Non-asymptotic Statistical and Computational Guarantees of Reinforcement Learning Algorithms

Yuejie Chi

Carnegie Mellon University

2021 Goldsmith Lecture
Special thanks to...

Dean Andrea Goldsmith
Princeton University
My wonderful collaborators

Shicong Cen  
CMU

Chen Cheng  
Stanford

Gen Li  
Princeton

Yuxin Chen  
Princeton

Yuting Wei  
UPenn

Laixi Shi  
CMU

Changxiao Cai  
UPenn

Wenhao Zhan  
Princeton

Jason Lee  
Princeton

Yuantao Gu  
Tsinghua

Many thanks to Y. Wei and Y. Chen for significant contributions to the slides.
Reinforcement learning (RL)

In RL, an agent learns by interacting with an environment.

- unknown environments
- maximize total rewards
- trial-and-error
- sequential and online

"Recalculating ... recalculating ..."
Recent successes in RL

RL holds great promise in the next era of artificial intelligence.
Challenges of RL

- explore or exploit: unknown or changing environments
- credit assignment problem: delayed rewards or feedback
- enormous state and action space
- nonconcavity in value maximization
Sample efficiency

Collecting data samples might be expensive or time-consuming

- clinical trials
- autonomous driving
- online ads
Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials  autonomous driving  online ads

Calls for design of sample-efficient RL algorithms!
Running RL algorithms might take a long time and space

*many* CPUs / GPUs / TPUs + computing hours
Computational efficiency

Running RL algorithms might take a long time and space

many CPUs / GPUs / TPUs + computing hours

Calls for computationally efficient RL algorithms!
From asymptotic to non-asymptotic analyses

Non-asymptotic analyses are key to understand sample and computational efficiency in modern RL.
This tutorial

- **Part I: backgrounds and basics**
  - Markov decision processes
  - Planning

- **Part II: statistical guarantees under the generative model**
  - minimax lower bound
  - Is model-based RL minimax optimal?
  - Is Q-learning minimax optimal?

- **Part III: computational guarantees of policy optimization**
  - (natural) policy gradient methods
  - finite-time rate of global convergence
  - entropy regularization and beyond

- **Part IV: concluding remarks and further pointers**
Part I: backgrounds and basics
Markov decision process (MDP)

- $S$: state space
- $\mathcal{A}$: action space

- $s_t$: state
- $a_t \sim \pi(\cdot|s_t)$: action selection rule
- $r_t = r(s_t, a_t)$: immediate reward
- $s_{t+1} \sim P(\cdot|s_t, a_t)$: next state

Diagram:
- Agent
- Environment

[Diagram of a Markov decision process with states and actions labeled]
Markov decision process (MDP)

- $S$: state space
- $A$: action space
- $r(s, a) \in [0, 1]$: immediate reward

The diagram illustrates the interaction between the agent and the environment. The agent observes the current state $s_t$, selects an action $a_t$ according to the policy $\pi(\cdot | s_t)$, and receives an immediate reward $r_t = r(s_t, a_t)$. The environment transitions to the next state $s_{t+1}$ according to the transition probabilities $P(\cdot | s_t, a_t)$.
Markov decision process (MDP)

- $S$: state space
- $A$: action space
- $r(s_t, a_t) \in [0, 1]$: immediate reward
- $\pi(s_t)$: policy (or action selection rule)
Markov decision process (MDP)

- **S**: state space
- **A**: action space
- \( r(s, a) \in [0, 1] \): immediate reward
- \( \pi(\cdot|s) \): policy (or action selection rule)
- \( P(\cdot|s, a) \): transition probabilities
Value function of policy $\pi$:

$$V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right]$$
Value function of policy $\pi$:

$$\forall s \in S : \quad V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right]$$

- $\gamma \in [0, 1)$ is the discount factor; $\frac{1}{1-\gamma}$ is effective horizon
- Expectation is w.r.t. the sampled trajectory under $\pi$
Q-function of policy $\pi$:

$$\forall (s, a) \in S \times A : \quad Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]$$

- $(a_0, s_1, a_1, s_2, a_2, \cdots)$: generated under policy $\pi$
Goal: find the optimal policy $\pi^*$ that maximize $V^\pi(s)$

- optimal value / Q function: $V^* := V^{\pi^*}$, $Q^* := Q^{\pi^*}$
- optimal policy $\pi^*(s) = \arg\max_{a \in A} Q^*(s, a)$
Planning: when the model is known

**Planning**: find the optimal policy $\pi^*$ given MDP specification
Policy evaluation: Bellman’s consistency equation

- \( V^\pi \) / \( Q^\pi \): value / action-value function under policy \( \pi \)

Bellman’s consistency equation

\[
V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s, a)]
\]

\[
Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[V^\pi(s')]
\]

- one-step look-ahead
- Let \( P^\pi \) be the state-action transition matrix induced by \( \pi \):

\[
Q^\pi = r + \gamma P^\pi Q^\pi \quad \implies \quad Q^\pi = (I - \gamma P^\pi)^{-1} r
\]
Bellman’s optimality principle

Bellman operator

\[ \mathcal{T}(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(.|s, a)} \left[ \max_{a' \in A} Q(s', a') \right] \]

- immediate reward
- next state’s value

- one-step look-ahead
Bellman’s optimality principle

**Bellman operator**

\[
T(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \max_{a' \in A} Q(s', a') \right]
\]

- immediate reward
- next state’s value

• one-step look-ahead

**Bellman equation:** \( Q^* \) is *unique* solution to

\[
T(Q^*) = Q^*
\]

\( \gamma \)-contraction of Bellman operator:

\[
\|T(Q_1) - T(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty
\]

Richard Bellman
Value iteration (VI)

For $t = 0, 1, \ldots,$

$$Q^{(t+1)} = \mathcal{T}(Q^{(t)})$$
Value iteration and policy iteration

**Value iteration (VI)**

For \( t = 0, 1, \ldots, \)

\[
Q^{(t+1)} = \mathcal{T}(Q^{(t)})
\]

**Policy iteration (PI)**

For \( t = 0, 1, \ldots, \)

\[
\pi^{(t)} = \text{Greedy}(Q^{(t-1)})
\]

\[
Q^{(t)} = Q^{\pi^{(t)}}
\]
Proposition (Linear convergence of policy/value iteration)

\[ \| Q^{(t)} - Q^* \|_\infty \leq \gamma^t \| Q^{(0)} - Q^* \|_\infty \]

Implications: to achieve \( \| Q^{(t)} - Q^* \|_\infty \leq \epsilon \), it takes no more than \( \frac{1}{1 - \gamma} \log \left( \frac{\| Q^{(0)} - Q^* \|_\infty}{\epsilon} \right) \) iterations.

Linear convergence at a dimension-free rate!
Proposition (Linear convergence of policy/value iteration)

\[ \|Q^{(t)} - Q^*\|_\infty \leq \gamma^t \|Q^{(0)} - Q^*\|_\infty \]

Implications: to achieve \( \|Q^{(t)} - Q^*\|_\infty \leq \epsilon \), it takes no more than

\[ \frac{1}{1 - \gamma} \log \left( \frac{\|Q^{(0)} - Q^*\|_\infty}{\epsilon} \right) \]

iterations.
Proposition (Linear convergence of policy/value iteration)

\[ \|Q^{(t)} - Q^*\|_\infty \leq \gamma^t \|Q^{(0)} - Q^*\|_\infty \]

Implications: to achieve \( \|Q^{(t)} - Q^*\|_\infty \leq \epsilon \), it takes no more than

\[ \frac{1}{1 - \gamma} \log \left( \frac{\|Q^{(0)} - Q^*\|_\infty}{\epsilon} \right) \]

iterations.

Linear convergence at a \textbf{dimension-free} rate!
Part II: statistical guarantees under the generative model
Two approaches to RL

Model-based approach ("plug-in")
1. build an empirical estimate $\hat{P}$ for $P$
2. planning based on empirical $\hat{P}$
Two approaches to RL

Model-based approach ("plug-in")
1. build an empirical estimate $\hat{P}$ for $P$
2. planning based on empirical $\hat{P}$

Model-free approach
— learning w/o constructing model explicitly
For each state-action pair \((s, a)\), collect \(N\) samples

\[
\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}
\]
For each state-action pair \((s, a)\), collect \(N\) samples

\[
\{(s, a, s'_i)\}_{1 \leq i \leq N}
\]

**Question:** How many samples are necessary and sufficient to solve the RL problem without worrying about exploration?
Theorem (minimax lower bound; Azar et al., 2013)

For all $\epsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

$$\tilde{\Omega} \left( \frac{|S||A|}{(1-\gamma)^3 \epsilon^2} \right)$$

to achieve $\|\hat{Q} - Q^*\|_\infty \leq \epsilon$, where $\hat{Q}$ is the output of any RL algorithm.
Minimax lower bound

**Theorem (minimax lower bound; Azar et al., 2013)**

For all $\epsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

$$\tilde{\Omega}\left(\frac{|S||A|}{(1 - \gamma)^3 \epsilon^2}\right)$$

to achieve $\|\hat{Q} - Q^*\|_\infty \leq \epsilon$, where $\hat{Q}$ is the output of any RL algorithm.

- holds for both finding the optimal Q-function and the optimal policy over the entire range of $\epsilon$
- much smaller than the model dimension $|S|^2|A|$
Is model-based RL minimax optimal?

Model-based approach ("plug-in")
1. build an empirical estimate $\hat{P}$ for $P$
2. planning based on empirical $\hat{P}$

Model-free approach
— learning w/o constructing model explicitly
Model estimation under the generative model

For each \((s, a)\), collect \(N\) ind. samples \(\{(s, a, s'(i))\}_{1 \leq i \leq N}\)
Model estimation under the generative model

For each $(s, a)$, collect $N$ ind. samples $\{(s, a, s'_i)\}_{1 \leq i \leq N}$

**Empirical estimates:** estimate $\hat{P}(s'|s, a)$ by $\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{s'_i = s'\}$
Model-based (plug-in) estimator

— Azar et al., 2013; Agarwal et al., 2019

Run planning algorithms based on the *empirical* MDP
Challenges in the sample-starved regime

truth: \( P \in \mathbb{R}^{|S||A| \times |S|} \)

empirical estimate: \( \hat{P} \)

- Can’t recover \( P \) faithfully if sample size \( \ll |S|^2 |A|! \)
Challenges in the sample-starved regime

truth:
\[ P \in \mathbb{R}^{|S|\cdot|A| \times |S|} \]

empirical estimate:
\[ \hat{P} \]

- Can’t recover \( P \) faithfully if sample size \( \ll |S|^2|A|! \)
- Can we trust our policy estimate when reliable model estimation is infeasible?
Theorem (Azar et al., 2013)

For any $0 < \epsilon \leq 1$, the optimal $Q$-function $\hat{Q}$ of the empirical MDP achieves

$$\| \hat{Q} - Q^* \|_\infty \leq \epsilon$$

with sample complexity at most $\tilde{O} \left( \frac{|S||A|}{(1-\gamma)^3 \epsilon^2} \right)$.

- matches with the minimax lower bound whenever $\epsilon \in (0, 1]$. 
Sample complexity of the plug-in estimator

Theorem (Azar et al., 2013)

For any $0 < \epsilon \leq 1$, the optimal Q-function $\hat{Q}$ of the empirical MDP achieves

$$\|\hat{Q} - Q^*\|_\infty \leq \epsilon$$

with sample complexity at most $\tilde{O} \left( \frac{|S||A|}{(1-\gamma)^3 \epsilon^2} \right)$.

- matches with the minimax lower bound whenever $\epsilon \in (0, 1]$.
- **Question**: Does it imply a near minimax-optimal policy $\hat{\pi}$?
Proposition (Singh and Yee, 1994)

Let the greedy policy w.r.t. $\hat{Q}$ be $\hat{\pi}$, then

$$V^* - V^{\hat{\pi}} \leq \frac{2}{1 - \gamma} \|Q^* - \hat{Q}\|_\infty.$$
From Q-function to policy

Proposition (Singh and Yee, 1994)

Let the greedy policy w.r.t. $\hat{Q}$ be $\hat{\pi}$, then

$$ V^* - V^{\hat{\pi}} \leq \frac{2}{1 - \gamma} \| Q^* - \hat{Q} \|_{\infty}. $$

\[ \| \hat{Q} - Q^* \|_{\infty} \leq \epsilon \]

\[ \hat{\pi} = \text{Greedy}(\hat{Q}) \]

\[ V^* - V^{\hat{\pi}} \leq \frac{\epsilon}{1 - \gamma} \]
From Q-function to policy

**Proposition (Singh and Yee, 1994)**

Let the greedy policy w.r.t. $\hat{Q}$ be $\hat{\pi}$, then

$$V^* - V^{\hat{\pi}} \leq \frac{2}{1 - \gamma} \|Q^* - \hat{Q}\|_\infty.$$ 

This error amplification has consequences in sample complexities.

- To reach $\epsilon$-optimality, the greedy policy of a minimax-optimal Q-function estimator needs
  $$\tilde{\Omega} \left( \frac{|S||A|}{(1 - \gamma)^5 \epsilon^2} \right)$$
  samples invoking the above naive argument.
Sample complexity of the plug-in estimator

**Theorem (Agarwal et al., 2019)**

For any $0 < \epsilon \leq \frac{1}{\sqrt{1-\gamma}}$, the optimal policy $\hat{\pi}^*$ of the empirical MDP achieves

$$\|V_{\hat{\pi}^*} - V^*\|_{\infty} \leq \epsilon$$

with sample complexity at most $\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^3\epsilon^2}\right)$.

- matches with the minimax lower bound whenever
  $$\epsilon \in (0, \frac{1}{\sqrt{1-\gamma}}].$$
- requires a sample size of at least $\frac{|S||A|}{(1-\gamma)^2}$. 
A benchmark of the prior art

All prior theory requires sample size $\gtrsim |S||A| (1 - \gamma)^2$.

Is it possible to close the gap?
A benchmark of the prior art

All prior theory requires sample size \( \gtrsim |S| |A| (1 - \gamma)^2 \).

Is it possible to close the gap?
A benchmark of the prior art

All prior theory requires sample size $\gtrsim \frac{|S||A|}{(1-\gamma)^2}$
A benchmark of the prior art

All prior theory requires sample size \( \gtrsim \frac{|S||A|}{(1 - \gamma)^2} \)

Is it possible to close the gap?
Our method: a perturbed plug-in estimator

— Li, Wei, Chi, Gu, Chen, 2020

Run planning algorithms based on the empirical MDP with slightly perturbed rewards

\[ r_p(s, a) = r(s, a) + \zeta(s, a), \quad \zeta(s, a) \sim \text{Unif}(0, \xi). \]
Sample complexity of a perturbed plug-in estimator

**Theorem (Li, Wei, Chi, Gu, Chen, 2020)**

For any $0 < \epsilon \leq \frac{1}{1-\gamma}$, the optimal policy $\hat{\pi}_p^*$ of the perturbed empirical MDP with $\xi \approx \frac{(1-\gamma)\epsilon}{|S|^5 |A|^5}$ achieves

$$V^* - V^{\hat{\pi}_p^*} \leq \epsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^3 \epsilon^2}\right).$$
Sample complexity of a perturbed plug-in estimator

**Theorem (Li, Wei, Chi, Gu, Chen, 2020)**

For any $0 < \epsilon \leq \frac{1}{1-\gamma}$, the optimal policy $\hat{\pi}_p^*$ of the perturbed empirical MDP with $\xi \sim \frac{(1-\gamma)\epsilon}{|S|^5|A|^5}$ achieves

$$V^* - V^{\hat{\pi}_p^*} \leq \epsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^3\epsilon^2}\right).$$

- $\hat{\pi}_p^*$: obtained by empirical VI or PI within $\tilde{O}\left(\frac{1}{1-\gamma}\right)$ iterations
Theorem (Li, Wei, Chi, Gu, Chen, 2020)

For any $0 < \epsilon \leq \frac{1}{1 - \gamma}$, the optimal policy $\hat{\pi}_p^*$ of the perturbed empirical MDP with $\xi \approx \frac{(1 - \gamma)\epsilon}{|S|^5|A|^5}$ achieves

$$V^* - V^{\hat{\pi}_p^*} \leq \epsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|S||A|}{(1 - \gamma)^3\epsilon^2}\right).$$

- $\hat{\pi}_p^*$: obtained by empirical VI or PI within $\tilde{O}\left(\frac{1}{1 - \gamma}\right)$ iterations
- **Minimax lower bound**: $\tilde{\Omega}\left(\frac{|S||A|}{(1 - \gamma)^3\epsilon^2}\right)$ (Azar et al.’13)
Close the gap

sample complexity

\[
\frac{|S| \cdot |A|}{(1 - \gamma)^3}
\]

Sidford et al. '18b

\[
\frac{|S| \cdot |A|}{(1 - \gamma)^2}
\]

Agarwal et al. '19

\[
\frac{|S| \cdot |A|}{1 - \gamma}
\]

our work

minimax lower bound

\[
\frac{1}{\varepsilon^2}
\]
A glimpse of the analysis: notation

- \( V^\pi \): true value function under policy \( \pi \)
  - Bellman equation: \( V^\pi = (I - P_\pi)^{-1} r \)
A glimpse of the analysis: notation

- $V^\pi$: true value function under policy $\pi$
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1}r$

- $\hat{V}^\pi$: estimate of value function under policy $\pi$
  - Bellman equation: $\hat{V}^\pi = (I - \hat{P}_\pi)^{-1}r$
A glimpse of the analysis: notation

- $V^\pi$: true value function under policy $\pi$
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1}r$

- $\hat{V}^\pi$: estimate of value function under policy $\pi$
  - Bellman equation: $\hat{V}^\pi = (I - \hat{P}_\pi)^{-1}r$

- $\pi^*$: optimal policy w.r.t. true value function

- $\hat{\pi}^*$: optimal policy w.r.t. empirical value function
A glimpse of the analysis: notation

- $V^\pi$: true value function under policy $\pi$
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1}r$

- $\hat{V}^\pi$: estimate of value function under policy $\pi$
  - Bellman equation: $\hat{V}^\pi = (I - \hat{P}_\pi)^{-1}r$

- $\pi^*$: optimal policy w.r.t. true value function

- $\hat{\pi}^*$: optimal policy w.r.t. empirical value function

- $V^* := V^{\pi^*}$: optimal values under true models

- $\hat{V}^* := \hat{V}^{\hat{\pi}^*}$: optimal values under empirical models
Proof ideas

Elementary decomposition:

\[ V^* - V^\hat{\pi}^* = (V^* - \hat{V}^\pi^*) + (\hat{V}^\pi^* - \hat{V}\hat{\pi}^*) + (\hat{V}\hat{\pi}^* - V^\pi^*) \]
Proof ideas

Elementary decomposition:

\[ V^* - V_{\hat{\pi}^*} = (V^* - \hat{V}_{\pi}^*) + (\hat{V}_{\pi}^* - \hat{V}_{\hat{\pi}}^*) + (\hat{V}_{\hat{\pi}}^* - V_{\hat{\pi}}^*) \]
\[ \leq (V_{\pi}^* - \hat{V}_{\pi}^*) + 0 + (\hat{V}_{\hat{\pi}}^* - V_{\hat{\pi}}^*) \]
Proof ideas

Elementary decomposition:

\[ V^* - V_{\hat{\pi}}^* = (V^* - \hat{V}_{\pi}^*) + (\hat{V}_{\pi}^* - \hat{V}_{\hat{\pi}}^*) + (\hat{V}_{\hat{\pi}}^* - V_{\pi}^*) \leq (V_{\pi}^* - \hat{V}_{\pi}^*) + 0 + (\hat{V}_{\hat{\pi}}^* - V_{\hat{\pi}}^*) \]

• **Step 1:** control \( V_{\pi}^* - \hat{V}_{\pi}^* \) for a fixed \( \pi \)
  
  *(Bernstein inequality + high-order decomposition)*
Proof ideas

Elementary decomposition:

\[ V^* - V^{\hat{\pi}^*} = (V^* - \hat{V}^{\pi^*}) + (\hat{V}^{\pi^*} - \hat{V}^{\hat{\pi}^*}) + (\hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*}) \]

\[ \leq (V^{\pi^*} - \hat{V}^{\pi^*}) + 0 + (\hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*}) \]

• **Step 1:** control \( V^{\pi} - \hat{V}^{\pi} \) for a fixed \( \pi \)
  
  *(Bernstein inequality + high-order decomposition)*

• **Step 2:** extend it to control \( \hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*} \) (\( \hat{\pi}^* \) depends on samples)
  
  *(decouple statistical dependency)*
Step 1: improved theory for policy evaluation

Model-based policy evaluation:
— given a fixed policy $\pi$, estimate $V^\pi$ via the plug-in estimate $\hat{V}^\pi$
Step 1: improved theory for policy evaluation

Model-based policy evaluation:
— given a fixed policy \( \pi \), estimate \( V^\pi \) via the plug-in estimate \( \hat{V}^\pi \)

A sample size barrier \( \frac{|S|}{(1-\gamma)^2} \) already appeared in prior work
(Agarwal et al. '19, Pananjady & Wainwright '19, Khamaru et al. '20)
Step 1: improved theory for policy evaluation

Model-based policy evaluation:
— given a fixed policy $\pi$, estimate $V^\pi$ via the plug-in estimate $\hat{V}^\pi$

Theorem (Li, Wei, Chi, Gu, Chen, 2020)

Fix any policy $\pi$. For $0 < \epsilon \leq \frac{1}{1-\gamma}$, the plug-in estimator $\hat{V}^\pi$ obeys

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \epsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|S|}{(1-\gamma)^3\epsilon^2}\right)$$
Model-based policy evaluation:
— given a fixed policy $\pi$, estimate $V^\pi$ via the plug-in estimate $\hat{V}^\pi$

**Theorem (Li, Wei, Chi, Gu, Chen, 2020)**

*Fix any policy $\pi$. For $0 < \epsilon \leq \frac{1}{1-\gamma}$, the plug-in estimator $\hat{V}^\pi$ obeys

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \epsilon$$

*with sample complexity at most

$$\tilde{O}\left(\frac{|S|}{(1 - \gamma)^3\epsilon^2}\right)$$

• Minimax optimal for all $\epsilon$ (Azar et al. '13, Pananjady & Wainwright '19)
Key idea 1: a peeling argument

First-order expansion:

\[ \hat{V}^\pi - V^\pi = \gamma (I - \gamma P^\pi)^{-1} (\hat{P}^\pi - P^\pi) \hat{V}^\pi \]  

Higher-order expansion \( \rightarrow \) tighter control:

\[ \hat{V}^\pi - V^\pi = \gamma (I - \gamma P^\pi)^{-1} (\hat{P}^\pi - P^\pi) V^\pi + \]
Key idea 1: a peeling argument

First-order expansion:

\[ \hat{V}^\pi - V^\pi = \gamma (I - \gamma P_\pi)^{-1} (\hat{P}_\pi - P_\pi) \hat{V}^\pi \]  

Higher-order expansion \(\rightarrow\) tighter control:

\[ \hat{V}^\pi - V^\pi = \gamma (I - \gamma P_\pi)^{-1} (\hat{P}_\pi - P_\pi) V^\pi + \gamma (I - \gamma P_\pi)^{-1} (\hat{P}_\pi - P_\pi) (\hat{V}^\pi - V^\pi) \]
Key idea 1: a peeling argument

First-order expansion:

\[
\hat{V}^\pi - V^\pi = \gamma (I - \gamma P^\pi)^{-1} (\hat{P}_\pi - P^\pi) \hat{V}^\pi
\]  

Higher-order expansion \(\rightarrow\) tighter control:

\[
\hat{V}^\pi - V^\pi = \gamma (I - \gamma P^\pi)^{-1} (\hat{P}_\pi - P^\pi) V^\pi + \\
+ \gamma^2 \left( (I - \gamma P^\pi)^{-1} (\hat{P}_\pi - P^\pi) \right)^2 V^\pi \\
+ \gamma^3 \left( (I - \gamma P^\pi)^{-1} (\hat{P}_\pi - P^\pi) \right)^3 V^\pi \\
+ \ldots
\]
Step 2: controlling $\hat{V}^\pi - V^\pi$

A natural idea: apply our policy evaluation theory + union bound
Step 2: controlling $\hat{V}^{\pi^*} - V^{\pi^*}$

A natural idea: apply our policy evaluation theory + union bound

• highly suboptimal! (there are exponentially many policies)
Key idea 2: leave-one-out analysis

Decouple dependency by introducing auxiliary state-action absorbing MDPs by dropping randomness for each \((s, a)\)

--- inspired by (Agarwal et al. 2019) but quite different . . .

Other leave-one-out analysis: (El Karoui, 2015; Javanmard, Montanari, 2015; Abbe et al., 2017; Zhong, Boumal, 2017; Ma et al., 2017; Pananjady, Wainwright, 2019)
Is model-free RL minimax optimal?

Model-based approach ("plug-in")
1. build an empirical estimate $\hat{P}$ for $P$
2. planning based on empirical $\hat{P}$

Model-free approach
— learning w/o modeling & estimating environment explicitly
Q-learning: a classical model-free algorithm

Stochastic approximation for solving the Bellman equation

Robbins & Monro, 1951

\[ Q = \mathcal{T}(Q) \]

where

\[ \mathcal{T}(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a), a' \in \mathcal{A}} \left[ \max_{a' \in \mathcal{A}} Q(s', a') \right]. \]
Q-learning: a classical model-free algorithm

Stochastic approximation for solving Bellman equation $Q = T(Q)$

$$Q_{t+1}(s, a) = (1 - \eta_t)Q_t(s, a) + \eta_t T_t(Q_t)(s, a), \quad t \geq 0$$

draw the transition $(s, a, s')$ for all $(s, a)$
Q-learning: a classical model-free algorithm

Stochastic approximation for solving Bellman equation \( Q = \mathcal{T}(Q) \)

\[
Q_{t+1}(s, a) = (1 - \eta_t)Q_t(s, a) + \eta_t \mathcal{T}_t(Q_t)(s, a), \quad t \geq 0
\]

draw the transition \((s, a, s')\) for all \((s, a)\)

\[
\mathcal{T}_t(Q)(s, a) = r(s, a) + \gamma \max_{a'} Q(s', a')
\]

\[
\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \max_{a'} Q(s', a') \right]
\]
Prior art: achievability

**Question:** How many samples are needed for $\|\hat{Q} - Q^*\|_{\infty} \leq \epsilon$?
Prior art: achievability

**Question:** How many samples are needed for \( \| \hat{Q} - Q^* \|_\infty \leq \epsilon \)?

<table>
<thead>
<tr>
<th>paper</th>
<th>sample complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even-Dar &amp; Mansour '03</td>
<td>( 2^{1-\gamma} \frac{</td>
</tr>
<tr>
<td>Beck &amp; Srikant '12</td>
<td>( \frac{</td>
</tr>
<tr>
<td>Wainwright '19</td>
<td>( \frac{</td>
</tr>
<tr>
<td>Chen et al. '20</td>
<td>( \frac{</td>
</tr>
</tbody>
</table>

All prior results require sample size of at least \( \frac{|S||A|}{(1-\gamma)^5 \epsilon^2} \)!
**Question:** How many samples are needed to ensure \(|\|\hat{Q} - Q^*\|_\infty \leq \epsilon|\)?

<table>
<thead>
<tr>
<th>paper</th>
<th>sample complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even-Dar &amp; Mansour '03</td>
<td>(2^{1-\gamma} \frac{</td>
</tr>
<tr>
<td>Beck &amp; Srikant '12</td>
<td>(\frac{</td>
</tr>
<tr>
<td>Wainwright '19</td>
<td>(\frac{</td>
</tr>
<tr>
<td>Chen et al. '20</td>
<td>(\frac{</td>
</tr>
</tbody>
</table>

All prior results require sample size of at least \(\frac{|S||A|}{(1-\gamma)^5 \epsilon^2}\)!

*Is Q-learning sub-optimal, or is it an analysis artifact?*
A sharpened sample complexity of Q-learning

**Theorem (Li, Cai, Chen, Gu, Wei, Chi, 2021)**

For any $0 < \epsilon \leq 1$, Q-learning yields

$$\|\hat{Q} - Q^*\|_\infty \leq \epsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|S||A|}{(1 - \gamma)^4\epsilon^2}\right).$$

- Improves dependency on effective horizon $\frac{1}{1 - \gamma}$
A sharpened sample complexity of Q-learning

**Theorem (Li, Cai, Chen, Gu, Wei, Chi, 2021)**

For any $0 < \epsilon \leq 1$, Q-learning yields

$$\|\hat{Q} - Q^*\|_\infty \leq \epsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^4\epsilon^2}\right).$$

- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Allows both constant and rescaled linear learning rate:

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}}$$
A curious numerical example

Numerical evidence: $\frac{|S||A|}{(1-\gamma)^4 \epsilon^2}$ samples seem necessary …

— observed in Wainwright ’19

$p = \frac{4\gamma - 1}{3\gamma}$

$r(0, 1) = 0, \quad r(1, 1) = r(1, 2) = 1$
Q-learning is not minimax optimal

**Theorem (Li, Cai, Chen, Gu, Wei, Chi, 2021)**

For any $0 < \epsilon \leq 1$, there exists an MDP such that to achieve $\|\hat{Q} - Q^*\|_\infty \leq \epsilon$, Q-learning needs at least a sample complexity of

$$\tilde{\Omega}\left(\frac{|S||A|}{(1 - \gamma)^4\epsilon^2}\right).$$

- Tight algorithm-dependent lower bound
- Holds for both constant and rescaled linear learning rates
Where we stand now

Q-learning requires a sample size of \( \frac{|S| |A|}{(1-\gamma)^4 \epsilon^2} \).
Why is Q-learning sub-optimal?

Over-estimation of Q-functions \(\text{(Thrun and Schwartz, 1993; Hasselt, 2010)}\):

- \(\max_{a \in A} \mathbb{E} X(a)\) tends to be over-estimated (high positive bias) when \(\mathbb{E} X(a)\) is replaced by its empirical estimates using a small sample size;
- often gets worse with a large number of actions \(\text{(Hasselt, Guez, Silver, 2015)}\).

Figure 1: The orange bars show the bias in a single Q-learning update when the action values are \(Q(s, a) = V_*(s) + \epsilon_a\) and the errors \(\{\epsilon_a\}_{a=1}^m\) are independent standard normal random variables. The second set of action values \(Q'\), used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.
Why is Q-learning sub-optimal?

Over-estimation of Q-functions (Thrun and Schwartz, 1993; Hasselt, 2010):

- $\max_{a \in A} \mathbb{E} X(a)$ tends to be over-estimated (high positive bias) when $\mathbb{E} X(a)$ is replaced by its empirical estimates using a small sample size;
- often gets worse with a large number of actions (Hasselt, Guez, Silver, 2015).

A provable fix: Q-learning with variance reduction (Wainwright 2019) is provably minimax optimal.
Part III: policy optimization
Policy optimization

\[ \text{maximize}_{\theta} \ \text{value}(\text{policy}(\theta)) \]

- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.
Theoretical challenges: non-concavity

Little understanding on the global convergence of policy gradient methods until very recently, e.g. (Fazel et al., 2018; Bhandari and Russo, 2019; Agarwal et al., 2019; Mei et al. 2020), and many more.

Our goal:

- understand finite-time convergence rates of popular heuristics;
- design fast-convergent algorithms that scale for finding policies with desirable properties.
Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy $\pi$ such that

$$\maximize_\pi \quad V^\pi(\rho) := \mathbb{E}_{s \sim \rho}[V^\pi(s)]$$
Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy $\pi$ such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]$$

softmax parameterization:

$$\pi_\theta(a|s) \propto \exp(\theta(s, a))$$
Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy $\pi$ such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]$$

softmax parameterization:

$$\pi_\theta(a|s) \propto \exp(\theta(s, a))$$

$$\text{maximize}_\theta \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s \sim \rho} [V^{\pi_\theta}(s)]$$
Policy gradient methods

Given an initial state distribution \( s \sim \rho \), find policy \( \pi \) such that

\[
\max_{\pi} \quad V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]
\]

softmax parameterization:

\[
\pi_\theta(a|s) \propto \exp(\theta(s, a))
\]

\[
\max_\theta \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s \sim \rho} [V^{\pi_\theta}(s)]
\]

**Policy gradient method (Sutton et al., 2000)**

For \( t = 0, 1, \ldots \)

\[
\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)
\]

where \( \eta \) is the learning rate.
Global convergence of the PG method?

- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
Global convergence of the PG method?

- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

- (Mei et al., 2020) Softmax PG converges to global opt in $O\left(\frac{1}{\epsilon}\right)$ iterations.
Global convergence of the PG method?

- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

- (Mei et al., 2020) Softmax PG converges to global opt in 
  \[ c(|S|, |A|, \frac{1}{1-\gamma}, \ldots) O\left(\frac{1}{\epsilon}\right) \text{ iterations} \]
Global convergence of the PG method?

- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

- (Mei et al., 2020) Softmax PG converges to global opt in

\[ c(|S|, |A|, \frac{1}{1-\gamma}, \cdots) O\left(\frac{1}{\epsilon}\right) \] iterations

Is the rate of PG good, bad or ugly?
A negative message

**Theorem (Li, Wei, Chi, Gu, Chen, 2021)**

There exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta} |S|^{2^{\Theta\left(\frac{1}{1-\gamma}\right)}} \text{ iterations}$$

to achieve $$\|V^{(t)} - V^*\|_{\infty} \leq 0.15.$$
A negative message

**Theorem (Li, Wei, Chi, Gu, Chen, 2021)**

There exists an MDP s.t. it takes softmax PG at least

\[
\frac{1}{\eta} |S|^2 \Theta \left( \frac{1}{1-\gamma} \right) \text{ iterations}
\]

to achieve \( \| V(t) - V^* \|_\infty \leq 0.15 \).

- Softmax PG can take (super)-exponential time to converge (in problems w/ large state space & long effective horizon)!

- Also hold for average sub-opt gap \( \frac{1}{|S|} \sum_{s \in S} [V(t)(s) - V^*(s)] \).
MDP construction for our lower bound

Key ingredients:
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,
Key ingredients: for $3 \leq s \leq H \approx \frac{1}{1-\gamma}$,

- $\pi(t)(a_{opt} \mid s)$ keeps decreasing until $\pi(t)(a_{opt} \mid s-2) \approx 1$
What is happening in our constructed MDP?

\[
\max \ V(s) = \mathbb{E}_{s \sim \pi} V(s)
\]

\[
\max \ V(s) \mid s \in S
\]

\[
\pi(t)(a_1 \mid 1)
\]

Does policy gradient (PG) method converge?

• (Agarwal et al. '19) Softmax PG converges as

• (Mei et al. '20) Softmax PG converges to global solution in \(|S|, |A|, \ldots\)

However, "asymptotic convergence" might mean "takes forever"
What is happening in our constructed MDP?

\[
\max \ V(s) : = \mathbb{E}_{s \sim \pi} V(s) \ 
\]

\[
\max \ V(\pi(s)): = \mathbb{E}_{s \sim \pi} V(\pi(s)) \ 
\]

\[
\pi(a|s) = \exp(\theta(s,a)) P_a \exp(\theta(s,a)) \ 
\]

Convergence time for state \( s \) grows geometrically as \( s \) increases:

\[
\text{convergence-time}(s) \gtrsim \text{convergence-time}(s - 2)^1.5 \ 
\]
What is happening in our constructed MDP?

Convergence time for state $s$ grows geometrically as $s$ increases
What is happening in our constructed MDP?

Convergence time for state $s$ grows geometrically as $s$ increases

$$\text{convergence-time}(s) \gtrsim \left( \text{convergence-time}(s - 2) \right)^{1.5}$$
“Seriously, lady, at this hour you’d make a lot better time taking the subway.”
Booster #1: natural policy gradient

Natural policy gradient (NPG) method (Kakade, 2002)

For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^{\theta})^\dagger \nabla_{\theta} V_{\pi_{\theta}}^{(t)}(\rho)$$

where $\eta$ is the learning rate and $\mathcal{F}_\rho^{\theta}$ is the Fisher information matrix:

$$\mathcal{F}_\rho^{\theta} := \mathbb{E} \left[ (\nabla_{\theta} \log \pi_{\theta}(a|s)) (\nabla_{\theta} \log \pi_{\theta}(a|s))^\top \right].$$
Natural policy gradient (NPG) method (Kakade, 2002)

For \( t = 0, 1, \ldots \)

\[
\theta^{(t+1)} = \theta^{(t)} + \eta (F^\theta_\rho)^\dagger \nabla_{\theta} V^{\pi^{(t)}_\theta}(\rho)
\]

where \( \eta \) is the learning rate and \( F^\theta_\rho \) is the \textit{Fisher information matrix}:

\[
F^\theta_\rho := \mathbb{E} \left[ (\nabla_{\theta} \log \pi_\theta(a|s)) (\nabla_{\theta} \log \pi_\theta(a|s))^\top \right].
\]

In fact, popular heuristic TRPO (Schulman et al., 2015) = NPG + line search.
NPG in the tabular setting

Natural policy gradient (NPG) method (Tabular setting)

For $t = 0, 1, \cdots$, NPG updates the policy via

$$
\pi^{(t+1)}(\cdot|s) \propto \pi^{(t)}(\cdot|s) \exp \left( \frac{\eta Q^{(t)}(s, \cdot)}{1 - \gamma} \right)
$$

where $Q^{(t)} := Q^{\pi^{(t)}}$ is the Q-function of $\pi^{(t)}$, and $\eta > 0$.

- invariant with the choice of $\rho$
- Reduces to policy iteration (PI) when $\eta = \infty$. 
Global convergence of NPG

**Theorem (Agarwal et al., 2019)**

Set $\pi^{(0)}$ as a uniform policy. For all $t \geq 0$, we have

$$V^{(t)}(\rho) \geq V^*(\rho) - \left(\frac{\log |A|}{\eta} + \frac{1}{(1 - \gamma)^2}\right) \frac{1}{t}.$$
Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ as a uniform policy. For all $t \geq 0$, we have

$$V^{(t)}(\rho) \geq V^*(\rho) - \left( \frac{\log |A|}{\eta} + \frac{1}{(1 - \gamma)^2} \right) \frac{1}{t}.$$ 

Implication: set $\eta \geq (1 - \gamma)^2 \log |A|$, we find an $\epsilon$-optimal policy within at most

$$\frac{2}{(1 - \gamma)^2 \epsilon}$$ iterations.
Global convergence of NPG

**Theorem (Agarwal et al., 2019)**

Set $\pi^{(0)}$ as a uniform policy. For all $t \geq 0$, we have

$$V^{(t)}(\rho) \geq V^*(\rho) - \left( \frac{\log |\mathcal{A}|}{\eta} + \frac{1}{(1 - \gamma)^2} \right) \frac{1}{t}.$$

**Implication:** set $\eta \geq (1 - \gamma)^2 \log |\mathcal{A}|$, we find an $\epsilon$-optimal policy within at most

$$\frac{2}{(1 - \gamma)^2 \epsilon}$$

iterations.

Global convergence at a sublinear rate independent of $|S|, |\mathcal{A}|$!
Booster #2: entropy regularization

To encourage exploration, promote the stochasticity of the policy using the "soft" value function (Williams and Peng, 1991):

$$\forall s \in S : \quad V^\pi_\tau(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t (r_t + \tau \mathcal{H}(\pi(\cdot|s_t))) \mid s_0 = s \right]$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.
Booster #2: entropy regularization

To encourage exploration, promote the stochasticity of the policy using the "soft" value function (Williams and Peng, 1991):

\[
\forall s \in S : \quad V_\tau^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t (r_t + \tau \mathcal{H}(\pi(\cdot | s_t))) \mid s_0 = s \right]
\]

where \( \mathcal{H} \) is the Shannon entropy, and \( \tau \geq 0 \) is the reg. parameter.

\[
\text{maximize}_\theta \quad V_\tau^{\pi_\theta}(\rho) := \mathbb{E}_{s \sim \rho} [V_\tau^{\pi_\theta}(s)]
\]
Entropy-regularized natural gradient helps!

**Toy example:** a bandit with 3 arms of rewards 1, 0.9 and 0.1.
Entropy-regularized natural gradient helps!

**Toy example:** a bandit with 3 arms of rewards 1, 0.9 and 0.1.

Can we justify the efficacy of entropy-regularized NPG?
Entropy-regularized NPG in the tabular setting

For $t = 0, 1, \cdots$, the policy is updated via

$$\pi^{(t+1)}(\cdot|s) \propto \pi^{(t)}(\cdot|s) \left(1 - \frac{\eta \tau}{1-\gamma}\right) \exp\left(\frac{Q^{(t)}_\tau(s, \cdot)}{\tau}\right)$$

where $Q^{(t)}_\tau := Q^{(t)}_\pi$ is the soft Q-function of $\pi^{(t)}$, and $0 < \eta \leq \frac{1-\gamma}{\tau}$.

- invariant with the choice of $\rho$
- Reduces to soft policy iteration (SPI) when $\eta = \frac{1-\gamma}{\tau}$. 
Exact oracle: perfect evaluation of $Q_{\tau}^{\pi(t)}$ given $\pi(t)$;

**Theorem (Cen, Cheng, Chen, Wei, Chi, 2020)**

For any learning rate $0 < \eta \leq (1 - \gamma)/\tau$, the entropy-regularized NPG updates satisfy

- **Linear convergence of soft Q-functions:**

  \[
  \|Q_{\tau}^{*} - Q_{\tau}^{(t+1)}\|_{\infty} \leq C_1 \gamma (1 - \eta \tau)^t
  \]

  for all $t \geq 0$, where $Q_{\tau}^{*}$ is the optimal soft Q-function, and

  \[
  C_1 = \|Q_{\tau}^{*} - Q_{\tau}^{(0)}\|_{\infty} + 2\tau \left(1 - \frac{\eta \tau}{1 - \gamma}\right) \|\log \pi_{\tau}^{*} - \log \pi^{(0)}\|_{\infty}.
  \]
Implications

To reach $\|Q^{*}_\tau - Q^{(t+1)}_\tau\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates** ($0 < \eta < \frac{1-\gamma}{\tau}$):

  $$\frac{1}{\eta \tau} \log \left( \frac{C_1 \gamma}{\epsilon} \right)$$

- **Soft policy iteration** ($\eta = \frac{1-\gamma}{\tau}$):

  $$\frac{1}{1 - \gamma} \log \left( \frac{\|Q^{*}_\tau - Q^{(0)}_\tau\|_\infty \gamma}{\epsilon} \right)$$
Implications

To reach $\|Q^*_\tau - Q^{(t+1)}_\tau\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates** ($0 < \eta < \frac{1-\gamma}{\tau}$):
  $$\frac{1}{\eta \tau} \log \left( \frac{C_1 \gamma}{\epsilon} \right)$$

- **Soft policy iteration** ($\eta = \frac{1-\gamma}{\tau}$):
  $$\frac{1}{1-\gamma} \log \left( \frac{\|Q^*_\tau - Q^{(0)}_\tau\|_\infty \gamma}{\epsilon} \right)$$

Global linear convergence of entropy-regularized NPG at a rate independent of $|S|$, $|A|$!
Comparisons with entropy-regularized PG

(Mei et al., 2020) showed entropy-regularized PG achieves

\[ V^*_\tau(\rho) - V^{(t)}_\tau(\rho) \leq \left( V^*_\tau(\rho) - V^{(0)}_\tau(\rho) \right) \]

\[ \cdot \exp \left( -\frac{(1 - \gamma)^4t}{(8/\tau + 4 + 8 \log |A|)|S|} \left\| \frac{d\pi^*_\tau}{\rho} \right\|^{-1} \min_s \rho(s) \left( \inf_{0 \leq k \leq t - 1} \min_{s,a} \pi^{(k)}(a|s) \right)^2 \right) \]

Much faster convergence of entropy-regularized NPG
at a **dimension-free** rate!
Comparison with unregularized NPG

Regularized NPG
\( \tau = 0.001 \)

Vanilla NPG
\( \tau = 0 \)

Linear rate: \( \frac{1}{\eta \tau} \log \left( \frac{1}{\epsilon} \right) \)
Ours

Sublinear rate: \( \frac{1}{\min\{\eta, (1-\gamma)^2\}} \epsilon \)
(Agarwal et al. 2019)
Comparison with unregularized NPG

Regularized NPG
\( \tau = 0.001 \)

Vanilla NPG
\( \tau = 0 \)

Linear rate: \( \frac{1}{\eta \tau} \log \left( \frac{1}{\epsilon} \right) \)

Ours

Sublinear rate: \( \frac{1}{\min\{\eta, (1-\gamma)^2\}} \epsilon \)

(Agarwal et al. 2019)

Entropy regularization enables fast convergence!
**Inexact oracle:** inexact evaluation of $Q^\pi_t$ given $\pi^t$, which returns $\hat{Q}^t_T$ that

$$\|\hat{Q}^t_T - Q^t_T\|_{\infty} \leq \delta,$$

e.g., using sample-based estimators (Williams, 1992).
Entropy-regularized NPG with inexact gradients

**Inexact oracle:** inexact evaluation of $Q^{\pi(t)}_\tau$ given $\pi(t)$, which returns $\hat{Q}^{(t)}_\tau$ that

$$
\| \hat{Q}^{(t)}_\tau - Q^{(t)}_\tau \|_{\infty} \leq \delta,
$$
e.g., using sample-based estimators (Williams, 1992).

**Inexact entropy-regularized NPG:**

$$
\pi^{(t+1)}(a|s) \propto (\pi^{(t)}(a|s))^{1-\frac{\gamma \tau}{1-\gamma}} \exp \left( \frac{\eta \hat{Q}^{(t)}_\tau(s, a)}{1 - \gamma} \right)
$$
Entropy-regularized NPG with inexact gradients

**Inexact oracle:** inexact evaluation of $Q_\pi^{(t)}$ given $\pi^{(t)}$, which returns $\hat{Q}_\tau^{(t)}$ that

$$\|\hat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty \leq \delta,$$

e.g., using sample-based estimators (Williams, 1992).

**Inexact entropy-regularized NPG:**

$$\pi^{(t+1)}(a|s) \propto (\pi^{(t)}(a|s))^{1-\frac{\eta_\tau}{1-\gamma}} \exp\left(\frac{\eta\hat{Q}_\tau^{(t)}(s, a)}{1-\gamma}\right)$$

**Question:** Robustness of entropy-regularized NPG?
Linear convergence with inexact gradients

**Theorem (Cen, Cheng, Chen, Wei, Chi ’20; improved)**

For any learning rate $0 < \eta \leq (1 - \gamma)/\tau$, the entropy-regularized NPG updates achieve the same iteration complexity as the exact case, as long as

\[
\delta \leq \frac{1 - \gamma}{\gamma} \cdot \min \left\{ \frac{\epsilon}{4}, \sqrt{\frac{\epsilon \tau}{2}} \right\}
\]
Theorem (Cen, Cheng, Chen, Wei, Chi ’20; improved)

For any learning rate $0 < \eta \leq (1 - \gamma) / \tau$, the entropy-regularized NPG updates achieve the same iteration complexity as the exact case, as long as

$$\delta \leq \frac{1 - \gamma}{\gamma} \cdot \min \left\{ \frac{\epsilon}{4}, \sqrt{\frac{\epsilon \tau}{2}} \right\}$$

- **Intuition:** assume $\tau = O(\epsilon)$, the per-iteration policy evaluation error is no larger than

$$\frac{\text{final error}}{\text{iteration complexity}} = \frac{\epsilon}{\tilde{O}( (1 - \gamma)^{-1} )} \approx (1 - \gamma) \epsilon.$$
Aside: statistical implication

**Question:** how many samples are sufficient to find an $\epsilon$-optimal policy of the unregularized MDP?
Aside: statistical implication

**Question:** how many samples are sufficient to find an $\epsilon$-optimal policy of the unregularized MDP?

**Recipe:**
- set $\tau = \frac{(1-\gamma)\epsilon}{\log |A|}$;
- use fresh samples for policy evaluation with a targeted accuracy $\delta \simeq \frac{(1-\gamma)^{1.5} \epsilon}{\gamma \sqrt{\log |A|}}$, e.g. using model-based plug-in estimators (Li et al., 2020).
Aside: statistical implication

**Question:** how many samples are sufficient to find an $\epsilon$-optimal policy of the unregularized MDP?

**Recipe:**
- set $\tau = \frac{(1-\gamma)\epsilon}{\log|A|}$;
- use fresh samples for policy evaluation with a targeted accuracy $\delta \approx \frac{(1-\gamma)^{1.5}\epsilon}{\gamma\sqrt{\log|A|}}$, e.g. using model-based plug-in estimators (Li et al., 2020).

**A crude answer:**

$$\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^7\epsilon^2}\right)$$ samples
A key lemma: monotonic performance improvement

\[
V^{(t+1)}_\tau(\rho) - V^{(t)}_\tau(\rho) = \mathbb{E}_{s \sim d^{(t+1)}_\rho} \left[ \left( \frac{1}{\eta} - \frac{\tau}{1 - \gamma} \right) \text{KL}(\pi^{(t+1)}(\cdot|s) \| \pi^{(t)}(\cdot|s)) + \frac{1}{\eta} \text{KL}(\pi^{(t)}(\cdot|s) \| \pi^{(t+1)}(\cdot|s)) \right]
\]

Implication: monotonic improvement of \(V^{(t)}_\tau(s)\) and \(Q^{(t)}_\tau(s,a)\).
A key lemma: monotonic performance improvement

\[ V^{(t+1)}(\rho) - V^{(t)}(\rho) = \mathbb{E}_{s \sim d^{(t+1)}_\rho} \left[ \left( \frac{1}{\eta} - \frac{\tau}{1 - \gamma} \right) KL \left( \pi^{(t+1)}(\cdot|s) \parallel \pi^{(t)}(\cdot|s) \right) \right. \]

\[ \left. + \frac{1}{\eta} KL \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right] \]

**Implication:** monotonic improvement of \( V_\tau(s) \) and \( Q_\tau(s, a) \).
A key operator: soft Bellman operator

**Soft Bellman operator**

\[ T_\tau(Q)(s, a) := \begin{cases} r(s, a) \\ \text{immediate reward} \end{cases} \]

\[ + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \max_{\pi(\cdot | s')} \mathbb{E}_{a' \sim \pi(\cdot | s')} \left[ Q(s', a') - \tau \log \pi(a' | s') \right] \right], \]

next state's value

entropy
A key operator: soft Bellman operator

Soft Bellman operator

\[ T_\tau(Q)(s, a) := r(s, a) \]

immediate reward

\[ + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \max_{\pi(\cdot | s')} \mathbb{E}_{a' \sim \pi(\cdot | s')} \left[ Q(s', a') - \tau \log \pi(a' | s') \right] \],

next state’s value

entropy

Soft Bellman equation: \( Q^*_\tau \) is unique solution to

\[ T_\tau(Q^*_\tau) = Q^*_\tau \]

\( \gamma \)-contraction of soft Bellman operator:

\[ \| T_\tau(Q_1) - T_\tau(Q_2) \|_\infty \leq \gamma \| Q_1 - Q_2 \|_\infty \]
Analysis of soft policy iteration \( (\eta = \frac{1-\gamma}{\tau}) \)

Policy iteration

\[ \pi^{(0)} \rightarrow Q_{\pi^{(0)}} \rightarrow \text{evaluate} \rightarrow Q_{\pi^{(0)}} \rightarrow \text{greedy} \rightarrow Q_{\pi^{(1)}} \rightarrow \text{evaluate} \rightarrow Q_{\pi^{(1)}} \rightarrow \text{greedy} \rightarrow Q_{\pi^{(2)}} \rightarrow \cdots \rightarrow Q_{\pi^*} \rightarrow \pi^* \]

Bellman operator
Analysis of soft policy iteration \( (\eta = \frac{1-\gamma}{\tau}) \)

**Policy iteration**

\[
\begin{align*}
\pi^{(0)} & \xrightarrow{\text{evaluate}} Q^{\pi^{(0)}} \\
\pi^{(1)} & \xrightarrow{\text{evaluate}} Q^{\pi^{(1)}} \\
\pi^{(2)} & \xrightarrow{\text{evaluate}} Q^{\pi^{(2)}} \\
& \vdots \\
\pi^* & \xrightarrow{} Q^* 
\end{align*}
\]

**Bellman operator**

**Soft policy iteration**

\[
\begin{align*}
\pi^{(0)} & \xrightarrow{\text{evaluate}} Q^{\pi^{(0)}} \\
\pi^{(1)} & \xrightarrow{\text{soft greedy}} Q^{\pi^{(1)}} \\
\pi^{(2)} & \xrightarrow{\text{soft greedy}} Q^{\pi^{(2)}} \\
& \vdots \\
\pi^*_\tau & \xrightarrow{} Q^*_\tau 
\end{align*}
\]

**Soft Bellman operator**
A key linear system: general learning rates

Let \( x_t := \begin{bmatrix} \| Q^{*}_T - Q^{(t)}_T \|_{\infty} \\ \| Q^{*}_T - \tau \log \xi^{(t)} \|_{\infty} \end{bmatrix} \) and \( y := \begin{bmatrix} \| Q^{(0)}_T - \tau \log \xi^{(0)} \|_\infty \\ 0 \end{bmatrix} \), where \( \xi^{(t)} \propto \pi^{(t)} \) is an auxiliary sequence, then
A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \|Q^*_\tau - Q^{(t)}_\tau\|_{\infty} \\ \|Q^*_\tau - \tau \log \xi^{(t)}\|_{\infty} \end{bmatrix}$ and $y := \begin{bmatrix} \|Q^{(0)}_\tau - \tau \log \xi^{(0)}\|_{\infty} \\ 0 \end{bmatrix}$, where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

$$x_{t+1} \leq A x_t + \gamma \left( 1 - \frac{\eta \tau}{1 - \gamma} \right)^{t+1} y,$$

where

$$A := \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\eta \tau}{1 - \gamma} & 1 - \frac{\eta \tau}{1 - \gamma} \end{bmatrix}$$

is a rank-1 matrix with a non-zero eigenvalue $\underbrace{1 - \eta \tau}_{\text{contraction rate!}}$. 
Beyond entropy regularization

Leverage regularization to promote structural properties of the learned policy.

- **cost-sensitive RL**
  - weighted 1-norm

- **sparse exploration**
  - Tsallis entropy

- **constrained and safe RL**
  - log-barrier
The regularized value function is defined as

\[ \forall s \in S : \quad V_\pi^\tau(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t (r_t - \tau h_{s_t}(\pi(\cdot|s_t))) \mid s_0 = s \right], \]

where \( h_s \) is convex (and possibly nonsmooth) w.r.t. \( \pi(\cdot|s) \).
The regularized value function is defined as

$$\forall s \in S : V^\pi_\tau(s) := \mathbb{E}\left[ \sum_{t=0}^{\infty} \gamma^t (r_t - \tau h_{st}(\pi(\cdot|s_t))) \mid s_0 = s \right],$$

where $h_s$ is convex (and possibly nonsmooth) w.r.t. $\pi(\cdot|s)$.

maximize$_\pi V^\pi_\tau(\rho) := \mathbb{E}_{s \sim \rho}[V^\pi_\tau(s)]$
Detour: a mirror descent view of entropy-regularized NPG

Entropy-reg. NPG = mirror descent with KL divergence:
(Lan, 2021; Shani et al., 2020)

\[
\pi^{(t+1)}(\cdot | s) = \arg\min_{p \in \Delta(A)} \left\langle -Q^{(t)}_\tau(s, \cdot), p \right\rangle - \tau \mathcal{H}(p) + \frac{1}{\eta} \text{KL}(p||\pi^{(t)}(\cdot | s))
\]

\[
\propto \pi^{(t)}(\cdot | s) \frac{1}{1+\eta\tau} \exp\left(\frac{Q^{(t)}_\tau(s, \cdot)}{\tau}\right) \frac{\eta\tau}{1+\eta\tau}
\]

for all \( s \in S \).
Generalized policy mirror descent (GPMD)

**Definition (Generalized Bregman divergence, Kiwiel 1997)**

The generalized Bregman divergence w.r.t. to a convex $h : \Delta(A) \mapsto \mathbb{R}$ is defined as:

$$D_h(p, q; g) = h(p) - h(q) - \langle g, p - q \rangle$$

$$= h(p) - h(q) - \langle g - c \cdot 1, p - q \rangle,$$

for $p, q \in \Delta(A)$, where $g \in \partial h(q)$ and $c \in \mathbb{R}$. 

A natural idea

For $t = 0, 1, \ldots$, $\pi(t+1) \cdot |s) = \arg\min_{p \in \Delta(A)} \langle -Q_{\tau}(s, \cdot), p \rangle + \tau h_{s}(p) + \eta D_{h_{s}}(p, \pi(t)(\cdot|s)); \partial h_{s}(\pi(t)(\cdot|s)))$.
Generalized policy mirror descent (GPMD)

**Definition (Generalized Bregman divergence, Kiwiel 1997)**

The generalized Bregman divergence w.r.t. to a convex $h : \Delta(A) \mapsto \mathbb{R}$ is defined as:

$$D_h(p, q; g) = h(p) - h(q) - \langle g, p - q \rangle$$

$$= h(p) - h(q) - \langle g - c \cdot 1, p - q \rangle,$$

for $p, q \in \Delta(A)$, where $g \in \partial h(q)$ and $c \in \mathbb{R}$.

**A natural idea**

*For $t = 0, 1, \cdots,*

$$\pi^{(t+1)}(\cdot|s) = \text{argmin}_{p \in \Delta(A)} \langle -Q_{\tau}(s, \cdot), p \rangle + \tau h_s(p)$$

$$+ \frac{1}{\eta} D_{h_s}(p, \pi^{(t)}(\cdot|s); \partial h_s(\pi^{(t)}(\cdot|s)))$$
PMD with Generalized Bregman Divergence (GPMD)

Plugging in a recursive surrogate \( \{ \xi^{(t)} \} \) of \( \partial h_s(\pi^{(t)}(\cdot|s)) \), we obtain the formal algorithm.

**Generalized policy mirror descent (GPMD) method**

For \( t = 0, 1, \cdots \), update

\[
\pi^{(t+1)}(\cdot|s) = \arg\min_{p \in \Delta(A)} \langle -Q_\tau(s, \cdot), p \rangle + \tau h_s(p) \\
+ \frac{1}{\eta} D_{h_s}(p, \pi^{(t)}(\cdot|s); \xi^{(t)}(s, \cdot))
\]

and

\[
\xi^{(t+1)}(s, \cdot) = \frac{1}{1 + \eta \tau} \xi^{(t)}(s, \cdot) + \frac{\eta}{1 + \eta \tau} Q_\tau^{(t)}(s, \cdot).
\]

The subproblem does not admit closed-form solution in general.
Linear convergence with exact gradient

**Exact oracle:** perfect evaluation of $Q_{\pi}^{(t)}$ given $\pi^{(t)}$; exact solution to subproblems.

— *Read our paper for the inexact case!*
Linear convergence with exact gradient

**Exact oracle:** perfect evaluation of $Q^\pi(t)$ given $\pi(t)$; exact solution to subproblems.

— Read our paper for the inexact case!

**Theorem (Zhan*, Cen*, Huang, Chen, Lee, Chi ’21)**

For any learning rate $\eta > 0$, the GPMD updates satisfy

- **Linear convergence of soft Q-functions:**

\[
\|Q^*_\tau - Q^{(t+1)}\|_\infty \leq C_1 \gamma \left(1 - \frac{\eta \tau (1 - \gamma)}{1 + \eta \tau}\right)^t
\]

where $C_1 = \|Q^*_\tau - Q^{(0)}\|_\infty + \frac{2}{1+\eta \tau} \|Q^*_\tau - \tau \xi^{(0)}\|_\infty$. 
Implications

To reach $\|Q_{\tau}^* - Q_{\tau}^{(t+1)}\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates ($\eta > 0$):**

$$\frac{1 + \eta\tau}{\eta\tau(1 - \gamma)} \log \left( \frac{C_1 \gamma}{\epsilon} \right)$$

- **Regularized policy iteration ($\eta = \infty$):**

$$\frac{1}{1 - \gamma} \log \left( \frac{\|Q_{\tau}^* - Q_{\tau}^{(0)}\|_\infty \gamma}{\epsilon} \right)$$
To reach $\|Q^*_\tau - Q^{(t+1)}_\tau\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates ($\eta > 0$):**

  $$\frac{1 + \eta \tau}{\eta \tau (1 - \gamma)} \log \left( \frac{C_1 \gamma}{\epsilon} \right)$$

- **Regularized policy iteration ($\eta = \infty$):**

  $$\frac{1}{1 - \gamma} \log \left( \frac{\|Q^*_\tau - Q^{(0)}_\tau\|_\infty \gamma}{\epsilon} \right)$$

Global linear convergence of GPMD at a **dimension-free** rate!
Comparison with PMD (Lan, 2021)

Policy mirror descent (PMD) method (Lan, 2021)

For $t = 0, 1, \cdots$

$$\pi^{(t+1)}(\cdot|s) = \arg\min_{p \in \Delta(A)} \left\langle -Q_\tau(s, \cdot), p \right\rangle + \tau h_s(p) + \frac{1}{\eta} \text{KL}(p||\pi^{(t)}(\cdot|s))$$

- Linear convergence is established only when $h_s$ is stronger than entropy regularization ($h_s + \mathcal{H}$ is convex).
- In contrast, GPMD converges linearly for general convex and nonsmooth $h_s$!
Numerical examples

\[ h_s = \text{Tsallis Entropy} \]

\[ h_s = \text{Log Barrier} \]

GPMD achieves faster convergence than PMD!
Numerical examples

$h_s = \text{Tsallis Entropy}$

$h_s = \text{Log Barrier}$

GPMD achieves faster convergence than PMD!
Part IV: concluding remarks and further pointers
Concluding remarks

Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Future directions:

- function approximation
- multi-agent RL
- offline RL
- many more...
Beyond the generative model

**Sampling under a behavior policy:** asynchronous/offline RL

(Bhandari et al, 2018; Srikant and Ying, 2019; Qu and Wierman, 2020; Li et al., 2020)

**Exploration under an adaptive policy:** minimize the regret against the optimal policy

(Azar et al., 2017; Jin et al., 2018; Li et al., 2021)
Beyond the tabular setting

- function approximation for dimensionality reduction
- Provably efficient RL algorithms under minimal assumptions

(Silver et al., 2016; Osband and Van Roy, 2014; Dai et al., 2018; Du et al., 2019; Jin et al., 2020)
Multi-agent RL

- **Competitive setting:** finding Nash equilibria for Markov games

- **Collaborative setting:** multiple agents jointly optimize the policy to maximize the total reward

(Zhang, Yang, and Basar, 2021; Cen, Wei, and Chi, 2021)
Can we design RL algorithms based on history data?
(Rashidinejad, Zhu, Ma, Jiao, and Russell, 2021)
Disclaimer: this straw-man list is by no means exhaustive (in fact, it is quite the opposite given the fast pace of the field), and biased towards materials most related to this tutorial; readers are invited to further delve into the references therein to gain a more complete picture.

Books and monographs:

Model-based RL:


**Value-based RL:**


• Beck and Srikant. “Error bounds for constant step-size Q-learning.”


Policy optimization:


Additional ad-hoc pointers:

Thanks!

https://users.ece.cmu.edu/~yuejiec/