Caltech

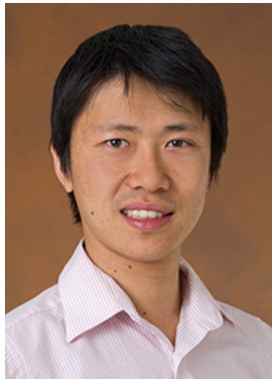
Jet Propulsion Laboratory
California Institute of Technology

Co-Training for Policy Learning


Jialin Song


Ravi Lanka


Yisong Yue

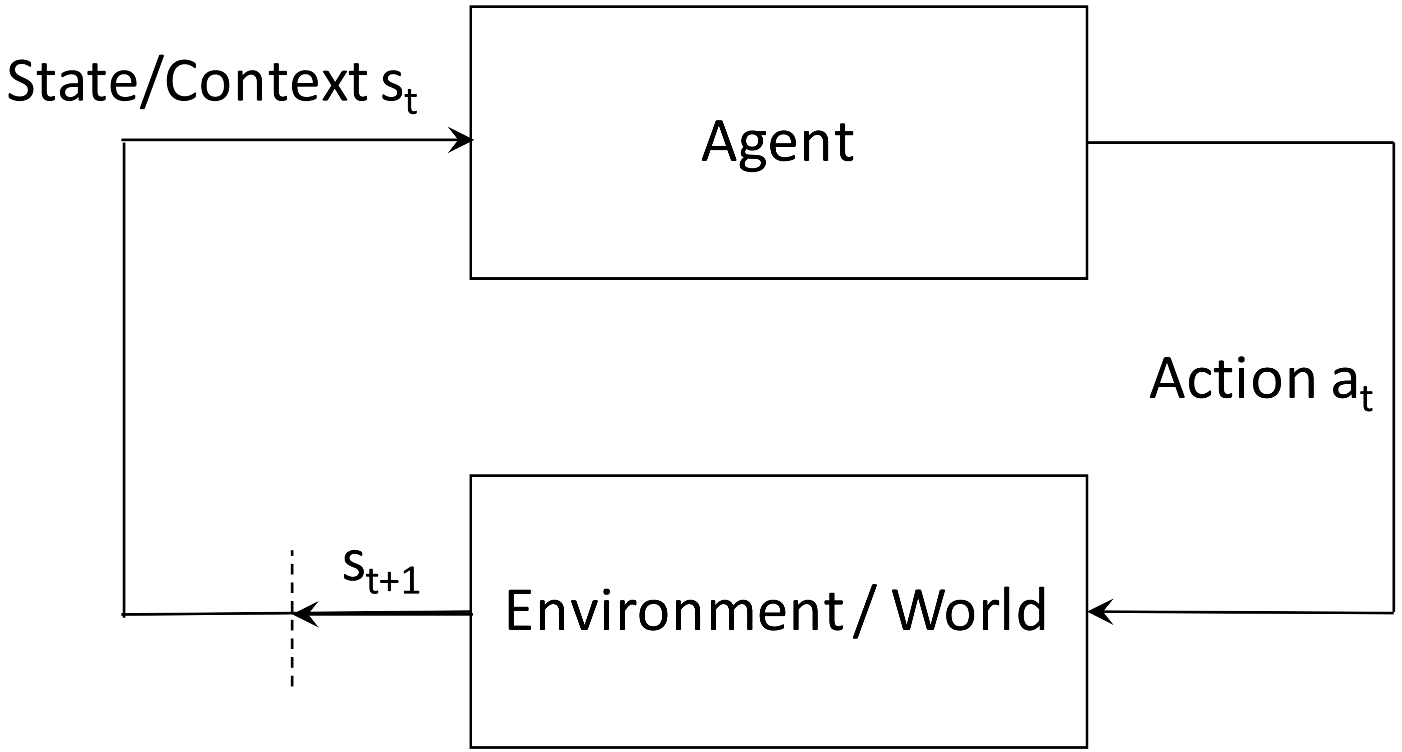

Hiro Ono

Policy Learning (Reinforcement & Imitation)

Goal: Find "Optimal" Policy

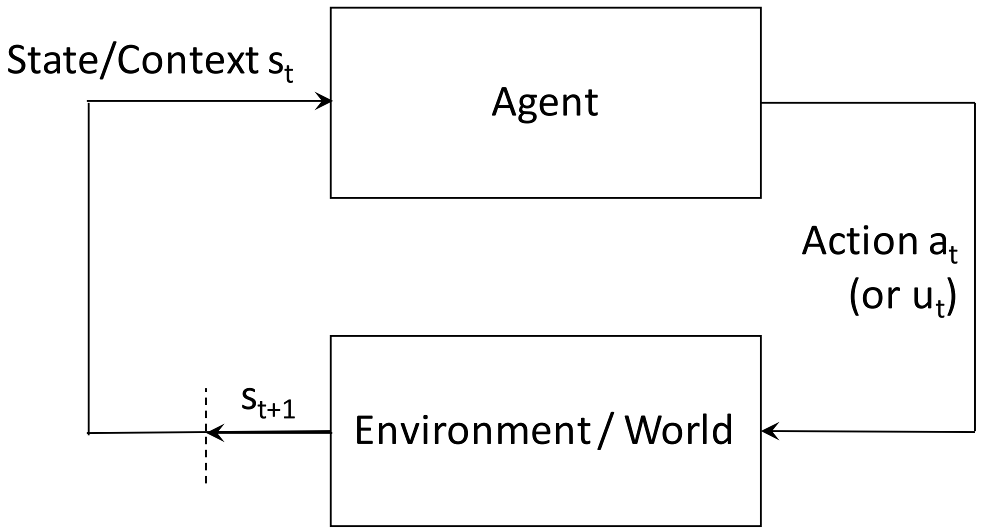
Imitation Learning:
Optimize imitation loss

Reinforcement Learning:
Optimize environmental reward



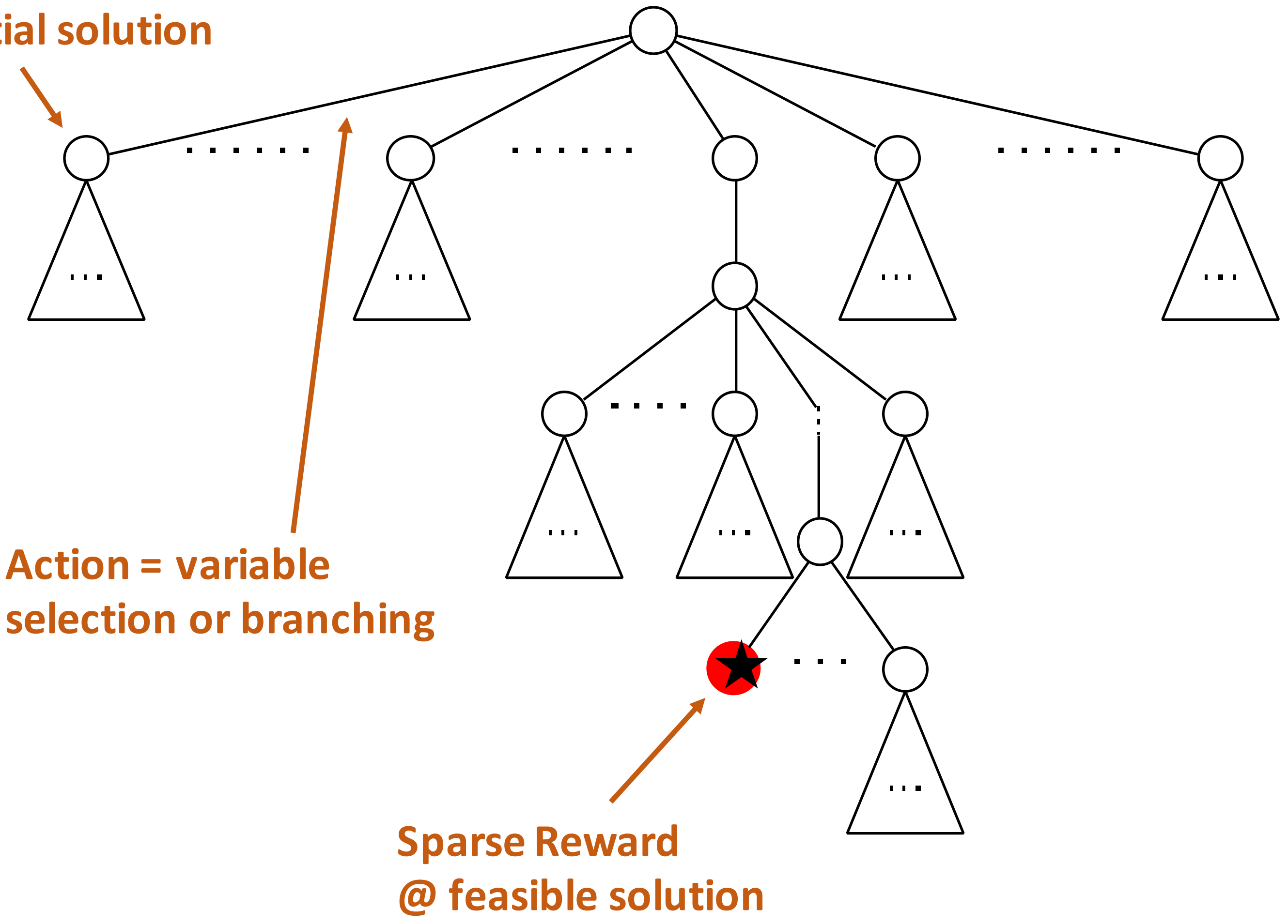
Policy Learning is Hard

- Long time horizons
- Sparse or expensive feedback
- Exponential in time horizon
  - (reinforcement learning)
- Infeasible to obtain sufficient demonstrations
  - (imitation learning)



Example: Learning to Search (Combinatorial Optimization)

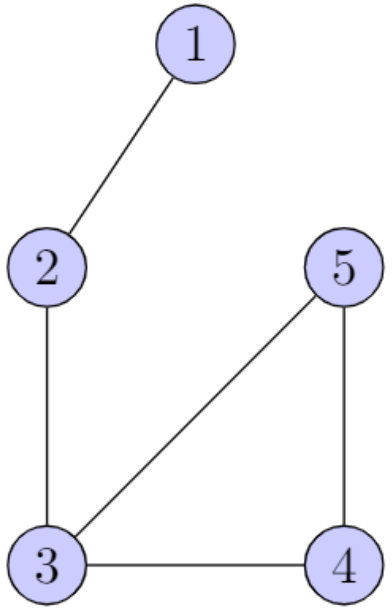
$$
\begin{aligned}
&\max \; -\sum_{i=1}^{5} x_i, \\
&\text{subject to:} \\
&x_1 + x_2 \geq 1, \\
&x_2 + x_3 \geq 1, \\
&x_3 + x_4 \geq 1, \\
&x_3 + x_5 \geq 1, \\
&x_4 + x_5 \geq 1, \\
&x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}
\end{aligned}
$$



[He et al., 2014] [Song et al., arXiv]

Learning from Multiple Views

Example: Minimum Vertex Cover



Graph View

[Khalil et al., 2017]

$$\max \; -\sum_{i=1}^{5} x_i,$$

subject to:

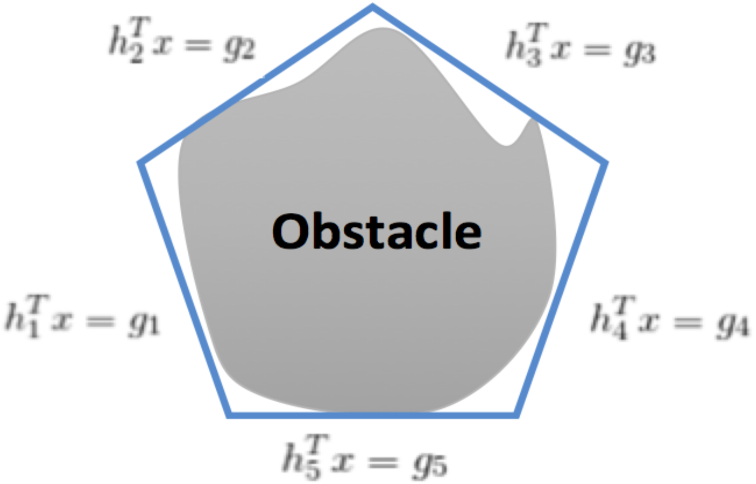
$$x_1 + x_2 \geq 1,$$
$$x_2 + x_3 \geq 1,$$
$$x_3 + x_4 \geq 1,$$
$$x_3 + x_5 \geq 1,$$
$$x_4 + x_5 \geq 1,$$
$$x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}$$

Integer Program View
(Branch & Bound View)

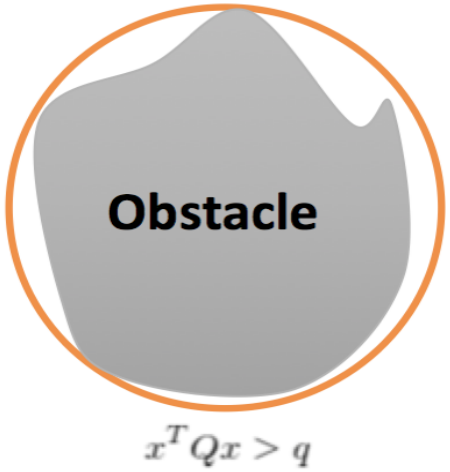
[He et al., 2014]

Learning from Multiple Views

Example: Different Types of Integer Programs



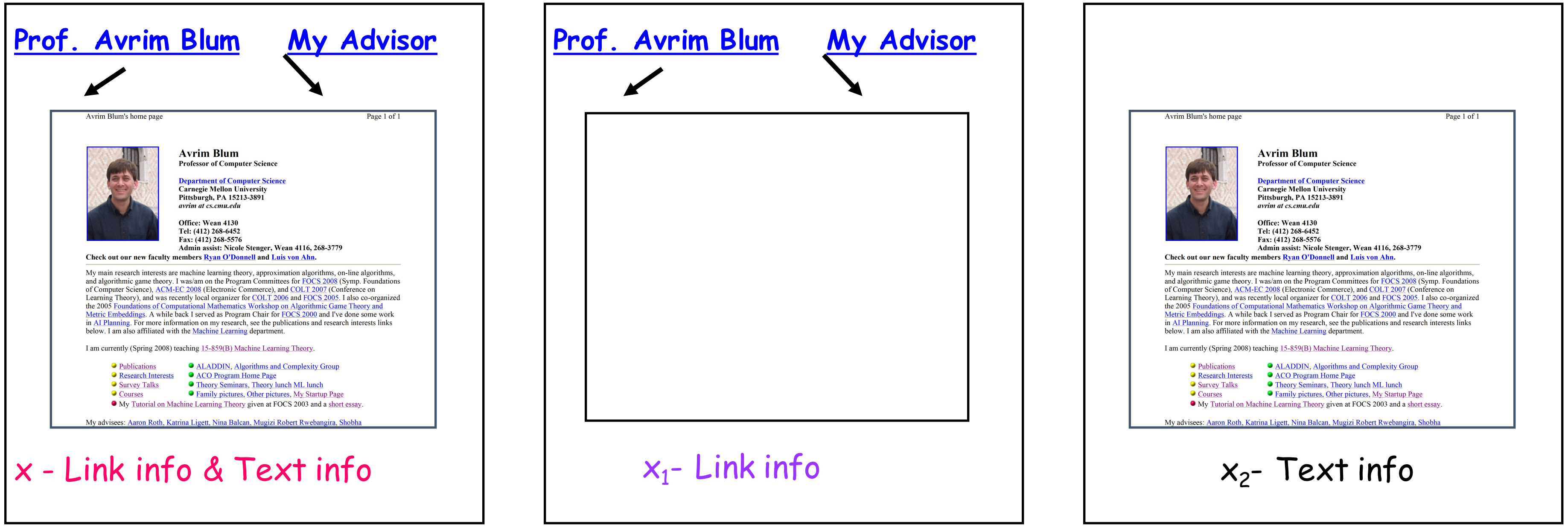
ILP



QCQP

Co-Training [Blum & Mitchell, 1998]

- Many learning problems have different sources of information
- Webpage Classification: Words vs Hyperlinks



(Taken from Avrim Blum's slides)

Semi-Supervised Regression with Co-Training

Zhi-Hua Zhou and **Ming Li**
National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{zhouzh, lim}@lamda.nju.edu.cn

Co-Training for Domain Adaptation

Minmin Chen, Kilian Q. Weinberger
Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, MO 63130
mc15,kilian@wustl.edu

John C. Blitzer
Google Research
1600 Amphitheatre Parkway
Mountain View, CA 94043
blitzer@google.com

Email Classification with Co-Training

Svetlana Kiritchenko and **Stan Matwin**
School of Information Technology and Engineering
University of Ottawa
Ottawa, ON, Canada
{svkir,stan}@site.uottawa.ca

A New Analysis of Co-Training

Wei Wang
Zhi-Hua Z
National Ke
DU.CN
EDU.CN

labeled D

Co-Training and Expansion: Towards Bridging Theory and Practice

Maria-Florina
Computer Scien
Carnegie Mell

Understanding the Behavior of Co-training

Kamal Nigam

Rayid Ghani
hool of Computer Science
arnegie Mellon University
Pittsburgh, PA 15213
rayid@cs.cmu.edu

Applying Co-Training methods to Statistical Parsing*

Anoop Sarkar
Dept. of Computer and Information Science
University of Pennsylvania
200 South 33rd Street,
Philadelphia, PA 19104-6389 USA

Applying Co-Training to Reference Resolution

ristoph Müller
ledia Laboratory GmbH
Villa Bosch
Volfsbrunnenweg 33
eidelberg, Germany
ml.villa-bosch.de

Stefan Rapp
Sony International (Europe) G
Advanced Technology Center S
Heinrich-Hertz-Straße 1
70327 Stuttgart, German

Michael Strube

Bayesian Co-Training

Shipeng Yu SHIPENG.YU@SIEMENS.COM
Balaji Krishnapuram BALAJI.KRISHNAPURAM@SIEMENS.COM
Business Intelligence and Analytics
Siemens Medical Solutions USA, Inc.
51 Valley Stream Parkway
Malvern, PA 19355, USA

Rómer Rosales ROMERR@YAHOO-INC.COM

Unsupervised Improvement of Visual Detectors using Co-Training

Anat Levin*
School of CS and Eng.
The Hebrew University
91904 Jerusalem, Israel
alevin@cs.huji.ac.il

Paul Viola
Microsoft Research
One Microsoft Way
Redmond, WA 98052
viola@microsoft.com

Yoav Freund
Computer Science Dept.
Columbia University
New York, NY 10027
freund@cs.columbia.edu

Reinforced Co-Training

William Yang Wang
Department of Computer Science
University of California
Santa Barbara, CA 93106 USA
william@cs.ucsb.edu

A Co-training Approach for Multi-view Spectral Clustering

Abhishek Kumar ABHISHEK@CS.UMD.EDU
Hal Daumé III HAL@UMIACS.UMD.EDU
Department of Computer Science, University of Maryland, College Park, MD 20742, USA

PAC Generalization Bounds for Co-training

Sanjoy Dasgupta
AT&T Labs–Research
dasgupta@research.att.com

Michael L. Littman
AT&T Labs–Research
mlittman@research.att.com

David McAllester
AT&T Labs–Research
dmac@research.att.com

Co-training for Policy Learning

Jialin Song[†] **Ravi Lanka**[‡] **Yisong Yue**[†] **Masahiro Ono**[‡]
[†] California Institute of Technology
[‡]Jet Propulsion Laboratory, California Institute of Technology

Semi-Supervised Regression with Co-Training

Zhi-Hua Zhou and **Ming Li**
National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{zhouzh, lim}@lamda.nju.edu.cn

Co-Training for Domain Adaptation

Minmin Chen, Kilian Q. Weinberger
Department of Computer Science and Engineering
Washington University in St. Louis

John C. Blitzer
Google Research
1600 Amphitheatre Parkway

Email Classification with Co-Training

Svetlana Kiritchenko and **Stan Matwin**
School of Information Technology and Engineering
University of Ottawa
Ottawa, ON, Canada
{svkir,stan}@site.uottawa.ca

A Co-training Approach for Multi-view Spectral Clustering

Abhishek Kumar ABHISHEK@CS.UMD.EDU
Hal Daumé III HAL@UMIACS.UMD.EDU
Department of Computer Science, University of Maryland, College Park, MD 20742, USA

PAC Generalization Bounds for Co-training

Sanjoy Dasgupta
AT&T Labs–Research
dasgupta@research.att.com

Michael L. Littman
AT&T Labs–Research
mlittman@research.att.com

David McAllester
AT&T Labs–Research
dmac@research.att.com

William Yang Wang
Department of Computer Science
University of California
Santa Barbara, CA 93106 USA
william@cs.ucsb.edu

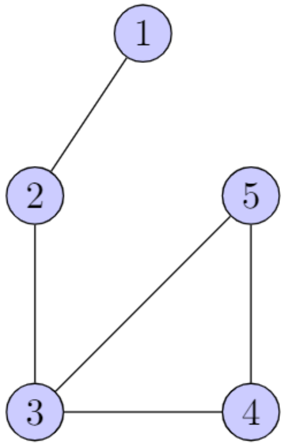
What's Different about Policy Co-Training?

- Sequential Decisions vs 1-Shot Decisions

- (Sparse) Environmental Feedback
  - Can collect more "labels"

(Not always applicable)

- Different Action Spaces
  - Graph vs Branch-and-Bound



$$\max \; -\sum_{i=1}^{5} x_i,$$

subject to:

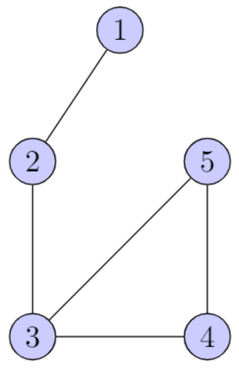
$$x_1 + x_2 \geq 1,$$
$$x_2 + x_3 \geq 1,$$
$$x_3 + x_4 \geq 1,$$
$$x_3 + x_5 \geq 1,$$
$$x_4 + x_5 \geq 1,$$
$$x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}$$

[1] "Learning combinatorial optimization algorithms over graphs" [Khalil et al., 2017]
[2] "Learning to Search in Branch and Bound Algorithms" [He et al., 2014]
[3] "Learning to Search via Retrospective Imitation" [Song et al., 2019]

Intuition

MVC Instance

E.g., [1]



E.g., [2,3]

$$\max \; -\sum_{i=1}^{5} x_i,$$

subject to:

$x_1 + x_2 \geq 1,$

$x_2 + x_3 \geq 1,$

$x_3 + x_4 \geq 1,$

$x_3 + x_5 \geq 1,$

$x_4 + x_5 \geq 1,$

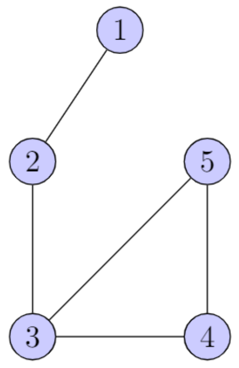
$x_i \in \{0,1\}, \forall i \in \{1, \cdots, 5\}$

[1] “Learning combinatorial optimization algorithms over graphs” [Khalil et al., 2017]
[2] “Learning to Search in Branch and Bound Algorithms” [He et al., 2014]
[3] “Learning to Search via Retrospective Imitation” [Song et al., 2019]

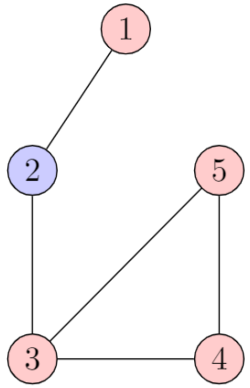
Intuition

MVC Instance

E.g., [1]



π^1



E.g., [2,3]

$$
\begin{aligned}
&\max \; -\sum_{i=1}^{5} x_i, \\
&\text{subject to:} \\
&x_1 + x_2 \geq 1, \\
&x_2 + x_3 \geq 1, \\
&x_3 + x_4 \geq 1, \\
&x_3 + x_5 \geq 1, \\
&x_4 + x_5 \geq 1, \\
&x_i \in \{0,1\}, \forall i \in \{1, \cdots, 5\}
\end{aligned}
$$

π^2

$$
\begin{aligned}
x_1 &= 0 \\
x_2 &= 1 \\
x_3 &= 1 \\
x_4 &= 1 \\
x_5 &= 0
\end{aligned}
$$

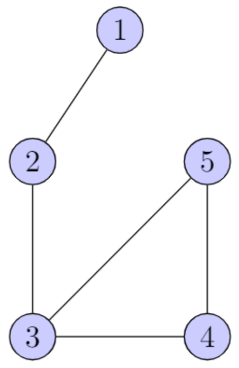
Better!

Intuition

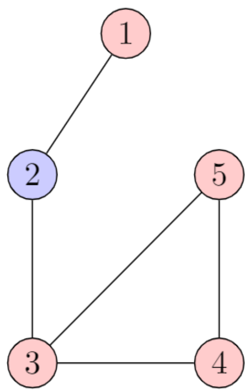
[1] “Learning combinatorial optimization algorithms over graphs” [Khalil et al., 2017]
[2] “Learning to Search in Branch and Bound Algorithms” [He et al., 2014]
[3] “Learning to Search via Retrospective Imitation” [Song et al., 2019]

MVC Instance

E.g., [1]



π^1



E.g., [2,3]

$$\max -\sum_{i=1}^{5} x_i,$$

subject to:

$$x_1 + x_2 \geq 1,$$
$$x_2 + x_3 \geq 1,$$
$$x_3 + x_4 \geq 1,$$
$$x_3 + x_5 \geq 1,$$
$$x_4 + x_5 \geq 1,$$
$$x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}$$

π^2

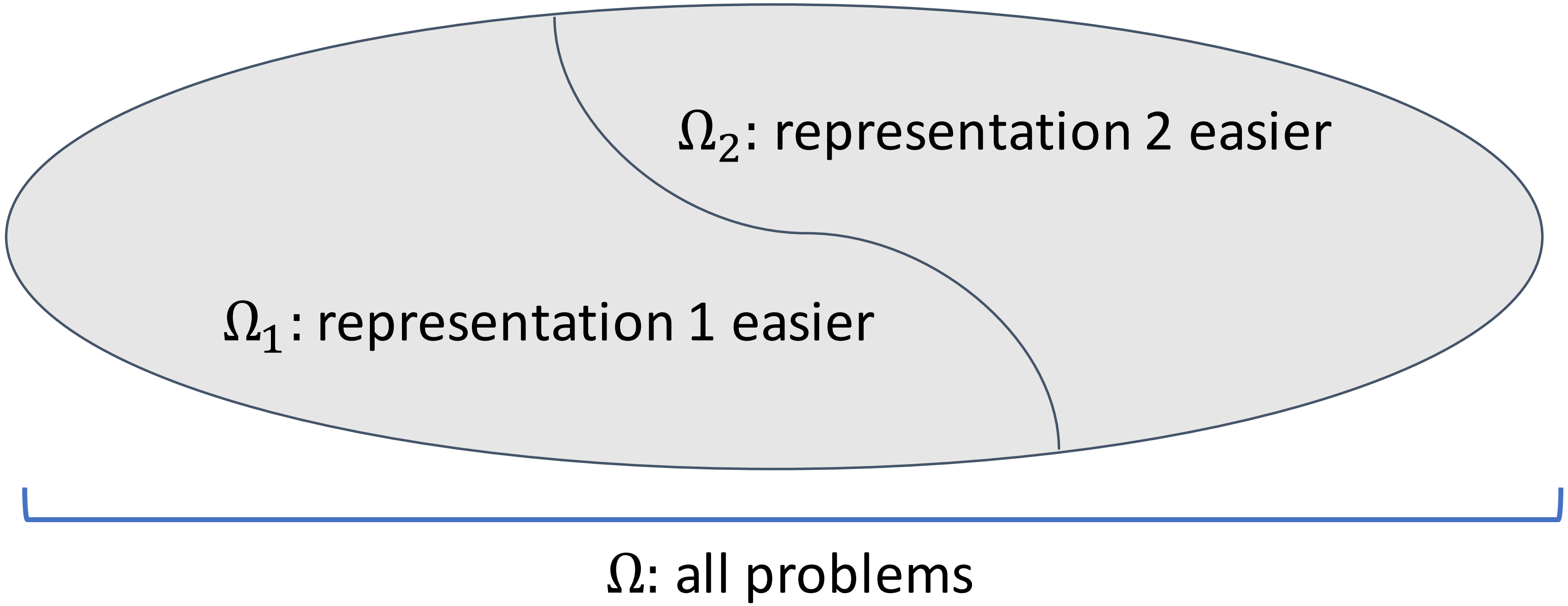
$$x_1 = 0$$
$$x_2 = 1$$
$$x_3 = 1$$
$$x_4 = 1$$
$$x_5 = 0$$

Demonstration

Better!

Theoretical Insight

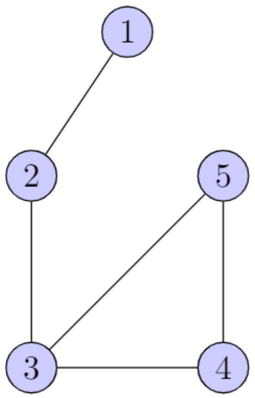
- Different representations differ in hardness
- Goal: quantify improvement



(Towards) a Theory of Policy Co-Training

- Two MDP "views": M^1 & M^2
  - $f^{1\to 2}(\tau^1) \implies \tau^2$ (and vice versa)

    "Trajectory" / "Rollout"
  - Realizing τ^1 on M^1 $\Longleftrightarrow$ realizing τ^2 on M^2



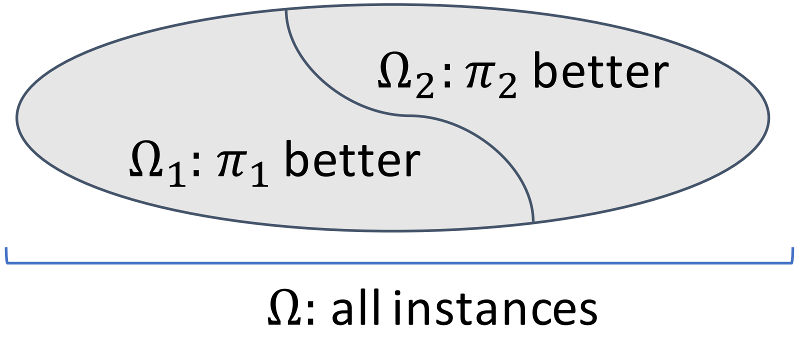
$$\max \; -\sum_{i=1}^{5} x_i,$$

subject to:

$$x_1 + x_2 \geq 1,$$
$$x_2 + x_3 \geq 1,$$
$$x_3 + x_4 \geq 1,$$
$$x_3 + x_5 \geq 1,$$
$$x_4 + x_5 \geq 1,$$
$$x_i \in \{0,1\}, \forall i \in \{1, \cdots, 5\}$$

- **Question:** when does having two views/policies help?
  - Policy Improvement (next slide)
    - Builds upon [Kang et al., ICML 2018]
  - Optimality Gap for Shared Action Spaces (in paper)
    - Builds upon [DasGupta et al., NeurIPS 2002]

Policy Improvement Bound



Standard for Policy Gradient

1-step suboptimality of π^1 on Ω

KL Divergence of π^1 vs π'^1 on Ω

JS Divergence of π^2 vs π^1 on Ω_2

Want to Minimize

1-step suboptimality of π^2 on Ω_2

$$J(\pi'^1) \geq J_{\pi^1}(\pi'^1) - \frac{2\gamma\left(\alpha^1_\Omega \varepsilon^1_\Omega + 4\beta^2_{\Omega_2}\varepsilon^2_{\Omega_2}\right)}{(1-\gamma)^2} + \delta^2_{\Omega_2}$$

Performance of new policy (either RL or IL)

Approximation by sampling from π^1

Discount

Performance Gap of π^2 over π^1 on Ω_2: $J(\pi^2|M\sim\Omega_2) - J(\pi^1|M\sim\Omega_2)$

Want to Maximize

Builds upon theoretical results from [Kang et al., ICML 2018]

Policy Improvement Bound (Summary)

$$J(\pi'^1) \geq J_{\pi^1}(\pi'^1) - \frac{2\gamma\left(\alpha_{\Omega}^1 \varepsilon_{\Omega}^1 + 4\beta_{\Omega_2}^2 \varepsilon_{\Omega_2}^2\right)}{(1-\gamma)^2} + \delta_{\Omega_2}^2$$

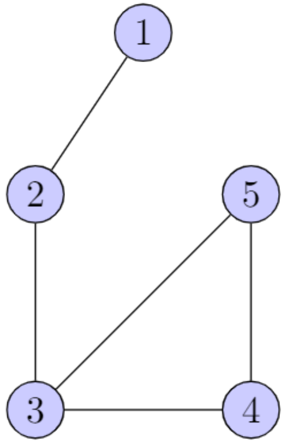
- Minimizing $\beta_{\Omega_2}^2$ → low disagreement between π^2 vs π^1
- Maximizing $\delta_{\Omega_2}^2$ → high performance gap π^2 over π^1 on some MDPs

CoPiEr Algorithm (Co-training for Policy Learning)

Sample $M \sim \Omega$

Rollout

Run $\pi^1 \rightarrow \tau^1$ Run $\pi^2 \rightarrow \tau^2$



M^1

$$\max -\sum_{i=1}^{5} x_i,$$

subject to:

$$x_1 + x_2 \geq 1,$$
$$x_2 + x_3 \geq 1,$$
$$x_3 + x_4 \geq 1,$$
$$x_3 + x_5 \geq 1,$$
$$x_4 + x_5 \geq 1,$$
$$x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}$$

M^2

Exchange (only showing 1 version)

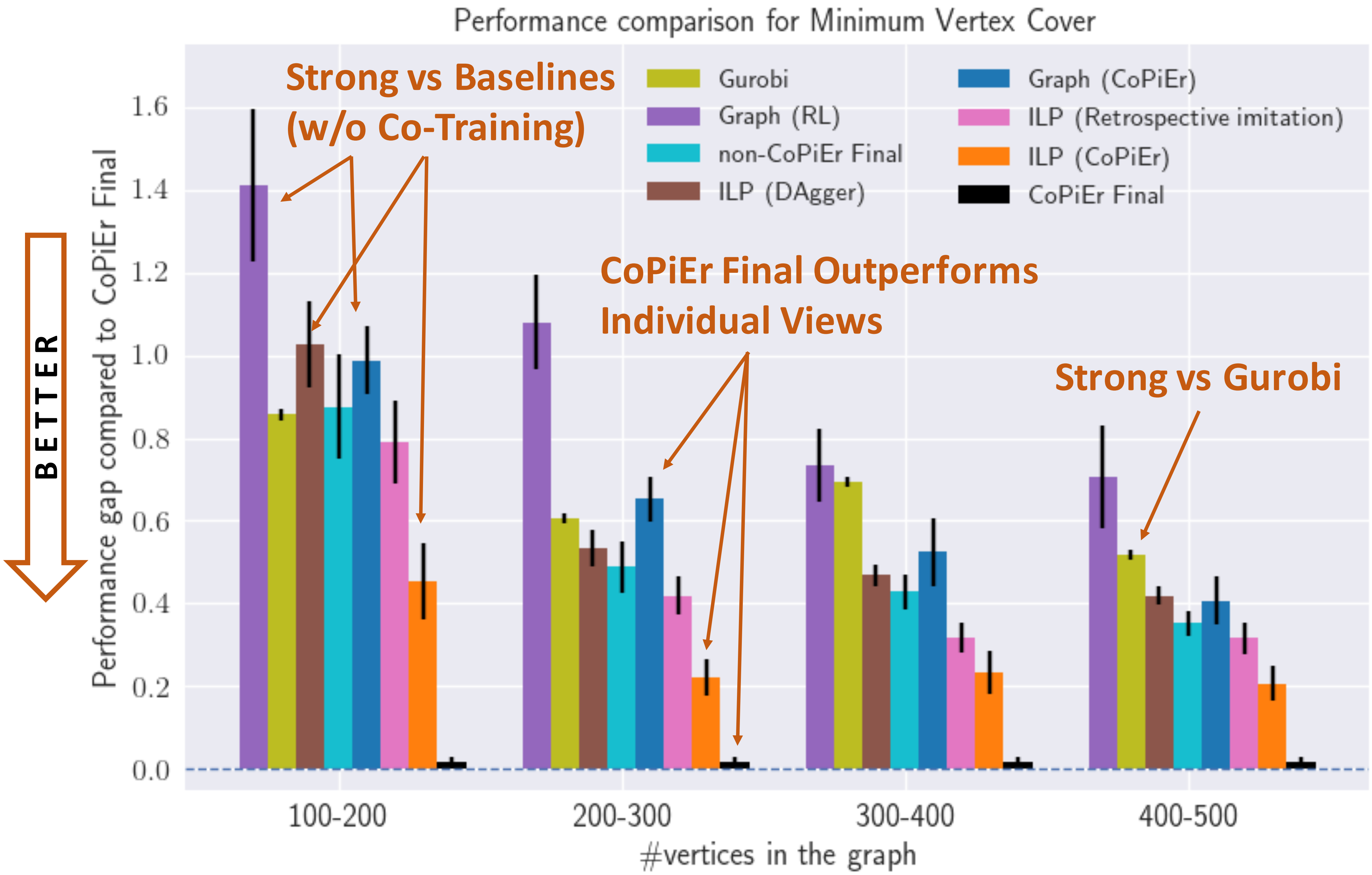
If π^1 better: $\tau'^2 = f^{1\rightarrow 2}(\tau^1)$, $\tau'^1 = \emptyset$

If π^2 better: $\tau'^1 = f^{2\rightarrow 1}(\tau^2)$, $\tau'^2 = \emptyset$

Update (only showing 1 view)

Augmented Obj: $\tilde{J}(\pi') = J_\pi(\pi') - \lambda L(\pi', \tau')$

Take gradient step



Erdős–Rényi
(100-500 vertices)

RL on Graph View
[Khalil et al., 2017]

IL on MILP View
[He et al., 2014]

More experiments
in paper

Co-Training for Policy Learning (summary)

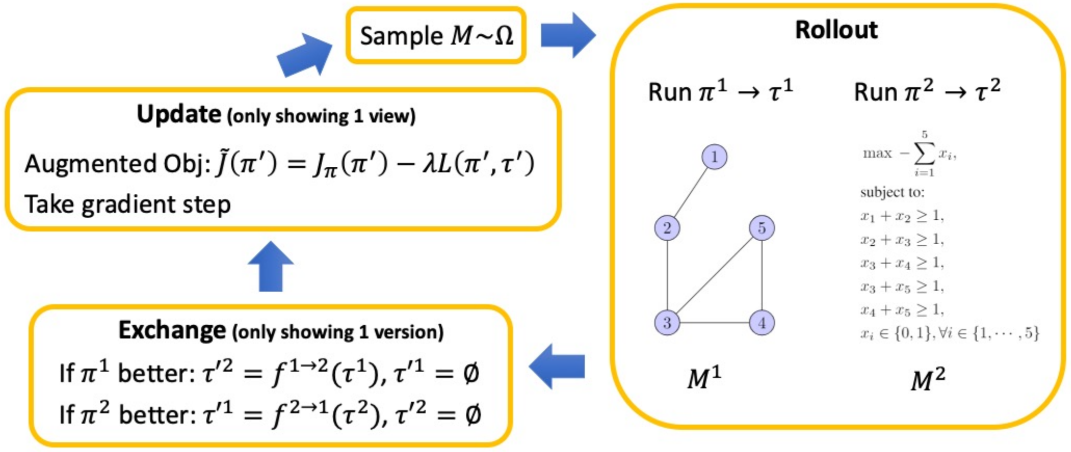
- First formal framework for policy co-training

- Novel theoretical insights

$$J(\pi'^1) \geq J_{\pi^1}(\pi'^1) - \frac{2\gamma\left(\alpha^1_\Omega \varepsilon^1_\Omega + 4\beta^2_{\Omega_2}\varepsilon^2_{\Omega_2}\right)}{(1-\gamma)^2} + \delta^2_{\Omega_2}$$

- Principled algorithm design

- Strong experimental results

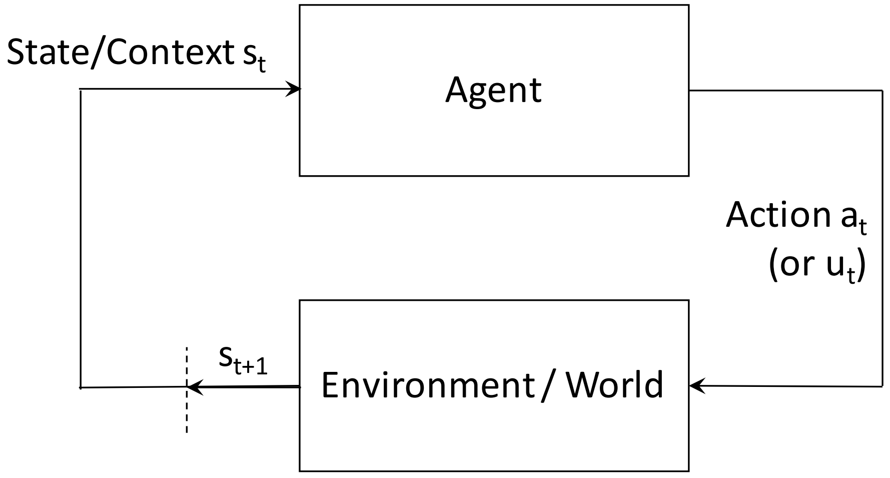


References

- **"Co-Training for Policy Learning,"** Jialin Song, Ravi Lanka, Yisong Yue, Masahiro Ono, UAI 2019
- **"Combining Labeled and Unlabeled data with Co-training,"** Avrim Blum, Tom Mitchell, COLT 1998
- **"PAC Generalization Bounds for Co-training,"** Sanjoy DasGupta, Michael Littman, David McAllester, NeurIPS 2002
- **"Policy Optimization with Demonstrations,"** Bingyi Kang, Zequn Jie, Jiashi Feng, ICML 2018
- **"Learning Combinatorial Optimization over Graphs,"** Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, Le Song, NeurIPS 2017
- **"Learning to Search in Branch and Bound Algorithms,"** He He, Hal Daume III, Jason Eisner, NeurIPS 2014
- **"Learning to Search via Retrospective Imitation,"** Jialin Song, Ravi Lanka, Albert Zhao, Aadyot Bhatnagar, Yisong Yue, Masahiro Ono, arXiv

Extra Slides

Notation



Transition P(s'|s,a)
Discount Factor

- MDP: $M = (S, A, P, r, \gamma, S_T)$

Terminal State
State Space
Action Space
Reward r(s,a)
(either RL or IL)

- Policy (or agent): $\pi(s) \implies a$
  - $V_\pi(s) = E_\tau\left[\sum_{t=0} \gamma^t r(s_t, a_t) | s_0 = s\right]$ Value Function
  - $Q_\pi(s, a) = E_\tau\left[\sum_{t=0} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right]$ Q Function
  - $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$ Advantage Function

- **Goal:** maximize $J(\pi) \equiv E_{M, s_0}[V_\pi(s_0)]$ (E.g., via Policy Gradient)

(Over Distribution of MDPs)

Policy Improvement Bound (detailed)

$$\max_{M \in supp(\Omega_2)} \max_{s,a} |A_{\pi^2}(s,a)|$$

$$E_{M\sim\Omega}\left[\max_s D_{KL}(\pi^1(s) \| \pi'^1(s))\right]$$

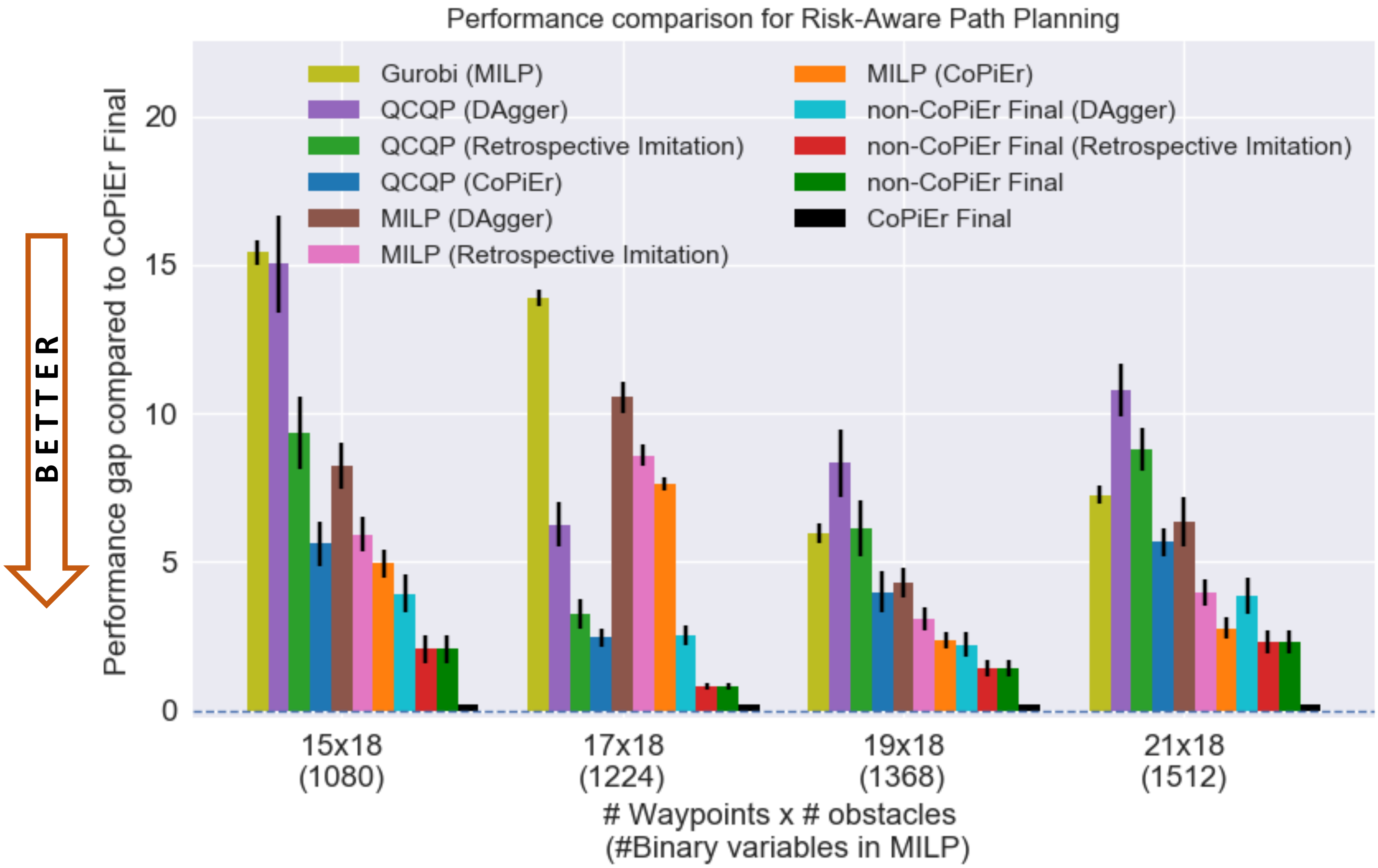
$$E_{M\sim\Omega_2}\left[\max_s D_{JS}(\pi^1(s) \| \pi^2(s))\right]$$

$$J(\pi'^1) \geq J_{\pi^1}(\pi'^1) - \frac{2\gamma\left(4\beta^2_{\Omega_2}\varepsilon^2_{\Omega_2} + \alpha^1_{\Omega}\varepsilon^1_{\Omega}\right)}{(1-\gamma)^2} + \delta^2_{\Omega_2}$$

New policy **Sampled via π^1** **Discount**

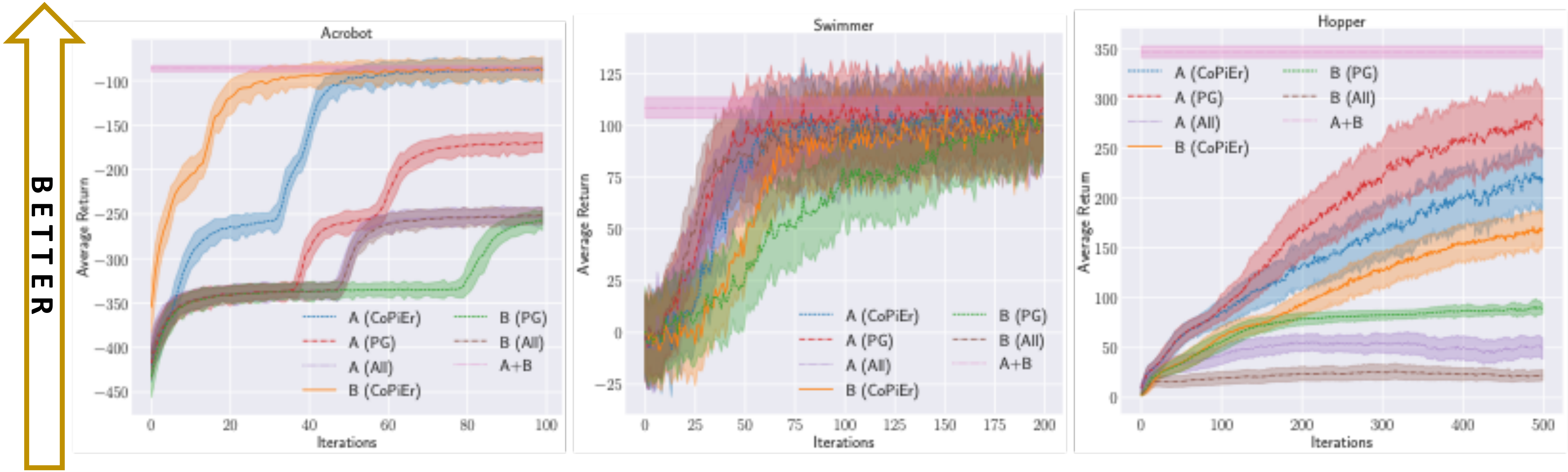
$$J(\pi^2 | M \sim \Omega_2) - J(\pi^1 | M \sim \Omega_2)$$

Builds upon theoretical results from [Kang et al., ICML 2018]



Chance-Constrained Path Planning [Ono & Williams, 2008]

IL on MILP
IL on QCQP
[He et al., 2014]
[Song et al., arXiv]



OpenAI Gym & Mujoco

Partitioned state space into two views

Shared action space

RL on both views