REINFORCEMENT LEARNING VIA LINEAR PROGRAMMING

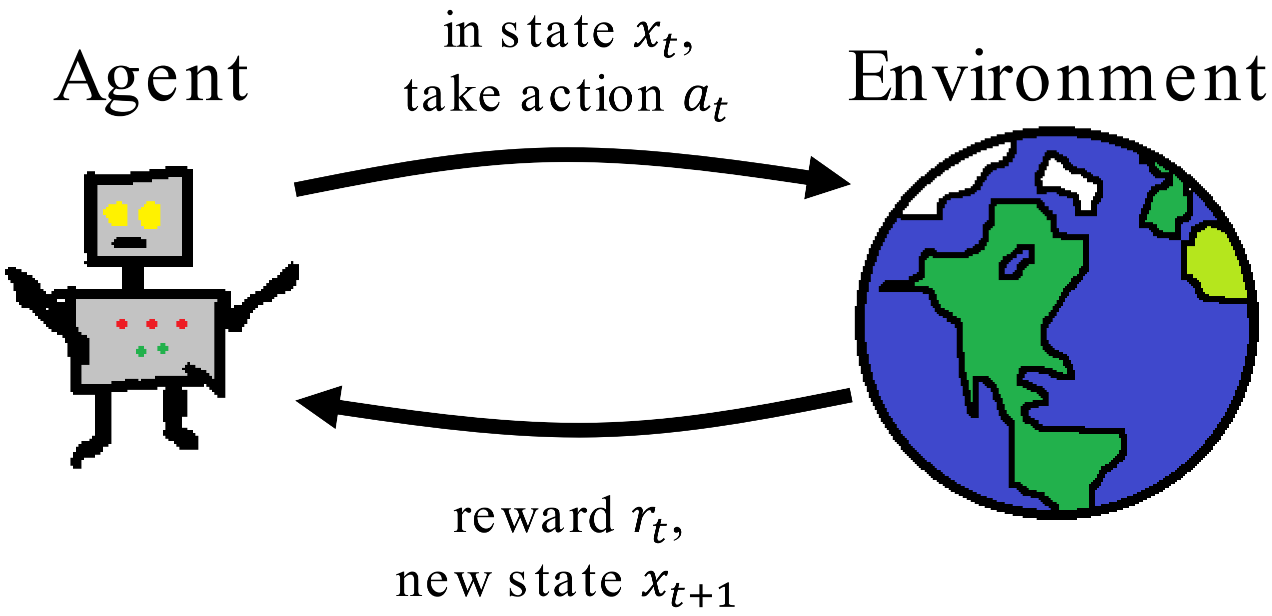
Gergely Neu

upf.
Universitat
Pompeu Fabra
Barcelona

OUTLINE

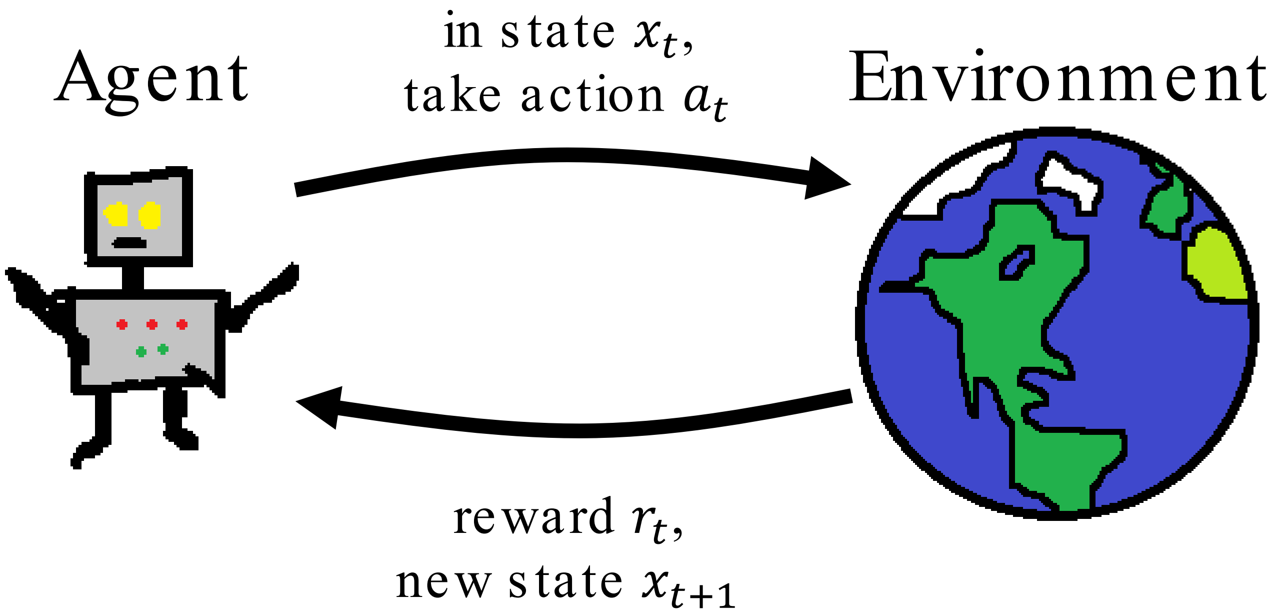
- Reinforcement Learning
- Markov Decision Processes and the Bellman equations
- Linear Programming for MDPs
- A new breed of RL algorithms
  - Relative entropy policy search
  - Primal-dual methods

REINFORCEMENT LEARNING



Goal: learn behaviors that maximize reward on the long run

REINFORCEMENT LEARNING

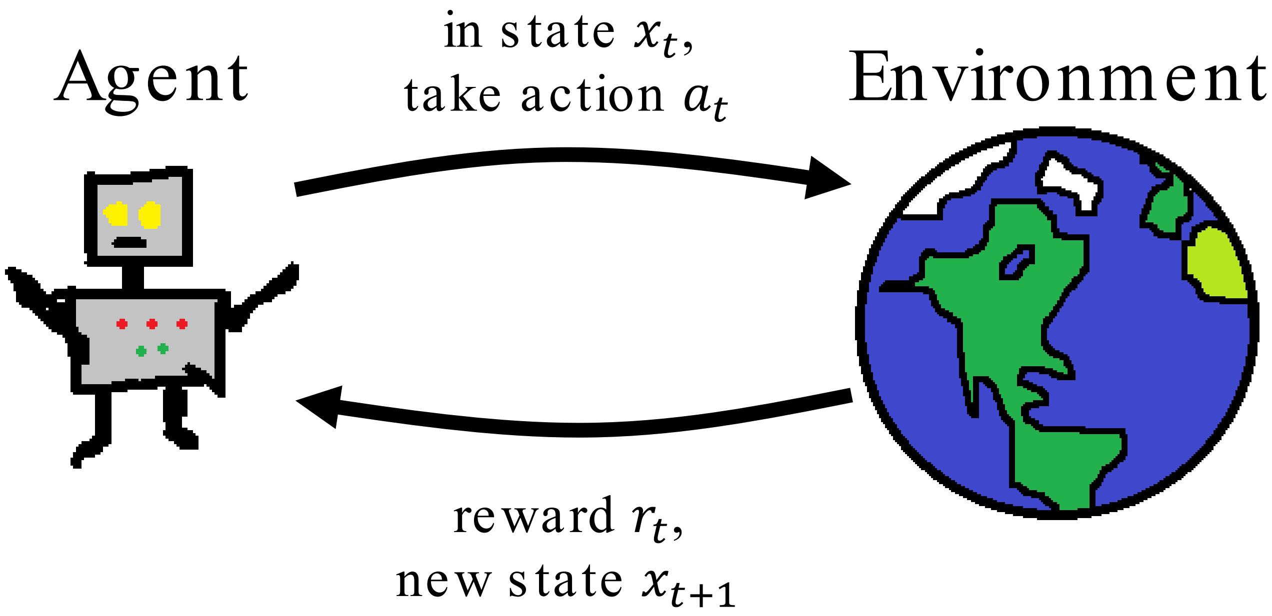


Why is this interesting?

- Model captures many important real-world problems!

Goal: learn behaviors that maximize reward on the long run

REINFORCEMENT LEARNING



Goal: learn behaviors that maximize reward on the long run

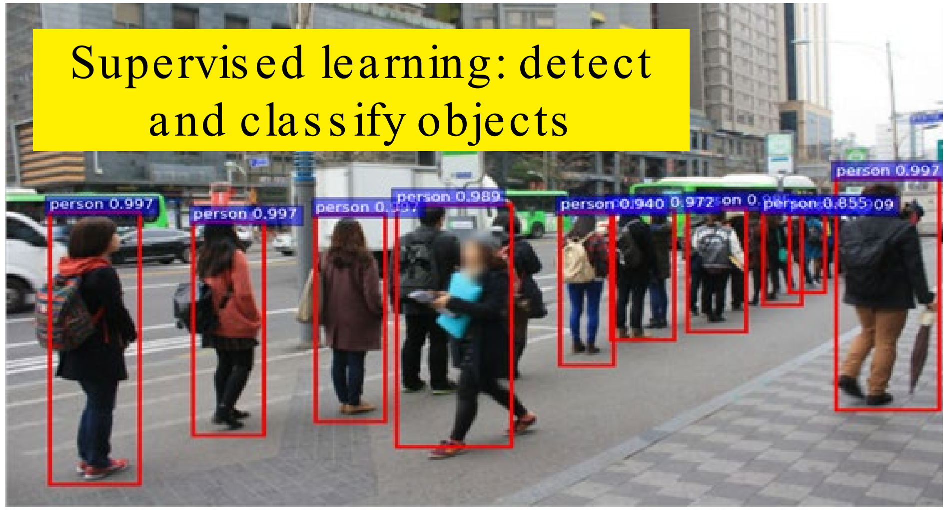
Why is this interesting?

- Model captures many important real-world problems!

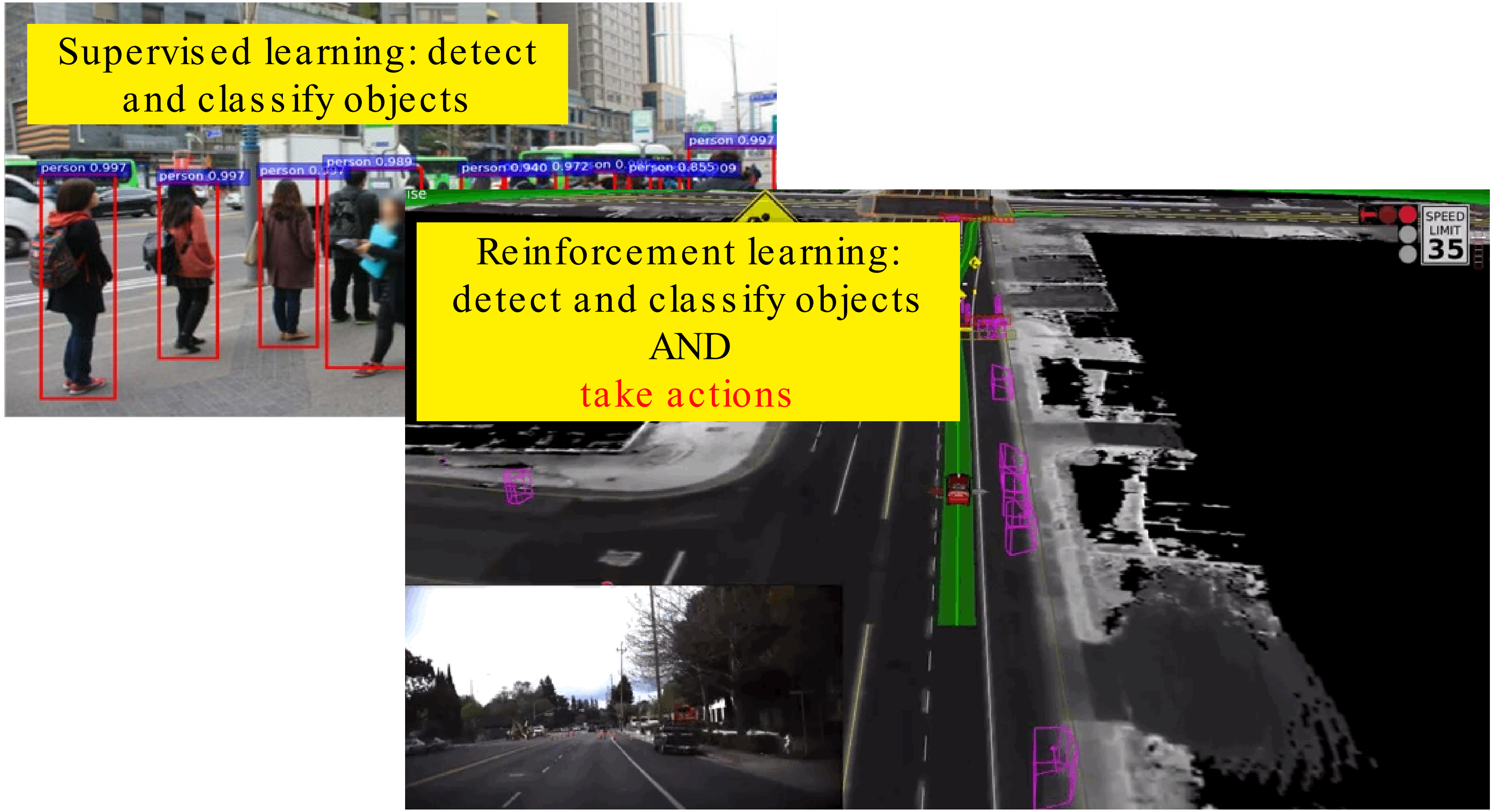
Why is this challenging?

- Environment dynamics typically unknown
- Actions influence long-term performance

REINFORCEMENT LEARNING VS. SUPERVISED LEARNING



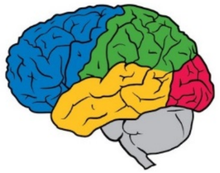
REINFORCEMENT LEARNING VS. SUPERVISED LEARNING



RL BREAKTHROUGHS

Superhuman performance in
- Atari (Mnih et al., 2013)
- Go (Silver et al., 2016, 2017)
- Starcraft (Silver et al., 2019)

Emerging applications in
- Robotics
- Autonomous driving
- Dialogue management
- Recommendation systems,...



Facebook AI Research

DeepMind

OpenAI

Microsoft Research

RL BREAKTHROUGHS

Superhuman performance in
- Atari (Mnih et al., 2013)
- Go (Silver et al., 2016, 2017)
- Starcraft (Silver et al., 2019)

Emerging applications in
- Robotics
- Autonomous driving
- Dialogue management
- Recommendation systems,...

This talk:
taking a fresh look at the foundations

Facebook AI Research

DeepMind

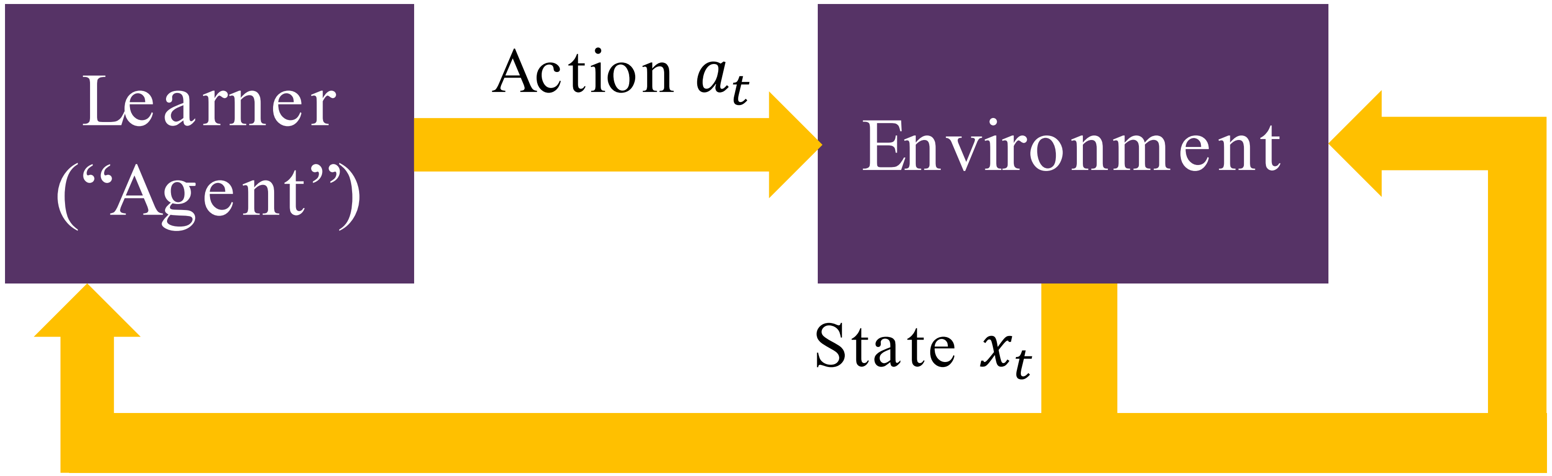
OpenAI

Microsoft Research

MARKOV DECISION PROCESSES

and the Bellman equations

MARKOV DECISION PROCESSES



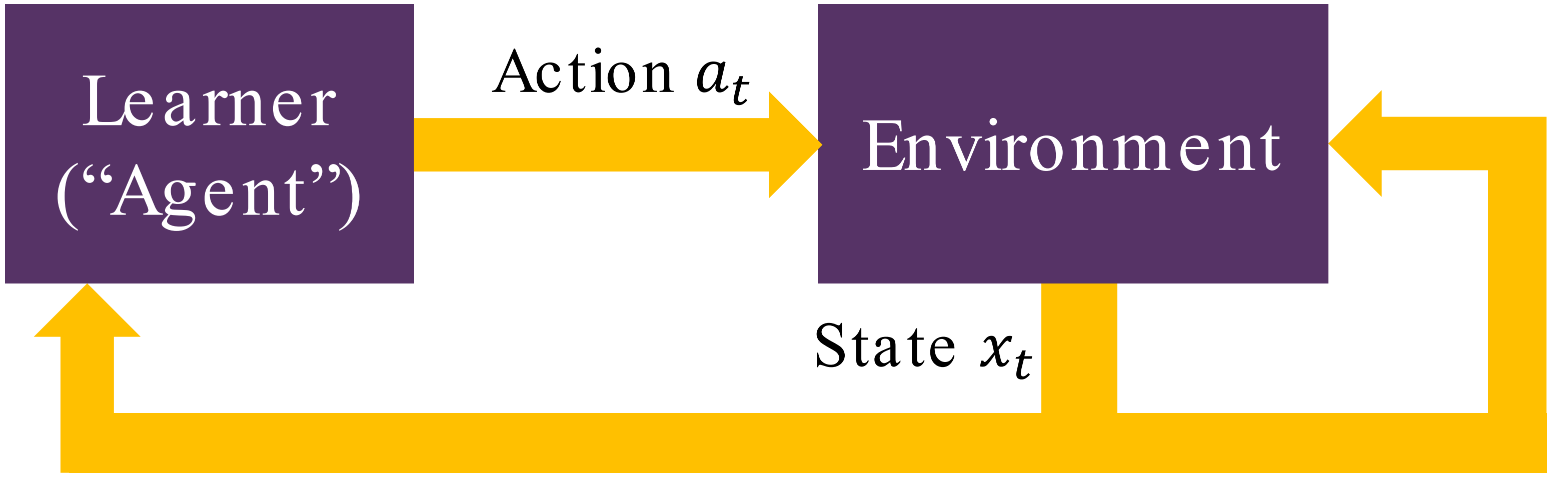
Learner:

- Observe state x_t, take action a_t
- Obtain reward $r(x_t, a_t)$

Environment:

- Generate next state $x_{t+1} \sim P(\cdot\,|x_t, a_t)$

MARKOV DECISION PROCESSES



Learner:

- Observe state x_t, take action a_t
- Obtain reward $r(x_t, a_t)$

Environment:

- Generate next state $x_{t+1} \sim P(\cdot \mid x_t, a_t)$

Goal:

maximize discounted return

$$R = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right]$$

from initial state $x_0 \sim \nu_0$

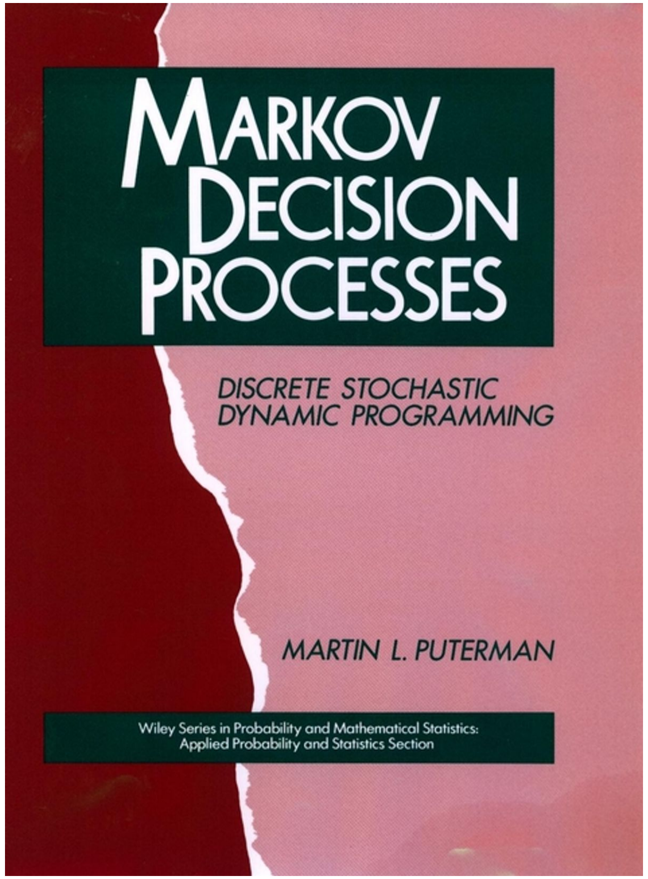
$\gamma \in (0,1)$

BASIC MDP FACTS

- Markov property: x_{t+1} only depends on (x_t, a_t)
- Stationarity: $P(\cdot\,|x_t, a_t)$ doesn't depend on t

$\Rightarrow$ enough to consider stationary policies
$$\pi(a|x) = \mathbb{P}[a_t = a | x_t = x]$$

- Many other beautiful properties:
  - There is a deterministic optimal policy
  - Simultaneous optimality regardless of ν_0
  - …



SOLVING MDPs: THE CANONICAL WAY

The Bellman optimality equations

$$Q^*(x,a) = r(x,a) + \gamma \mathbb{E}[\max_{a'} Q^*(x',a') \,|x,a]$$

SOLVING MDPs: THE CANONICAL WAY

The Bellman optimality equations

$$Q^*(x,a) = r(x,a) + \gamma \mathbb{E}[\max_{a'} Q^*(x',a')\,|x,a]$$

value of taking action a in state x

immediate reward

expected future value

SOLVING MDPs: THE CANONICAL WAY

The Bellman optimality equations

$$Q^*(x,a) = r(x,a) + \gamma \mathbb{E}[\max_{a'} Q^*(x',a') \,|x,a]$$

value of taking action a in state x

immediate reward

expected future value

Optimal policy can be extracted as:

$$\pi(a|x) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q^*(x,a') \\ 0 & \text{otherwise} \end{cases}$$

SOLVING MDPs: THE CANONICAL WAY

The Bellman optimality equations

$$Q^*(x,a) = r(x,a) + \gamma \mathbb{E}[\max_{a'} Q^*(x',a') \,|x,a]$$

value of taking action a in state x

immediate reward

expected future value

Optimal policy can be extracted as:

$$\pi(a|x) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q^*(x,a') \\ 0 & \text{otherwise} \end{cases}$$

Richard Bellman (1954):
Solution can be found via
"Dynamic Programming"

SOLVING MDPs: THE CANONICAL WAY

The Bellman optimality equations

$$Q^*(x,a) = r(x,a) + \gamma \mathbb{E}[\max_{a'} Q^*(x',a') \,|x,a]$$

value of taking action a in state x

immediate reward

expected future value

Challenges for reinforcement learning:

- Expectation over next state x' cannot be computed explicitly when transition dynamics P are unknown!
- No hope of finding exact solution when state space is large!

The Recipe for Modern RL Algorithms

- Parametrize a set of Q-functions: $Q_\theta : \theta \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$
  (e.g., via neural networks)
- Find a Q-function that approximately solves the Bellman equations, e.g., by minimizing the "squared Bellman error":

$$\mathcal{L}(Q) = \mathbb{E}_{(x,a)\sim\mu}\left[\left(r(x,a) + \gamma\mathbb{E}[\max_{a'} Q(x',a')\,|x,a] - Q(x,a)\right)^2\right]$$

- Add lots of heuristics to stabilize training
- Add lots of computational resources and bake on 1000 GPUs until ready

The Recipe for Modern RL Algorithms

- Parametrize a set of Q functions: $Q_\theta : \mathcal{A} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$

Don't try this at home!

- This objective is
  - non-convex
  - non-smooth
  - impossible to evaluate
- Does this process converge anywhere at all?
- If it converges, does it lead to a good policy??

- Add lots of computational resources and ... 1000 CPUs until ready

$$\mathcal{L}(Q) = \mathbb{E}_{(x,a)\sim\mu}\left[\left(r(x,a) + \gamma\mathbb{E}[\max_{a'} Q(x',a') \,|x,a] - Q(x,a)\right)^2\right]$$

LINEAR PROGRAMMING FOR MDPs

A LINEAR REFORMULATION

Observe: the discounted return of policy π is

$$R_\gamma^\pi = \mathbf{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)]$$

A LINEAR REFORMULATION

Observe: the discounted return of policy π is

$$R_\gamma^\pi = \mathbf{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right]$$
$$= \sum_{t=0}^{\infty} \gamma^t\, \mathbf{E}_\pi\left[r(x_t, a_t)\right]$$

A LINEAR REFORMULATION

Observe: the discounted return of policy π is

$$\begin{aligned} R_\gamma^\pi &= \mathbf{E}_\pi[\textstyle\sum_{t=0}^\infty \gamma^t r(x_t, a_t)] \\ &= \textstyle\sum_{t=0}^\infty \gamma^t \, \mathbf{E}_\pi[r(x_t, a_t)] \\ &= \textstyle\sum_{t=0}^\infty \gamma^t \sum_{x,a} \mathbf{P}_\pi[x_t = x, a_t = a] r(x, a) \end{aligned}$$

A LINEAR REFORMULATION

Observe: the discounted return of policy π is

$$
\begin{aligned}
R_\gamma^\pi &= \mathbf{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r(x_t, a_t)\right] \\
&= \sum_{t=0}^\infty \gamma^t\, \mathbf{E}_\pi[r(x_t, a_t)] \\
&= \sum_{t=0}^\infty \gamma^t \sum_{x,a} \mathbf{P}_\pi[x_t = x, a_t = a]\, r(x, a) \\
&= \sum_{x,a} \sum_{t=0}^\infty \gamma^t \mathbf{P}_\pi[x_t = x, a_t = a]\, r(x, a)
\end{aligned}
$$

A LINEAR REFORMULATION

Observe: the discounted return of policy π is

$$
\begin{aligned}
R_\gamma^\pi &= \mathbf{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r(x_t, a_t)\right] \\
&= \sum_{t=0}^\infty \gamma^t \, \mathbf{E}_\pi[r(x_t, a_t)] \\
&= \sum_{t=0}^\infty \gamma^t \sum_{x,a} \mathbf{P}_\pi[x_t = x, a_t = a] r(x, a) \\
&= \sum_{x,a} \underbrace{\sum_{t=0}^\infty \gamma^t \mathbf{P}_\pi[x_t = x, a_t = a]}_{\stackrel{\text{def}}{=} \mu_\pi(x,a)} r(x, a)
\end{aligned}
$$

discounted occupancy measure of π

A LINEAR REFORMULATION

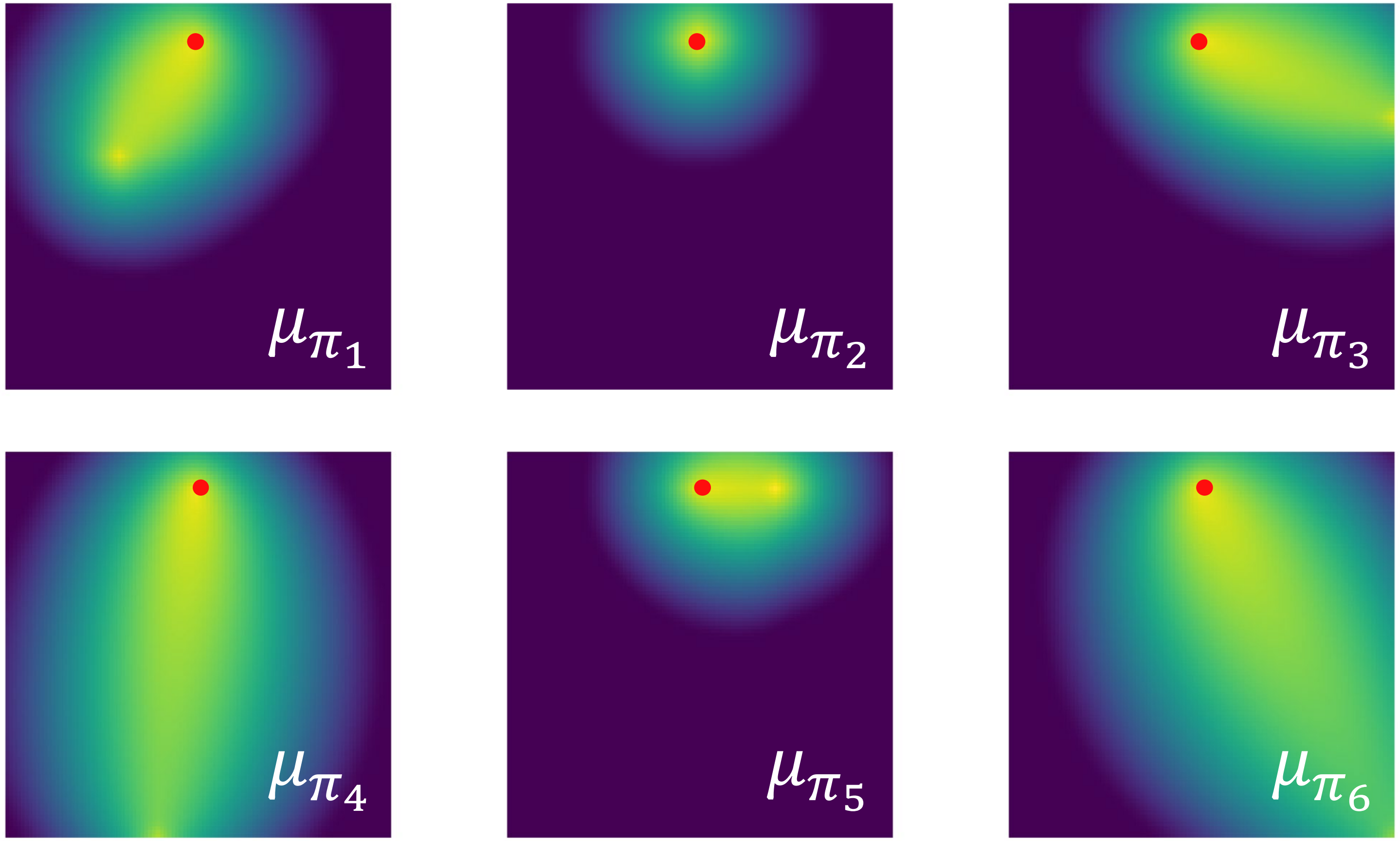
Observe: the discounted return of policy π is

$$
\begin{aligned}
R_\gamma^\pi &= \mathbf{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \, \mathbf{E}_\pi[r(x_t, a_t)] \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{x,a} \mathbf{P}_\pi[x_t = x, a_t = a] r(x,a) \\
&= \sum_{x,a} \underbrace{\sum_{t=0}^{\infty} \gamma^t \mathbf{P}_\pi[x_t = x, a_t = a]}_{\stackrel{\text{def}}{=} \mu_\pi(x,a)} r(x,a)
\end{aligned}
$$

Discounted return is linear in μ_π:

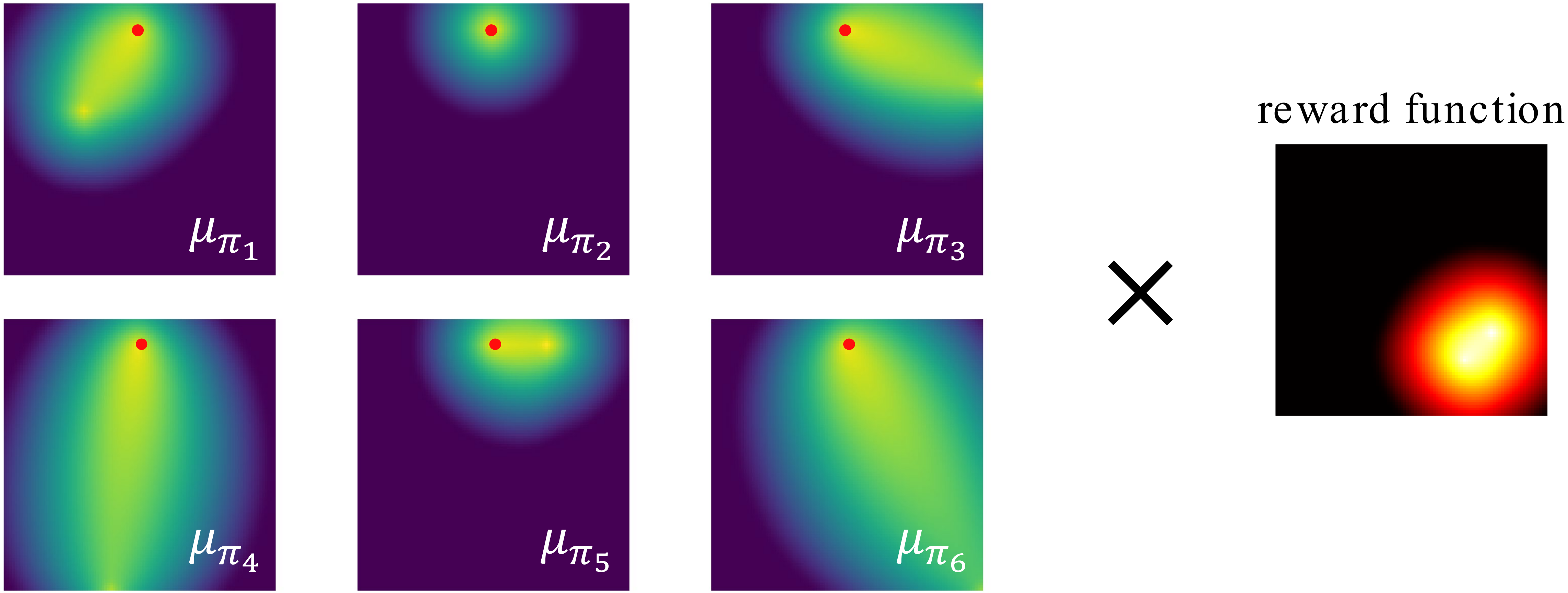
$$R_\gamma^\pi = \langle \mu_\pi, r \rangle \stackrel{\text{def}}{=} \sum_{x,a} \mu_\pi(x,a) r(x,a)$$

EXAMPLE: 2D STATE SPACE



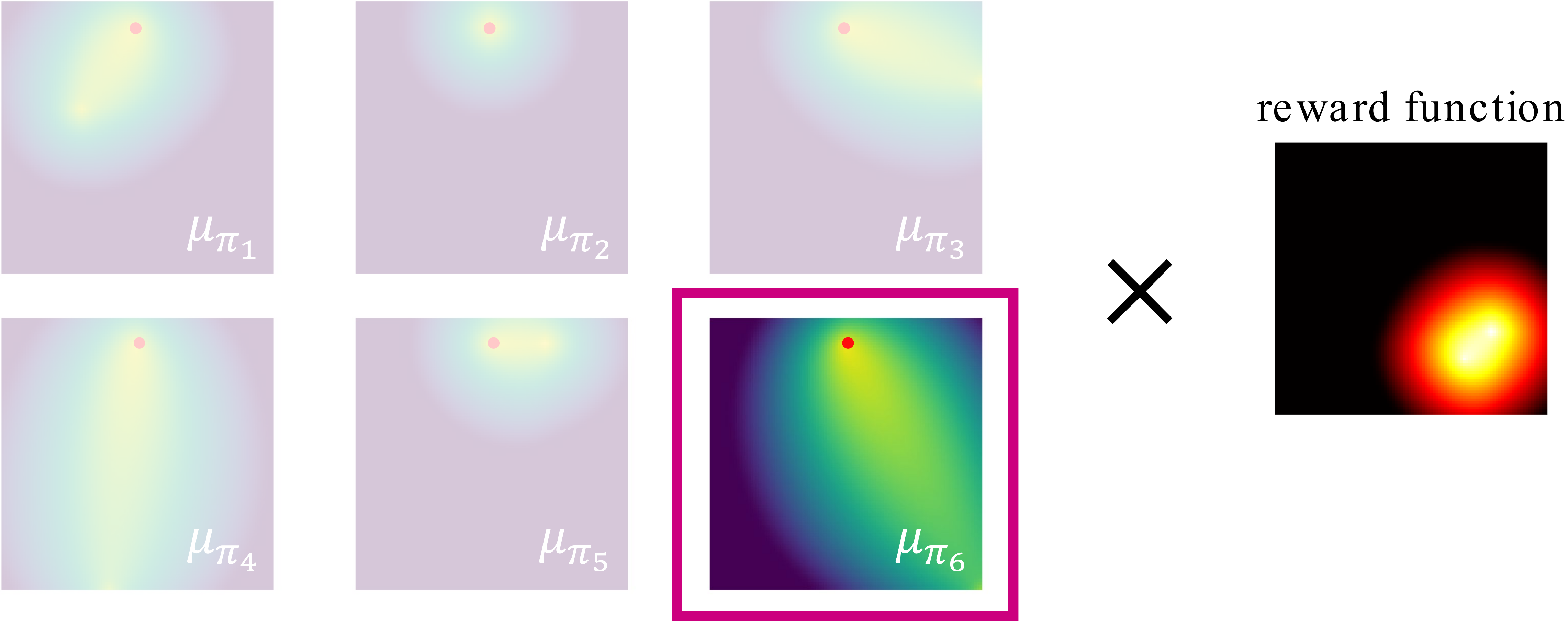
● =initial state

EXAMPLE: 2D STATE SPACE



• =initial state

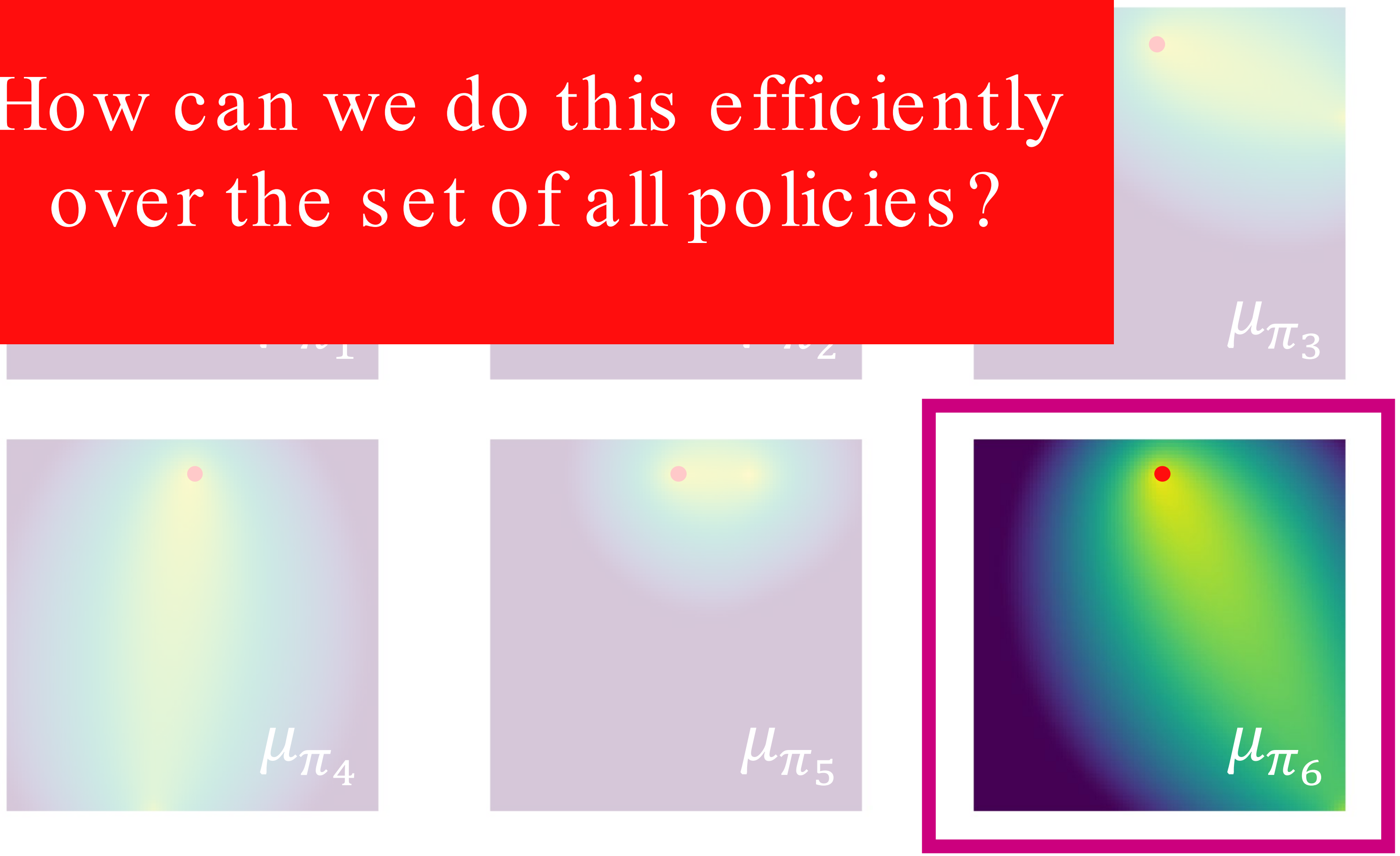
EXAMPLE: 2D STATE SPACE



● =initial state

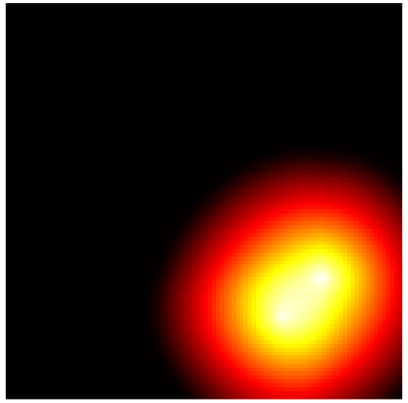
EXAMPLE: 2D STATE SPACE

How can we do this efficiently over the set of all policies?



$\times$

reward function



• =initial state

THE SET OF OCCUPANCY MEASURES

For any policy π, the occupancy measure satisfies

$$\sum_a \mu_\pi(x,a) = \nu_0(x) + \gamma(P\mu_\pi)(x)$$

THE SET OF OCCUPANCY MEASURES

For any policy π, the occupancy measure satisfies

$$\sum_a \mu_\pi(x,a) = \nu_0(x) + \gamma(P\mu_\pi)(x)$$

occupancy of state
$X_t = x$

initial state distribution

occupancy of next state
$X_{t+1} = x$

THE SET OF OCCUPANCY MEASURES

For any policy π, the occupancy measure satisfies

$$\sum_a \mu_\pi(x,a) = \nu_0(x) + \gamma(P\mu_\pi)(x)$$

occupancy of state
$X_t = x$

initial state distribution

occupancy of next state
$X_{t+1} = x$

Theorem (Manne 1960)

μ is a valid occupancy measure if and only if it satisfies

$$E\mu = \gamma P\mu + \nu_0$$

"Bellman flow constraints"

THE LP FORMULATION

Linear Programming for MDPs

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle \\ \text{subject to} \quad & E^\top \mu = \gamma P^\top \mu + \nu_0 \end{aligned}$$

THE LP FORMULATION

Linear Programming for MDPs

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle \\ \text{subject to} \quad & E^\top \mu = \gamma P^\top \mu + \nu_0 \end{aligned}$$

- Optimal policy π^* can be extracted from solution μ^* as
$$\pi^*(a|x) = \frac{\mu^*(x,a)}{\sum_a' \mu^*(x,a')}$$
- Basic solutions correspond to deterministic policies

THE LP FORMULATION

Linear Programming for MDPs

$$\begin{array}{ll} \text{maximize} & \langle \mu, r \rangle \\ \text{subject to} & E^\top \mu = \gamma P^\top \mu + \nu_0 \end{array}$$

- Optimal policy π^* can be extracted from solution μ^* as

$$\pi^*(a|x) = \frac{\mu^*(x,a)}{\sum_a' \mu^*(x,a')}$$

- Basic solutions correspond to deterministic policies

Dual Linear Program for MDPs

$$\begin{array}{ll} \text{minimize} & \langle \nu_0, V \rangle \\ \text{subject to} & EV \geq r + \gamma P V \end{array}$$

- Dual solution related to Bellman eqns as

$$Q^* = r + \gamma P V^*$$

PROS AND CONS

Why is this useful?

- Defining optimality is very simple: no value functions, no fixed-point equations, no nonlinearity... just a single numerical objective!
- Easily comprehensible with an optimization background
- Powerful tool for developing algorithms

PROS AND CONS

Why is this useful?

- Defining optimality is very simple: no value functions, no fixed-point equations, no nonlinearity… just a single numerical objective!
- Easily comprehensible with an optimization background
- Powerful tool for developing algorithms

“Why don’t they teach this in school?!?”

- Need to ensure $\mu^*(x, a) > 0$ to extract policy :’(
- Temporal aspect is a bit abstract
- Number of variables and constraints is large

A BIT OF HISTORY

- Manne (1960), de Ghellinck (1960), Denardo (1970)
  - Formulated the primal LP and showed equivalence to Bellman eqns.
- Schweitzer & Seidmann (1982)
  - Proposed a relaxation to reduce the number of constraints
  - (also proposed the squared Bellman error objective!)
- De Farias & Van Roy (2003)
  - Analyzed the reduction of [SS82]
  - Inspired some follow-up work in RL [dFvR05,PZ09,PTPZ10,DFM12,LBS17]

A BIT OF HISTORY

- Manne (1960), de Ghellinck (1960), Denardo (1970)
  - Formulated the primal LP and showed equivalence to Bellman eqns.
- Schweitzer & Seidmann (1982)
  - Proposed a relaxation to reduce the number of constraints
  - (also proposed the squared Bellman error objective!)
- De Farias & Van Roy (2003)
  - Analyzed the reduction of [SS82]
  - Inspired some follow-up work in RL [dFvR05,PZ09,PTPZ10,DFM12,LBS17]

Common theme:
analyze quality of approximate solution &
solve the LP with generic solver

A BIT OF HISTORY

- Manne (1960), de Ghellinck (1960), Denardo (1970)
  - Formulated the primal
- Schweitzer & Seidma
  - Proposed a relaxation
  - (also proposed the sq
- De Farias & Van Roy (2003)
  - Analyzed the reduction of [SS82]
  - Inspired some follow-up work in RL [dFvR05,PZ09,PTPZ10,DFM12,LBS17]

Is this the best we can do?

Common theme:
analyze quality of approximate solution &
solve the LP with generic solver

A NEW BREED OF RL ALGORITHMS

RELATIVE ENTROPY POLICY SEARCH

Peters, Mülling, Altün (2010)

Linear Program for MDPs

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle \\ \text{subject to} \quad & E^\top \mu = \gamma P^\top \mu + \nu_0 \end{aligned}$$

- add regularization for tractable solution
- relax constraints like [SS85]

RELATIVE ENTROPY POLICY SEARCH

Peters, Mülling, Altün (2010)

REPS (primal form)

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu | \mu_0) \\ \text{subject to} \quad & \Psi^\top E^\top \mu = \gamma \Psi^\top P^\top \mu + \Psi^\top \nu_0 \end{aligned}$$

- add regularization for tractable solution
- relax constraints like [SS85]

RELATIVE ENTROPY POLICY SEARCH

Peters, Mülling, Altün (2010)

REPS (primal form)

$$\text{maximize} \quad \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu | \mu_0)$$

$$\text{subject to} \quad \Psi^\top E^\top \mu = \gamma \Psi^\top P^\top \mu + \Psi^\top \nu_0$$

- add regularization for tractable solution
- relax constraints like [SS85]

Lagrangian duality

REPS (dual form)

- $\theta^* = \min_\theta \frac{1}{\eta} \log \mathbb{E}_{x,a \sim \mu_0} \left[e^{\eta (r(x,a) + \gamma \, \mathbb{E}[\Psi\theta(x') | x,a] - \Psi\theta(x))} \right]$
- $\mu^* = \mu_0 \circ e^{\eta (r + \gamma P \Psi \theta^* - E \Psi \theta^*)}$

RELATIVE ENTROPY POLICY SEARCH

Peters, Mülling, Altün (2010)

REPS (primal form)

maximize $\quad \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu | \mu_0)$

subject to $\quad \Psi^\top E^\top \mu = \gamma \Psi^\top P^\top \mu + \Psi^\top \nu_0$

- add regularization for tractable solution
- relax constraints like [SS85]

Lagrangian duality

REPS (dual form)

Unconstrained convex optimization problem!

- $\theta^* = \min_\theta \frac{1}{\eta} \log \mathbb{E}_{x,a \sim \mu_0} \left[e^{\eta(r(x,a) + \gamma\, \mathbb{E}[\Psi\theta(x') | x,a] - \Psi\theta(x))} \right]$
- $\mu^* = \mu_0 \circ e^{\eta(r + \gamma P \Psi \theta^* - E \Psi \theta^*)}$

RELATIVE ENTROPY POLICY SEARCH

Peters, Mülling, Altün (2010)

REPS (primal form)

$$\text{maximize} \quad \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu | \mu_0)$$

$$\text{subject to} \quad \Psi^\top E^\top \mu = \gamma \Psi^\top P^\top \mu + \Psi^\top \nu_0$$

- add regularization for tractable solution
- relax constraints like [SS85]

Lagrangian duality

REPS (dual form)

Unconstrained convex optimization problem!

- $\theta^* = \min_\theta \frac{1}{\eta} \log \mathbb{E}_{x,a \sim \mu_0} \left[e^{\eta (r(x,a) + \gamma \, \mathbb{E}[\Psi\theta(x') | x,a] - \Psi\theta(x))} \right]$
- $\mu^* = \mu_0 \circ e^{\eta (r + \gamma P \Psi \theta^* - E \Psi \theta^*)}$

Intractable due to unknown P in exponent!

RELATIVE ENTROPY POLICY SEARCH

Peters, Mülling, Altün (2010)

REPS (primal form)

maximize $\quad \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu | \mu_0)$

subject to $\quad \Psi^\top E^\top \mu = \gamma \Psi^\top P^\top \mu + \Psi^\top \nu_0$

Lagrangian duality

Can we do better?

Unconstrained convex optimization problem!

REPS (dual form)

- $\theta^* = \min_\theta \frac{1}{\eta} \log \mathbb{E}_{x,a \sim \mu_0} \left[e^{\eta(r(x,a) + \gamma\, \mathbb{E}[\Psi\theta(x') | x,a] - \Psi\theta(x))} \right]$
- $\mu^* = \mu_0 \circ e^{\eta(r + \gamma P \Psi \theta^* - E \Psi \theta^*)}$

Intractable due to unknown P in exponent!

LOGISTIC Q-LEARNING

Bas-Serrano, Curi, Krause & Neu (2021)

Q-REPS (primal form)

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu | \mu_0) - \frac{1}{\alpha} H(u | u_0) \\ \text{subject to} \quad & E^\top \mu = \gamma P^\top u + \nu_0 \\ & \Phi^\top \mu = \Phi^\top u \end{aligned}$$

- Lagrangian decomposition to introduce "Q"
- Composite regularization

LOGISTIC Q-LEARNING

Bas-Serrano, Curi, Krause & Neu (2021)

Q-REPS (primal form)

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle - \frac{1}{\eta} \mathrm{KL}(\mu|\mu_0) - \frac{1}{\alpha} H(u|u_0) \\ \text{subject to} \quad & E^\top \mu = \gamma P^\top u + \nu_0 \\ & \Phi^\top \mu = \Phi^\top u \end{aligned}$$

- Lagrangian decomposition to introduce "Q"
- Composite regularization

Lagrangian duality

Q-REPS (dual form)

- $\theta^* = \min_\theta \frac{1}{\eta} \log \mathbb{E}_{x,a \sim \mu_0} \left[e^{\eta (r(x,a) + \gamma \, \mathbb{E}[V_\theta(x') | x,a] - \Phi\theta(x,a))} \right]$
- $\pi^* = \pi_0 \circ e^{\eta \left(\Phi\theta^* - V_{\theta^*} \right)}$

LOGISTIC Q-LEARNING

Bas-Serrano, Curi, Krause & Neu (2021)

Q-REPS (primal form)

$$\begin{aligned} \text{maximize} \quad & \langle \mu, r \rangle - \frac{1}{\eta}\mathrm{KL}(\mu|\mu_0) - \frac{1}{\alpha} H(u|u_0) \\ \text{subject to} \quad & E^\top \mu = \gamma P^\top u + \nu_0 \\ & \Phi^\top \mu = \Phi^\top u \end{aligned}$$

- Lagrangian decomposition to introduce “Q”
- Composite regularization

Lagrangian duality

Q-REPS (dual form)

Unconstrained convex optimization problem!

- $\theta^* = \min_\theta \frac{1}{\eta} \log \mathbb{E}_{x,a\sim\mu_0}\left[e^{\eta(r(x,a)+\gamma\,\mathbb{E}[V_\theta(x')|x,a]-\Phi\theta(x,a))}\right]$
- $\pi^* = \pi_0 \circ e^{\eta\left(\Phi\theta^* - V_{\theta^*}\right)}$

Explicit, tractable policy update!!

A PRINCIPLED LOSS FUNCTION

Bas-Serrano, Curi, Krause & Neu (2021)

The Logistic Bellman Error (LBE)

$$\mathcal{G}(Q) = \frac{1}{\eta} \log \mathbb{E}_{(x,a)\sim\mu_0}\left[e^{\eta(r(x,a)+\gamma\,\mathbb{E}[V_Q(x')|x,a]-Q(x,a))}\right]$$

- Convex and smooth (composition of two monotone convex functions that are smooth)
- 2-Lipschitz w.r.t. ℓ_∞-norm:

$$\left\|\nabla_Q \mathcal{G}_k(Q)\right\|_1 \leq 2$$

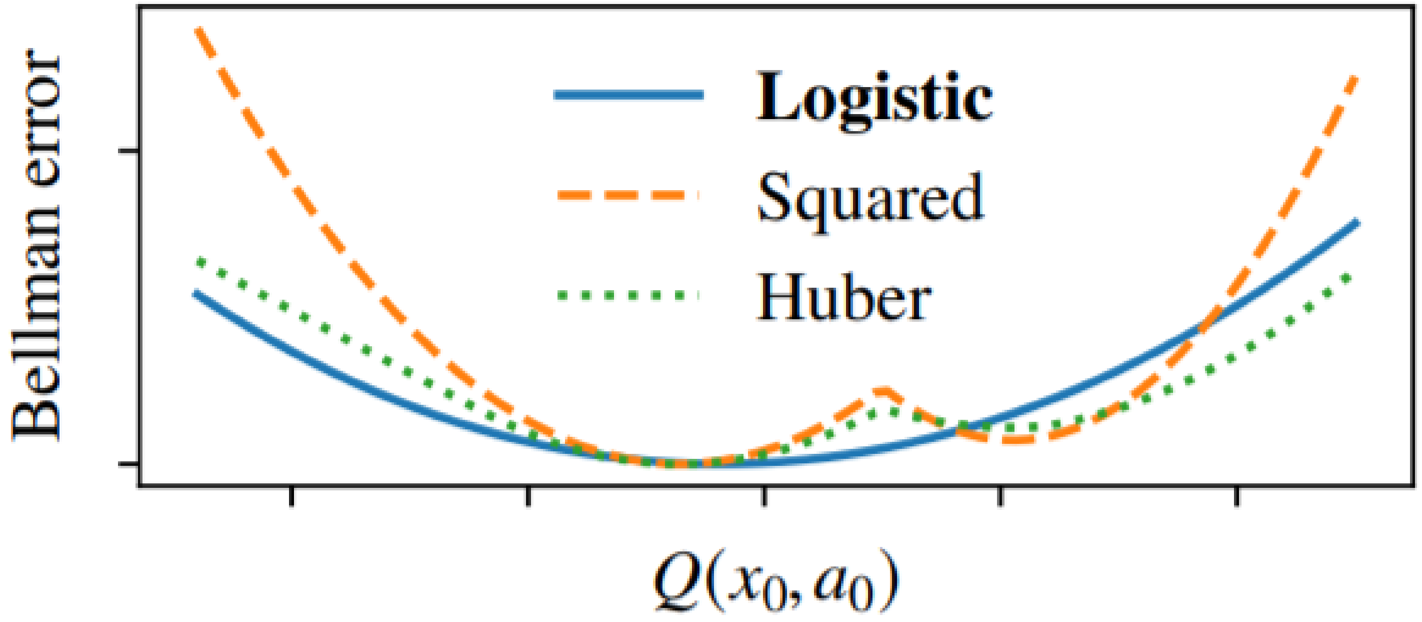
- Easy to estimate reliably using sample transitions

A PRINCIPLED LOSS FUNCTION

Bas-Serrano, Curi, Krause & Neu (2021)

The Logistic Bellman Error (LBE)

$$\mathcal{G}(Q) = \frac{1}{\eta} \log \mathbb{E}_{(x,a)\sim\mu_0}\left[e^{\eta(r(x,a)+\gamma\,\mathbb{E}[V_Q(x')|x,a]-Q(x,a))}\right]$$



STRONG GUARANTEES!

Bas-Serrano, Curi, Krause & Neu (2021)

"Theorem"

$$\left|\mathcal{G}_k(\theta) - \hat{\mathcal{G}}_k(\theta)\right| = O(\eta)$$

"LBE can be estimated with small bias"

Impossible for squared BE!

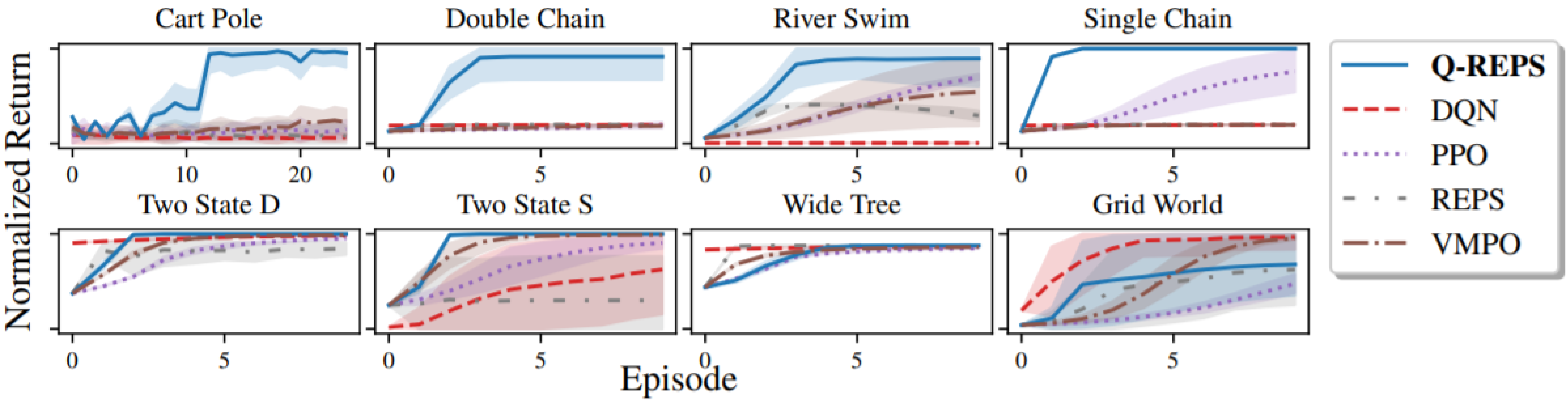
"Theorem"

$$\mathrm{err}_K \leq O\left(\frac{1}{K}\sum_{k=1}^{K}\left(\varepsilon_k + \sqrt{\eta\varepsilon_k}\right)\right)$$

"Optimization errors ε_k have moderate long-term impact"

Comparable with best results for SBE!

AND IT WORKS!!!



Cart Pole
Double Chain
River Swim
Single Chain
Two State D
Two State S
Wide Tree
Grid World
Normalized Return
Episode
0
10
20
5
Q-REPS
DQN
PPO
REPS
VMPO

OTHER LP-BASED METHODS

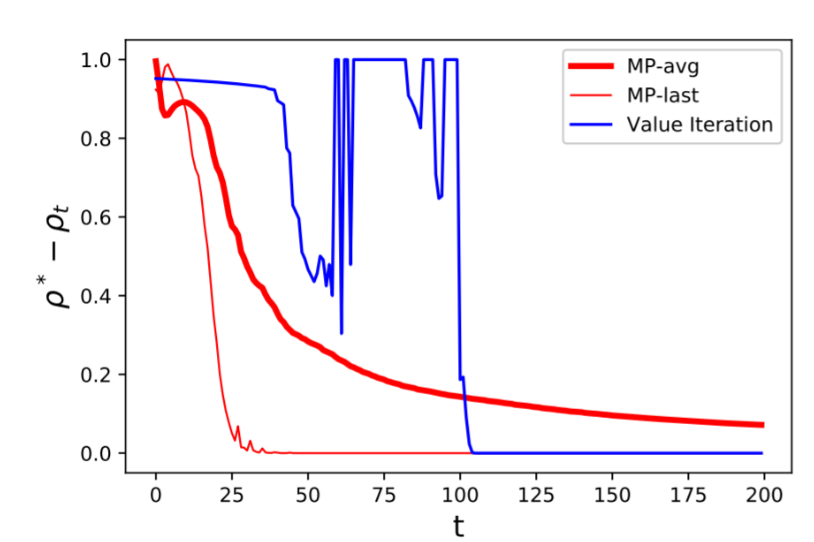
- Primal-dual methods:
  - consider equivalent saddle-point problem
$$\max_{\mu} \min_{V} \langle \mu, r + \gamma PV - EV \rangle + \langle \nu_0, V \rangle$$
  - solve with primal-dual gradient descent
  - scale up by parametrizing $\mu = U\lambda$ and $V = \Psi\theta$

OTHER LP-BASED METHODS

- Primal-dual methods:
  - consider equivalent saddle-point problem
$$\max_{\mu} \min_{V} \langle \mu, r + \gamma P V - E V \rangle + \langle \nu_0, V \rangle$$
  - solve with primal-dual gradient descent
  - scale up by parametrizing $\mu = U\lambda$ and $V = \Psi\theta$

- Implementable with only sample access to P
- State of the art method for small MDPs
- When features Φ and Ψ are chosen well:
  - guaranteed convergence to optimum
  - excellent empirical performance

Wang (2017), Chen, Li & Wang (2018), Bas-Serrano & Neu (2019)



OTHER LP-BASED METHODS 2

- Off-policy RL: fixed data set sampled from μ_0
- “DualDICE” reparametrization of primal variables:

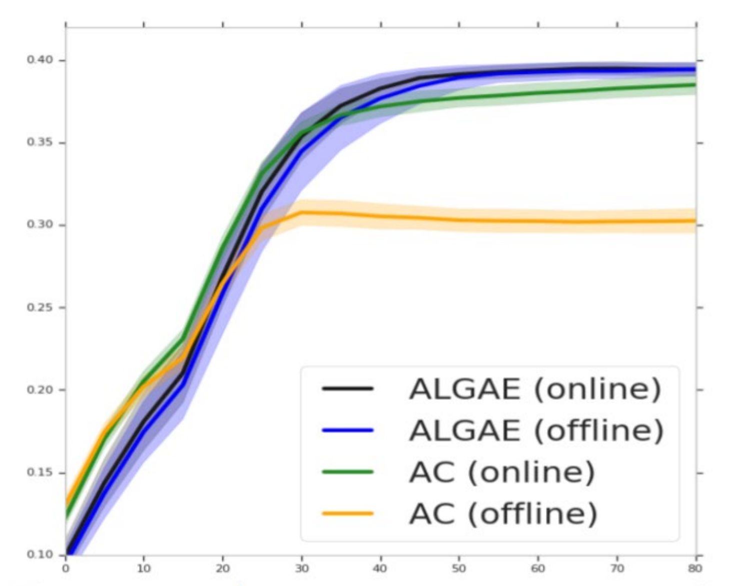
$$\xi(x, a) = \mu(x, a)/\mu_0(x, a)$$

- Leads to new primal-dual and REPS-like algorithms

OTHER LP-BASED METHODS 2

- Off-policy RL: fixed data set sampled from μ_0
- "DualDICE" reparametrization of primal variables:
$$\xi(x,a) = \mu(x,a)/\mu_0(x,a)$$
- Leads to new primal-dual and REPS-like algorithms

- Incredibly practical methods for off-policy value estimation!
- Even works without knowledge of μ_0!!



Nachum et al. (2019a,2019b), Nachum & Dai (2020), Zhang et al. (2020), Dai et al. (2020)

SUMMARY

- LP formulation is currently obscure but holds huge potential!
- Solid alternative to fixed-point computation
- Historical limitations are mostly due to rigid interpretation
- Useful for deriving new algorithms & analyzing existing ones
- Lots of work left to do!
  - Room for improvement both in theory & practice
  - Existing toolbox not as well-developed as for other RL approaches

SUMMARY

- LP formulation is currently obscure but holds huge potential!
- Solid alternative to fixed-point computation
- Historical limitations are mostly due to rigid interpretation
- Useful for deriving new algorithms & analyzing existing ones
- Lots of work left to do!
  - Room for improvement both in theory & practice
  - Existing toolbox not as well-developed as for other RL approaches

THANKS!!

PRIMAL-DUAL METHODS

Primal LP for MDPs	Dual LP for MDPs
maximize $\langle \mu, r \rangle$	minimize $\langle \nu_0, V \rangle$
subject to $E^\top \mu = \gamma P^\top \mu + \nu_0$	subject to $EV \geq r + \gamma PV$

PRIMAL-DUAL METHODS

Primal LP for MDPs

maximize $\langle \mu, r \rangle$

subject to $E^\top \mu = \gamma P^\top \mu + \nu_0$

Dual LP for MDPs

minimize $\langle \nu_0, V \rangle$

subject to $EV \geq r + \gamma PV$

Equivalent via Lagrangian duality

Primal-dual formulation for MDPs

$$\max_{\mu} \min_{V} \langle \mu, r + \gamma PV - EV \rangle + \langle \nu_0, V \rangle$$

SADDLE-POINT OPTIMIZATION

Primal-dual formulation for MDPs

$$\max_{\mu} \min_{V} \langle \mu, r + \gamma PV - EV \rangle + \langle \nu_0, V \rangle$$

Can be solved via iterative updates:

- $V_{k+1} = V_k - \eta\big((\gamma P - E)^\top \mu_k + \nu_0\big)$
- $\mu_{k+1} = \mu_k \circ e^{\eta(r+\gamma PV_k - EV_k)}$

SADDLE-POINT OPTIMIZATION

Primal-dual formulation for MDPs

$$\max_{\mu} \min_{V} \langle \mu, r + \gamma P V - E V \rangle + \langle \nu_0, V \rangle$$

Can be solved via iterative updates:

- $V_{k+1} = V_k - \eta\big((\gamma P - E)^\top \mu_k + \nu_0\big)$
- $\mu_{k+1} = \mu_k \circ e^{\eta(r + \gamma P V_k - E V_k)}$

- Gradients are expectations under μ_k
  $\Rightarrow$ efficient stochastic implementation

SADDLE-POINT OPTIMIZATION

Primal-dual formulation for MDPs

$$\max_{\mu} \min_{V} \langle \mu, r + \gamma PV - EV \rangle + \langle \nu_0, V \rangle$$

Can be solved via iterative updates:

- $V_{k+1} = V_k - \eta\big((\gamma P - E)^\top \mu_k + \nu_0\big)$
- $\mu_{k+1} = \mu_k \circ e^{\eta(r + \gamma P V_k - E V_k)}$

- Gradients are expectations under μ_k
  $\Rightarrow$ efficient stochastic implementation

State of the art sample complexity for solving "small" MDPs!
(Wang 2017)

SCALING UP

- Problem: intractable for large state spaces due to large number of constraints & variables!
- Idea: parametrize μ and V via linear functions!
  - $\mu_\lambda = \Psi\lambda$ for some feature matrix $\Psi \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times n}$
  - $V_\theta = \Phi\theta$ for some feature matrix $\Phi \in \mathbb{R}^{\mathcal{X}\times m}$

Primal-dual formulation for MDPs

$$\max_{\mu} \min_{V} \langle \mu, r + \gamma PV - EV \rangle + \langle \nu_0, V \rangle$$

SCALING UP

- Problem: intractable for large state spaces due to large number of constraints & variables!
- Idea: parametrize μ and V via linear functions!
  - $\mu_\lambda = \Psi\lambda$ for some feature matrix $\Psi \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times n}$
  - $V_\theta = \Phi\theta$ for some feature matrix $\Phi \in \mathbb{R}^{\mathcal{X}\times m}$

Relaxed primal-dual formulation for MDPs

$$\max_{\lambda}\min_{\theta} \langle \lambda, \Phi^\top(r + \gamma P\Psi\theta - E\Psi\theta)\rangle + \langle \nu_0, \Psi\theta\rangle$$

SCALING UP

- Problem: intractable for large state spaces due to large number of constraints & variables!
- Idea: parametrize μ and V via linear functions!
  - $\mu_\lambda = \Psi\lambda$ for some feature matrix $\Psi \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times n}$
  - $V_\theta = \Phi\theta$ for some feature matrix $\Phi \in \mathbb{R}^{\mathcal{X}\times m}$

Relaxed primal-dual formulation for MDPs

$$\max_\lambda \min_\theta \langle \lambda, \Phi^\top (r + \gamma P\Psi\theta - E\Psi\theta)\rangle + \langle \nu_0, \Psi\theta\rangle$$

- $\theta_{k+1} = \theta_k - \eta\big((\gamma P\Psi - E\Psi)^\top \Phi\lambda_k + \Psi^\top \nu_0\big)$
- $\lambda_{k+1} = \lambda_k \circ e^{\eta\Phi^\top (r+\gamma P\Psi\theta_k - E\Psi\theta_k)}$

SCALING UP

- Implementable with only sample access to transition function P
- When features Φ and Ψ are chosen well:
  - guaranteed convergence to optimum
  - excellent empirical performance

Chen, Li & Wang (2018), Bas-Serrano & Neu (2019)

- $\theta_{k+1} = \theta_k - \eta\left((\gamma P\Psi - E\Psi)^\top \Phi\lambda_k + \Psi^\top \nu_0\right)$
- $\lambda_{k+1} = \lambda_k \circ e^{\eta\Phi^\top(r+\gamma P\Psi\theta_k - E\Psi\theta_k)}$

- $\theta_{k+1} = \theta_k - \eta\left((\gamma P\Psi - E\Psi)^\top \Phi\lambda_k + \Psi^\top \nu_0\right)$
- $\lambda_{k+1} = \lambda_k \circ e^{\eta\Phi^\top(r+\gamma P\Psi\theta_k - E\Psi\theta_k)}$

OFF-POLICY LEARNING

- What if we can't sample from μ_k?
- Off-policy RL: fixed data set sampled from μ_0

OFF-POLICY LEARNING

- What if we can't sample from μ_k?
- Off-policy RL: fixed data set sampled from μ_0
- DualDICE reparametrization (Nachum & Dai, 2020):

  rewrite primal variables as $\xi(x,a) = \mu(x,a)/\mu_0(x,a)$

OFF-POLICY LEARNING

- What if we can't sample from μ_k?
- Off-policy RL: fixed data set sampled from μ_0
- DualDICE reparametrization (Nachum & Dai, 2020):

rewrite primal variables as $\xi(x,a) = \mu(x,a)/\mu_0(x,a)$

Primal-dual formulation for MDPs

$$\max_{\mu} \min_{V} \langle \mu, r + \gamma PV - EV \rangle + \langle \nu_0, V \rangle$$

OFF-POLICY LEARNING

- What if we can't sample from μ_k?
- Off-policy RL: fixed data set sampled from μ_0
- DualDICE reparametrization (Nachum & Dai, 2020):

  rewrite primal variables as $\xi(x,a) = \mu(x,a)/\mu_0(x,a)$

DualDICE formulation for MDPs

$$\max_{\xi} \min_{V} \langle \xi, \mu_0 \circ (r + \gamma PV - EV) \rangle + \langle \nu_0, V \rangle$$

OFF-POLICY LEARNING

- What if we can't sample from μ_k?
- Off-policy RL: fixed data set sampled from μ_0
- DualDICE reparametrization (Nachum & Dai, 2020):

  rewrite primal variables as $\xi(x,a) = \mu(x,a)/\mu_0(x,a)$

DualDICE formulation for MDPs

$$\max_{\xi} \min_{V} \langle \xi, \mu_0 \circ (r + \gamma PV - EV) \rangle + \langle \nu_0, V \rangle$$

Incredibly practical methods for off-policy value estimation!
Even works without knowledge of μ_0!!