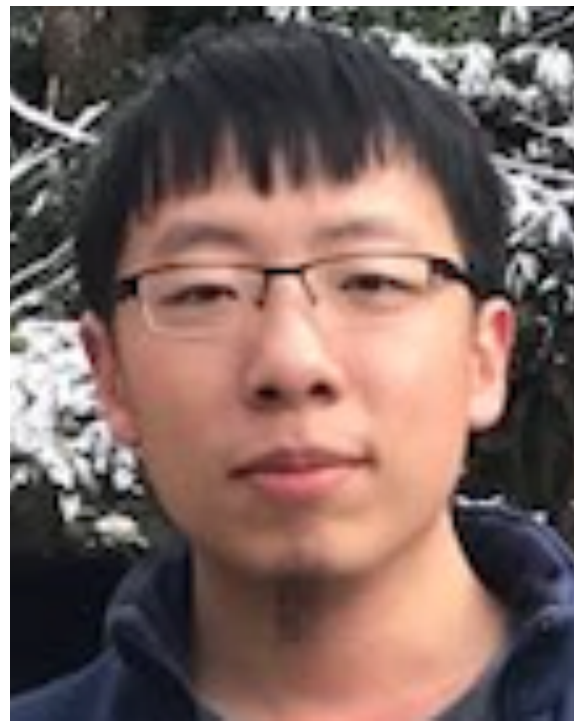
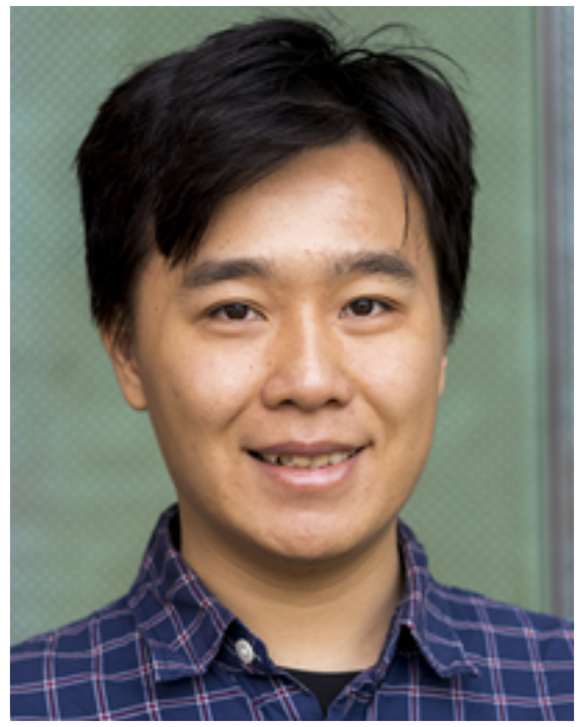
Information-Theoretic Considerations in Batch RL





Jinglin Chen, Nan Jiang

University of Illinois at Urbana Champaign

What we study: theory of batch RL (ADP)—backbone for "deep RL"

What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

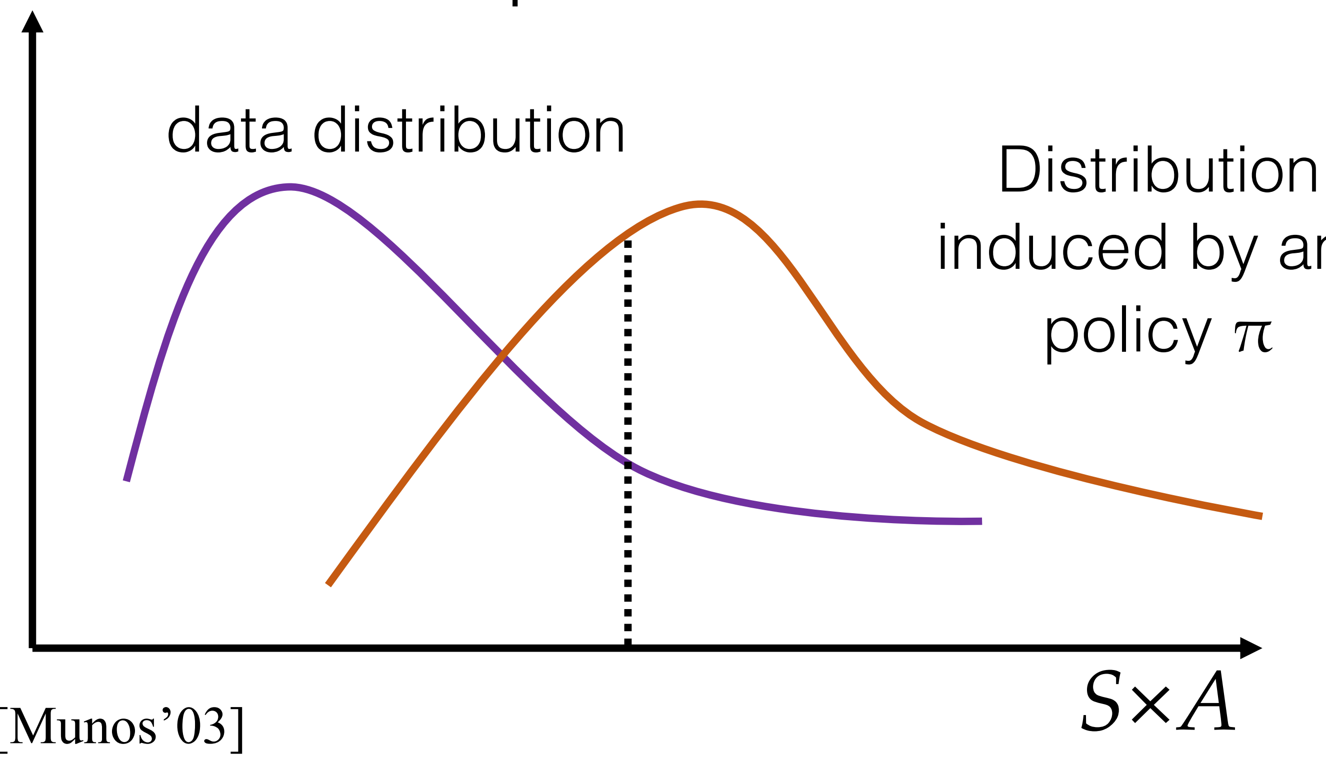
What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Assumption on data



[Munos'03]

Assumption on F

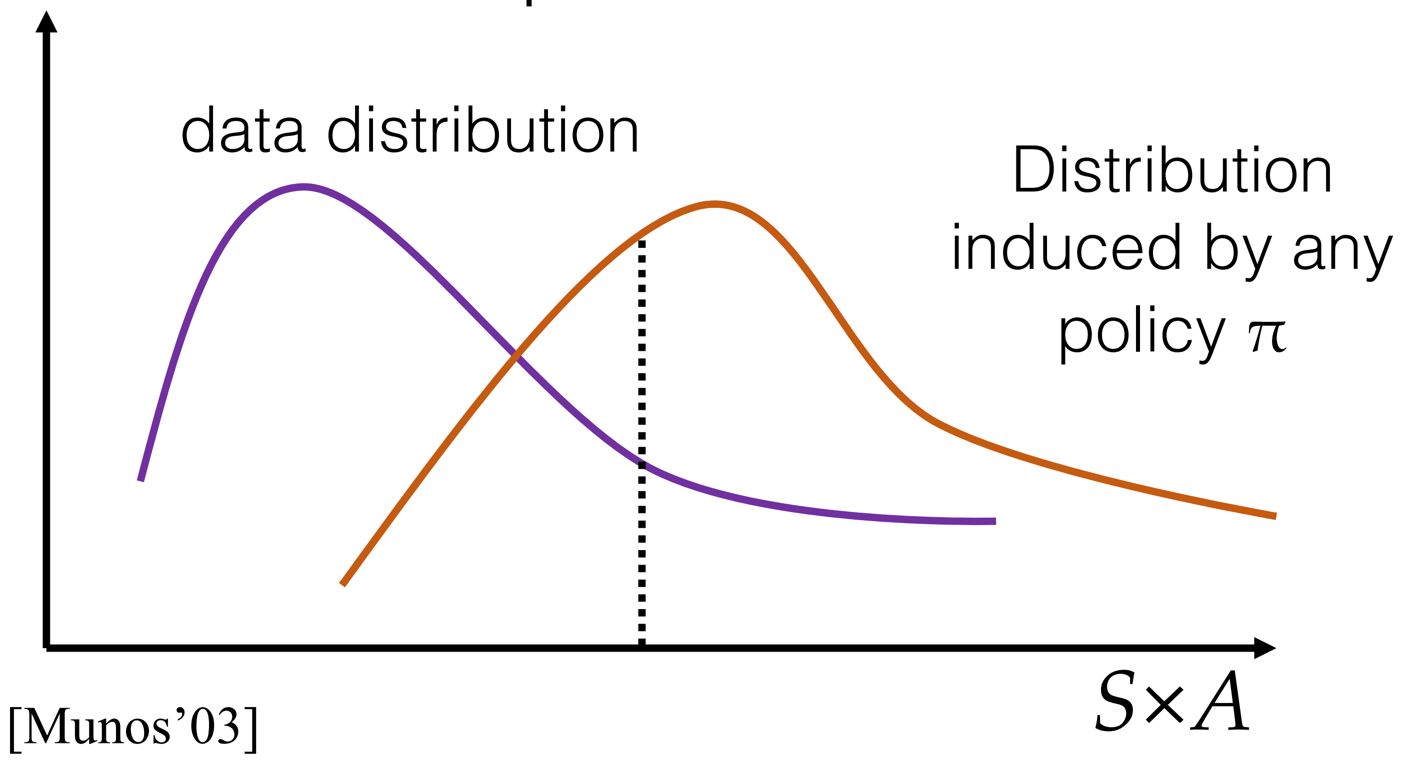
What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

⇩

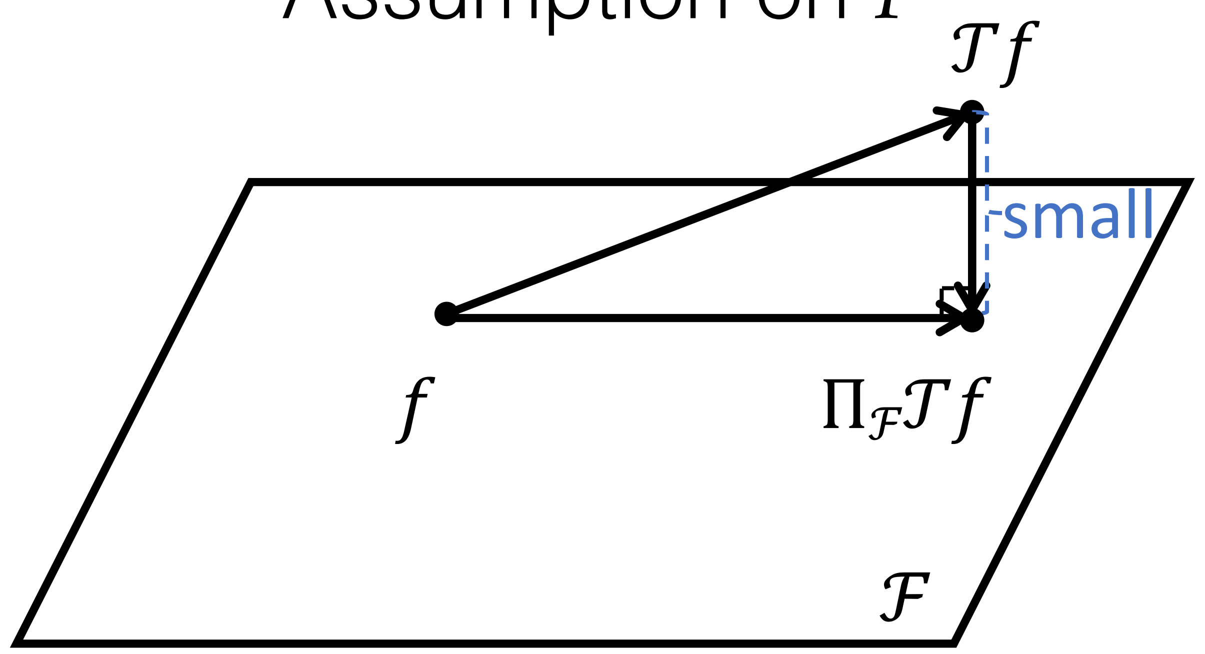
Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Assumption on data



[Munos'03]

Assumption on F



[Munos & Szepesvari '05]

What we study: theory of batch RL (ADP)—backbone for "deep RL"

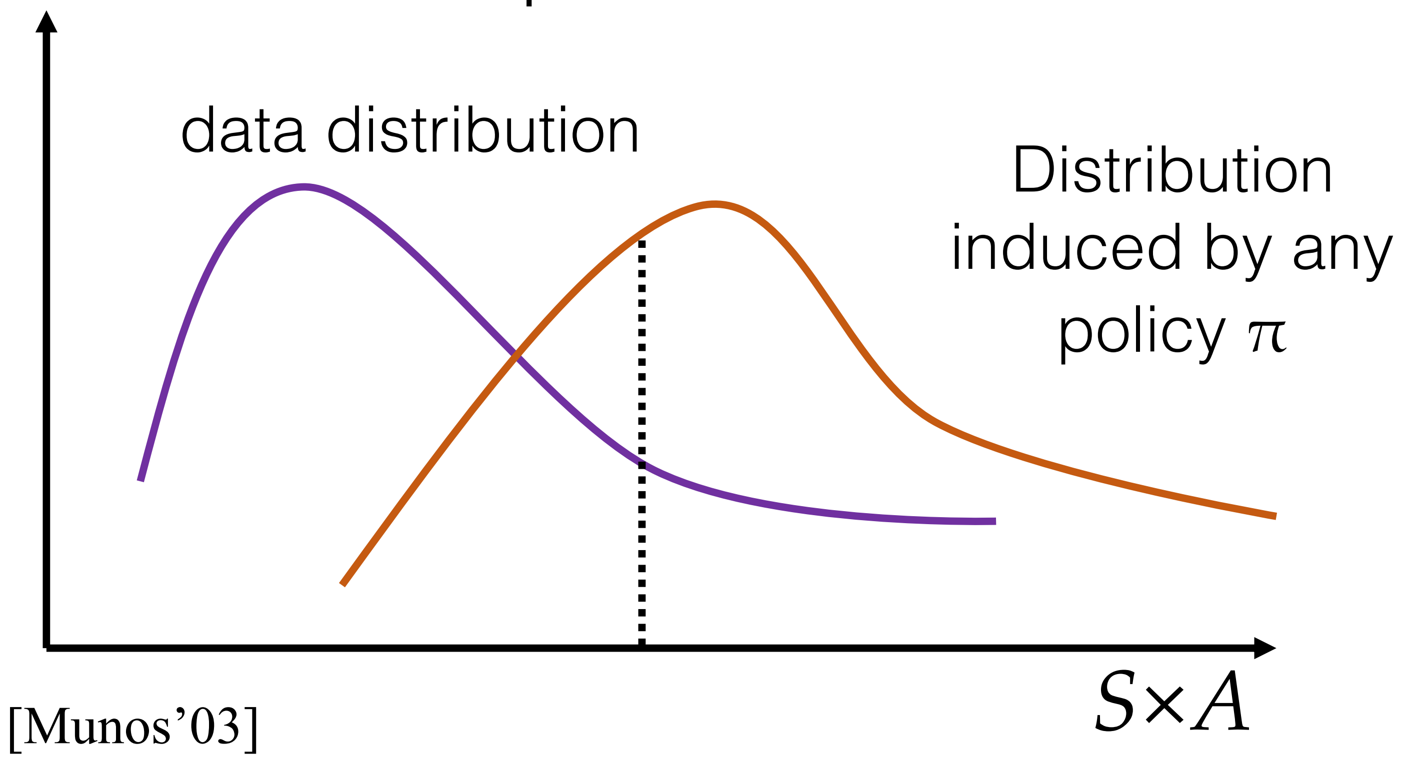
Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

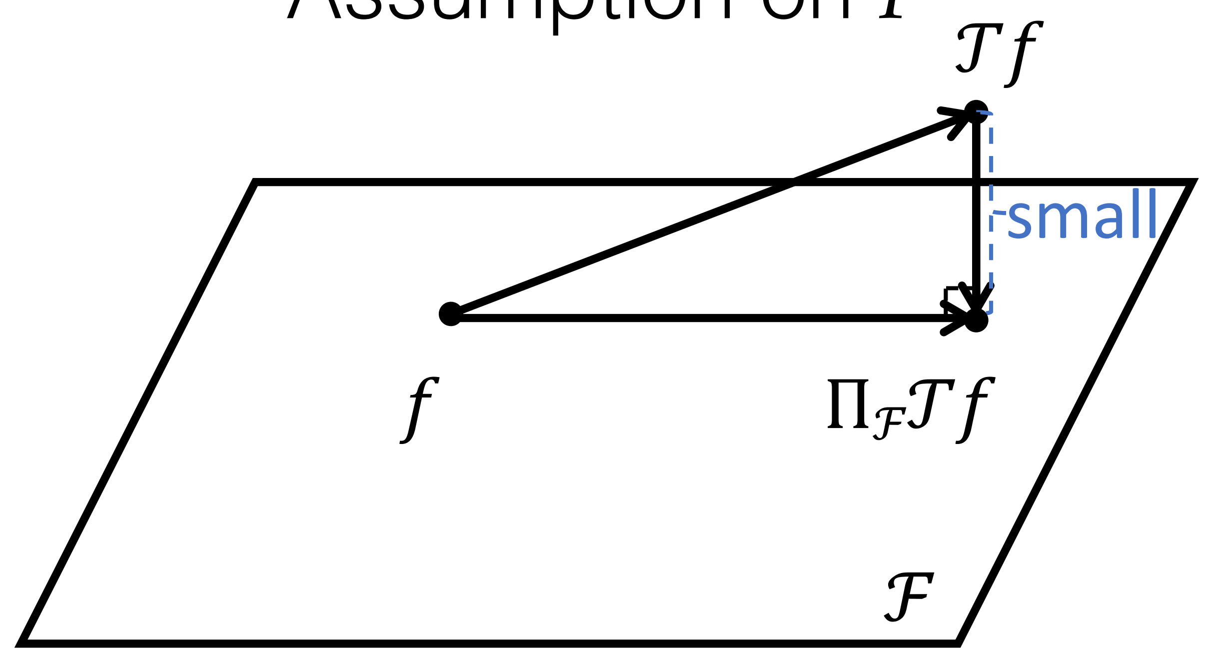
	Are they necessary? (hardness results)	*Do they hold in interesting scenarios?*
Assumption on data		
Assumption on F		

Assumption on data



[Munos'03]

Assumption on F



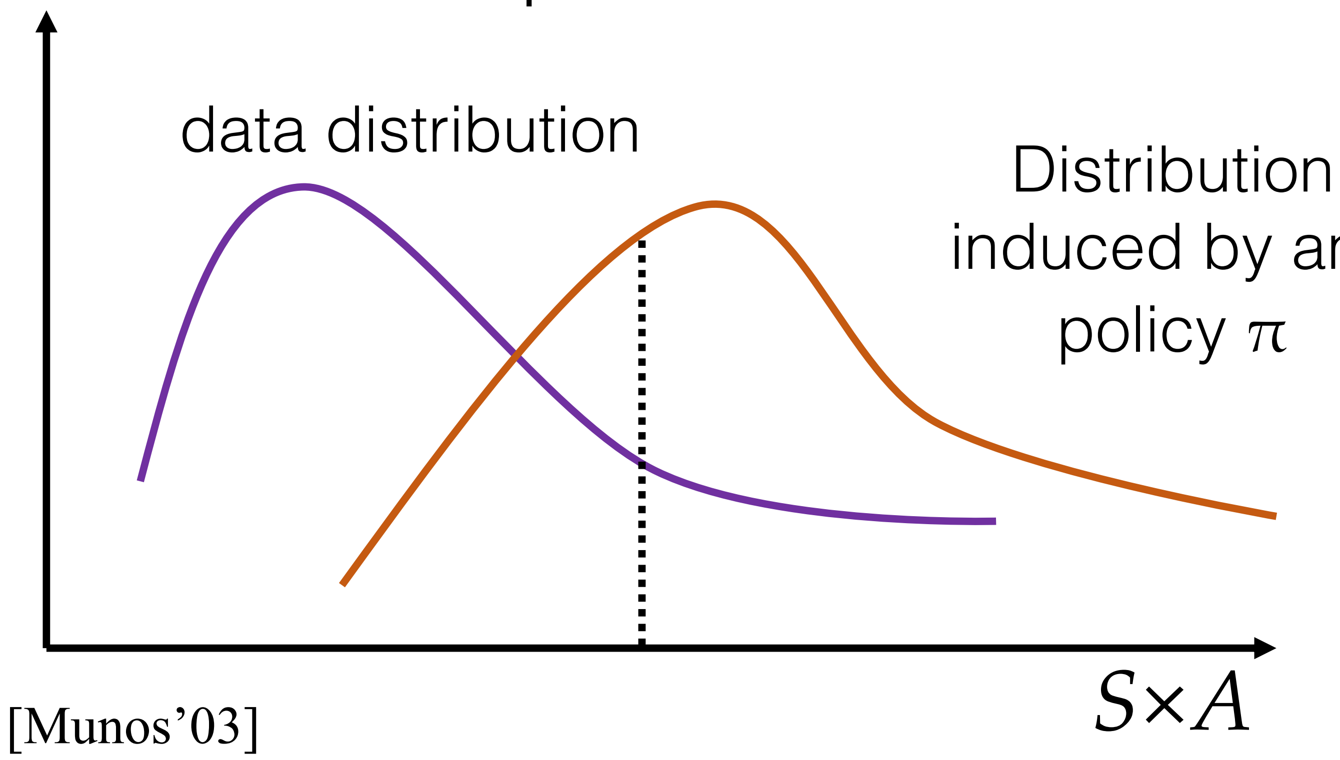
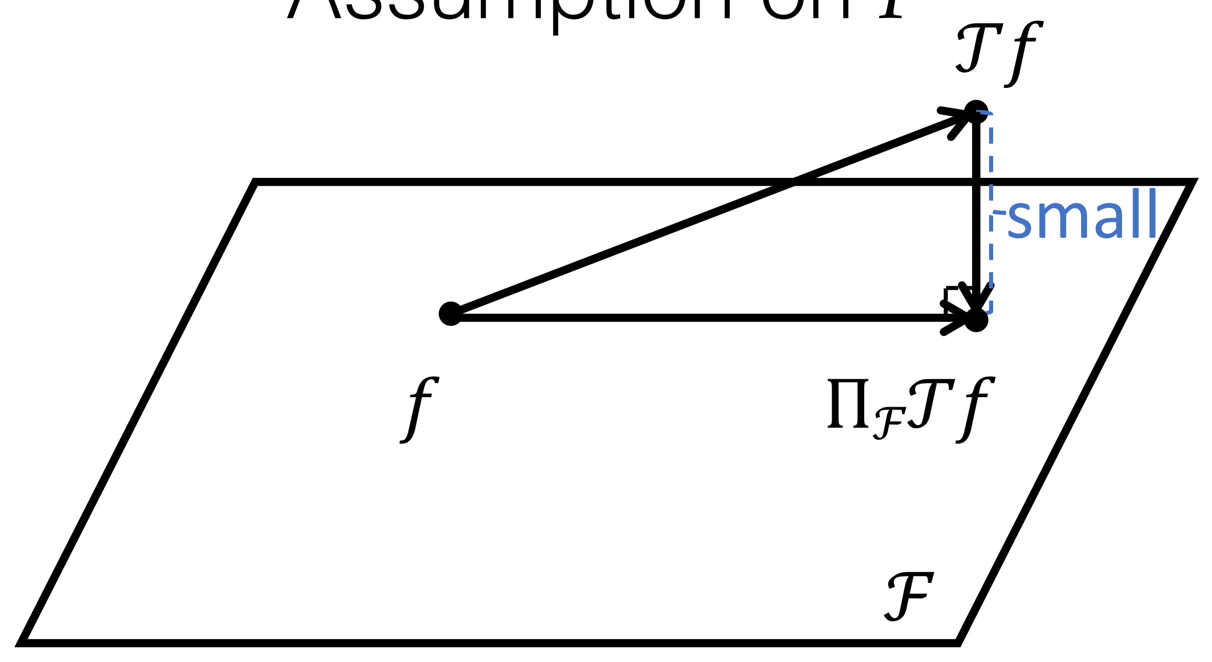
[Munos & Szepesvari '05]

What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

	Are they necessary? (hardness results)	*Do they hold in interesting scenarios?*
Assumption on data  [Munos’03]	• Intuition: data should be exploratory	
Assumption on F  [Munos & Szepesvari ’05]		

What we study: theory of batch RL (ADP)—backbone for "deep RL"

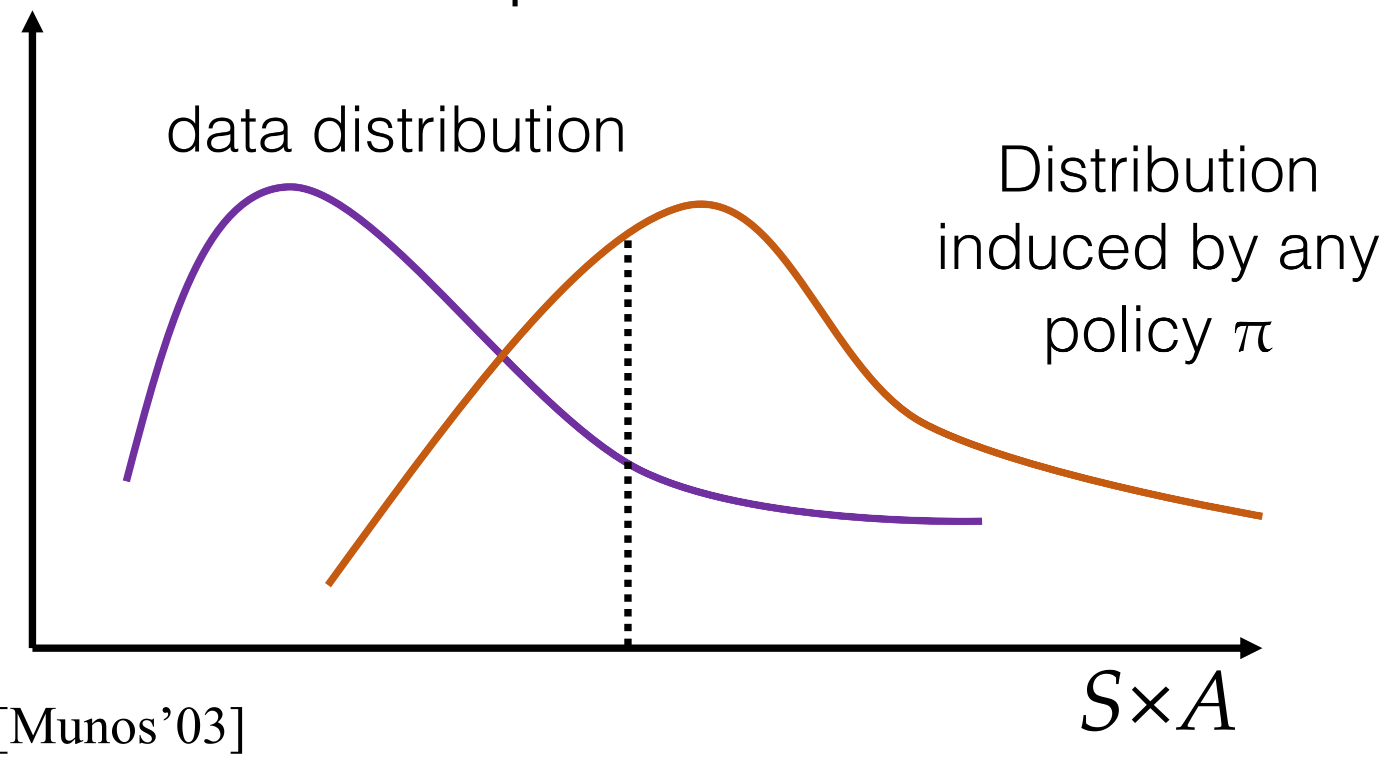
Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

⇩

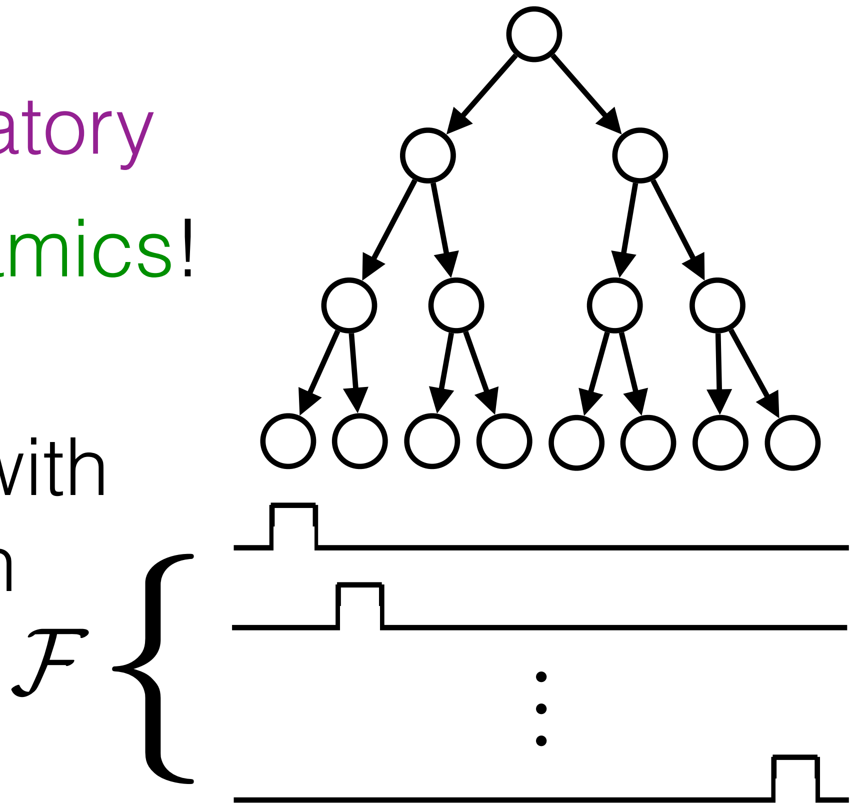
Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

	Are they necessary? (hardness results)	*Do they hold in interesting scenarios?*
Assumption on data	• Intuition: data should be exploratory • We show: also about MDP dynamics! • Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution	
Assumption on F		

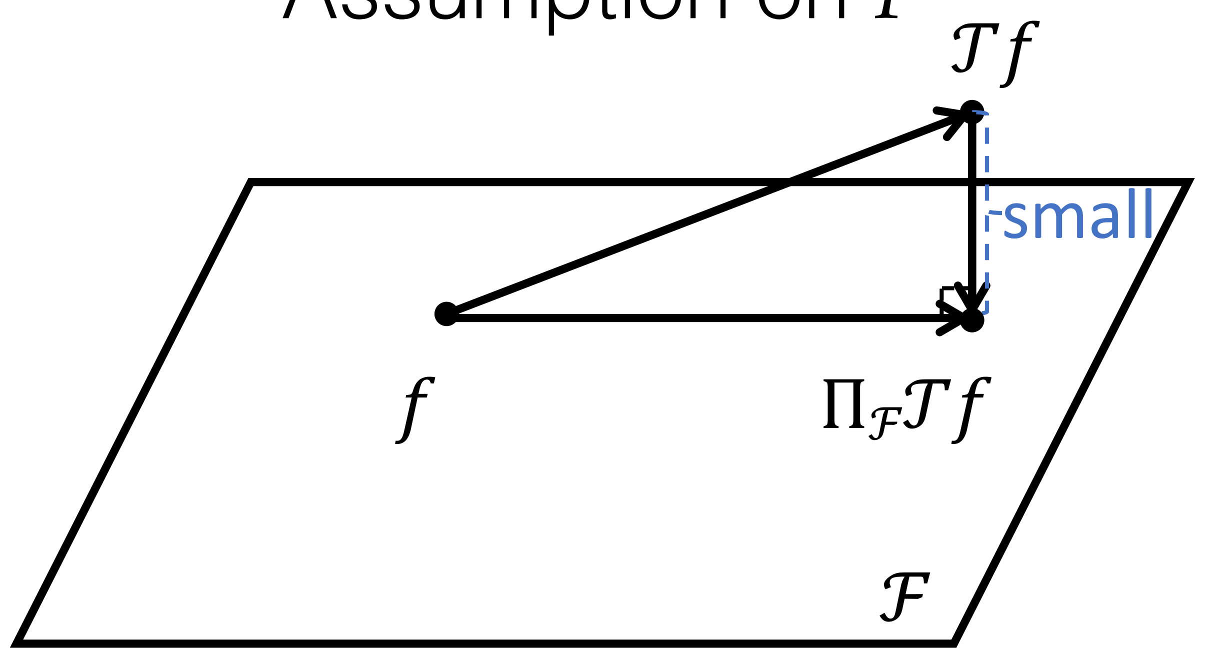
Assumption on data



[Munos'03]



Assumption on F



[Munos & Szepesvari '05]

What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

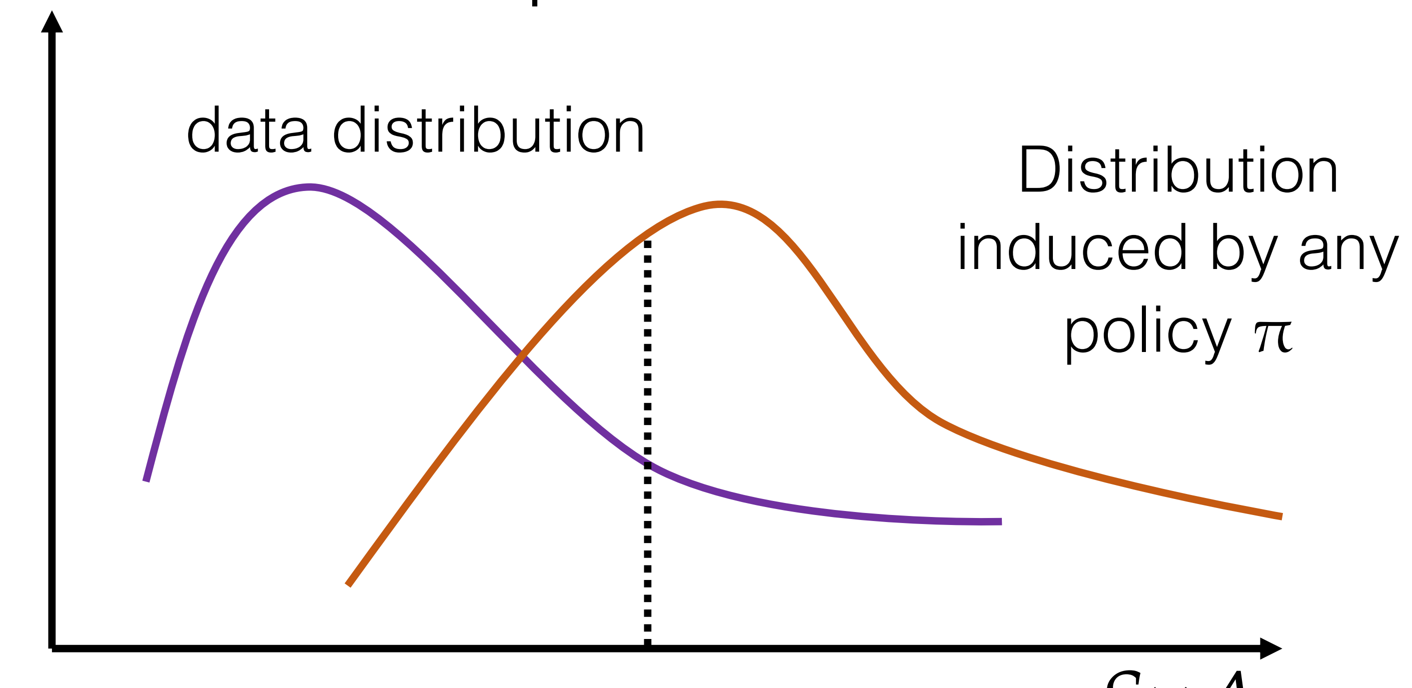
⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Are they necessary? (hardness results)

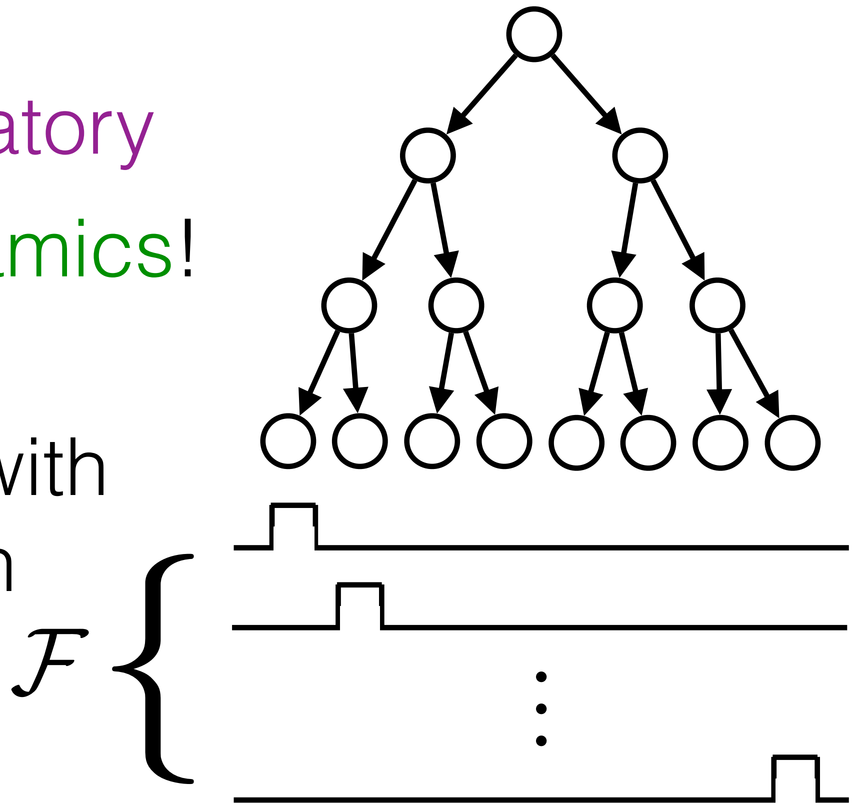
Do they hold in interesting scenarios?

Assumption on data

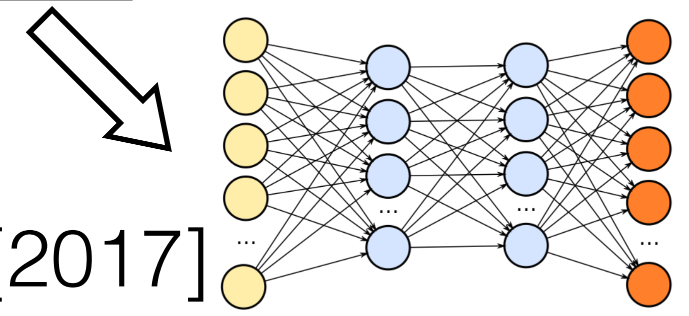


[Munos'03]

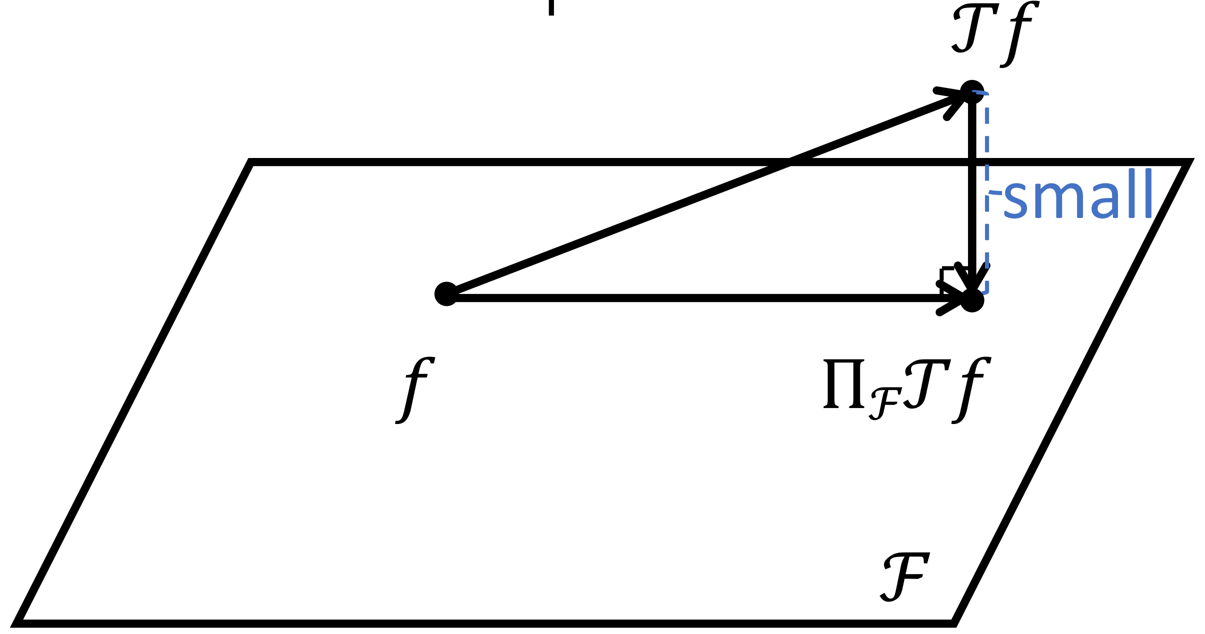
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution



Similar to Jiang et al [2017]



Assumption on F



[Munos & Szepesvari '05]

What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

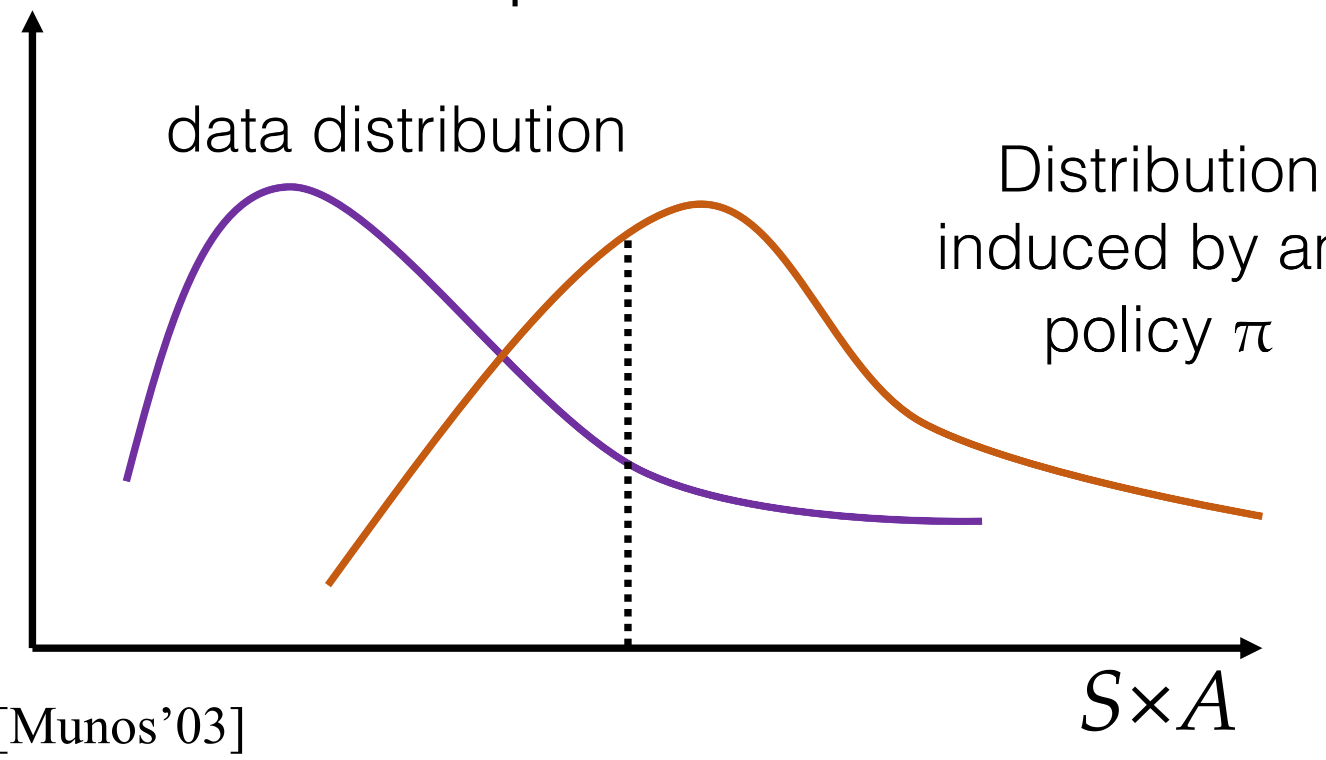
⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Are they necessary? (hardness results)

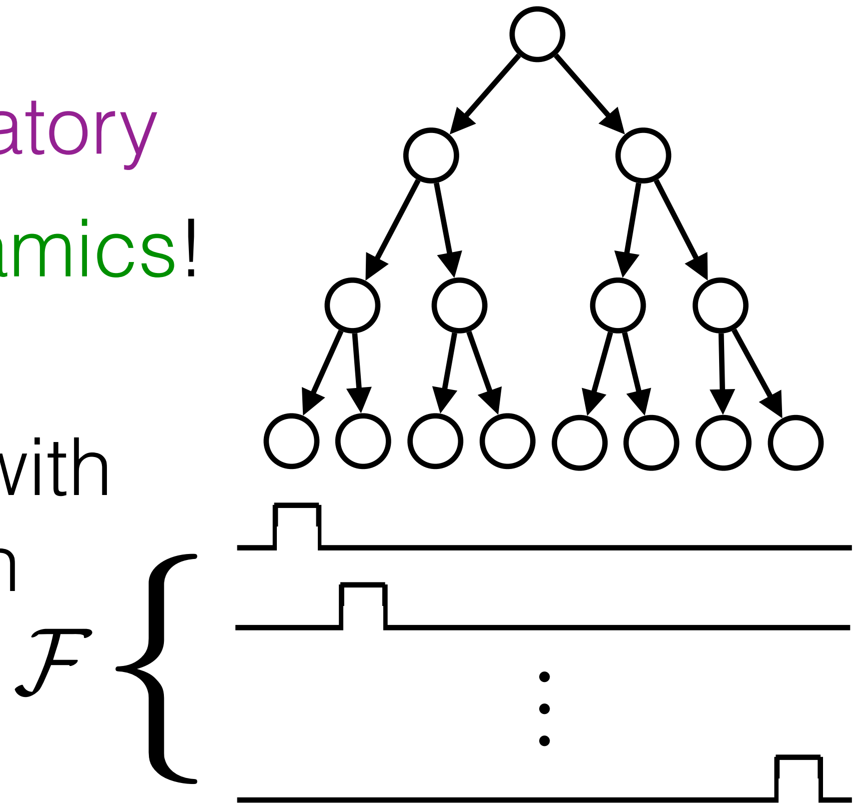
Do they hold in interesting scenarios?

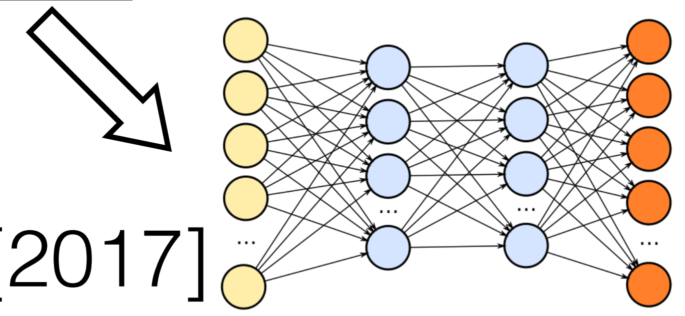
Assumption on data



[Munos'03]

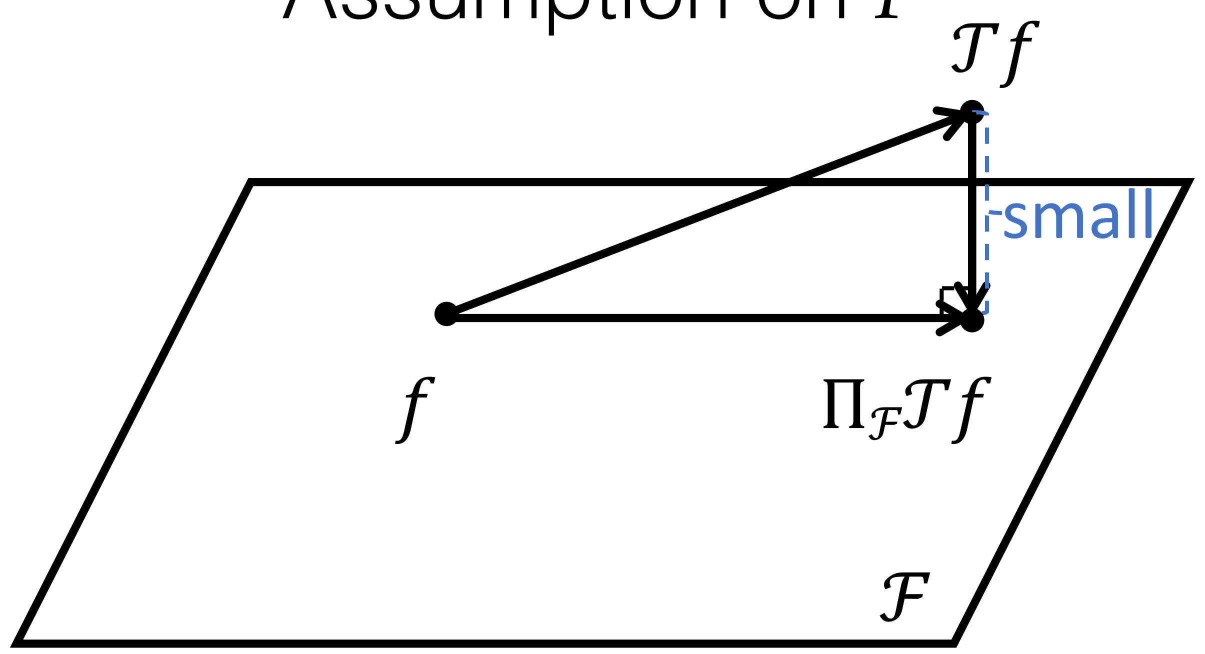
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution





Similar to Jiang et al [2017]

Assumption on F



[Munos & Szepesvari '05]

- Conjecture: realizability alone is insufficient

What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

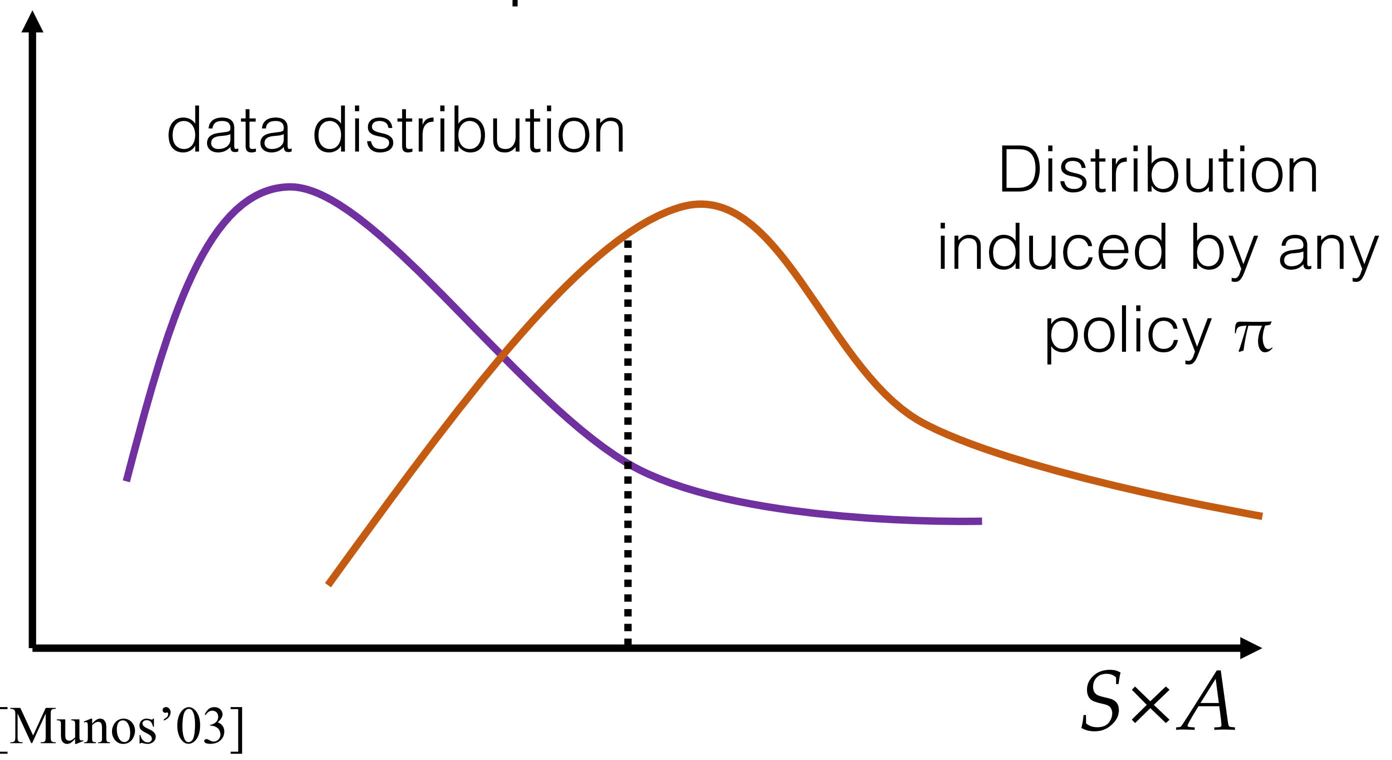
⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Are they necessary? (hardness results)

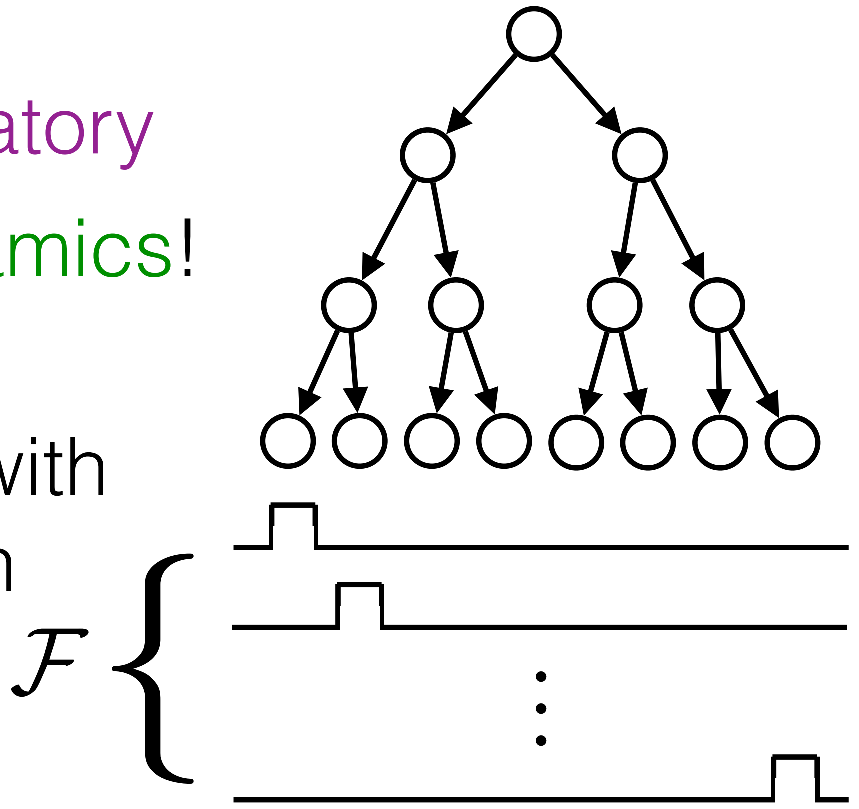
Do they hold in interesting scenarios?

Assumption on data

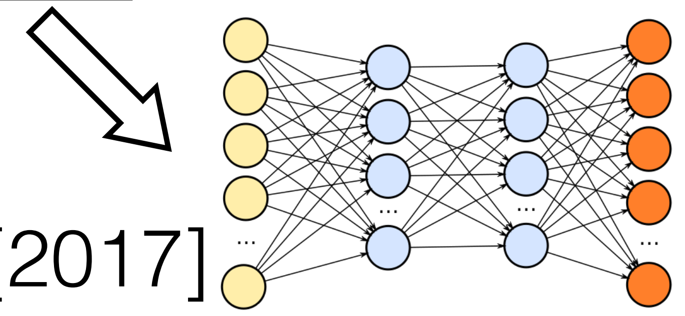


[Munos'03]

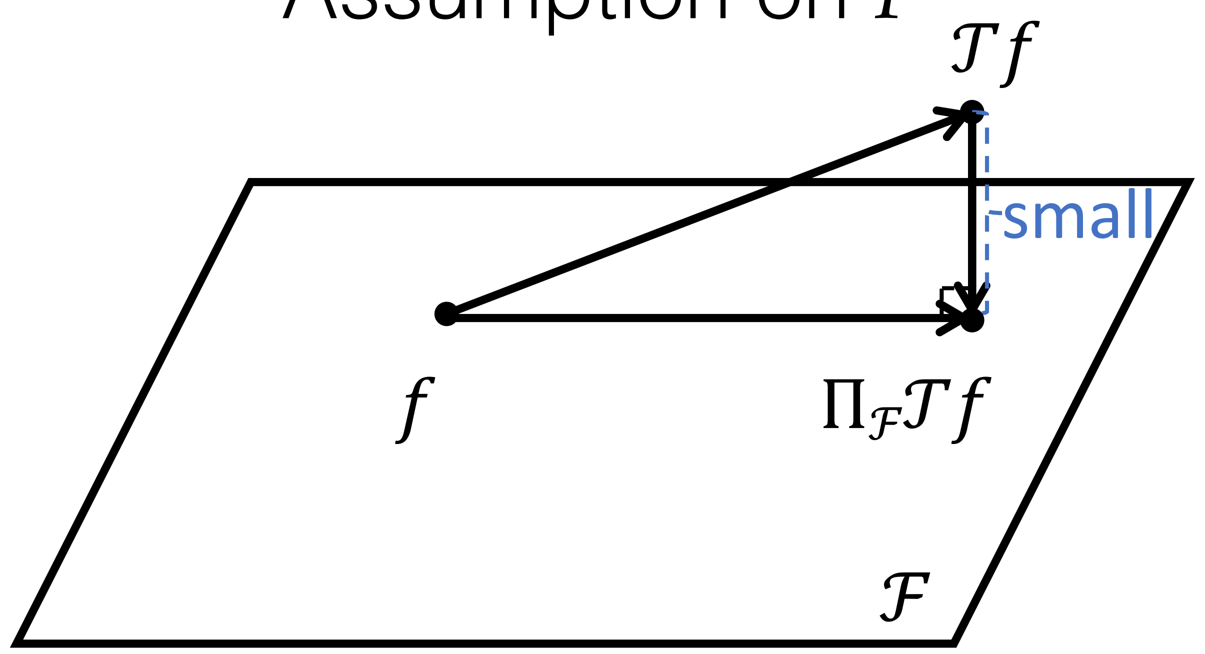
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution



Similar to Jiang et al [2017]



Assumption on F



[Munos & Szepesvari '05]

- Conjecture: realizability alone is insufficient
- Alg-specific lower bound exists for decades
- *Info-theoretic*?

?

What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

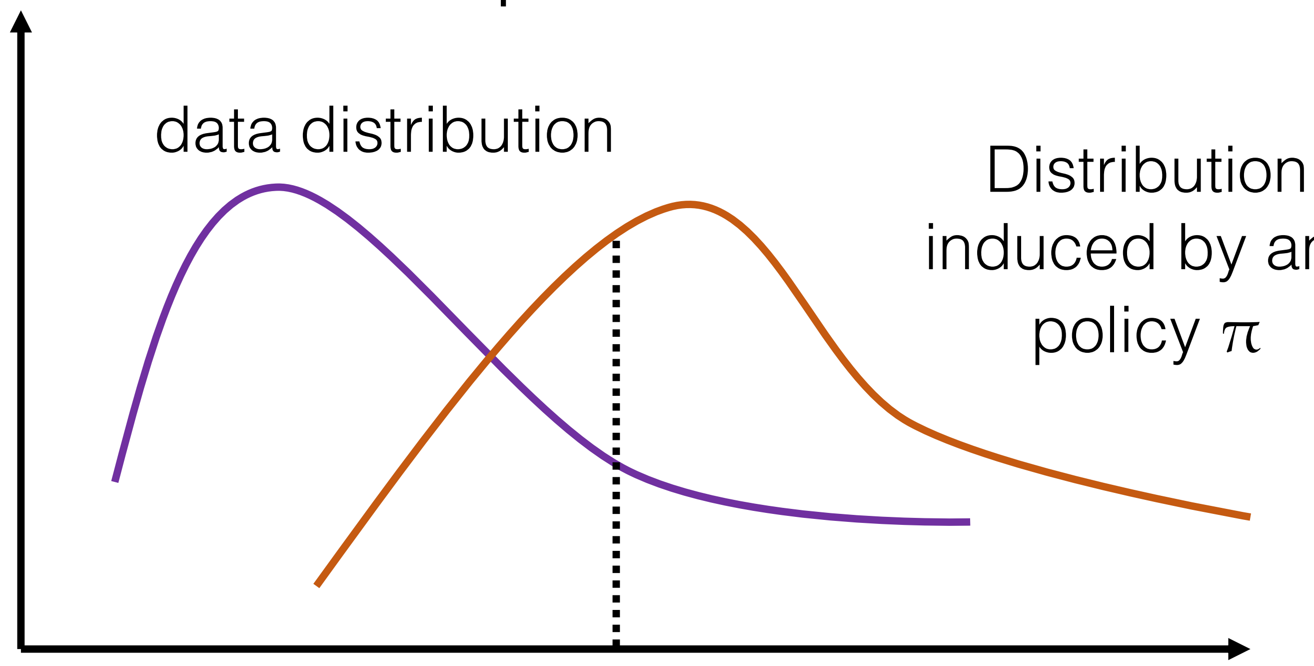
⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Are they necessary? (hardness results)

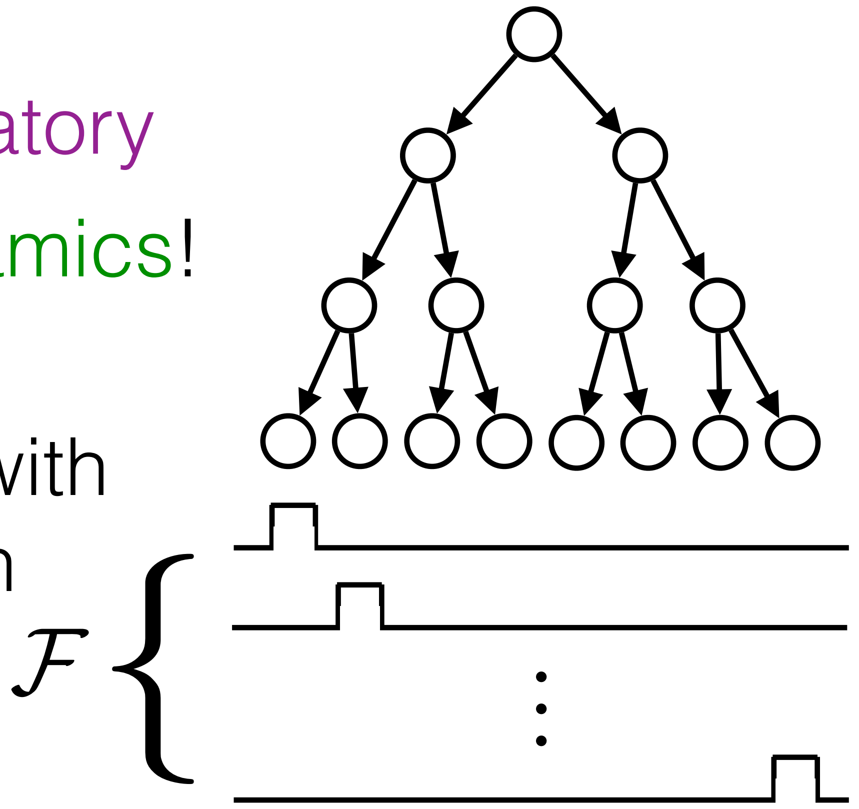
Do they hold in interesting scenarios?

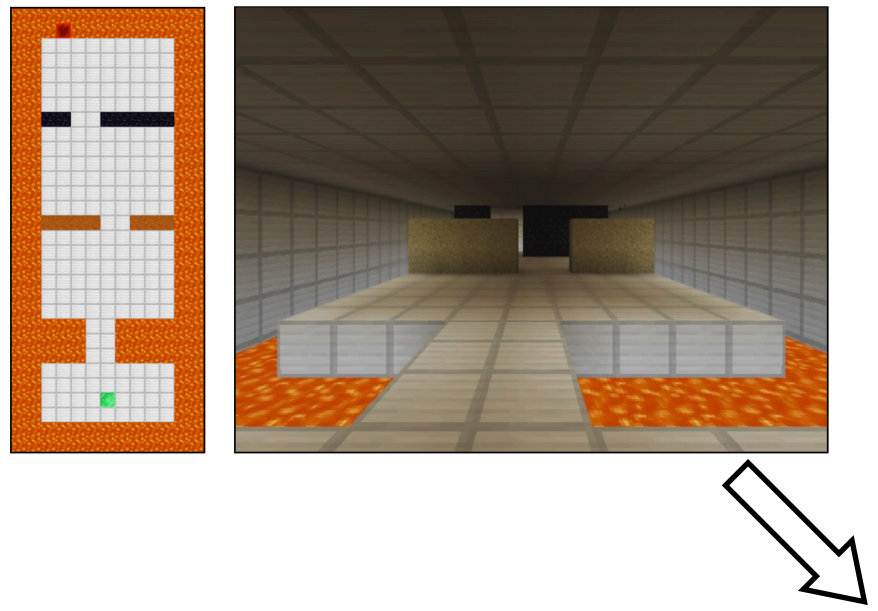
Assumption on data



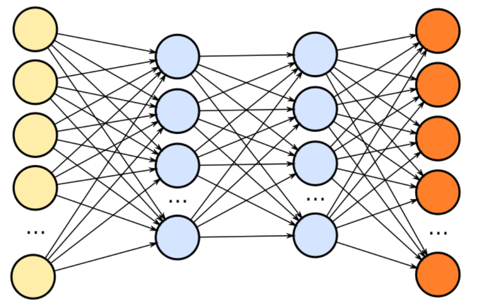
[Munos'03]

- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

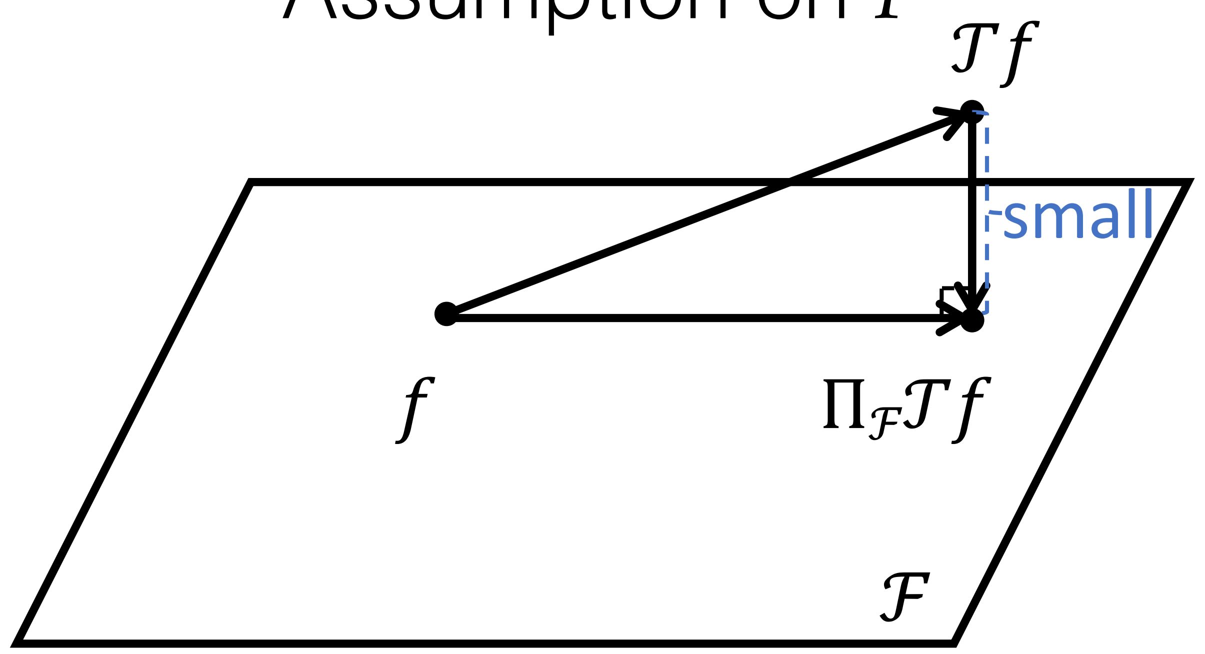




Similar to Jiang et al [2017]



Assumption on F



[Munos & Szepesvari '05]

- Conjecture: realizability alone is insufficient
- Alg-specific lower bound exists for decades
- *Info-theoretic*?
  - Negative results: two general proof styles excluded
  - e.g., construct an exponentially large MDP family => fail!

?

What we study: theory of batch RL (ADP)—backbone for "deep RL"

Setting: learn good policy from batch data $\{(s, a, r, s')\}$ + value-function approximator F (model Q^*)

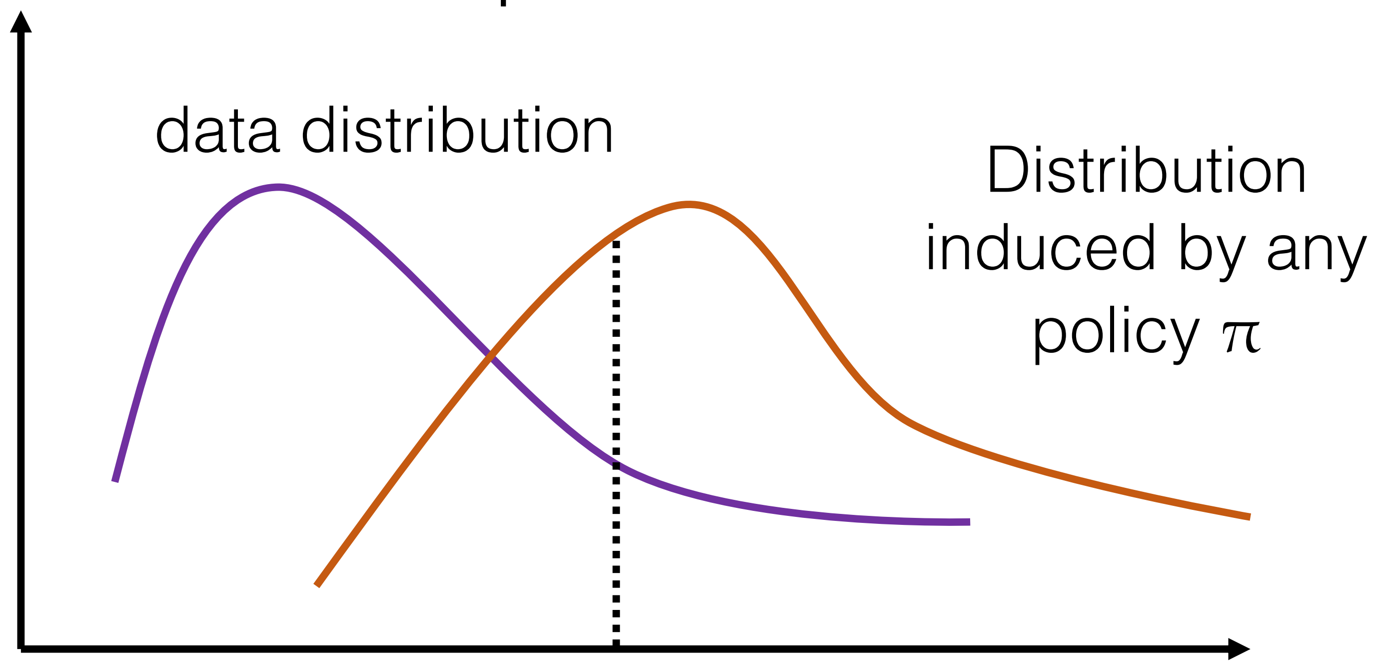
⇩

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?

Are they necessary? (hardness results)

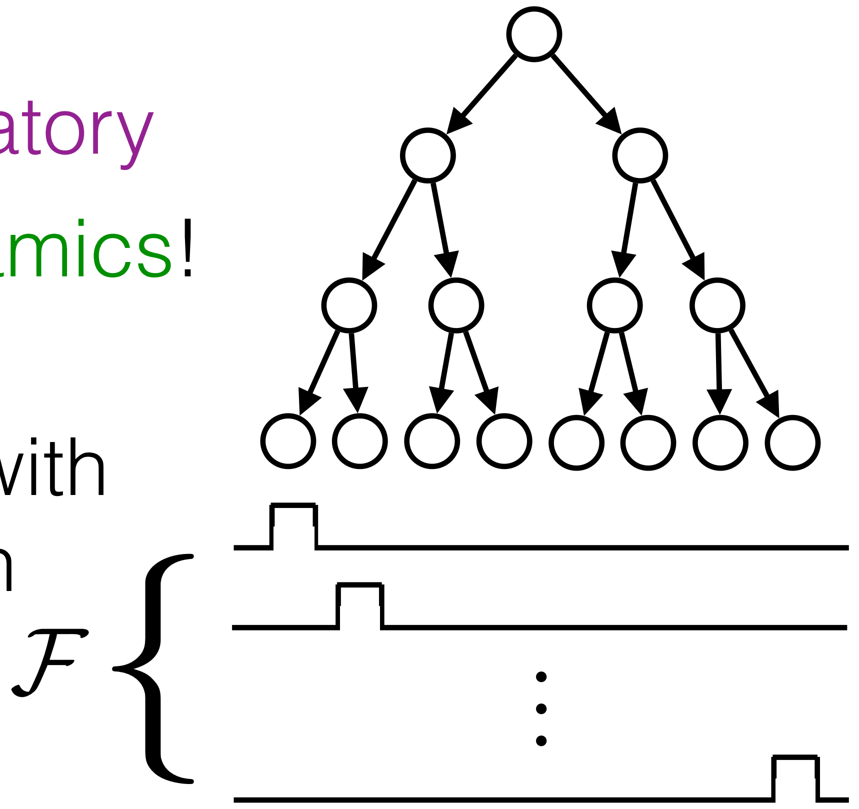
Do they hold in interesting scenarios?

Assumption on data

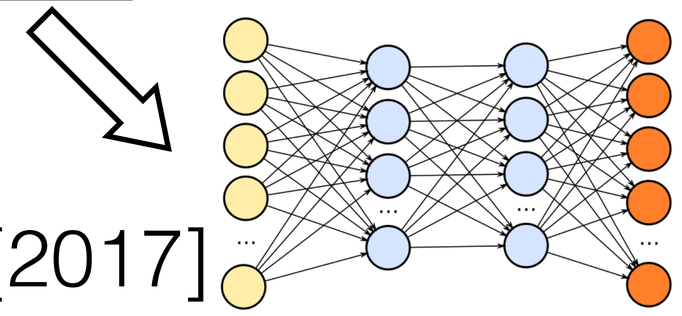


[Munos'03]

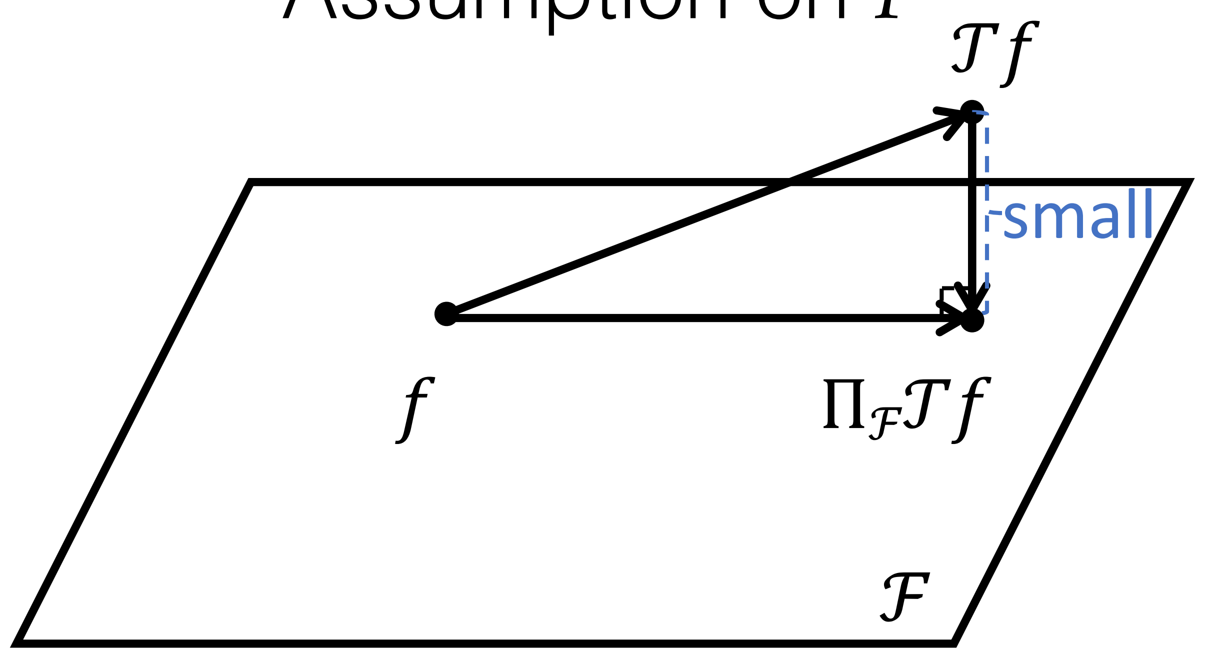
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution



Similar to Jiang et al [2017]



Assumption on F



[Munos & Szepesvari '05]

- Conjecture: realizability alone is insufficient
- Alg-specific lower bound exists for decades
- *Info-theoretic*?
  - Negative results: two general proof styles excluded
  - e.g., construct an exponentially large MDP family => fail!

?

F piece-wise constant

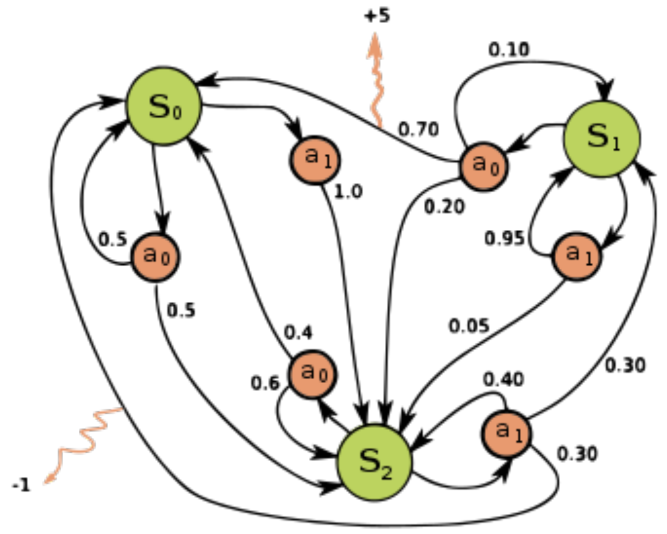
+

F closed under Bellman update

⇔ bisimulation [Givan et al'03]

Implications and the Bigger Picture

Tabular RL



Batch

RL with function approximation tractable

Nice dynamics & exploratory data
+ realizability + ???

Gap?

Gap?

Nice dynamics & exploratory data
+ realizability

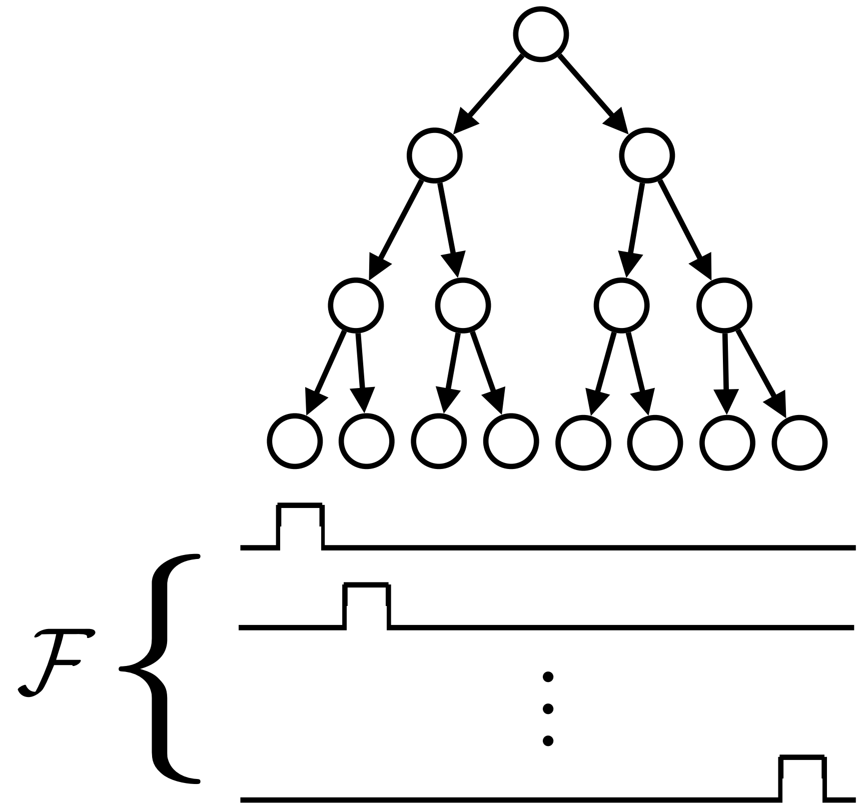
Online (exploration)

RL intractable

Nice dynamics
(low Bellman rank; Jiang et al'17)
+ realizability

Gap confirmed

Nice dynamics
(low witness rank; Sun et al'18)
+ realizability



Poster: Tue Evening
Pacific Ballroom #209

value-based

model-based