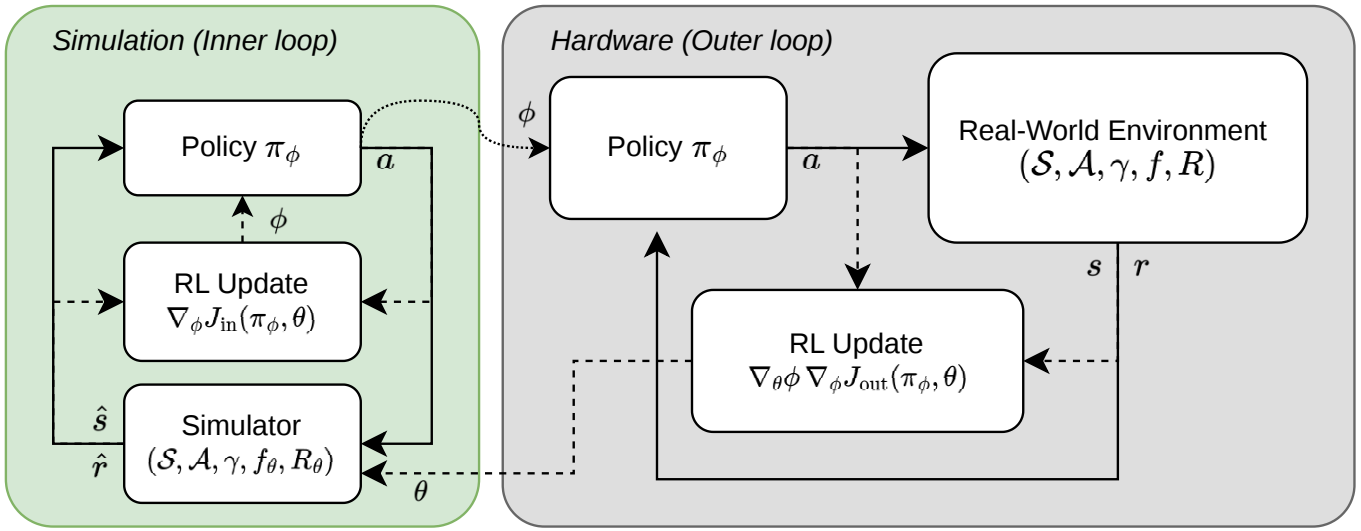
Idea: compute $\frac{\mathrm{d}\boldsymbol{\nu}^\star}{\mathrm{d}\boldsymbol{\theta}}$ from differentiating the in-sim actor-critic w.r.t. $\frac{\partial}{\partial \boldsymbol{\nu}}$ and $\frac{\partial}{\partial \boldsymbol{\theta}}$

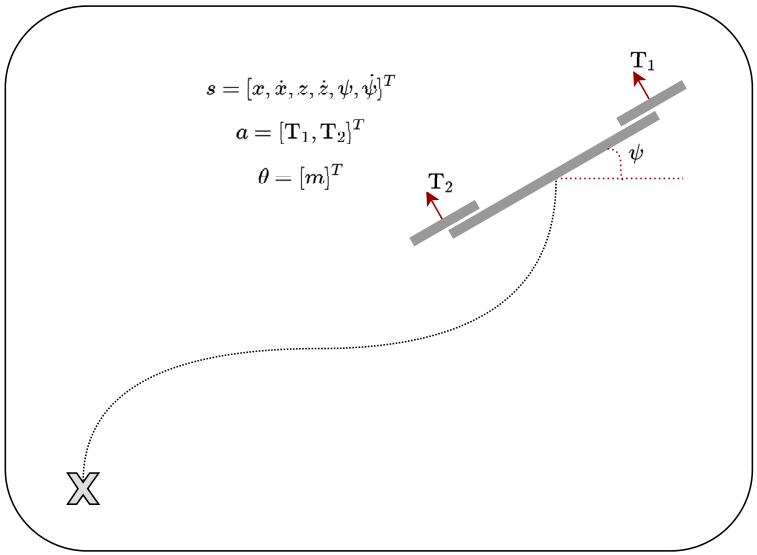
Proposed first algorithm to do that.
Computationally heavy, but we have "missed" some simplifications, testing further

Example - In-Real RL over In-Sim RL

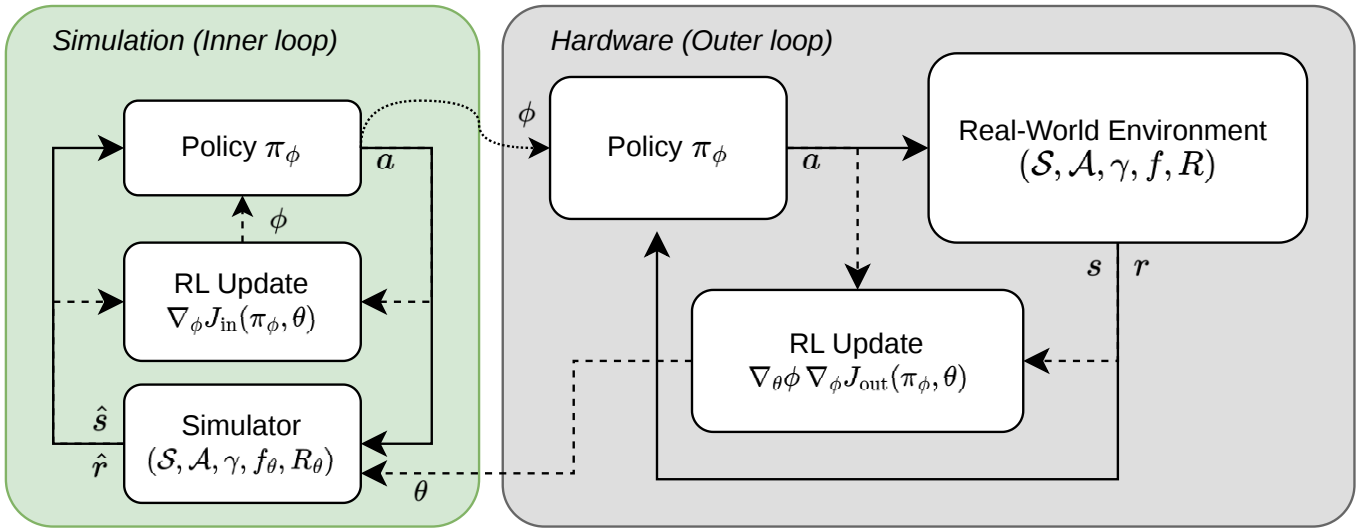


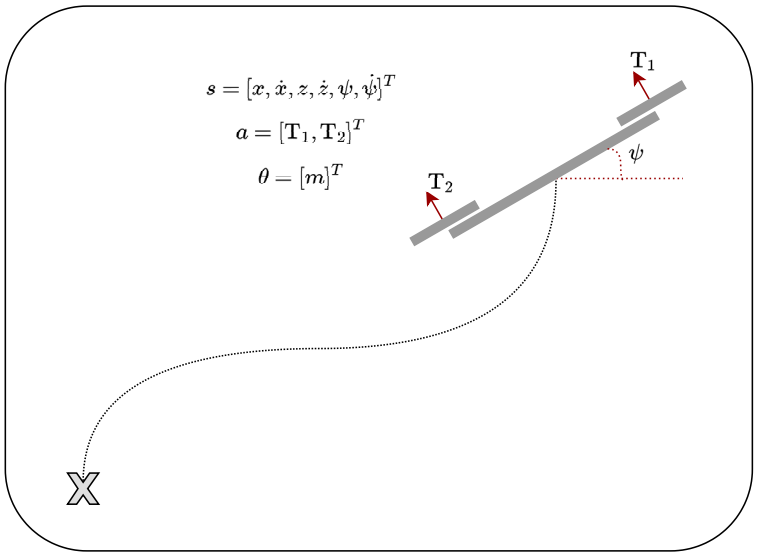
Example - In-Real RL over In-Sim RL

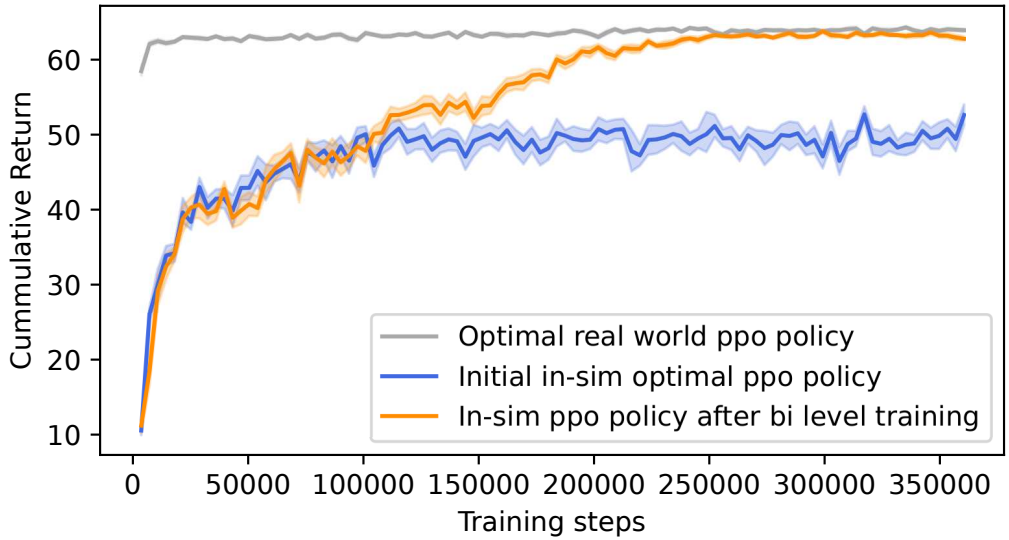




Example - In-Real RL over In-Sim RL







Norwegian Center on AI for Decisions

Email

- Start-up phase
- Recruiting soon!
- We can apply for funding to bring strong international postdocs

Thank you!