Combining parametric and nonparametric models for off-policy evaluation

Omer Gottesman[1], Yao Liu[2], Scott Sussex[1], Emma Brunskill[2], Finale Doshi-Velez[1]

[1]Paulson School of Engineering and Applied Science, Harvard University

[2]Department of Computer Science, Stanford University

Introduction

Off-Policy Evaluation –

We wish to estimate the value of a sequential decision making *evaluation policy* from batch data, collected using a *behavior policy* we do not control

Introduction

Model Based vs. Importance Sampling –

Importance sampling methods provide unbiased estimates of the value evaluation policy, but tend to require a huge amount of data to achieve reasonably low variance. When data is limited, model based methods tend to perform better.

In this work we focus on improving model based methods.

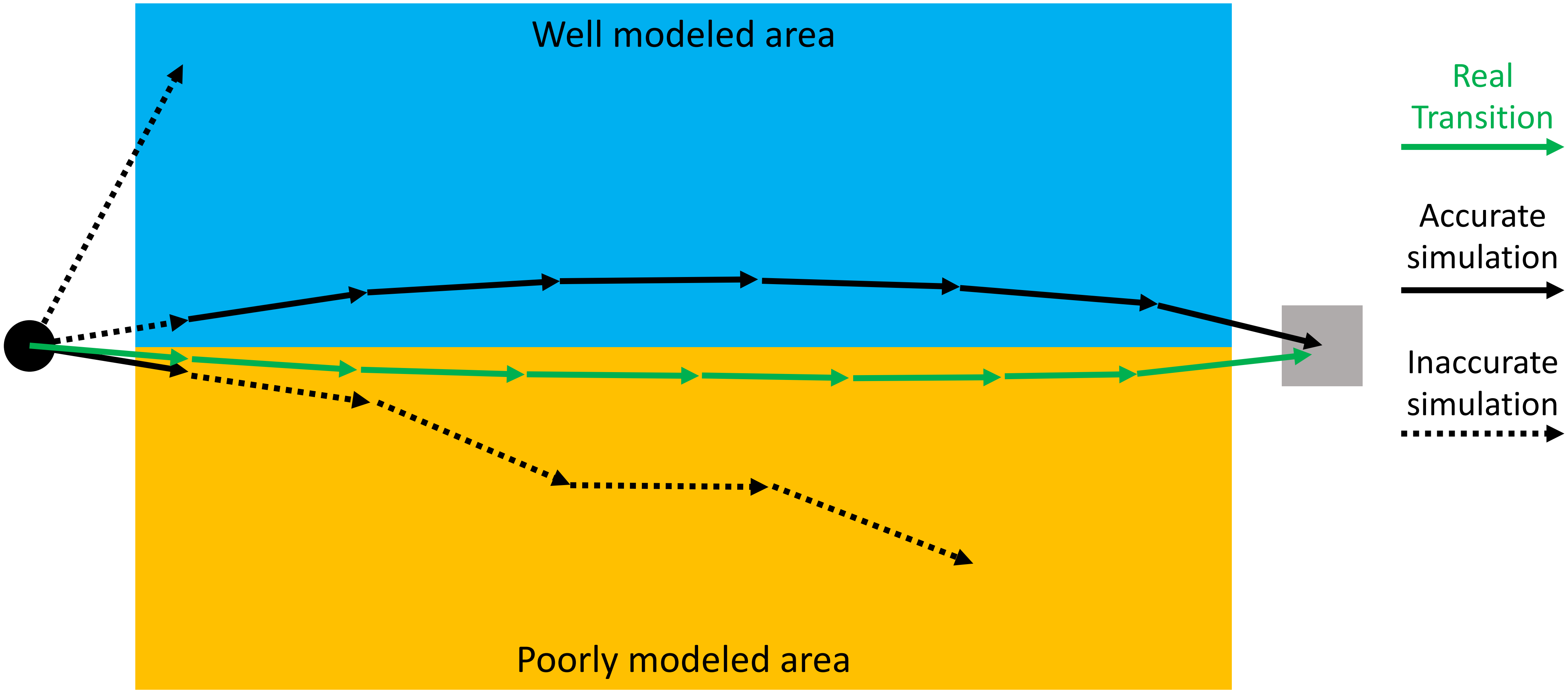
Combining multiple models

Challenge: Hard for one model to be good enough for the entire domain.

Question: If we had multiple models, with different strengths, could we combine them to get better estimates?

Approach: Use a **planner** to decide when to use each model to get the most accurate reward estimate over **entire** trajectories.

Balancing short vs. long term accuracy



Balancing short vs. long term accuracy

$$|g_T - \hat{g}_T| \leq L_r \sum_{t=0}^{T} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \varepsilon_t(t - t' - 1) + \sum_{t=0}^{T} \gamma^t \varepsilon_r(t)$$

Total return error	Error due to state estimation	Error due to reward estimation

$L_{t/r}$ - Lipschitz constants of transition/reward functions
$\varepsilon_{t/r}(t)$ - Bound on model errors for transition/reward at time t
T - Time horizon
γ - Reward discount factor
$g_T \equiv \sum_{t=0}^{T} \gamma^t r(t)$ - Return over entire trajectory

Closely related to bound in - *Asadi, Misra, Littman. "Lipschitz Continuity in Model-based Reinforcement Learning." (ICML 2018).*

Planning to minimize the estimated return error over entire trajectories

We use Monte Carlo Tree Search (MCTS) planning algorithm to minimize the return error bound over entire trajectories.

	Agent	Planner
State:	x_t	(x_t, a_t)
Action:	a_t	Model to use
Reward:	r_t	$-(r_t - \hat{r}_t)$

Parametric vs. Nonparametric Models

Nonparametric models –

Predicting the dynamics for a given state-action pair based on similarity to neighbors.

Nonparametric models can be very accurate in regions of state space where data is abundant.

Parametric Models –

Any parametric regression model or hand coded model incorporating domain knowledge.

Parametric models will tend to generalize better to situations very different from the ones observed in the data.

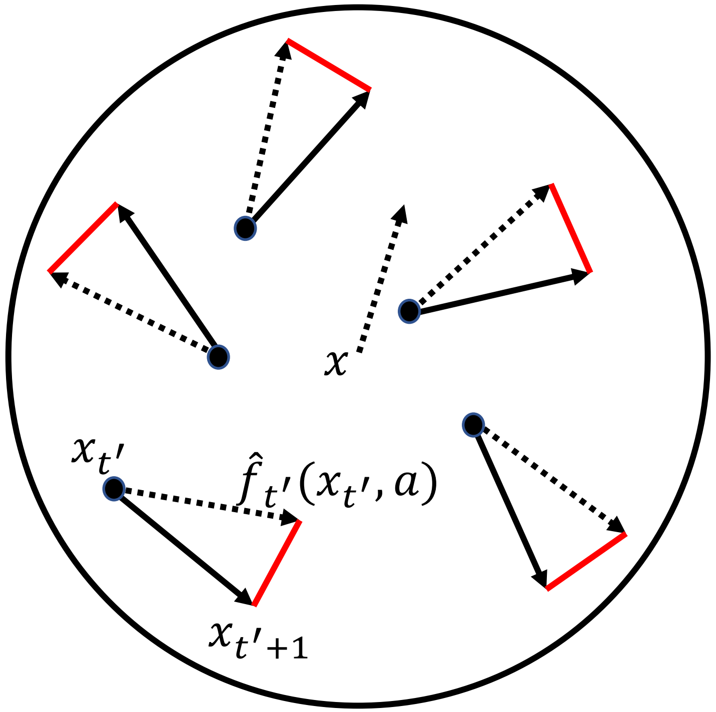
Estimating bounds on model errors

$$|g_T - \hat{g}_T| \leq L_r \sum_{t=0}^{T} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \varepsilon_t(t - t' - 1) + \sum_{t=0}^{T} \gamma^t \varepsilon_r(t)$$

$L_{t/r}$ - Lipschitz constants of transition/reward functions
$\varepsilon_{t/r}(t)$ - Bound on model errors for transition/reward at time t
T - Time horizon
γ - Reward discount factor
$g_T \equiv \sum_{t=0}^{T} \gamma^t r(t)$ - Return over entire trajectory

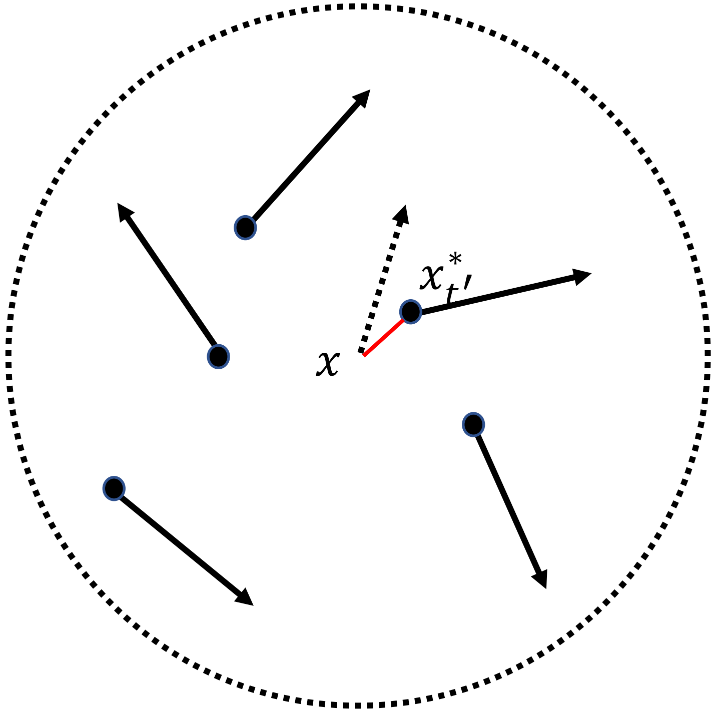
Estimating bounds on model errors

Parametric



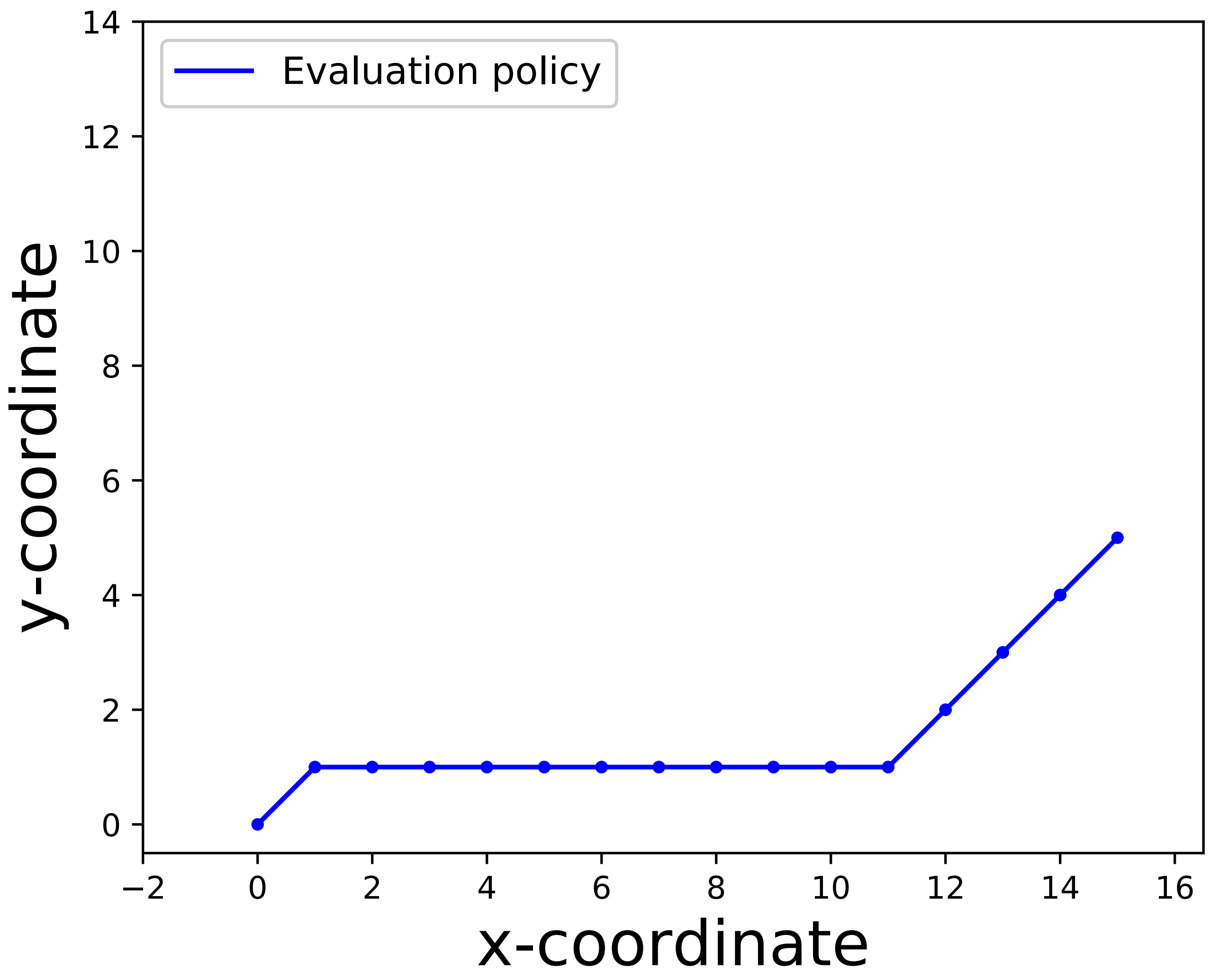
$$\hat{\varepsilon}_{t,p} \approx \max \Delta(x_{t'+1}, \hat{f}_t(x_{t'}, a))$$

Nonparametric

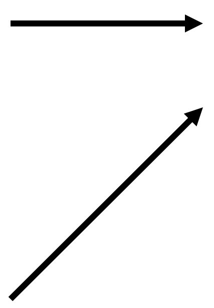


$$\hat{\varepsilon}_{t,np} \approx L_t \cdot \Delta(x, x^*_{t'})$$

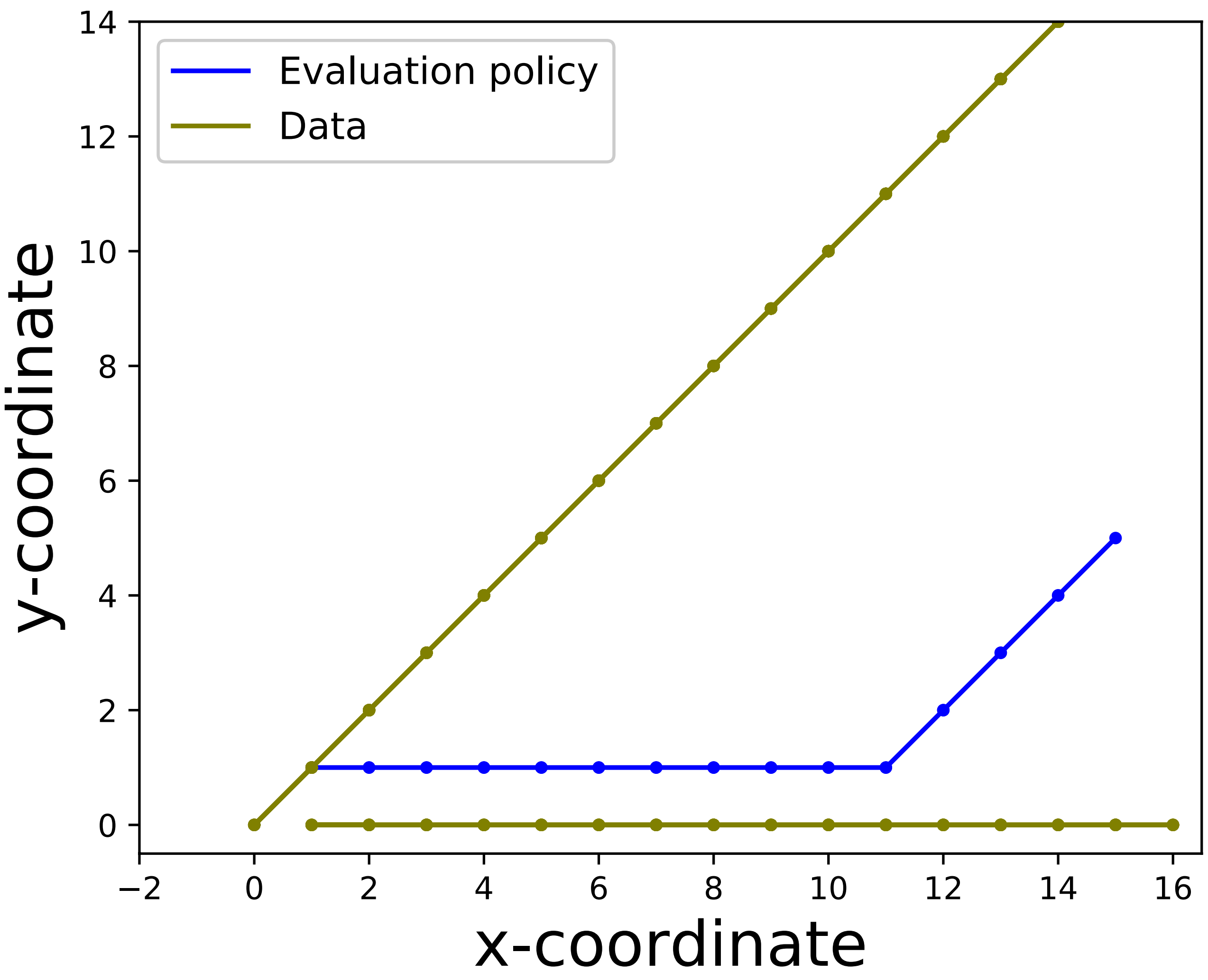
Demonstration on a toy domain



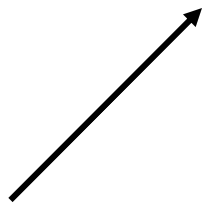
Possible actions



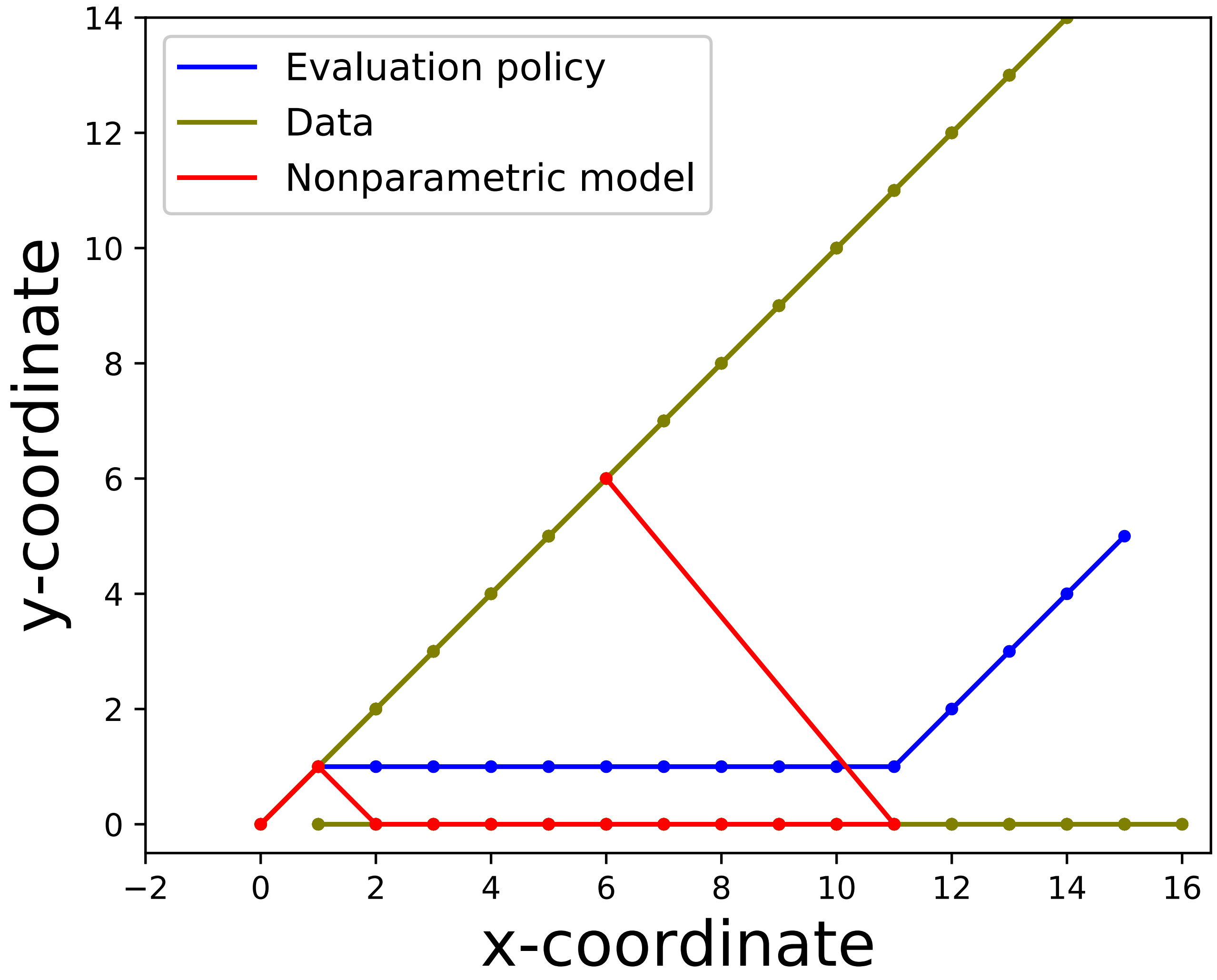
Demonstration on a toy domain



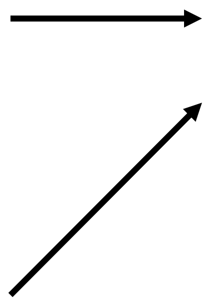
Possible actions



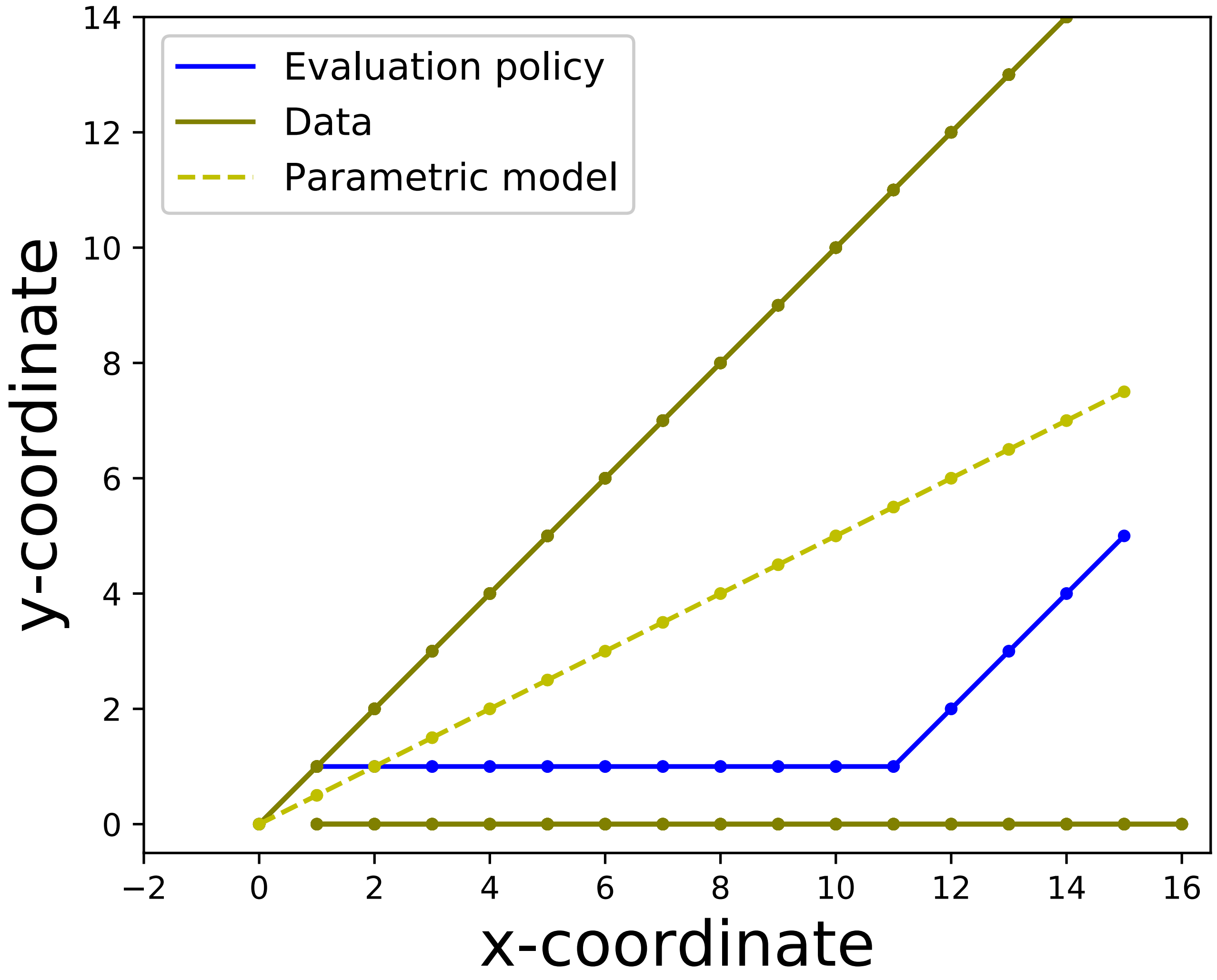
Demonstration on a toy domain



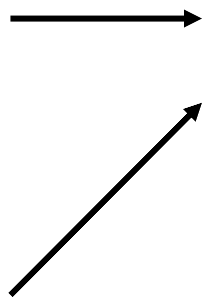
Possible actions



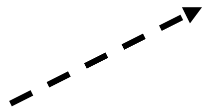
Demonstration on a toy domain



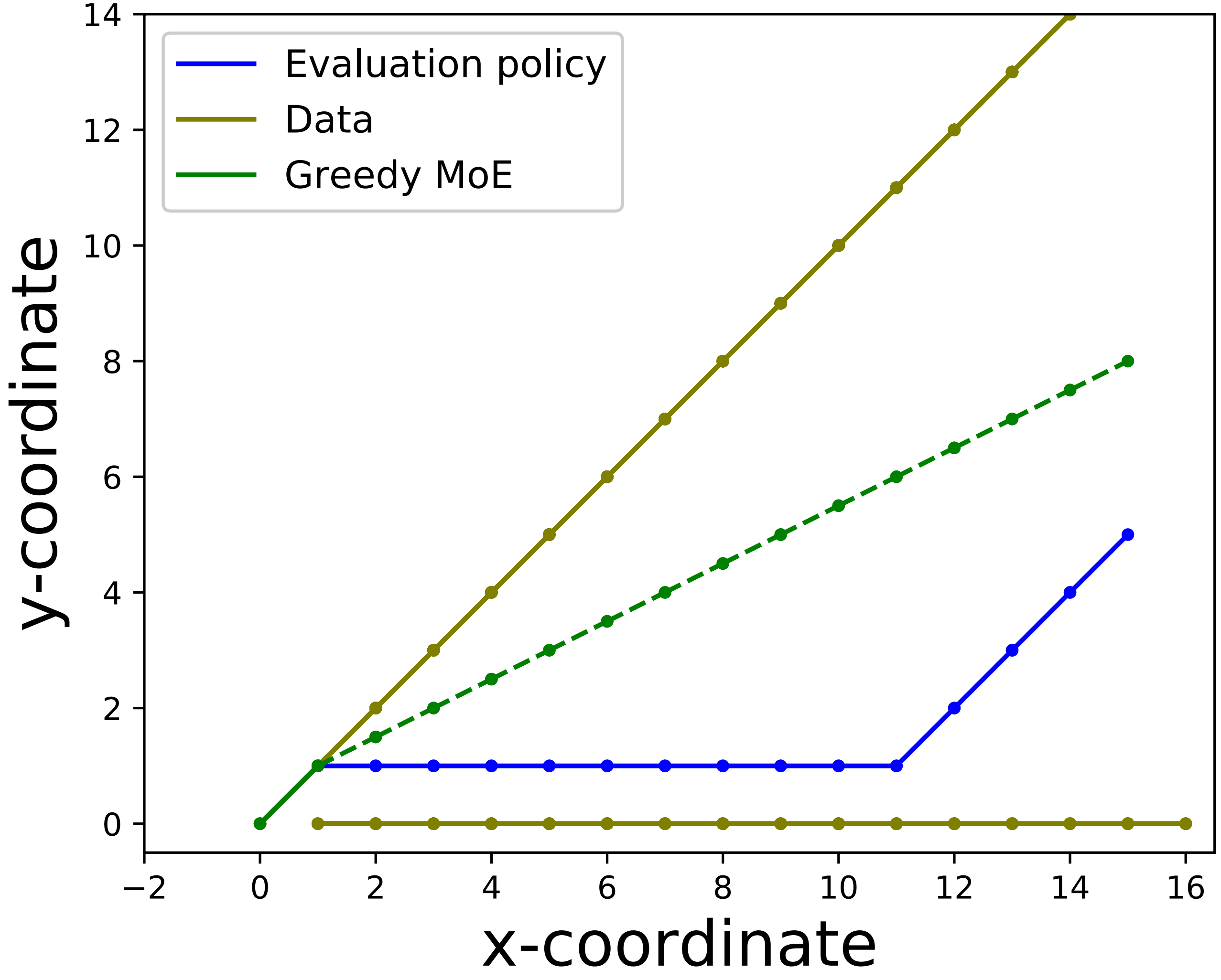
Possible actions

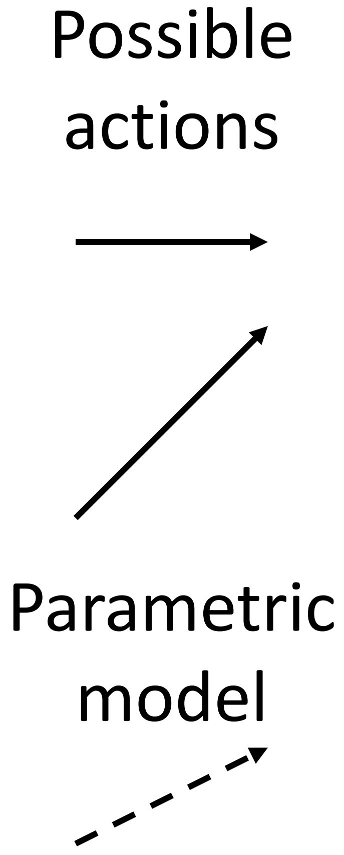


Parametric model

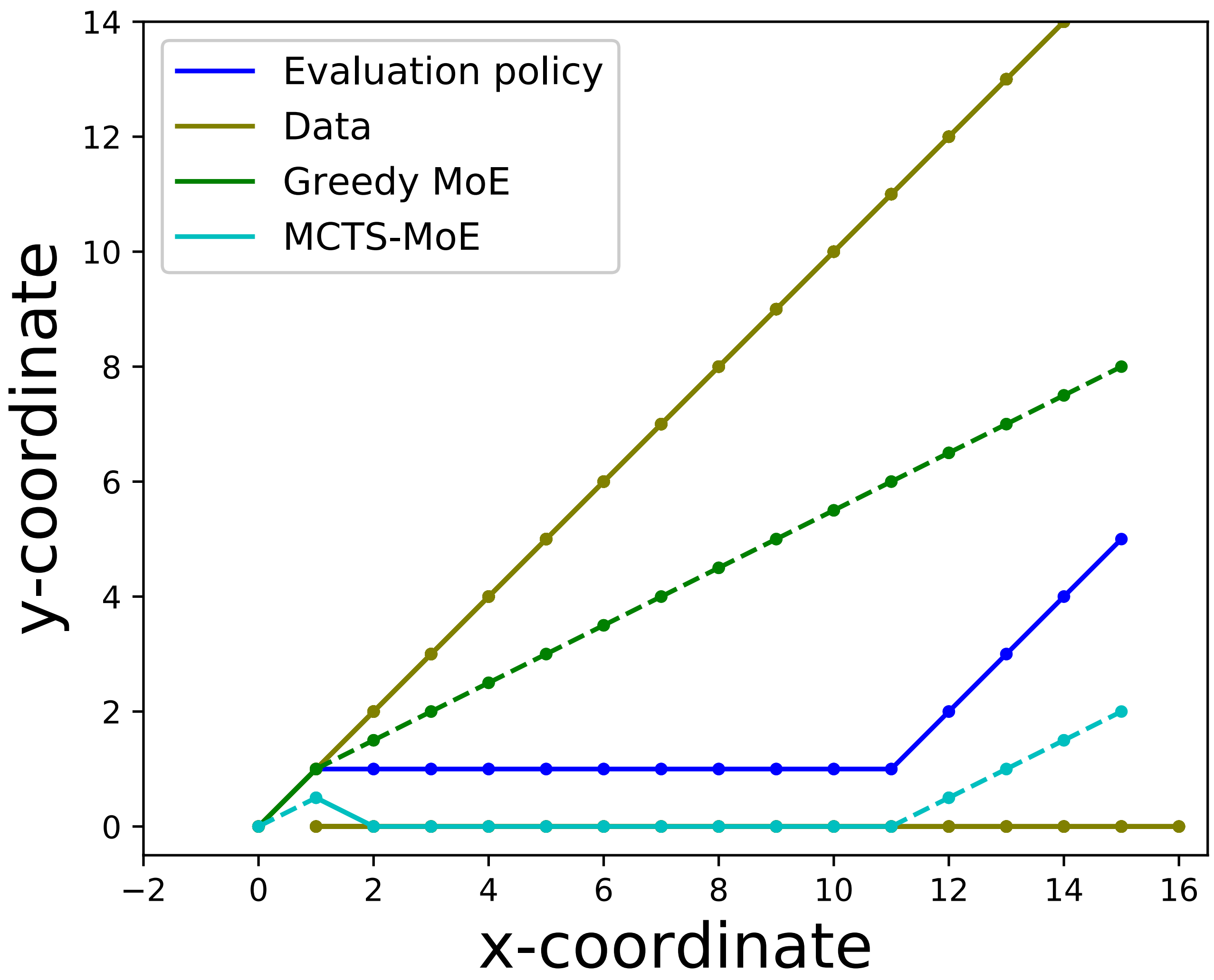


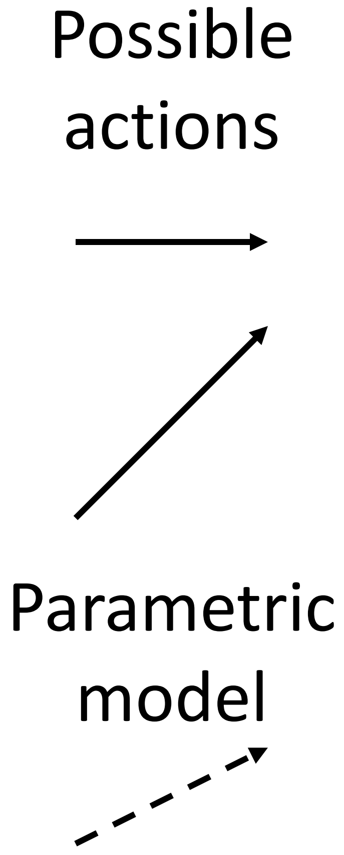
Demonstration on a toy domain





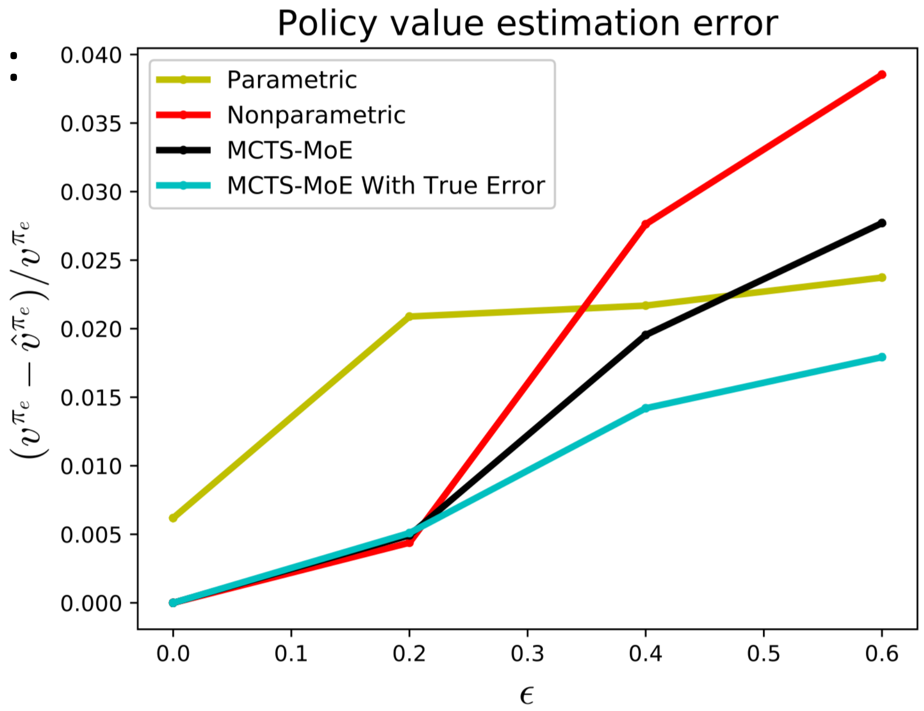
Demonstration on a toy domain



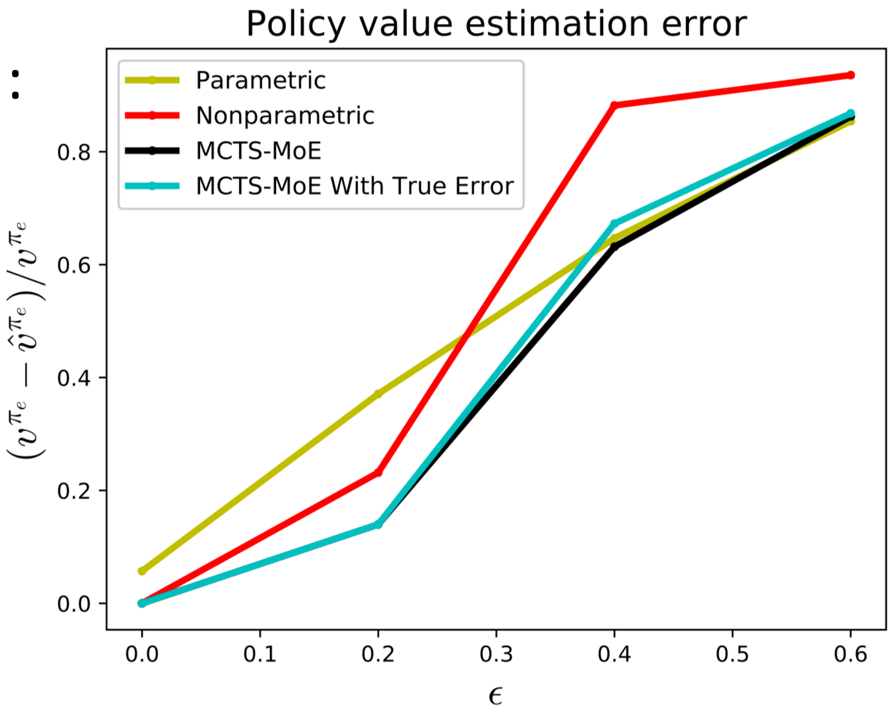


Performance on medical simulators

Cancer :



HIV :



- MCTS-MoE tends to outperforms both the parametric and nonparametric models
- With access to the true model errors, the performance of the MCTS-MoE could be improved even further
- For these domains, all importance sampling methods result in errors which are order of magnitudes larger than any model based method

Summary and Future Directions

- We provide a general framework for combining multiple models to improve off-policy evaluation.

- Improvements via individual models, error estimation or combining multiple models.
- Extension to stochastic domains is conceptually straight-forward but requires estimating distances between distributions rather than states.
- Identifying particularly loose or tight error bounds.