DDA 6201 Online Decision-Making Lecture 12

Application 2: Valued-Based RL

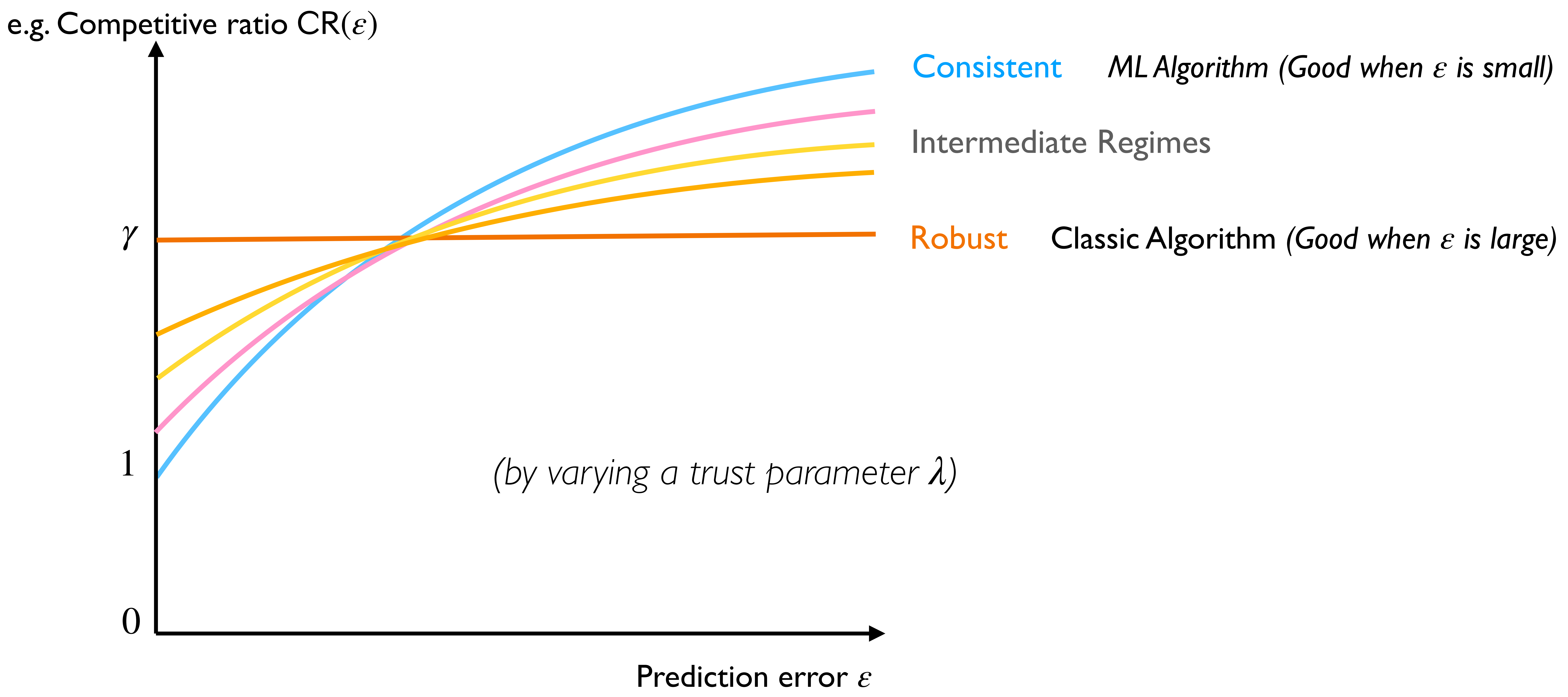
Tongxin Li

School of Data Science

The Chinese University of Hong Kong (Shenzhen)

Revisit: Learning-Augmented Algorithms

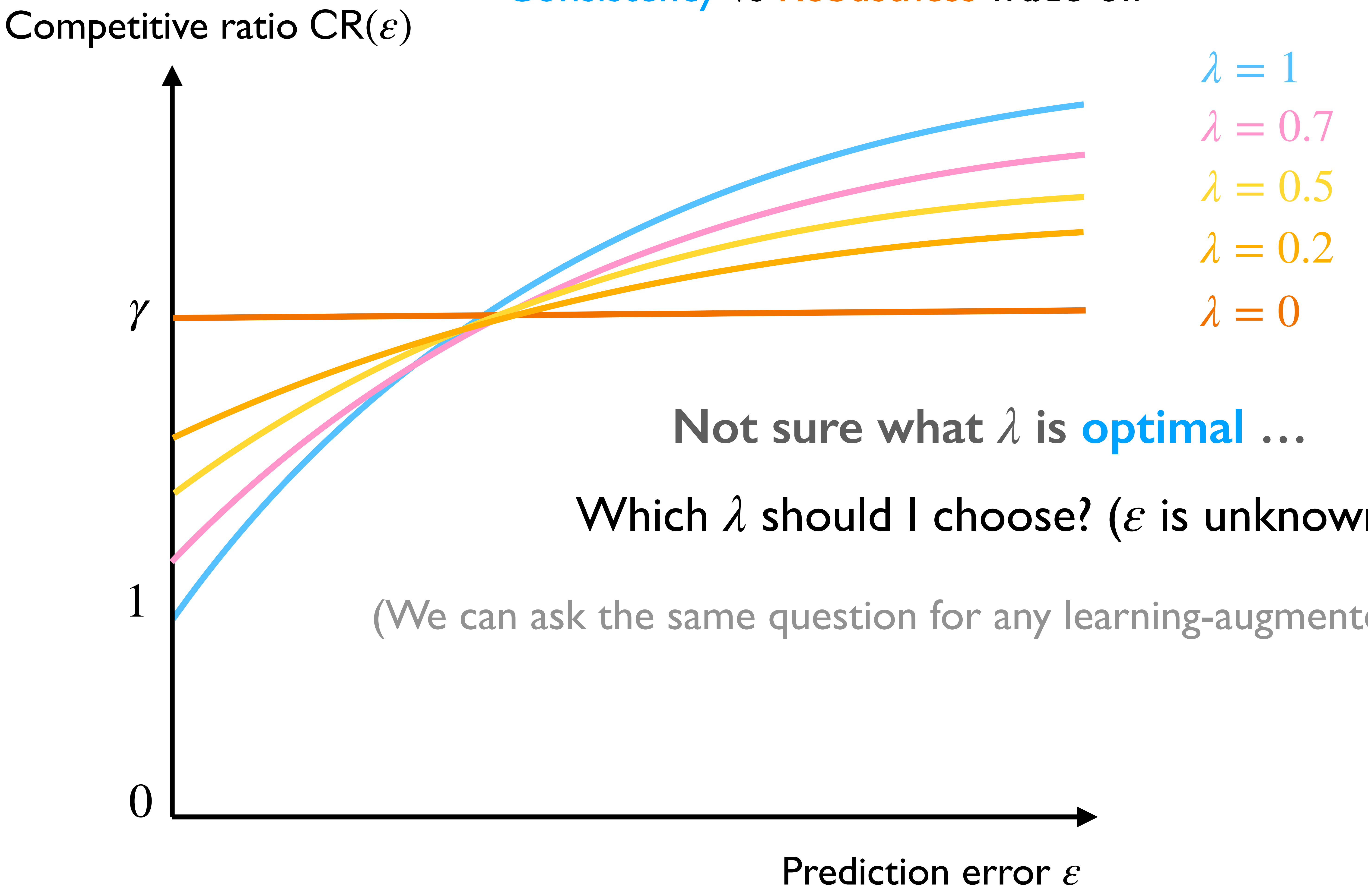
Performance Benchmark   Meta-algorithms   Consistency vs Robustness Trade-off



First Limitation

General Goal of Learning-Augmented Algorithms

Consistency vs Robustness Trade-off



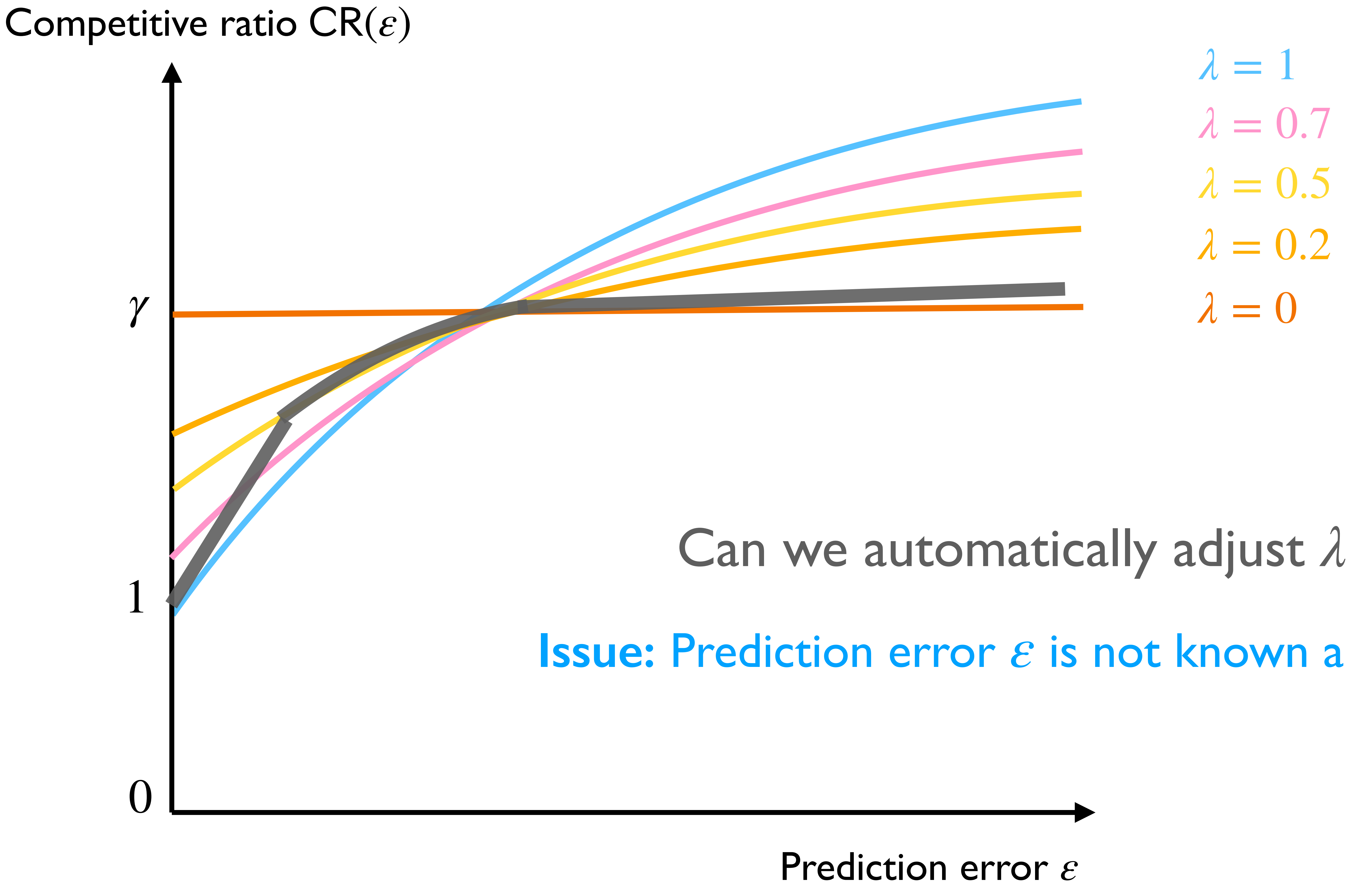
Not sure what λ is optimal …

Which λ should I choose? (ε is unknown)

(We can ask the same question for any learning-augmented online algorithms)

First Limitation

Goal: Find an online algorithm with good Competitive Ratio **CR** regardless of **prediction error ε**



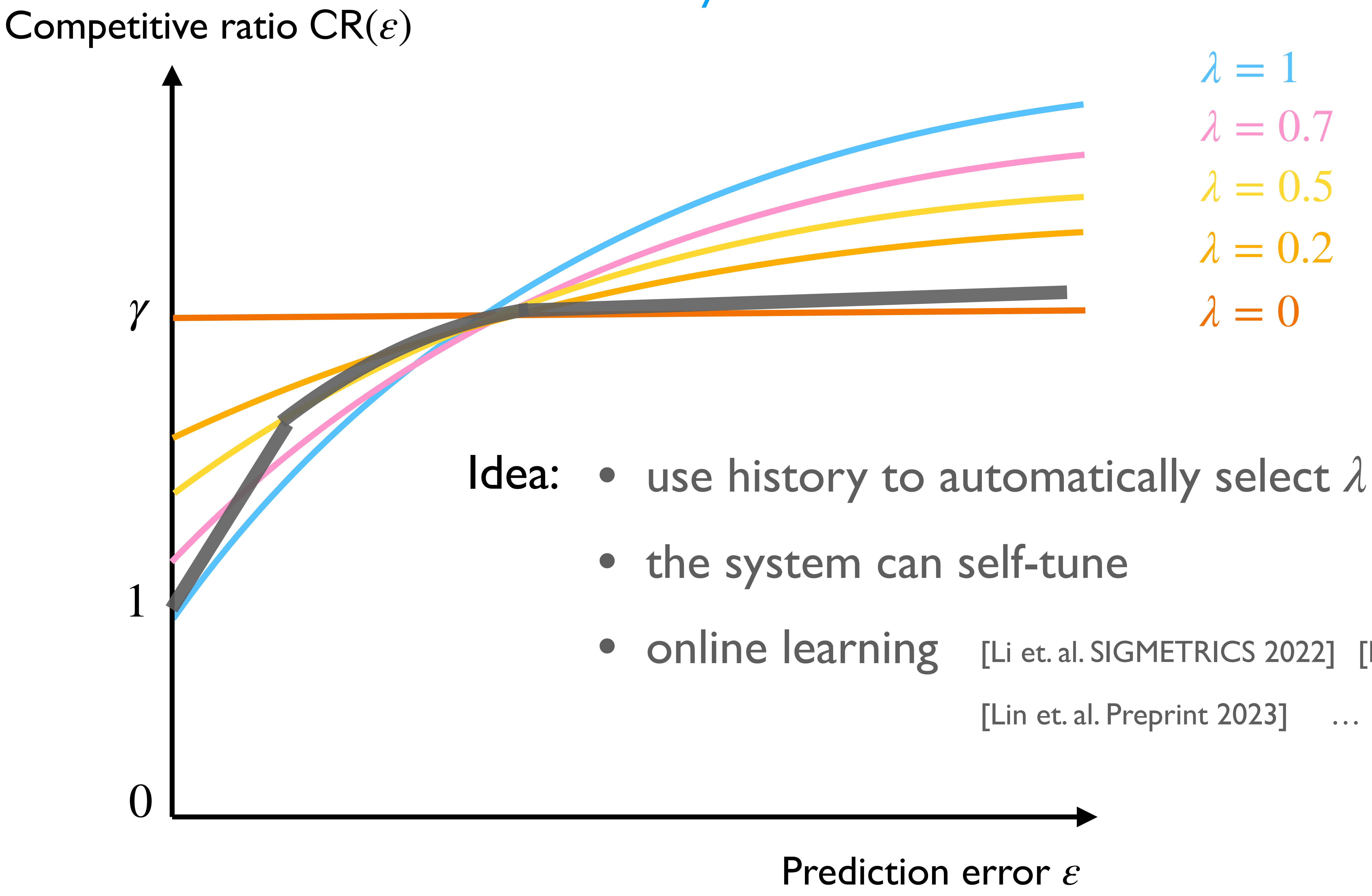
Can we automatically adjust λ ?

Issue: Prediction error ε is not known a priori

One Solution: **Online Learning**

General Goal of Learning-Augmented Algorithms

Consistency vs Robustness Trade-off



Idea:
- use history to automatically select λ
- the system can self-tune
- online learning [Li et. al. SIGMETRICS 2022] [Khodak et. al. NeurIPS 2022] [Lin et. al. Preprint 2023] … [Li et. al. NeurIPS 2024]

A Real-World Problem

Image: Paired Power

The UC San Diego/EVgo project

EV Charging with Uncertainties

Main Sources of Uncertainties → Data

Use RL for scheduling?

- Tons of existing policies

Question: Do they work well in practice?

Question: If so, why is it hard to see them being used?

Ideally, they work well, but …

Electricity price varies

Solar generation

Charging behavior

The UC San Diego/EVgo project

Large-Scale Adaptive Charging Network

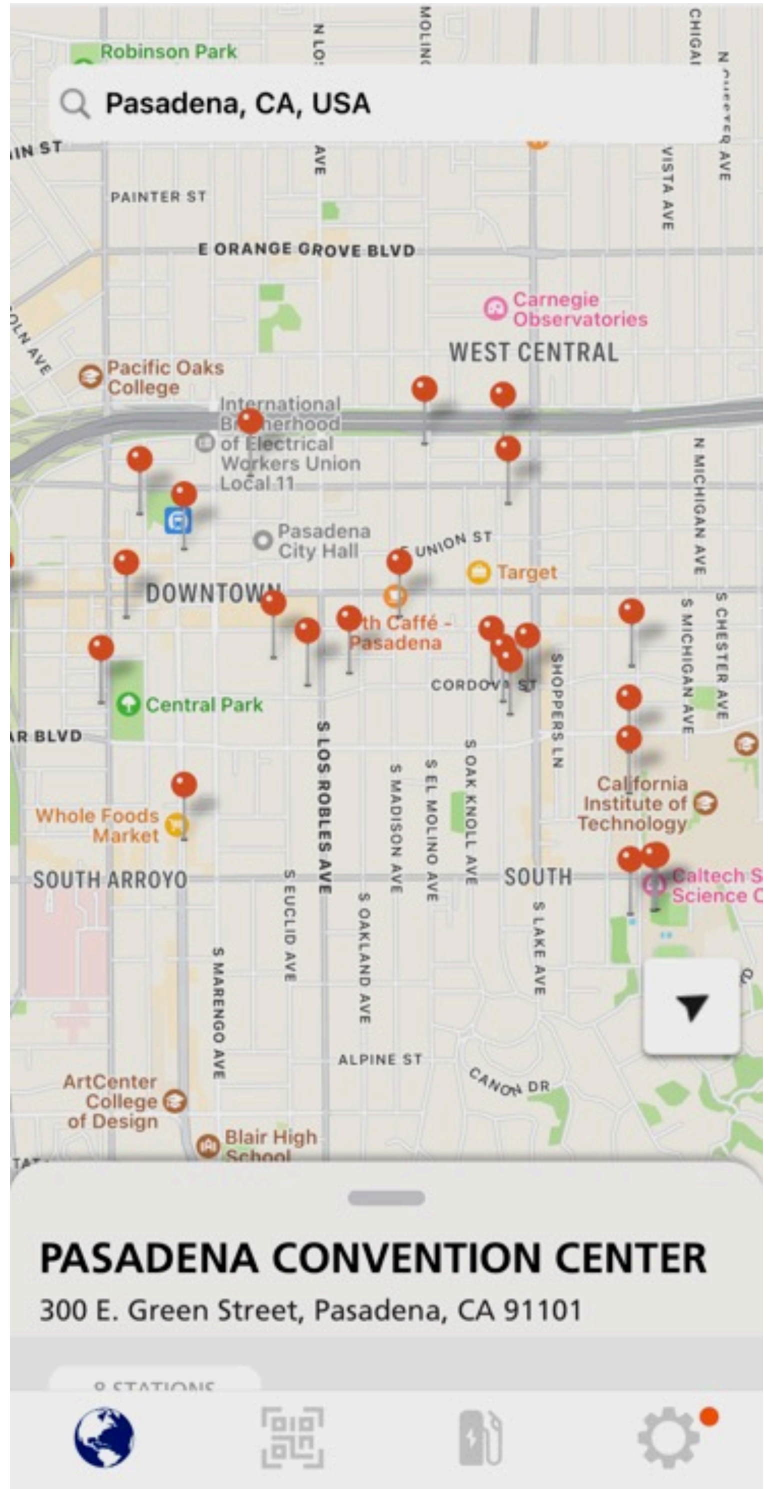
Adaptive Charging Network (ACN@Caltech)

- A Parking lot with 54 chargers
- How to schedule EV charging is challenging

Caltech

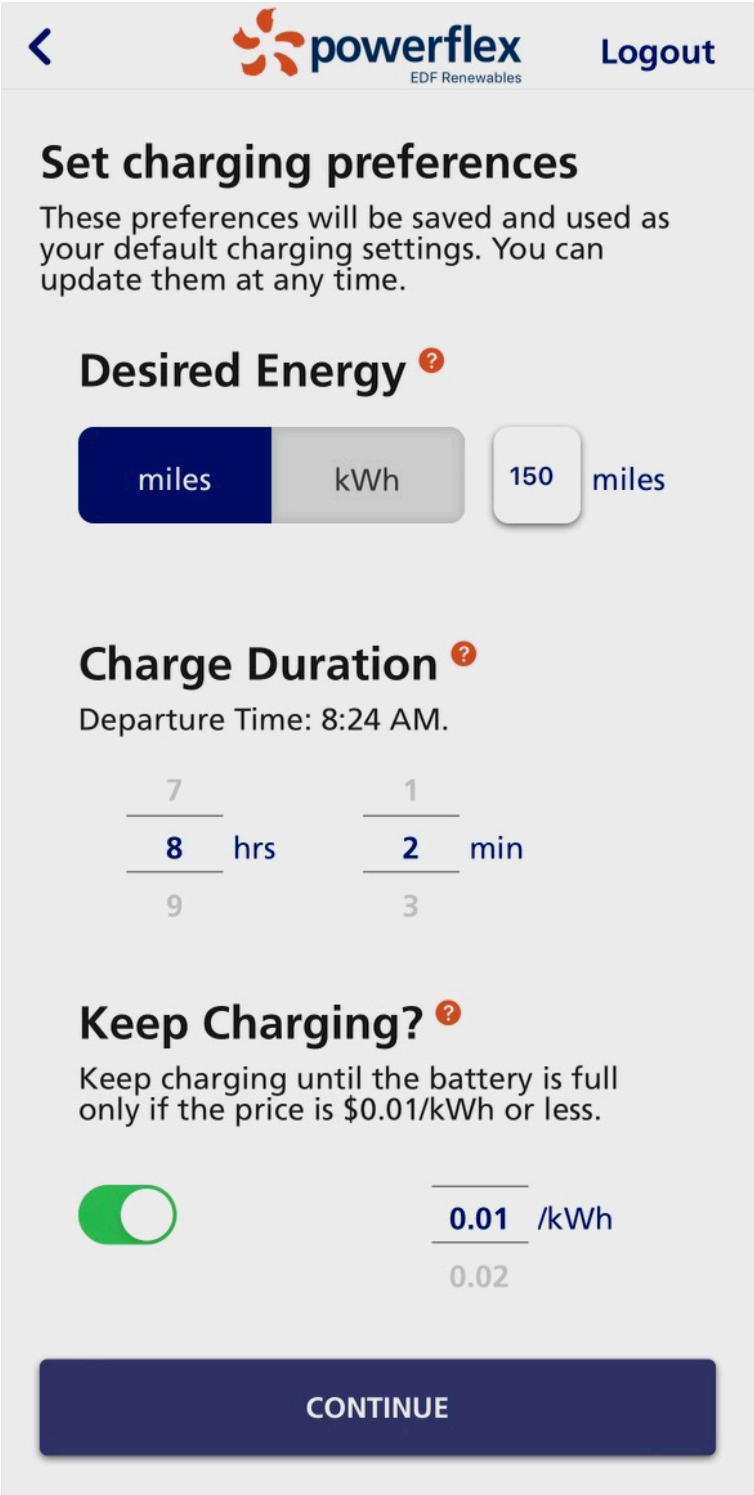
Large-Scale Workplace EV Charging

Caltech

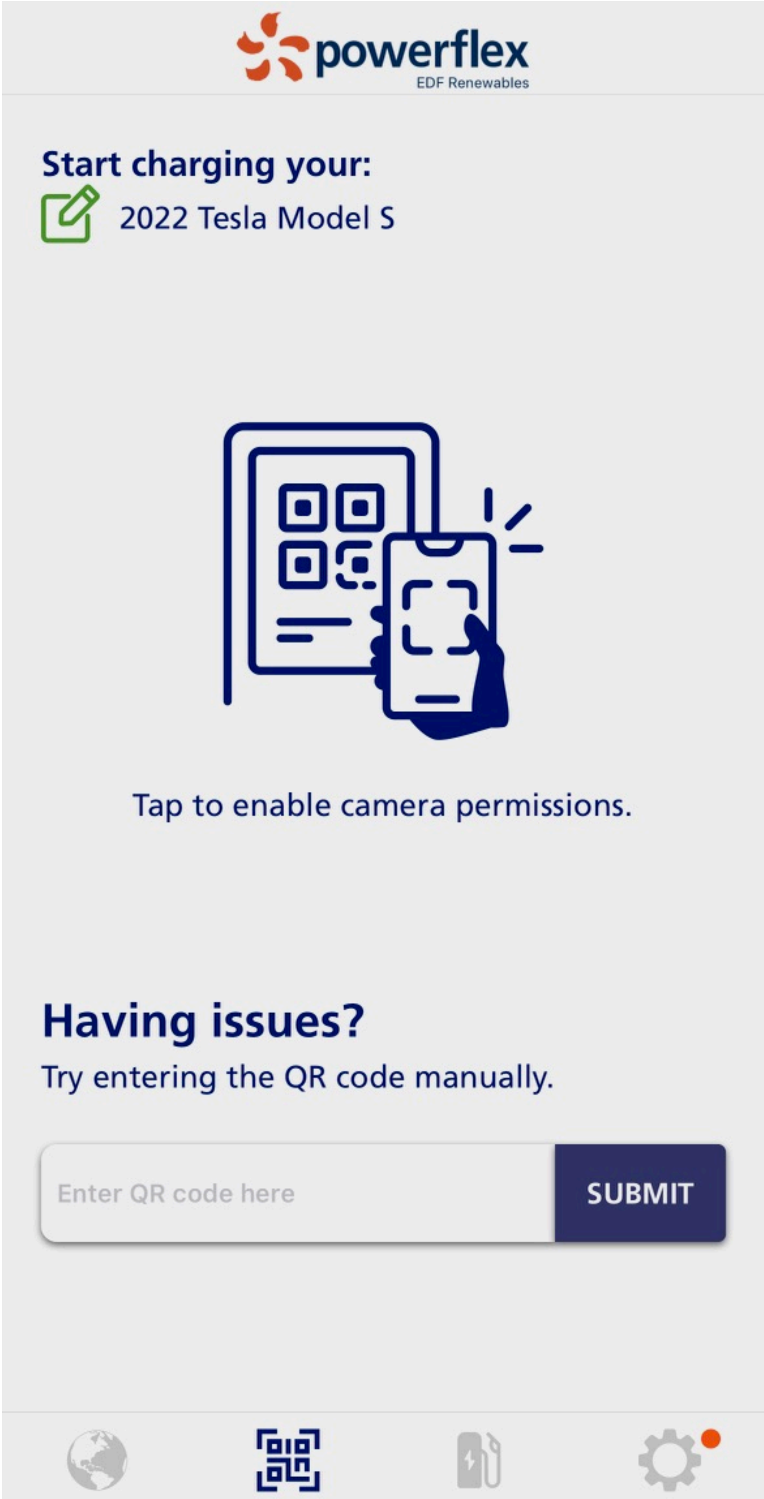


Fina a Location

Choose Model



Select Energy



Scan and Charge

Commercialized Version 2.7.39 / 2022

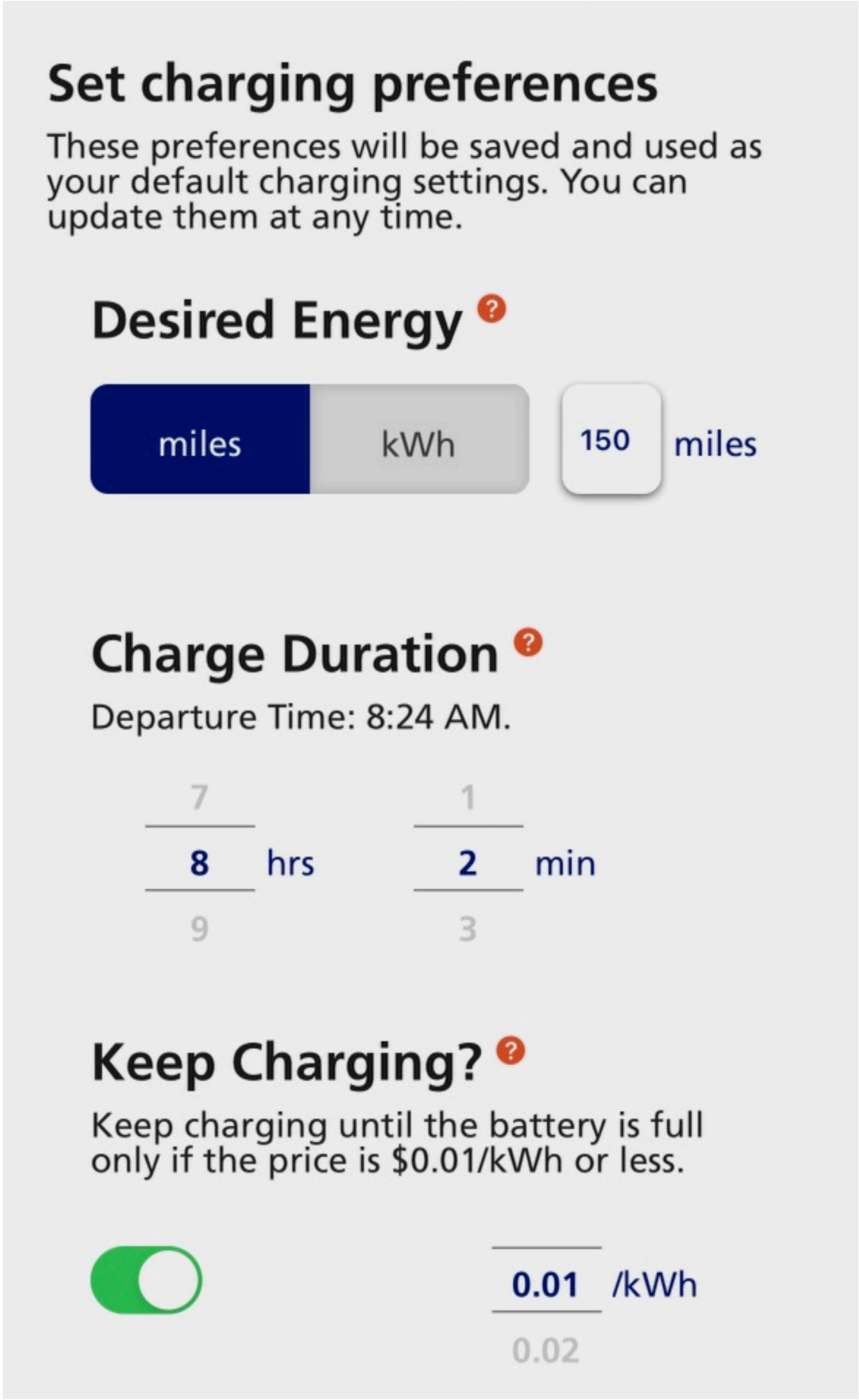
Large-Scale Workplace EV Charging

Classic Scheduling Algorithms

- Least laxity first (LLF)
- Earliest deadline first (EDF)
- Model predictive control (MPC)

(Currently used in Caltech ACNs)

…



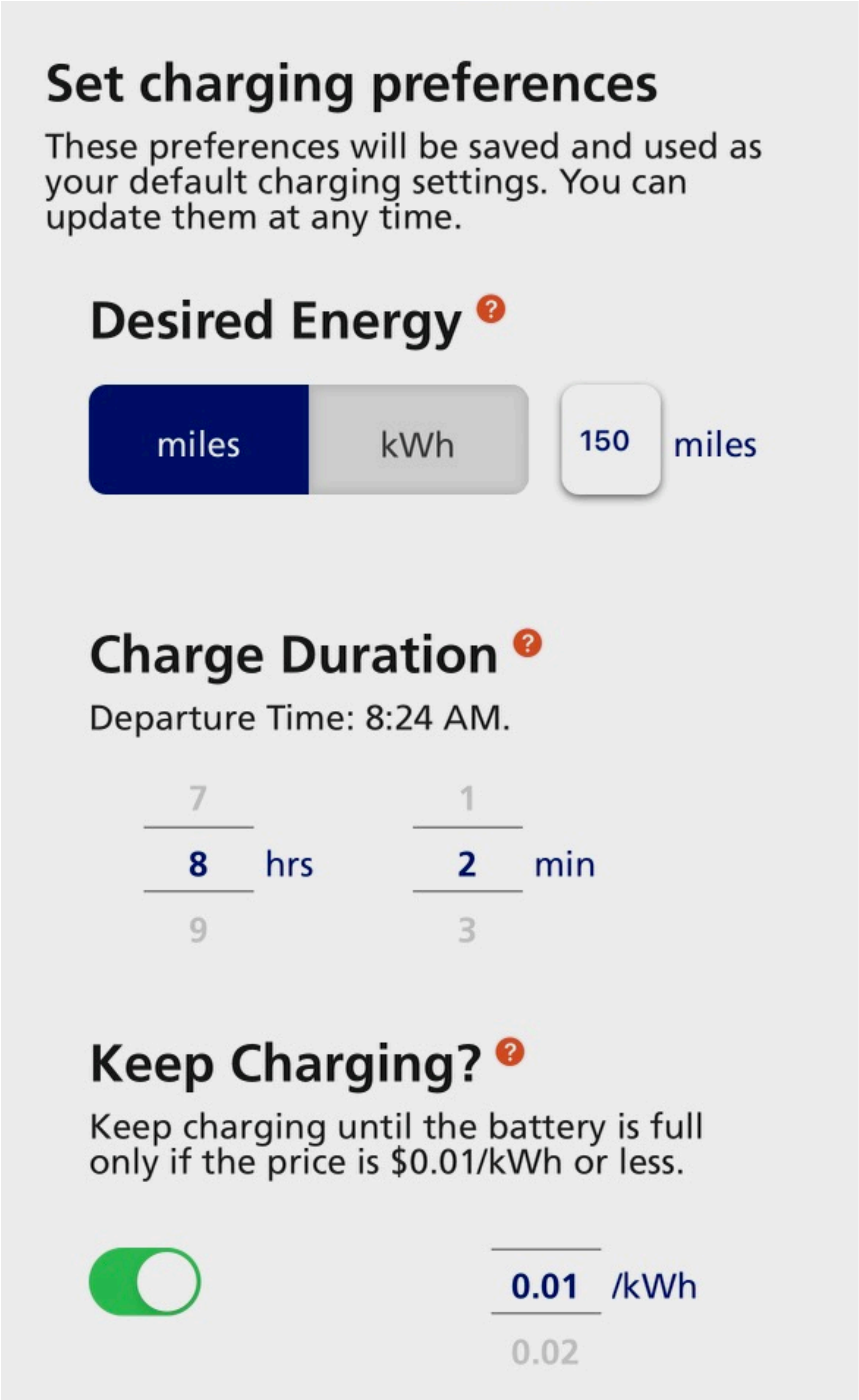
Large-Scale Workplace EV Charging

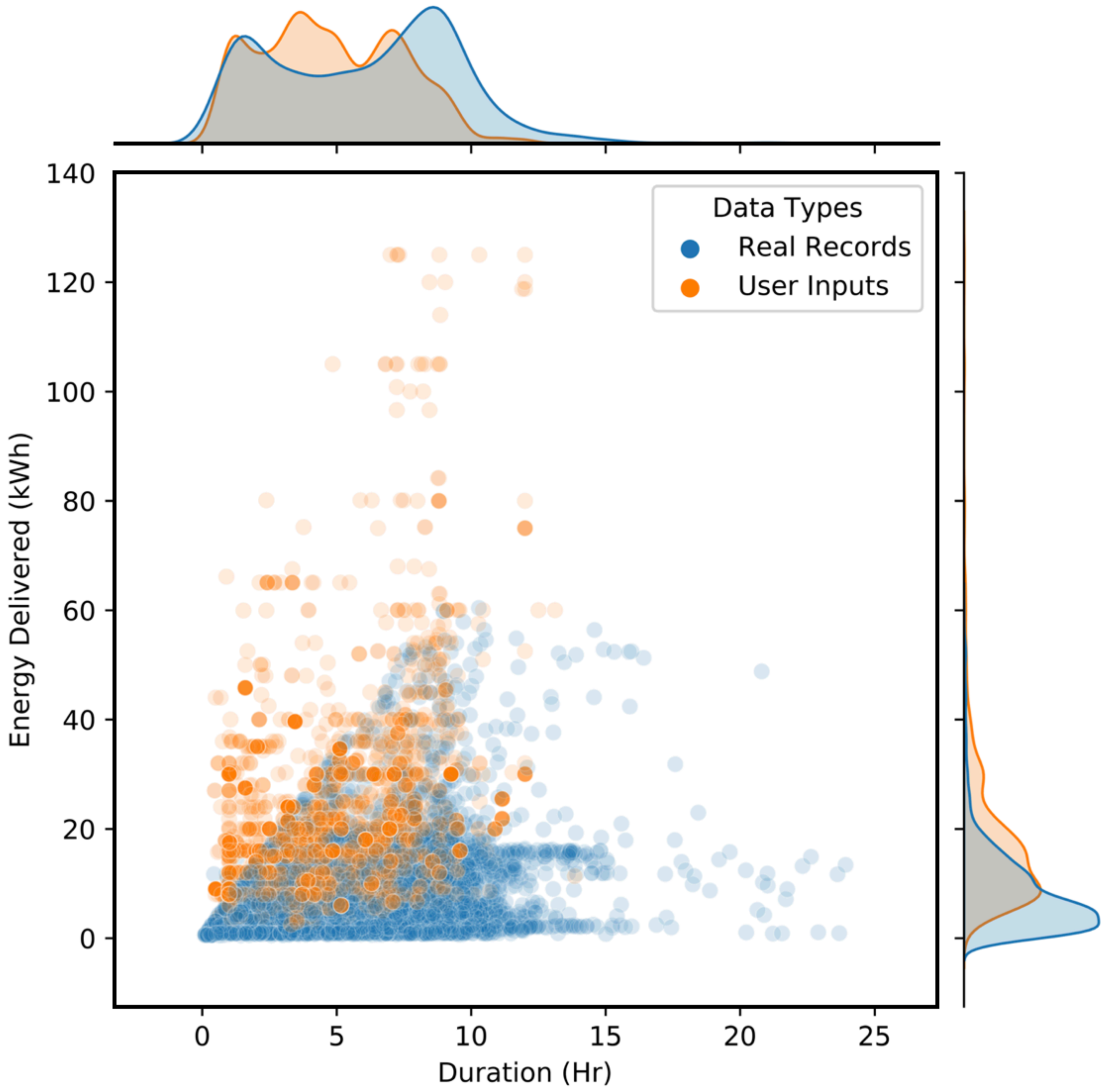
Classic Scheduling Algorithms

- Least laxity first (LLF)
- Earliest deadline first (EDF)
- Model predictive control (MPC)

(Currently used in Caltech ACNs)

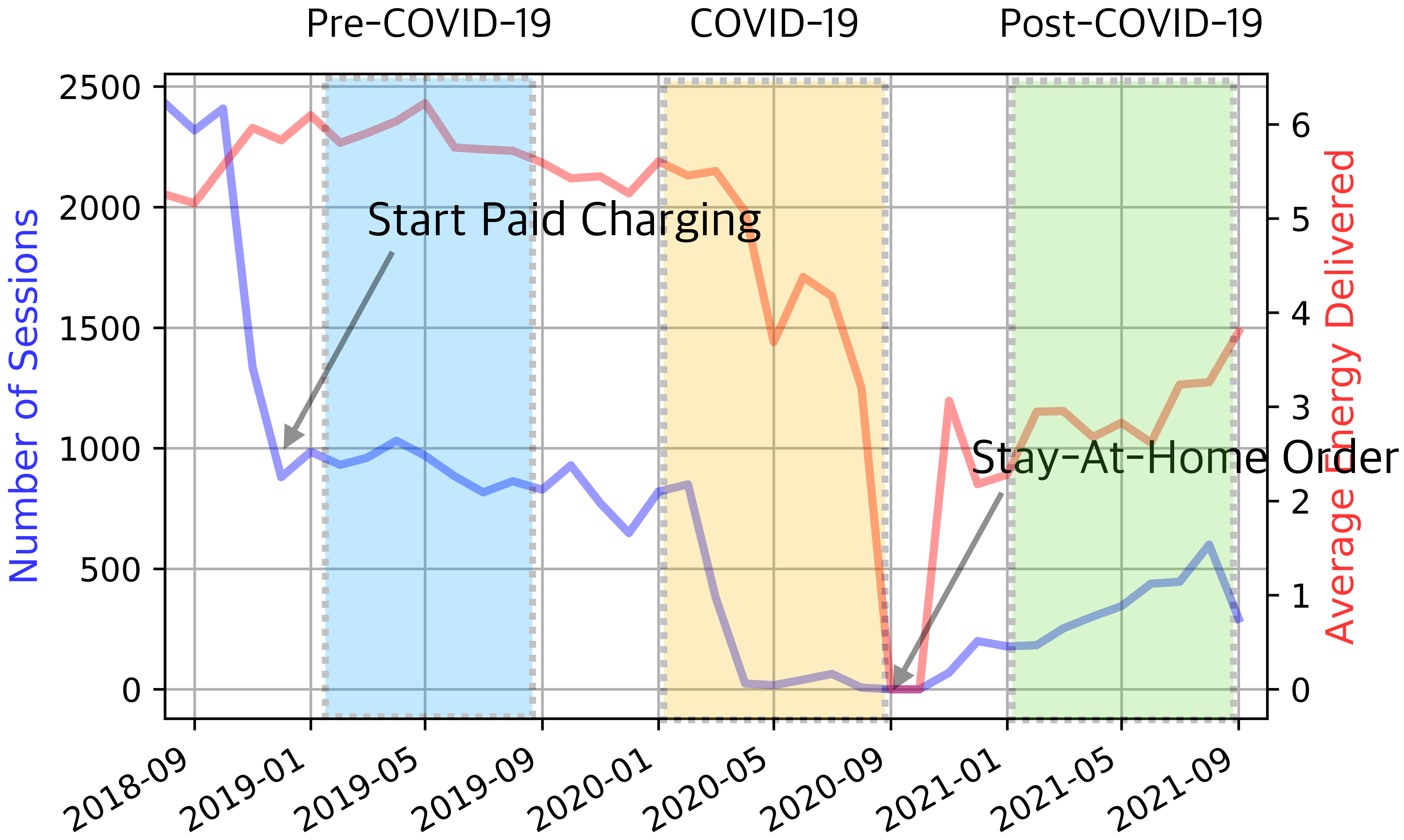
…



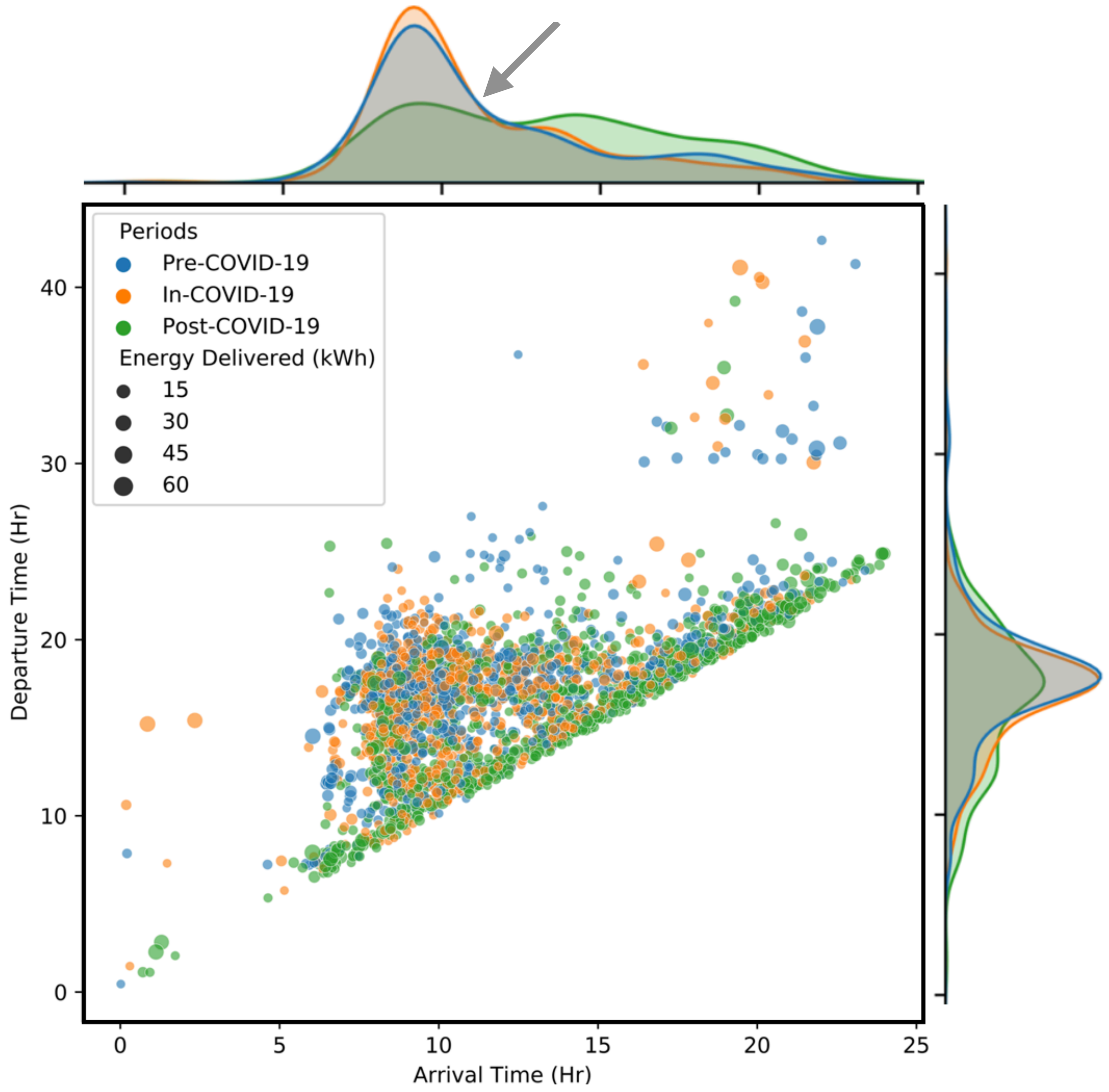


COVID-19 Caused Dataset Shift

Statistical distributions shifted (Data from the real Caltech system) **[e-Energy ’19]**



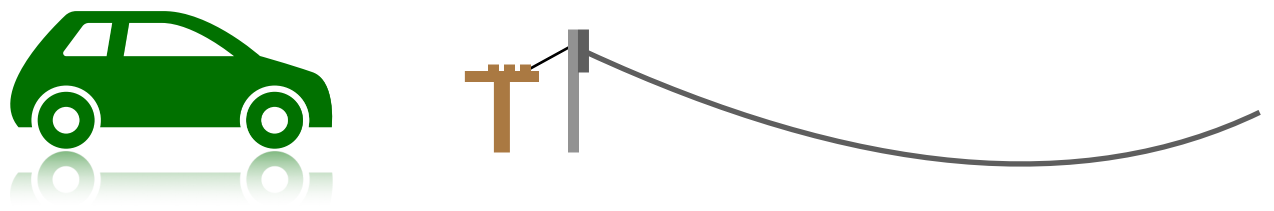
Flattened Arriving times



RL policies trained on out-of-distribution data can perform poorly

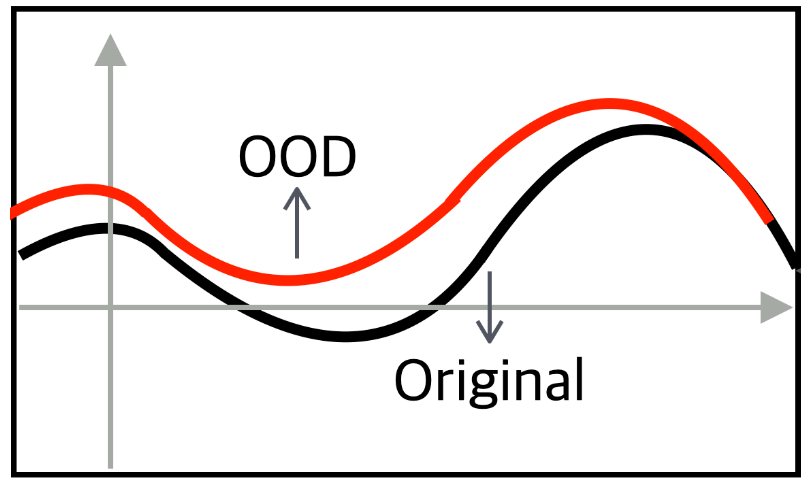
Combing MPC and RL Scheduling

EV arrives at α_j



The adaptive charging network at Caltech. source: https://ev.caltech.edu

EV departs at δ_j



Distribution shifts of
Solar and User Behaviors

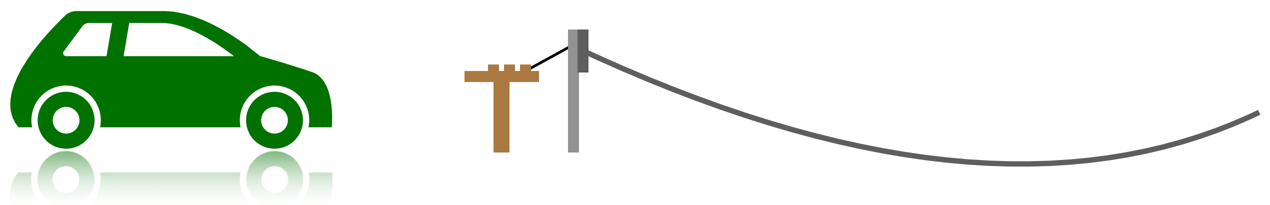
$$(e_{t+1}\|b_{t+1}) = s_{t+1} = g_{\mathcal{S}}\Big[\underbrace{A_t s_t + B_t g_{\mathscr{A}}(a_t)} + \underbrace{\ell'_t - \Delta h'_t}\Big],\ t \geq 0$$

Battery Dynamics (Uncertain) **Behavior/Solar Perturbations**

$\left(\alpha_j, \delta_j, \kappa_j, i\right)$ is a **charging session**: At time α_j, EV j arrives at charger i, with an EV battery capacity κ_j, and departs at time δ_j

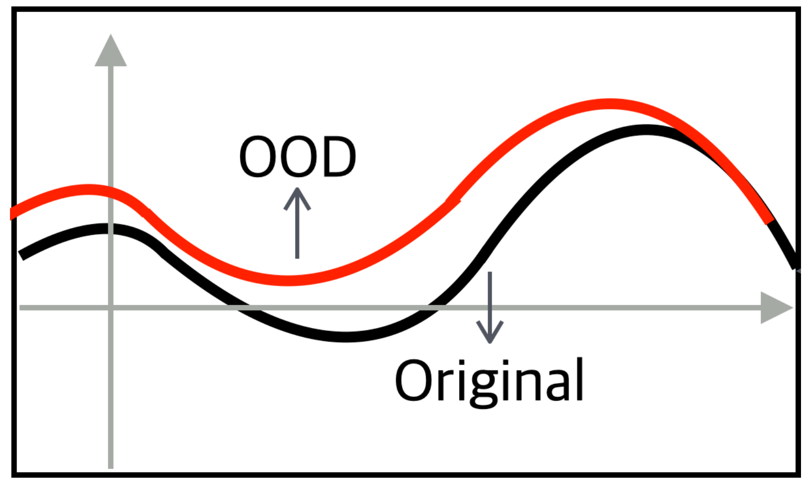
Combing MPC and RL Scheduling

EV arrives at α_j



The adaptive charging network at Caltech. source: https://ev.caltech.edu

EV departs at δ_j



Distribution shifts of Solar and User Behaviors

$$(e_{t+1}\|b_{t+1}) = s_{t+1} = g_{\mathcal{S}}\Big[\underbrace{A_t s_t + B_t g_{\mathcal{A}}(a_t)}_{\textbf{Battery Dynamics}} + \underbrace{\ell'_t - \Delta h'_t}_{\text{(Uncertain) }\textbf{Behavior/Solar Perturbations}}\Big],\ t \geq 0$$

Battery SoC

Charging Rates

Change of Charging Rates

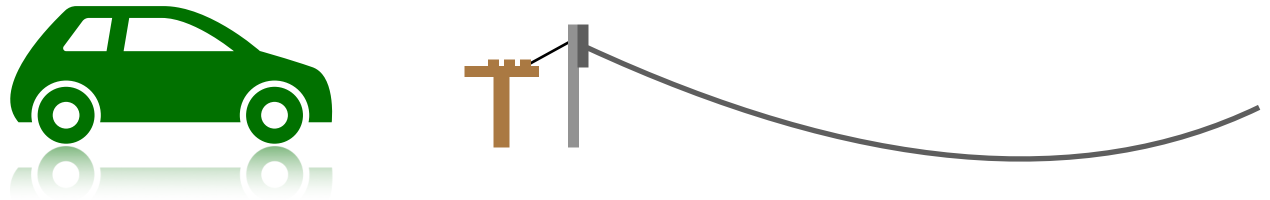
Human behaviors arrival/departure

Solar generation

$\left(\alpha_j, \delta_j, \kappa_j, i\right)$ is a **charging session**: At time α_j, EV j arrives at charger i, with an EV battery capacity κ_j, and departs at time δ_j

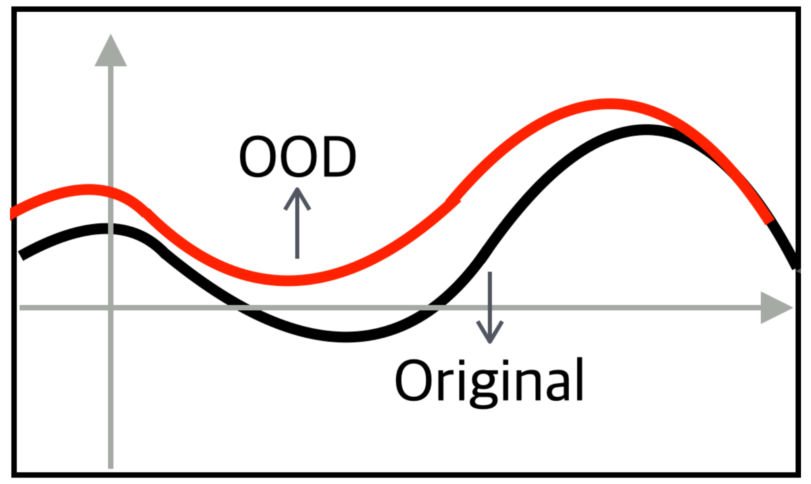
Combing MPC and RL Scheduling

EV arrives at α_j



The adaptive charging network at Caltech. source: https://ev.caltech.edu

EV departs at δ_j



Distribution shifts of Solar and User Behaviors

$$(e_{t+1} \| b_{t+1}) = s_{t+1} = g_{\mathcal{S}}\Big[\underbrace{A_t s_t + B_t g_{\mathcal{A}}(a_t)}_{\textbf{Battery Dynamics}} + \underbrace{\ell'_t - \Delta h'_t}_{\text{(Uncertain) }\textbf{Behavior/Solar Perturbations}}\Big],\ t \geq 0$$

Projections $g_{\mathcal{S}}$ and $g_{\mathcal{A}}$ capture network constraints, such as line constraints by the Kirchhoff's Current Law:

$$\Big| \sum_{i=1}^{N} D_{ij} b_t(i) e^{j\phi_i} \Big| \leq \gamma, \quad \forall t \in \mathscr{T}\$$$

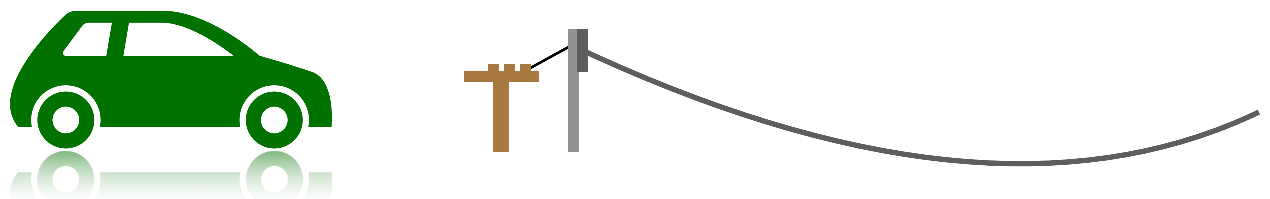
Formed by circuit analysis

Phase angle of current phasor

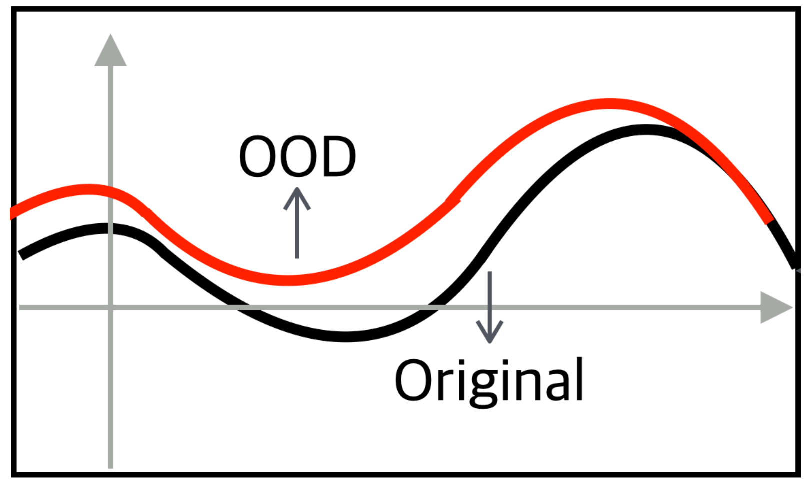
Current magnitude limit

Combing MPC and RL Scheduling

EV arrives at α_j



EV departs at δ_j



The adaptive charging network at Caltech. source: https://ev.caltech.edu

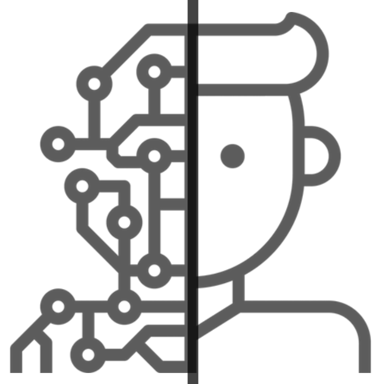
Distribution shifts of Solar and User Behaviors

$$(e_{t+1}\|b_{t+1}) = s_{t+1} = g_{\mathcal{S}}\Big[\underbrace{A_t s_t + B_t g_{\mathcal{A}}(a_t)}_{\textbf{Battery Dynamics}} + \underbrace{\ell'_t - \Delta h'_t}_{\text{(Uncertain) }\textbf{Behavior/Solar Perturbations}}\Big],\ t \geq 0$$

- **Robustness** Classic algorithm (MPC) depends on battery dynamics and user inputs
- **Consistency** RL policy can better learn uncertain residuals when they are not out-of-distribution
- This is a general paradigm in many real-world applications

ML in Real-World Decision-Making …

Digital Applications	Physical World
Text Generation	Control Power Systems
Computer Vision	Autonomous Driving
Play Go …	Embodied AI …
Successful	Challenging



Machine-learned policies have the advantage of utilizing data

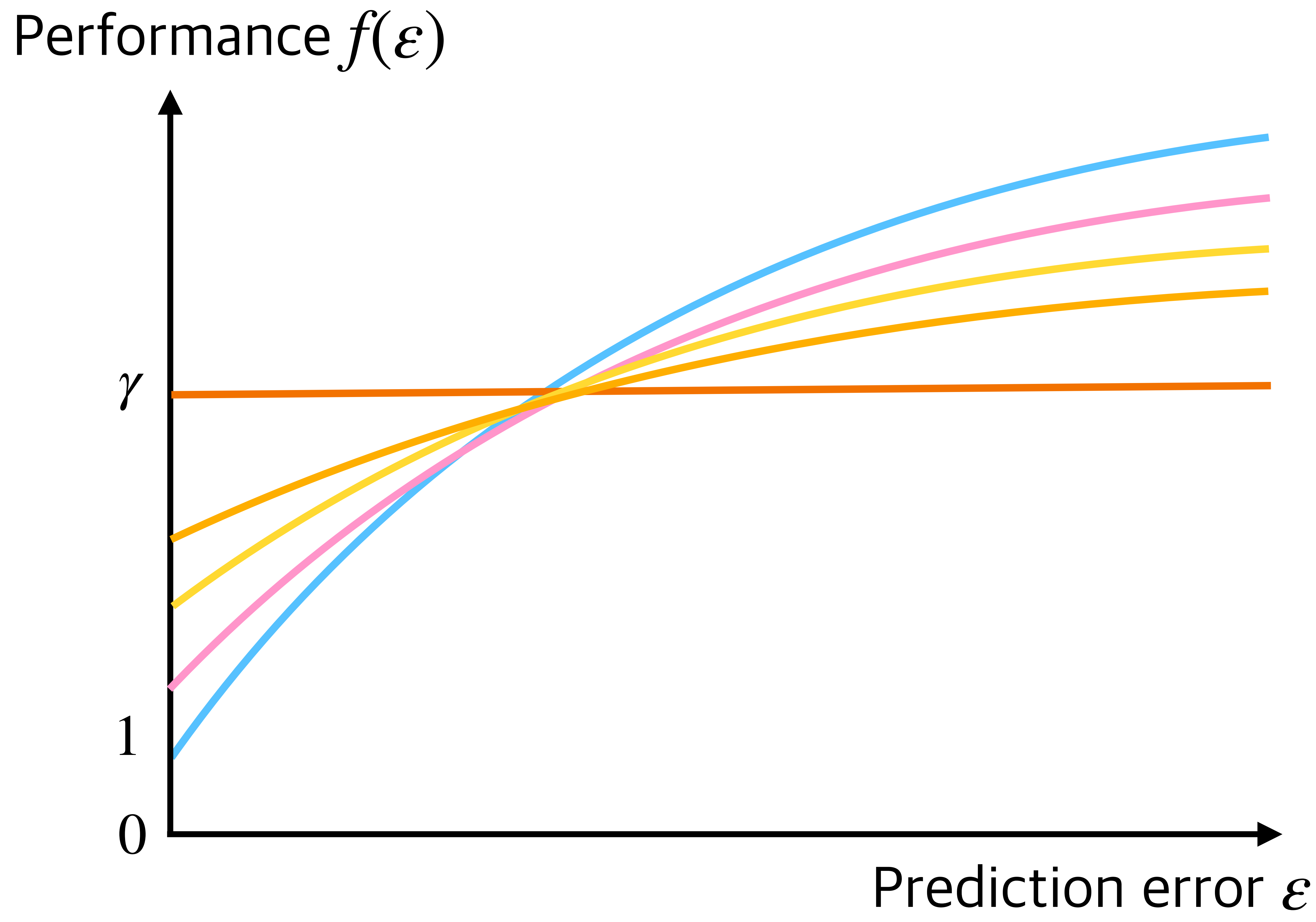
Existing well-established **classic methods** that are hard to be replaced entirely

On average near-optimal performance

Worst-case guarantee

Consistency vs Robustness Trade-off

On average good performance vs Worst-case guarantee



Consistent ML Policy (Good when ε is small)

Intermediate Regimes

Robust Classic Policy (Good when ε is large)

$f(0)$ -consistent

$\sup_{\varepsilon \geq 0} f(\varepsilon)$-robust

Today's Topic: Value-Based RL

Nonlinear Model is Harder if the ML Agent is a Black-Box

System Model	Classic Agent	ML Agent	Remarks	Tradeoffs
Linear Dynamics	LQR	MPC+Perturbation Predictions	Convex Combination	Consistency vs Robustness
NonLinear Dynamics	LQR	Black-Box RL	Switching	Consistency vs Stability
MDP	Robust Policy	?	?	?

Moving to general MDP …

- Linear combination of two stabilizing controllers can be unstable
- Many learning-augmented online algorithms consider black-box predictions or advice
- What if we move beyond black-box advice?

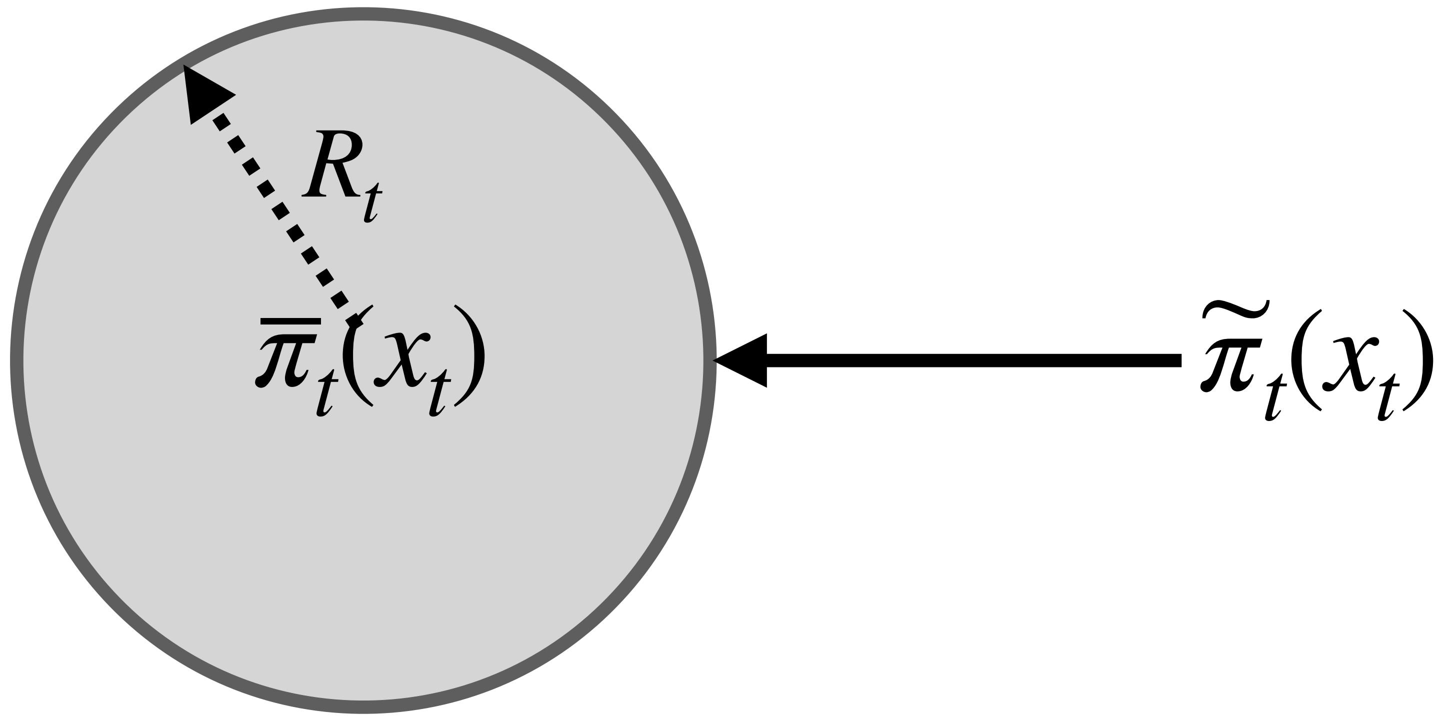
In General MDP ⋯

- Linear combination of two stabilizing controllers can be unstable
- Many learning-augmented online algorithms consider black-box predictions or advice
- What if we move beyond black-box advice?
- Need to consider more structural information, i.e., grey-box agents

Value-Based RL $\widetilde{\pi} : \mathsf{X} \to \mathsf{U}$

$$\widetilde{u}_t = \mathrm{arginf}_{u \in \mathsf{U}} \widetilde{Q}_t(u, x_t)$$

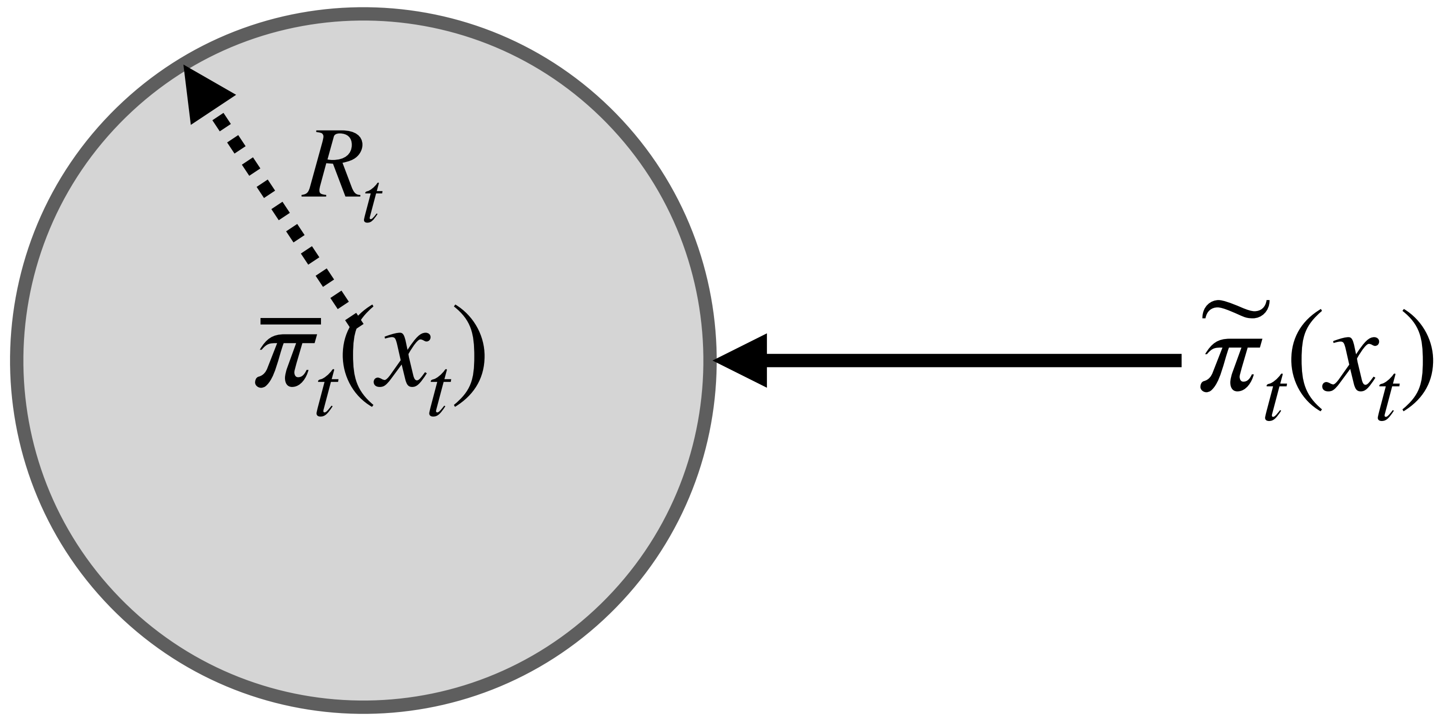
Q-value functions contain useful information



In General MDP ⋯

- Linear combination of two stabilizing controllers can be unstable
- Many learning-augmented online algorithms consider black-box predictions or advice
- What if we move beyond black-box advice?
- Need to consider more structural information, i.e., grey-box agents

How to select R_t?



In General MDP ···

Classic Agent

$\overline{\pi} : \mathsf{X} \rightarrow \mathsf{U}$

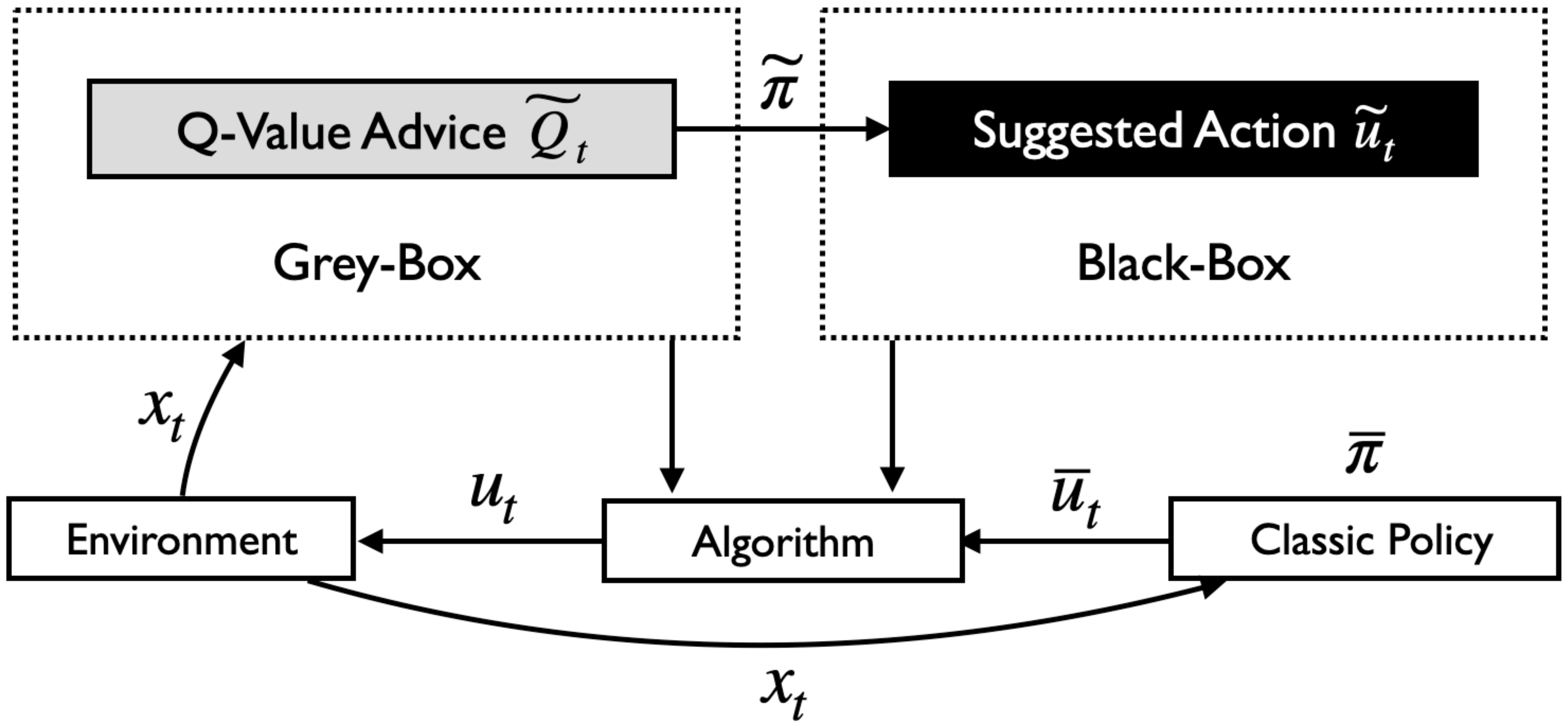
State Space: X

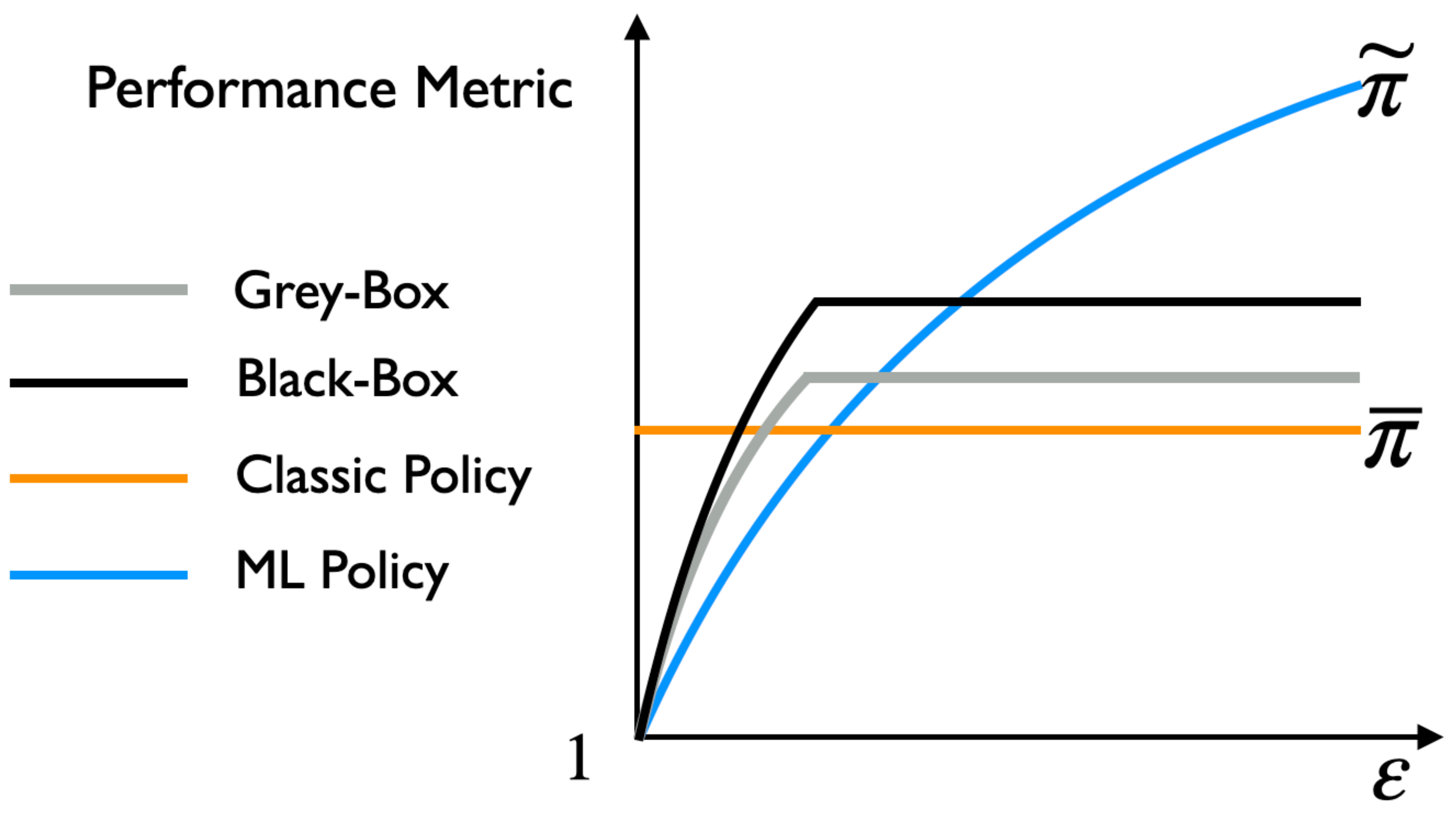
Action Space: U

ML Agent

$\widetilde{\pi} : \mathsf{X} \rightarrow \mathsf{U}$

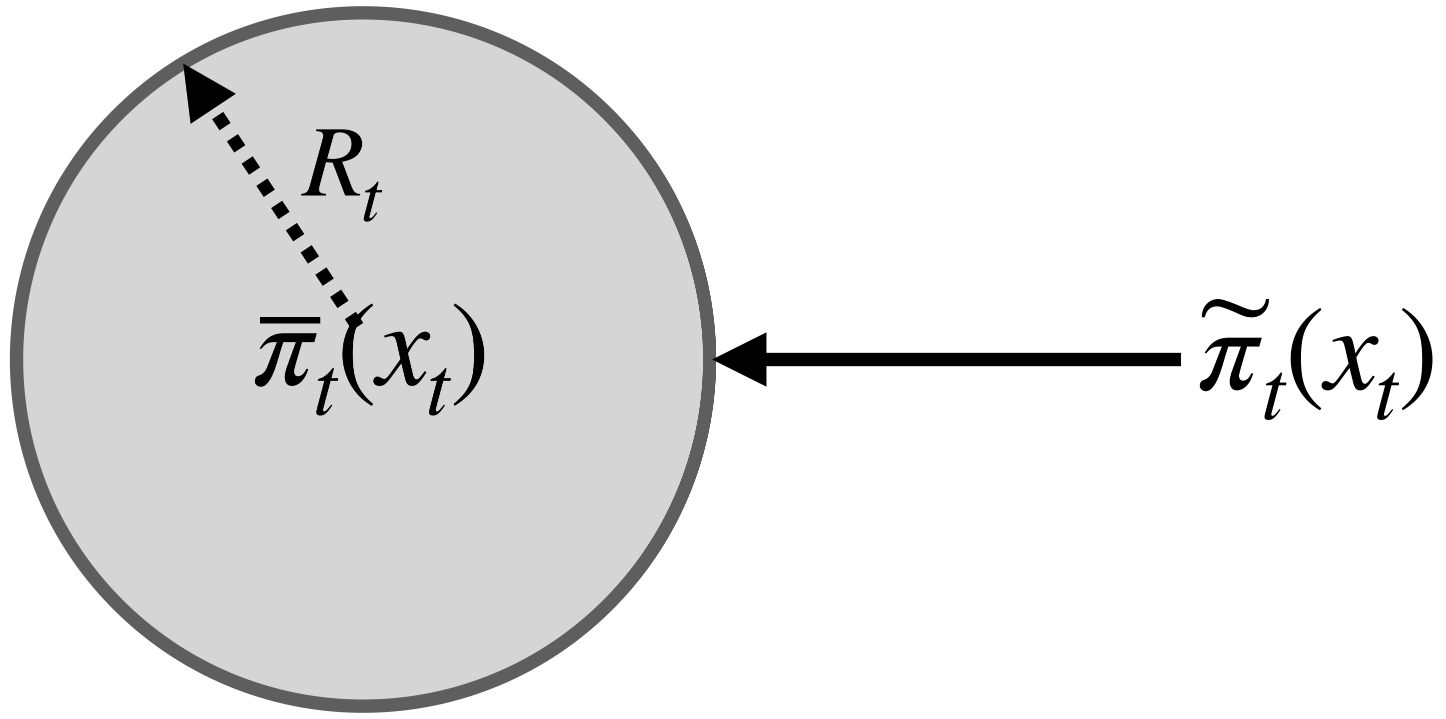
Value-Based RL $\widetilde{u}_t = \mathrm{arginf}_{u \in \mathsf{U}} \widetilde{Q}_t(u, x_t)$





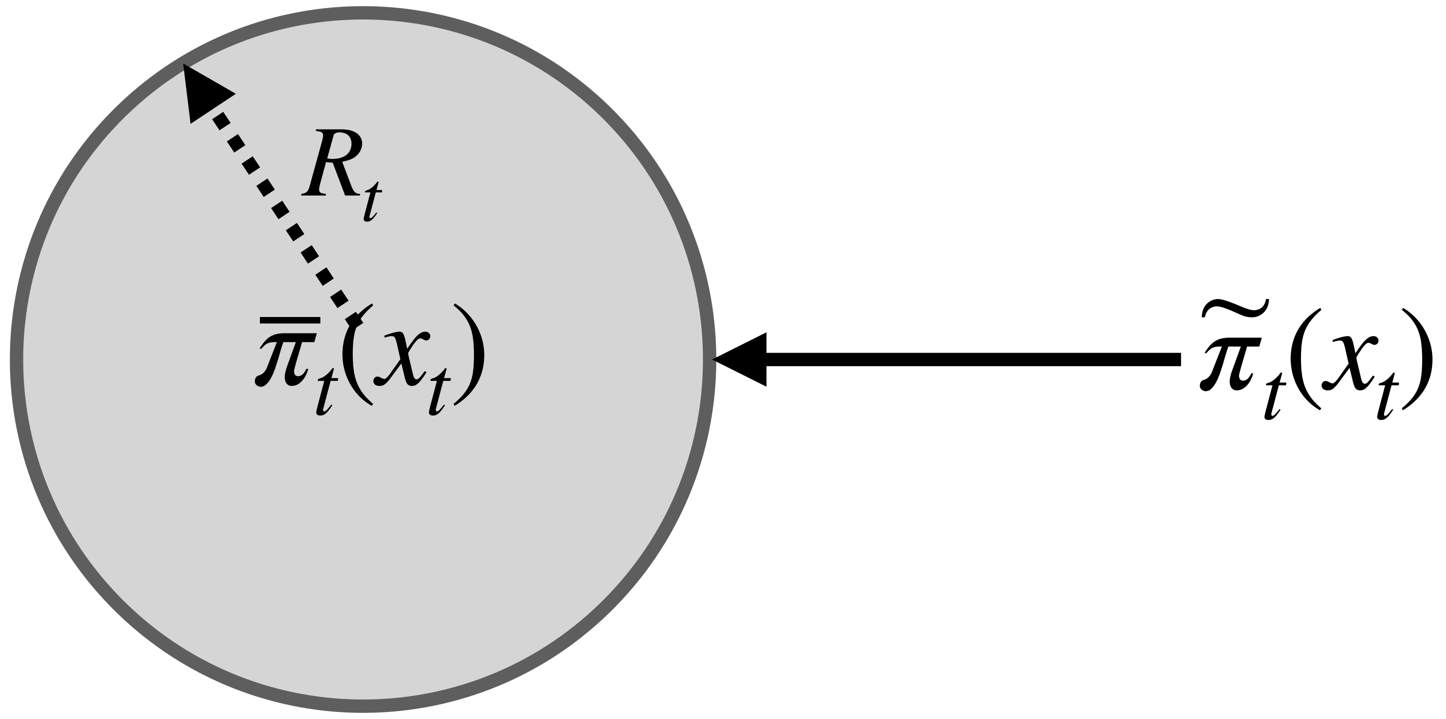
In the General MDP setting, can **Q-value advice** provide a better consistency vs robustness tradeoff?

Idea: Use Temporal Difference (TD) Error



Temporal Difference (TD)-Error: $\mathrm{TD}_t = c_{t-1} + \mathbb{P}_{t-1}\widetilde{V}_t - \widetilde{Q}_{t-1}$

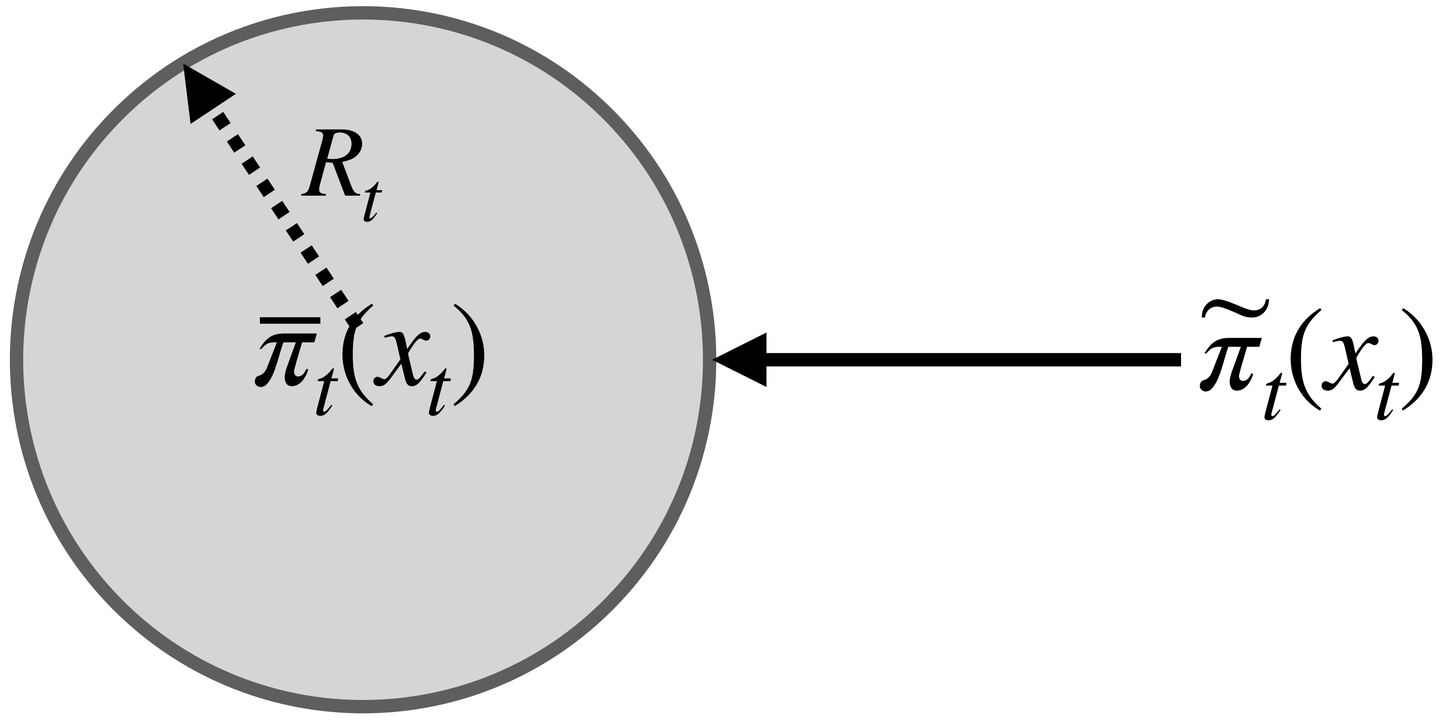
Idea: Use Temporal Difference (TD) Error



Temporal Difference (TD)-Error: $\mathrm{TD}_t = c_{t-1} + \mathbb{P}_{t-1}\widetilde{V}_t - \widetilde{Q}_{t-1}$

Hard to compute since we don't know $\mathbb{P}$

Idea: Use Temporal Difference (TD) Error



Temporal Difference (TD)-Error: $\mathrm{TD}_t = c_{t-1} + \mathbb{P}_{t-1}\widetilde{V}_t - \widetilde{Q}_{t-1}$

Approximate TD-Error: $\delta_t\left(x_t, x_{t-1}, u_{t-1}\right) := c_{t-1}\left(x_{t-1}, u_{t-1}\right) + \inf_{v \in \mathcal{U}} \widetilde{Q}_t\left(x_t, v\right) - \widetilde{Q}_{t-1}\left(x_{t-1}, u_{t-1}\right)$

$$R_t := \left[\underbrace{\|\tilde{\pi}_t\left(x_t\right) - \bar{\pi}_t\left(x_t\right)\|_{\mathsf{U}}}_{\text{Decision Discrepancy}\,\eta_t} - \frac{\beta}{L_Q}\sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}} \right]^+$$

β is a hyper-parameter

Lipschitz constant of costs/rewards

Robust Baseline $\overline{\pi}$

Not all classic policies can be used ⋯

We need to regulate the behaviors of the **classic policies** so they become baselines (to guarantee worst-case performance)

Robust Baseline $\overline{\pi}$

Definition (r-locally p-Wasserstein robustness)

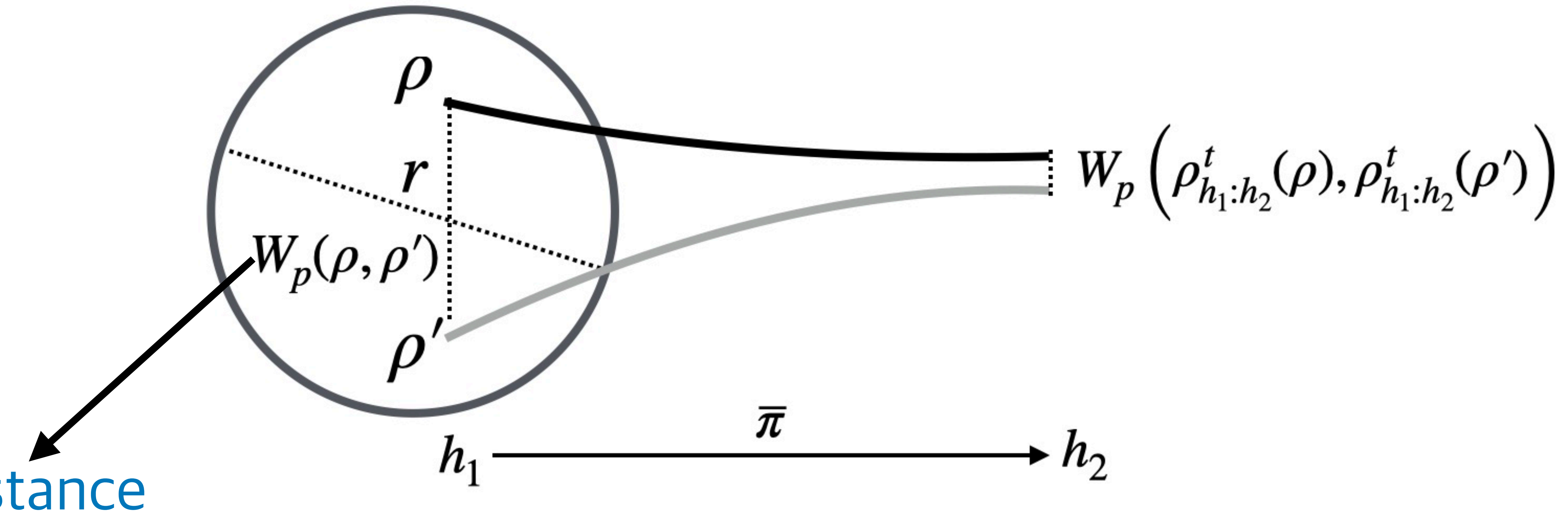
A policy $\overline{\pi} = (\pi_t : t \in [T])$ is r-locally p-Wasserstein-robust if for any $0 \leq t_1 \leq t_2 < T$ and state-action distributions ρ, ρ' such that $W_p(\rho, \rho') \leq r$, for some radius $r > 0$,

$$W_p\left(\rho_{t_1:t_2}(\rho), \rho_{t_1:t_2}(\rho')\right) \leq s(t_2 - t_1) W_p\left(\rho, \rho'\right)$$

for some function $s : [T] \to \mathbb{R}_+$ such that $\sum_{t \in [T]} s(t) \leq C_s$ for some constants $C_s > 0$.

Robust Policy

A general class of **robust** classic policies



Wasserstein Distance
of state-action distributions ρ and ρ'

Wasserstein Distance
of state-action distributions ρ and ρ'
after applying a robust policy $\overline{\pi}$
from h_1 to h_2

Many practical instances:

- Discrete MDP: Any Policy that Induced a Regular Markov Chain
- Time-varying LQR: MPC with Robust Predictions
- Extends a contraction property in [Lin 2022]

Lin, Y., Hu, Y., Qu, G., Li, T. and Wierman, A., 2022. Bounded-regret mpc via perturbation analysis: Prediction error, constraints, and nonlinearity. NeurIPS 2022.

PROjection Pursuit Policy (PROP)

Algorithm PROjection Pursuit (**PROP**)

Initialize: $\widetilde{\pi} = \left(\widetilde{\pi}_t : t \in [T]\right)$ and $\overline{\pi} = \left(\overline{\pi}_t : t \in [T]\right)$

for $t = 0, \ldots, T-1$

Get R_t using approximate TD-error

Take $u_t = \mathrm{Proj}_{\overline{\mathsf{U}}_t}(\widetilde{u}_t)$ where $\overline{\mathsf{U}}_t := \left\{u \in \mathscr{U} : \|u - \overline{\pi}_t\left(x_t\right)\|_{\mathsf{U}} \leq R_t\right\}$

Sample next state $x_{t+1} \sim \mathbb{P}_t(x_t, u_t)$

OOD-Aware EV Charging

Algorithm OOD-Aware EV **Charging (OOD-Charging)**

Initialize: $\widetilde{\pi} = \left(\widetilde{\pi}_t : t \in [T]\right)$ and $\overline{\pi} = \left(\overline{\pi}_t : t \in [T]\right)$

MPC Procedure with user inputs and estimated state

for $t = 0,\ldots,T-1$

NN that updates every t

Approximately Wasserstein robust

Receive user inputs

Get R_t using approximate TD-error

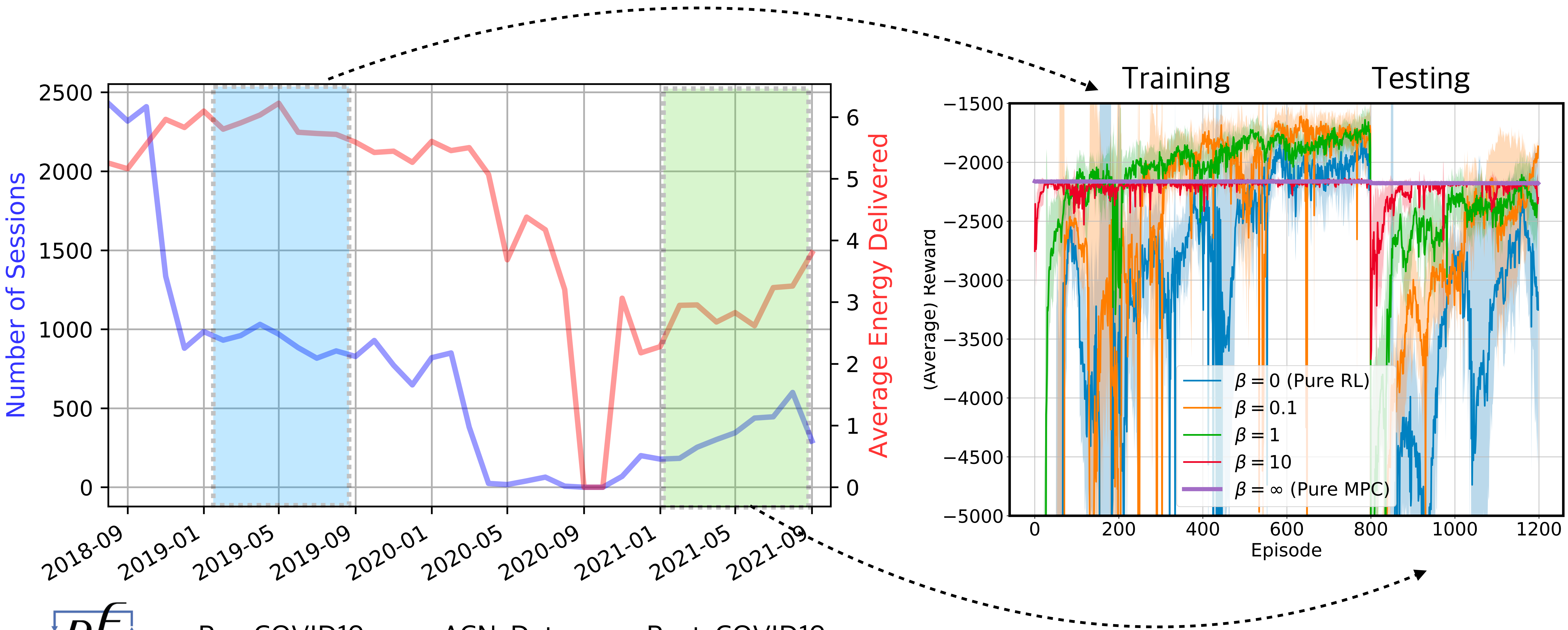
Take $u_t = \mathrm{Proj}_{\overline{\mathrm{U}}_t}(\widetilde{u}_t)$ where $\overline{\mathrm{U}}_t := \left\{u \in \mathscr{U} : \|u - \overline{\pi}_t(x_t)\|_{\mathrm{U}} \leq R_t\right\}$

Sample next state $x_{t+1} \sim \mathbb{P}_t(x_t, u_t)$

Estimate previous state $\widetilde{x}_t$

Update replay buffer and retrain $\widetilde{\pi}$

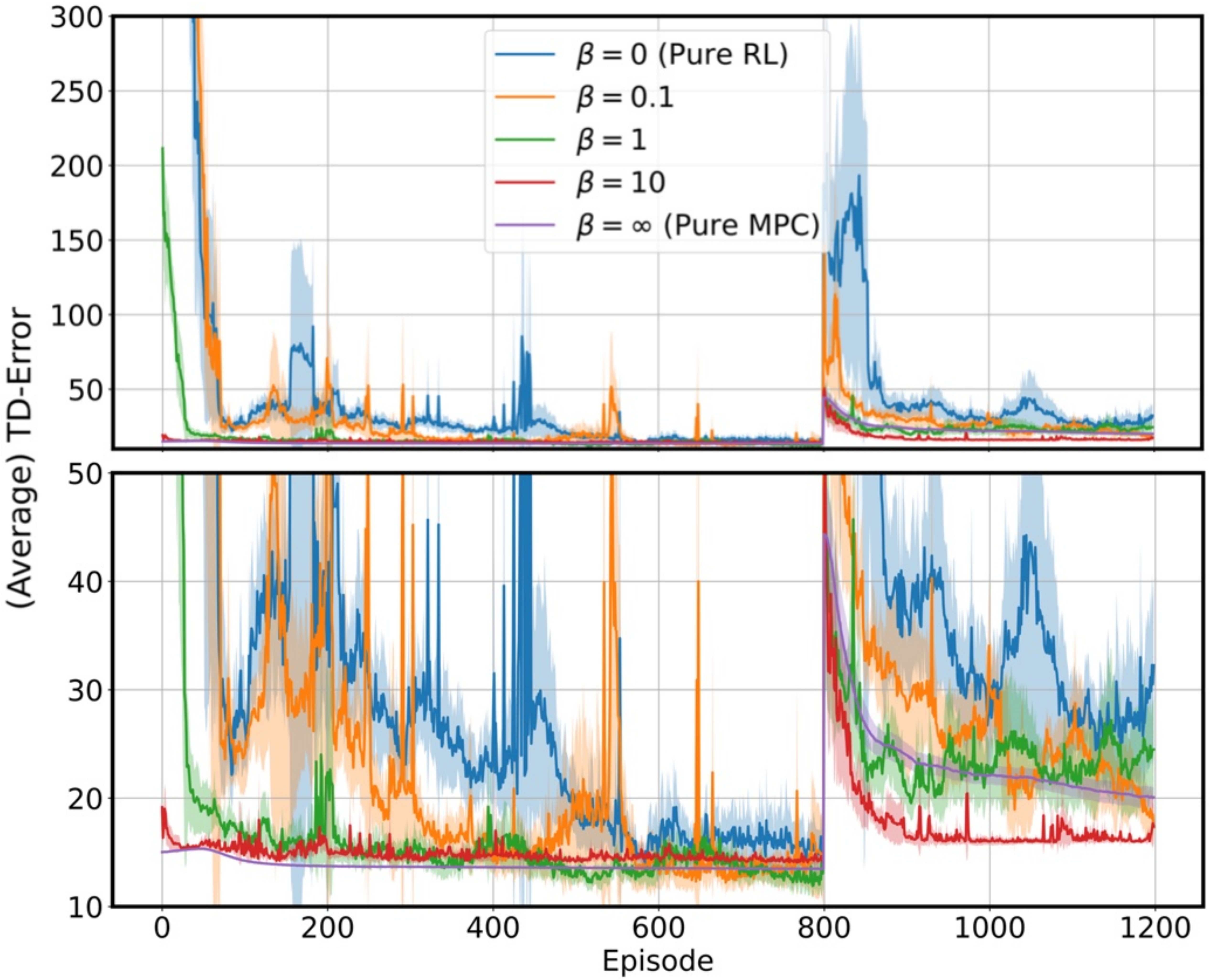
Out-of_Distribution EV Charging



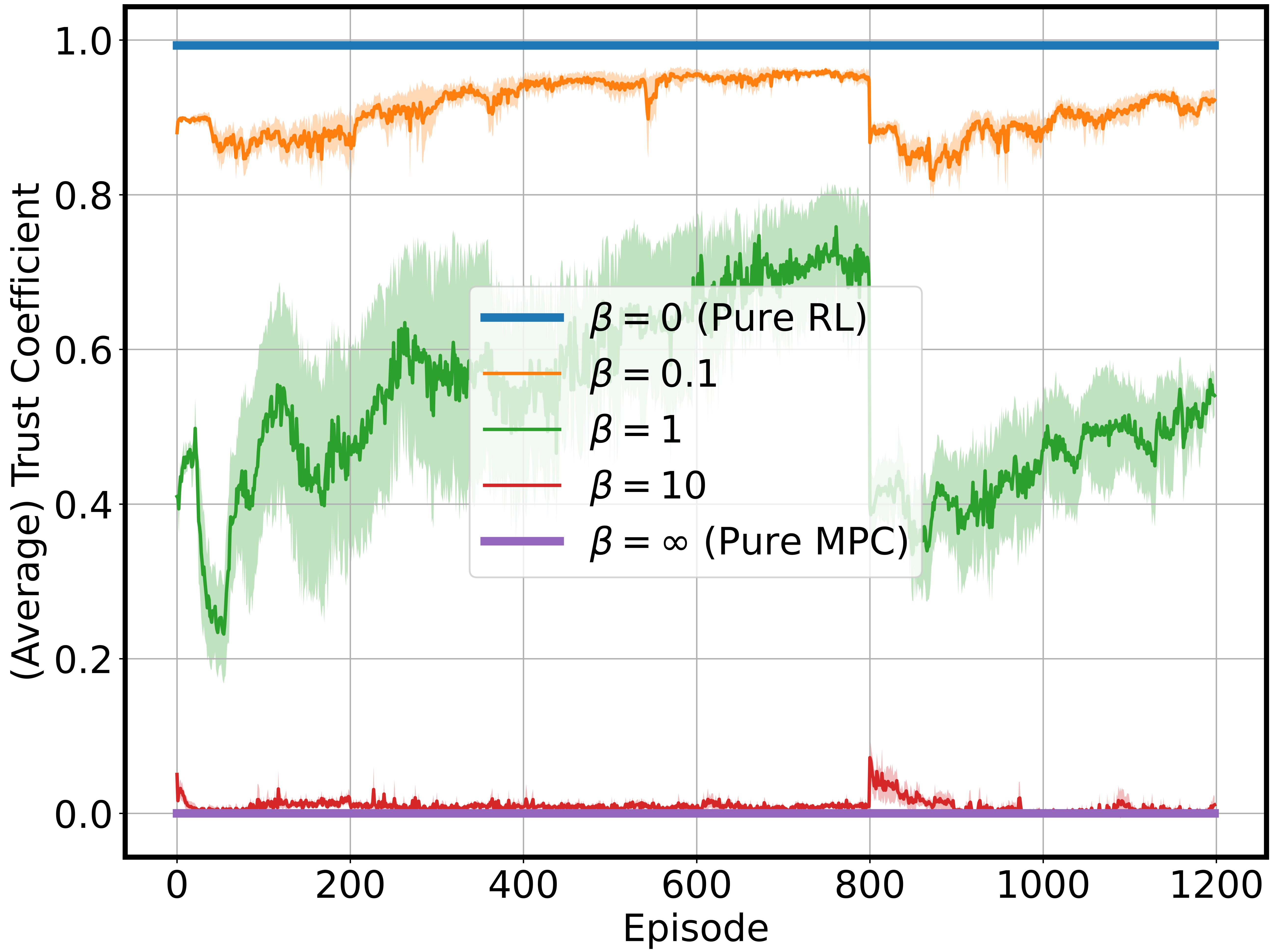
Pre-COVID19 ACN-Data Post-COVID19

Caltech

Out-of_Distribution EV Charging



Out-of_Distribution EV Charging



Trust Coefficient

$$\lambda(R_t) = \min\left\{1, R_t / \|\widetilde{\pi}_t(x_t) - \overline{\pi}_t(x_t)\|_2\right\}$$

Theoretical Guarantees

Consistency and Robustness

k-**Consistency**: Ratio of Expectations (RoE) satisfies $\mathrm{RoE}(\varepsilon) \leq k$ for $\varepsilon = 0$

l-**Robustness**: Ratio of Expectations (RoE) satisfies $\mathrm{RoE}(\varepsilon) \leq k$ for any ε

where $\quad \varepsilon := \sum_{t \in [T]} \left(\| \widetilde{Q}_t - Q_t^\star \|_\infty + \| \inf_{v \in \mathcal{U}} \widetilde{Q}_t - \inf_{v \in \mathcal{U}} Q_t^\star \|_\infty \right)$

(can be generalize to:)

$$\varepsilon(p, \rho) := \sum_{t \in [T]} \left(\| \widetilde{Q}_t - Q_t^\star \|_{p, \rho_t} + \left\| \inf_{v \in \mathcal{U}} \widetilde{Q}_t - \inf_{v \in \mathcal{U}} Q_t^\star \right\|_{p, \phi_t} \right)$$

Black-Box Impossibility

Theorem (Informal)

There exists an algorithm with a black-box agent that is $\big(1+\mathcal{O}((1-\lambda)\gamma)\big)$-consistent and $(\text{ROB}+\mathcal{O}(\lambda\gamma))$-robust where $0 \le \lambda \le 1$ is a hyper-parameter.

(ROB is a ratio of expectation upper bound for the robust baseline)

Theorem (Informal) Impossibility

Any algorithm with a black-box agent cannot be both $\big(1+o((1-\lambda)\gamma)\big)$-consistent and $(\text{ROB}+o(\lambda\gamma))$-robust for any $0 \le \lambda \le 1$.

Proof Highlights

Theorem (Informal) Impossibility

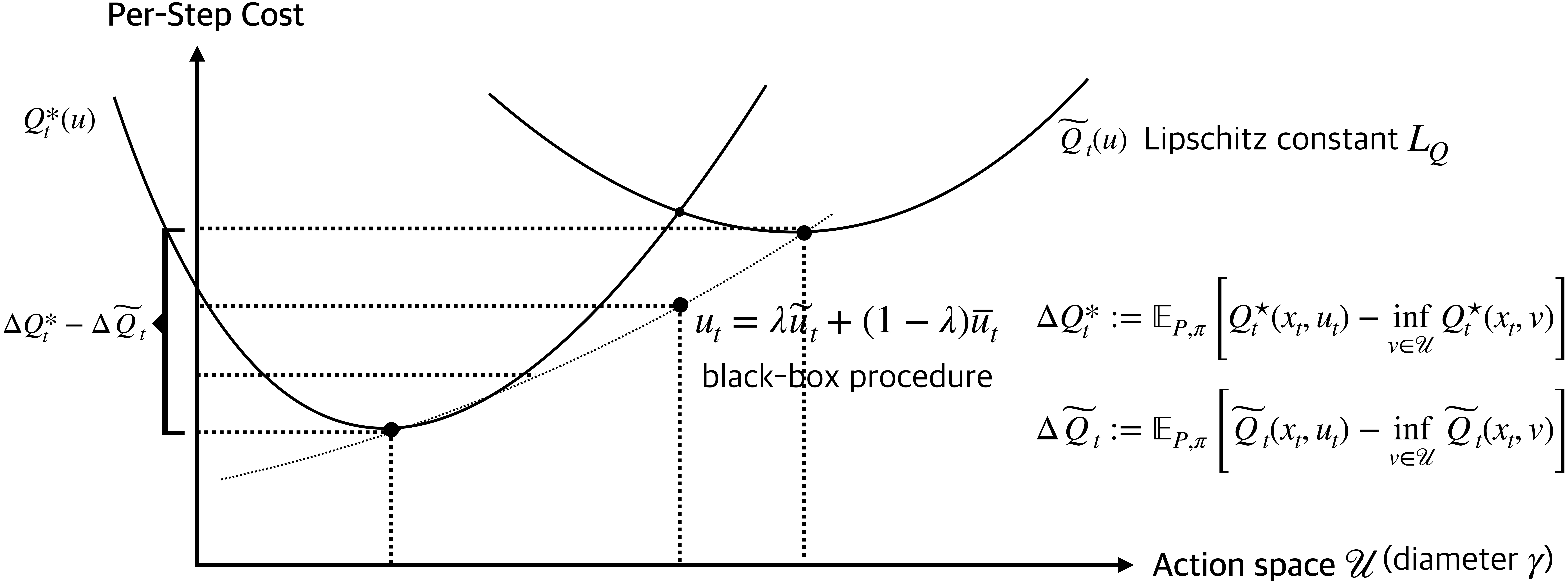
Any algorithm with a black-box agent cannot be both $\left(1 + o((1-\lambda)\gamma)\right)$-consistent and $(\text{ROB} + o(\lambda\gamma))$-robust for any $0 \leq \lambda \leq 1$.

Proof Idea:

Construct a special case (satisfying all model assumptions) with decoupled and identical cost at each t

Then argue with fixed λ, can separate Q^* and $\widetilde{Q}$ so that a lower bound can be derived

Proof Highlights



$$\Delta Q_t^* := \mathbb{E}_{P,\pi}\left[Q_t^\star(x_t, u_t) - \inf_{v \in \mathcal{U}} Q_t^\star(x_t, v)\right]$$

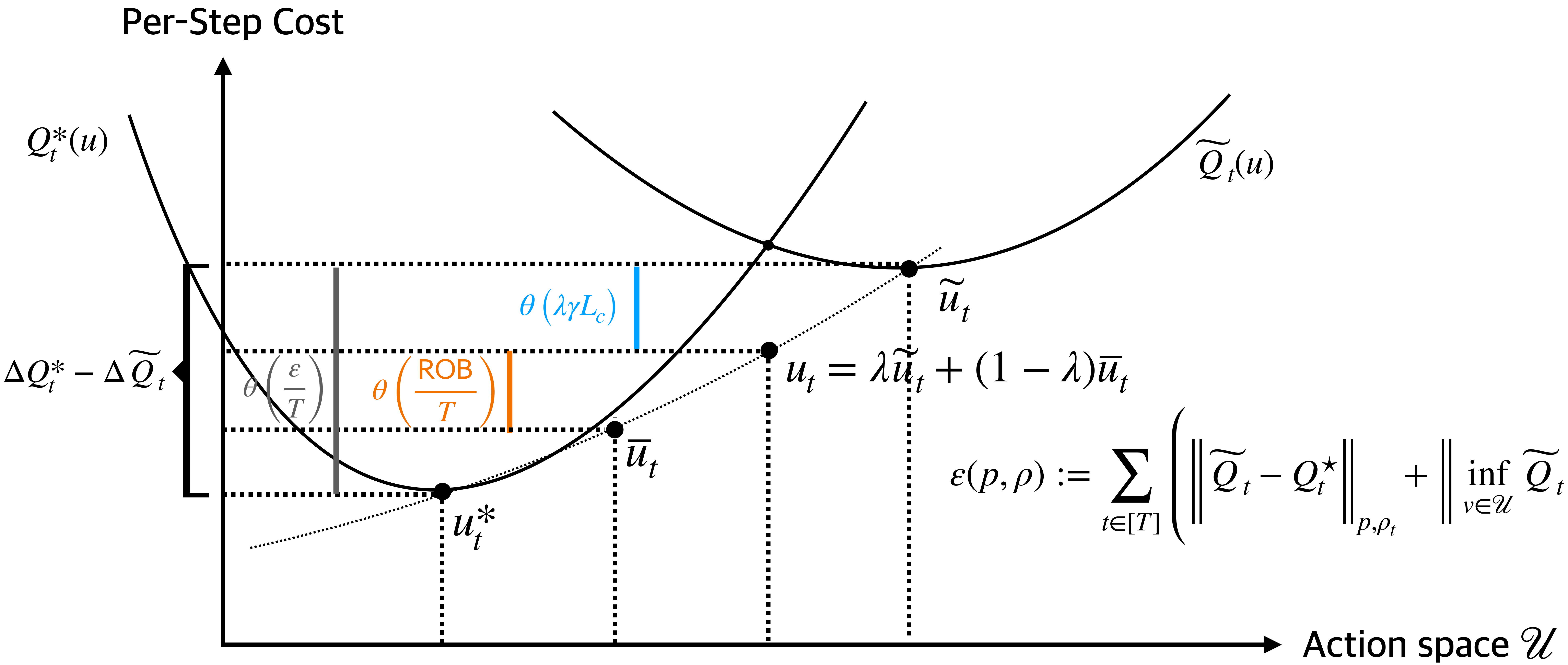
$$\Delta \widetilde{Q}_t := \mathbb{E}_{P,\pi}\left[\widetilde{Q}_t(x_t, u_t) - \inf_{v \in \mathcal{U}} \widetilde{Q}_t(x_t, v)\right]$$

Dynamic regret:

$$\text{DR}\left(\text{PROP(Black-Box)}\right) = \sum_{t \in [T]} \mathbb{E}_{P,\pi}\left[Q_t^\star\left(x_t, u_t\right) - \inf_{v \in \mathcal{U}} Q_t^\star\left(x_t, v\right)\right]$$

$$\text{DR}\left(\text{PROP(Black-Box)}\right) \geq \sum_{t \in [T]} \left(\Delta Q_t^\star(P, \pi) - \Delta \widetilde{Q}_t(P, \pi) + (1-\lambda) L_Q \gamma\right)$$

Proof Highlights



$$\varepsilon(p,\rho) := \sum_{t\in[T]}\left(\left\|\widetilde{Q}_t - Q_t^{\star}\right\|_{p,\rho_t} + \left\|\inf_{v\in\mathcal{U}}\widetilde{Q}_t - \inf_{v\in\mathcal{U}}Q_t^{\star}\right\|_{p,\phi_t}\right)$$

$$\frac{\sum_{t\in[T]}\left(\Delta Q_t^{\star}(P,\pi) - \Delta\widetilde{Q}_t(P,\pi)\right)}{\text{OPT}} \geq \Omega\left(\text{ROB} + \frac{\lambda\gamma L_c}{\text{OPT}}T\right) \Longrightarrow \text{RoE}(\text{PROP}(\text{Black-Box})) = 1 + \Omega\Big((1-\lambda)L_Q\gamma + \min\{\varepsilon, \lambda\gamma L_c + \text{ROB}\}\Big)$$

(cannot be both $\left(1+o((1-\lambda)\gamma)\right)$-consistent and $(\text{ROB}+o(\lambda\gamma))$-robust for any $0\leq\lambda\leq 1$)

Grey-Box Setting

Theorem (Informal)

PROP with a grey-box agent that is 1-consistent and $(\text{ROB} + o(1))$-robust for some $\beta > 0$.

Take-Aways: Grey-box information can grant nontrivial improvements on the consistency and robustness tradeoff

Out-of-Distribution EV Charging

Constant

Theorem (Informal)

OOD-Charging with a grey-box agent that is $1 + \mathcal{O}(\overline{W})$-consistent and $(\text{MPC} + \mathcal{O}(\overline{W}) + o(1))$-robust for some $\beta > 0$.

$$\text{MPC} \leq \frac{2\xi C^2(1 + C^2)(1 + \overline{A}^2 + \overline{B}^2)}{\mu(1 - \overline{\lambda})^2} \qquad \overline{\lambda} := \left((\overline{\sigma} - \underline{\sigma})/(\overline{\sigma} + \underline{\sigma})\right)^{\frac{1}{2}} \qquad C := \frac{4(\xi + 1 + \overline{A} + \overline{B})}{\underline{\sigma}^2 \cdot \lambda}$$

$$\underline{\sigma} := \min\{\mu, 1\}(\overline{A} + \overline{B} + 1)\left(\xi/(2\mu\xi + \mu\sigma^2)\right)^{\frac{1}{2}}$$

$$\overline{\sigma} := \sqrt{2}(\xi + \overline{A} + \overline{B} + 1)$$

State estimation error: $\left\| \ell_t' - \Delta h_t' \right\| \leq \overline{W}$

Standard assumptions: $\|A_t\| \leq \overline{A} \quad \|B_t\| \leq \overline{B} \quad \mu I_n \preceq Q_t \preceq \xi I_n, \ \mu I_m \preceq R_t \preceq \xi I_m, \ \mu I_n \preceq P \preceq \xi I_n$

Proof Highlights

Theorem (Informal)

PROP with a grey-box agent that is 1-consistent and (ROB $+\, o(1)$)-robust for some $\beta > 0$.

Proof Idea: A general bound on DR, therefore RoE:

p-Wasserstein robustness

$$\mathrm{DR(PROP)} \leq \sum_{t\in[T]} \min\left\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \right\}$$

$\|\widetilde{u}_t - \overline{u}_t\|$

Assume $c_t > 0$ for all t, we can bound RoE

Here, $\mu_t := \zeta_t^V - \zeta_t^Q$

$$\zeta_t^Q\left(x_t, u_t\right) := \widetilde{Q}_t\left(x_t, u_t\right) - Q_t^\star\left(x_t, u_t\right)$$ Q-value error

$$\zeta_t^V\left(x_t\right) := \inf_{v\in\mathcal{U}} \widetilde{Q}_t\left(x_t, v\right) - \inf_{v\in\mathcal{U}} Q_t^\star\left(x_t, v\right)$$ V-value error

Proof Highlights

Fix a choice of projection radii $(R_t : t \in [T])$.

$$\mathrm{DR(PROP)} \leq \sum_{t \in [T]} \min \left\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \right\}$$

Applying a projection lemma

$$J(\pi) - J(\bar{\pi}) = \sum_{t \in [T]} \mathbb{E}_{(x,u)\sim\rho_t}\left[c_t(x,u)\right] - \mathbb{E}_{(x,x)\sim\bar{\rho}_t}\left[c_t(x,u)\right]$$

Applying the Kantorovich-Rubinstein duality theorem

$$\leq L_C \sum_{t \in [T]} \sum_{\tau=0}^{t-1} s(\tau) \mathbb{E}_{P,\pi}\left[\left(R_{t-\tau}\right)^p\right]^{1/p}$$

Applying the Wasserstein robustness definition

$$\leq L_C C_s \sum_{t \in [T]} \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}$$

Proof Highlights

$$\mathrm{DR(PROP)} \leq \sum_{t\in[T]} \min \left\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \right\}$$

Consistency: Let $\varepsilon = 0$, $\mathbb{E}_{P,\pi}\left[\mu_t\right] = 0$ and the consistency bound becomes

$$\mathbb{E}_{P,\pi}\left[\eta_t - R_t\right] \leq \mathbb{E}_{P,\pi}\left[\frac{\beta}{L_Q}\sum_{s=1}^{t}\delta_s\right] = 0$$

Applying the radius update rule:

$$R_t := \left[\underbrace{\left\| \widetilde{\pi}_t\left(x_t\right) - \overline{\pi}_t\left(x_t\right) \right\|_{\mathrm{U}}}_{\text{Decision Discrepancy } \eta_t} - \frac{\beta}{L_Q}\sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}} \right]^+$$

Proof Highlights

$$\text{DR(PROP)} \leq \sum_{t\in[T]} \min \left\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \right\}$$

Robustness:

$$R_t := \left[\underbrace{\left\| \widetilde{\pi}_t\left(x_t\right) - \overline{\pi}_t\left(x_t\right) \right\|_{\mathcal{U}}}_{\text{Decision Discrepancy } \eta_t} - \frac{\beta}{L_Q} \sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}} \right]^+$$

TD-Error: $\text{TD}_t = c_{t-1} + \mathbb{P}_{t-1}\widetilde{V}_t - \widetilde{Q}_{t-1}$

Approximate TD-Error: $\delta_t\left(x_t, x_{t-1}, u_{t-1}\right) := c_{t-1}\left(x_{t-1}, u_{t-1}\right) + \inf_{v\in\mathcal{U}} \widetilde{Q}_t\left(x_t, v\right) - \widetilde{Q}_{t-1}\left(x_{t-1}, u_{t-1}\right)$

Key observation: $\mu_t - \delta_t = \zeta_{t-1}^Q - \zeta_t^Q \qquad (\zeta_{-1}^Q = 0)$

$$\implies \sum_{s=0}^{t}\left(\mu_s - \delta_s\right) = \sum_{s=0}^{t}\left(\zeta_{s-1}^Q - \zeta_s^Q\right) = \zeta_t^Q$$

Proof Highlights

$$\text{DR(PROP)} \leq \sum_{t\in[T]} \min\left\{\underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}}\right\}$$

Robustness:

$$R_t := \left[\underbrace{\left\| \widetilde{\pi}_t\left(x_t\right) - \overline{\pi}_t\left(x_t\right) \right\|_{\mathsf{U}}}_{\text{Decision Discrepancy } \eta_t} - \frac{\beta}{L_Q} \sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}}\right]^+$$

$\exists\ \Delta = o(T)$ such that $|\zeta_t^Q| \leq \Delta$ for all $t \in [T]$ (model assumption)

Consider two cases:

Case I $\sum_{t\in[T]} \mu_t \leq \Delta$ Automatically obtain $\text{RoE}(\varepsilon) \leq \text{ROB} + o(1)$ by the consistency bound

Case II $\sum_{t\in[T]} \mu_t > \Delta$ (Cont.)

Proof Highlights

$$\text{DR(PROP)} \leq \sum_{t\in[T]} \min \left\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \right\}$$

Robustness:

$$R_t := \left[\underbrace{\left\| \widetilde{\pi}_t\left(x_t\right) - \overline{\pi}_t\left(x_t\right) \right\|_{\mathrm{U}}}_{\text{Decision Discrepancy } \eta_t} - \frac{\beta}{L_Q} \sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}} \right]^+$$

Case II $\sum_{t\in[T]} \mu_t > \Delta$

$\Longrightarrow$ $\sum_{t\in[T]} \delta_t > 0$ (Applying $\sum_{s=0}^{t}\left(\mu_s - \delta_s\right) = \sum_{s=0}^{t}\left(\zeta_{s-1}^{Q} - \zeta_s^{Q}\right) = \zeta_t^{Q}$)

$\Longrightarrow$ There exists $\beta > 0$ such that $R_t = 0$

(The action space is compact, discrepancy η_t is bounded)

Proof Highlights

$$\mathrm{DR}(\mathrm{PROP}) \leq \sum_{t\in[T]} \min \Bigg\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \Bigg\}$$

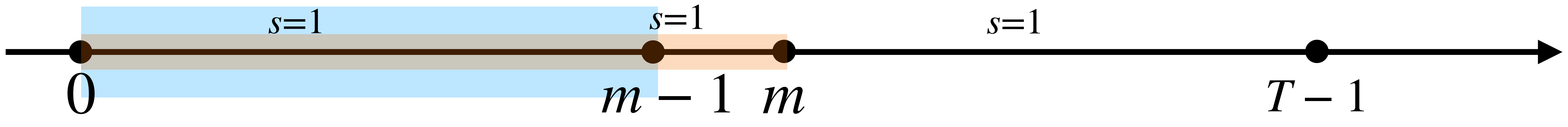
Robustness:

$$R_t := \Bigg[\underbrace{\left\| \widetilde{\pi}_t\left(x_t\right) - \overline{\pi}_t\left(x_t\right) \right\|_{\mathsf{U}}}_{\text{Decision Discrepancy } \eta_t} - \frac{\beta}{L_Q} \sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}} \Bigg]^{+}$$

Case II $\quad \displaystyle\sum_{t\in[T]} \mu_t > \Delta$

Let m be the latest time index such that $\displaystyle\sum_{s=1}^{m} \mu_s > \Delta$ and $\displaystyle\sum_{s=1}^{m-1} \mu_s \leq \Delta$

$$\sum_{s=1}^{m-1} \mu_s \leq \Delta \qquad \sum_{s=1}^{m} \mu_s > \Delta \qquad \sum_{s=1}^{k} \mu_s > \Delta, \forall k \geq m$$



Proof Highlights

$$\text{DR(PROP)} \leq \sum_{t\in[T]} \min \left\{ \underbrace{\mathbb{E}_{P,\pi}\left[\mu_t\right] + L_Q \mathbb{E}_{P,\pi}\left(\eta_t\left(x_t\right) - R_t\right)}_{\text{Consistency Bound}}, \underbrace{\varphi_t + L_C C_s \mathbb{E}_{P,\pi}\left[\left(R_t\right)^p\right]^{1/p}}_{\text{Robustness Bound}} \right\}$$

Robustness:

$$R_t := \left[\underbrace{\left\| \widetilde{\pi}_t\left(x_t\right) - \overline{\pi}_t\left(x_t\right) \right\|_{\mathsf{U}}}_{\text{Decision Discrepancy } \eta_t} - \frac{\beta}{L_Q} \sum_{s=1}^{t} \underbrace{\delta_s\left(x_s, x_{s-1}, u_{s-1}\right)}_{\text{Approximate TD-Error}} \right]^+$$

Case II $\sum_{t\in[T]} \mu_t > \Delta$ (Cont.)

Per-step robustness bound

$$\Longrightarrow \text{DR(PROP(Grey-Box))} \leq \sum_{s=1}^{m-1} \mu_s + \sum_{s=m}^{T-1} \varphi_s + O(\eta_m) \leq \sum_{s=1}^{m-1} \mu_s + \sum_{s=m}^{T-1} \varphi_s + O(\gamma)$$

(Applying the Wasserstein robustness definition)

$$\Longrightarrow \text{RoE}(\varepsilon) \leq \text{ROB} + o(1)$$

Comparison

	Noah & Moitra	This Work
MDP Model	Finite Action/State Spaces **Episodic** setting	Finite or Continuous **Single-trajectory** setting
Assumption on $\widetilde{Q}$	Approximate distillation for stronger results	Lipschitz continuous $\widetilde{Q} - Q* = o(T)$
Robust Policies	N/A	Wasserstein Robustness
Main Results	Regret bound	Consistency-robustness Tradeoff

Golowich, Noah, and Ankur Moitra. "Can Q-learning be Improved with Advice?." Conference on Learning Theory. PMLR, 2022.

Summary: Learning-Augmented Decision-Making

Systems	Untrusted AI		Performance Guarantees
Nonlinear Dynamics	Black-box Policy	[Li et. al. OJCSYS 2023]	Stability + CR
Anytime-Constrained MDP	Black-box Policy	[Yang et. al. NeurIPS 2023]	Regret
MDP	Grey-Box Policy	[Li et. al. NeurIPS 2023]	Ratio of Expectations
Two-Controller System	Black-box Advice	[Li et. al. e-Energy 2020] [Li et. al. TSG 2021] [Li et. al. SIGMETRICS 2021]	Feasibility + CR
Linear Quadratic System	Perturbation Predictions	[Li et. al. SIGMETRICS 2022]	CR
Linear Quadratic System	Disentangled Perturbations	Ongoing	
General Sum Games	Black-box Bayesian probability	Ongoing	

Scenario II : Battery Scheduling

Scenario III : Remand Response

Scenario I : Linear Control

References

(1) **Tongxin Li**, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low.
“Robustness and Consistency in Linear Quadratic Control with Untrusted Predictions"
SIGMETRICS Performance Evaluation Review 50.1 (2022): 107-108.

(2) Jianyi Yang, Pengfei Li, **Tongxin Li**, Adam Wierman, and Shaolei Ren.
“Anytime-competitive reinforcement learning with policy prior"
Advances in Neural Information Processing Systems 36 (2024).

(3) Yejia Liu, Jianyi Yang, Pengfei Li, **Tongxin Li**, and Shaolei Ren.
“Building Socially-Equitable Public Models"
Advances in Neural Information Processing Systems 36 (2024).

(4) **Tongxin Li**, Yiheng Lin, Shaolei Ren, Adam Wierman.
"Beyond Black-Box Advice: Learning-Augmented Algorithms for MDP with Value-Based Policies”
Advances in Neural Information Processing Systems 36 (2024).

(5) **Tongxin Li**, Chenxi Sun.
"Out-of-Distribution-Aware Electric Vehicle Charging"
IEEE Transactions on Transportation Electrification, 2024

(6) **Tongxin Li**, Hao Liu, Yisong Yue.
"Disentangling Linear Quadratic Control with Untrusted ML Predictions"
Preprint, 2024