Policy Optimization in Reinforcement Learning:
A Tale of Preconditioning and Regularization

Yuejie Chi

Carnegie Mellon University

Instituto Superior Técnico, June 2021
My wonderful collaborators

Shicong Cen
CMU

Chen Cheng
Stanford

Yuxin Chen
Princeton

Yuting Wei
CMU

Gen Li
Princeton

Wenhao Zhan
Princeton

Jason Lee
Princeton

Yuantao Gu
Tsinghua
In RL, an agent learns by interacting with an environment.

- unknown environments
- delayed feedback or rewards
- trial-and-error
- sequential and online

“Recalculating ... recalculating ...”
Recent successes in RL

Policy optimization is a major driver to these successes.
Policy optimization in practice

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

- 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.
Backgrounds: policy optimization in tabular Markov decision processes
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

![Diagram of agent and environment interactions in an MDP with state $s_t$ and action $a_t$.]
Markov decision process (MDP)

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

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

- **Q-function**
  $$\forall (s, a) \in S \times A : \quad Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]$$
Value function and Q-function of policy $\pi$:

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

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

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

Goal: find the optimal policy $\pi^*$ that maximize $V^\pi(s)$

- optimal value / Q function: $V^* := V^{\pi^*}$, $Q^* := Q^{\pi^*}$
Given an initial state distribution $s \sim \rho$, find policy $\pi$ such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]$$
Given an initial state distribution $s \sim \rho$, find policy $\pi$ such that

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

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

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

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 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 method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

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

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

\[ c(|S|, |A|, \frac{1}{1-\gamma}, \cdots) O(\frac{1}{\varepsilon}) \] 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

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

Is the rate of PG good, bad or ugly?
Softmax PG can take exponential time to converge
Theorem (Li, Wei, Chi, Gu, Chen, 2021)

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

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

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*

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

*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:
• $s, a$

$\log(\max) = a^2 = |a^2|$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

$\sum s, a$

$\exp(\cdot) = \sum s, a$

$\log(a^2) = a^2$

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

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

\[
\max_{\pi} V(\pi) = \mathbb{E}_{s \sim \pi} [V(s)] = \max_{\pi} \mathbb{E}_{s \sim \pi} [\pi(a|s) \cdot \exp(\phi(s,a))]
\]

\[
\max_{\pi} \pi(t)(a_1 | 1)
\]

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?

Does policy gradient (PG) method converge?

However, “asymptotic convergence” might mean “takes forever”
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

$$\text{convergence-time}(s) \gtrsim (\text{convergence-time}(s - 2))^{1.5}$$
“ Seriously, lady, at this hour you’d make a lot better time taking the subway. ”
Accelerating the convergence via preconditioning and regularization

Shicong Cen
CMU

Chen Cheng
Stanford

Yuxin Chen
Princeton

Yuting Wei
CMU
Natural policy gradient (NPG) method (Kakade, 2002)

For $t = 0, 1, \cdots$

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

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 $t = 0, 1, \cdots$

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

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

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

In fact, popular heuristic TRPO (Schulman et al., 2015) = NPG + line search.
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.
Booster #2: entropy regularization

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

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

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

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

where \( Q^{(t)}_{\tau} := Q^{(t)}_{\pi^{(t)}} \) 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 \( \eta = \frac{1-\gamma}{\tau} \).
Linear convergence with exact gradient

**Exact oracle:** perfect evaluation of $Q_{\pi}^{(t)}$ given $\pi^{(t)}$;

---

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

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 $\|Q^*_{t} - Q^{(t+1)}_{\tau}\|_{\infty} \leq \epsilon$, 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

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

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

can be exponential in \(|S|\) and \(\frac{1}{1-\gamma}\)

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

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

Ours

(Agarwal et al. 2019)
Comparison with unregularized NPG

**Regularized NPG**

\( \tau = 0.001 \)

**Vanilla NPG**

\( \tau = 0 \)

**Linear rate:**

\[
\frac{1}{\eta \tau} \log \left( \frac{1}{\epsilon} \right)
\]

**Sublinear rate:**

\[
\frac{1}{\min\{\eta, (1-\gamma)^2\}} \epsilon
\]

Ours

(Agarwal et al. 2019)

Entropy regularization enables fast convergence!
Inexact oracle: inexact evaluation of $Q^{\pi(t)}_T$ 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 $Q_{\pi}^{(t)}$ 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**:

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

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

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

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

**Inexact entropy-regularized NPG:**

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

**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

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

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

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

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

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

$$\tilde{O} \left( \frac{|S| |A|}{(1 - \gamma)^7 \epsilon^2} \right)$$ samples

Recipe: set $\tau = \frac{(1-\gamma)\epsilon}{\log |A|}$; use fresh samples for policy evaluation (Li et al., 2020).
A glimpse of the analysis
A key lemma: monotonic performance improvement

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

KL divergence

discounted state visitation distribution
A key lemma: monotonic performance improvement

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

Implication: monotonic improvement of \( V_{\tau}(s) \) and \( Q_{\tau}(s, a) \).
Recall: Bellman’s optimality principle

Bellman operator

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

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

Bellman operator

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

- immediate reward
- next state’s value

• one-step look-ahead

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

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

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

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

Richard Bellman
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
A key operator: soft Bellman operator

**Soft Bellman operator**

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

immediate reward

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

next state's value

\[ - \tau \log \pi(a'|s') \]

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**

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

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

**Policy iteration**

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

**Soft policy iteration**

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

**Bellman operator**

**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 \( x_t := \begin{bmatrix} \| Q^* - Q^{(t)}_\tau \|_\infty \\ \| Q^* - \tau \log \xi^{(t)}_\tau \|_\infty \end{bmatrix} \) and \( y := \begin{bmatrix} \| Q^{(0)}_\tau - \tau \log \xi^{(0)}_\tau \|_\infty \\ 0 \end{bmatrix} \),

where \( \xi^{(t)} \propto \pi^{(t)} \) is an auxiliary sequence, then

\[
x_{t+1} \leq Ax_t + \gamma \left( 1 - \frac{\eta \tau}{1 - \gamma} \right)^{t+1} y,
\]

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 \( 1 - \eta \tau \).
Beyond entropy regularization

Shicong Cen  
CMU

Wenhao Zhan  
Princeton

Yuxin Chen  
Princeton

Jason Lee  
Princeton
The ever-important role of 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

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

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

Regularized RL in general form
The regularized value function is defined as
\[
\forall s \in S : \quad V_\tau^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t (r_t - \tau h_{s_t}(\pi(\cdot|s_t))) \left| s_0 = s \right. \right],
\]
where \( h_s \) is convex (and possibly nonsmooth) w.r.t. \( \pi(\cdot|s) \).

Maximize \( \pi \)
\[
\max_{\pi} V_\tau^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V_\tau^\pi(s)]
\]
Entropic-regularized NPG = mirror descent with KL divergence (Lan, 2021; Shani et al., 2020):

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

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

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 $c \in \mathbb{R}$.
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* $t = 0, 1, \cdots$,

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

$$+ \frac{1}{\eta} D_{h_s}(p, \pi^{(t)}(\cdot | s); \partial h_s(\pi^{(t)}(\cdot | s)))$$
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* \( t = 0, 1, \ldots \), *update*

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

*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 $Q_{\tau}^{\pi(t)}$ given $\pi(t)$; exact solution to subproblems.

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

Exact oracle: perfect evaluation of $Q^\pi_T(t)$ 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:

$$\|Q^*_T - Q^T(t+1)\|_\infty \leq C_1 \gamma \left(1 - \frac{\eta_T (1 - \gamma)}{1 + \eta_T} \right)^t$$

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

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 ($\eta = \infty$):**

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

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 ($\eta = \infty$):**

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

Global linear convergence of GPMD at a **dimension-free** rate!
Policy mirror descent (PMD) method (Lan, 2021)

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

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

- 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!
Bonus track: entropy-regularized games

Shicong Cen
CMU

Yuting Wei
CMU
Zero-sum entropy-regularized two-player matrix game

Quantal response equilibrium (McKelvey and Palfrey, 1995)

$$\max_{\mu \in \Delta(A)} \min_{\nu \in \Delta(B)} \mu^\top A \nu + \tau H(\mu) - \tau H(\nu)$$

- Basic building block for solving value iteration in zero-sum two-player Markov games.
Motivation: an implicit update method

Implicit update (IU) method

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

\[
\begin{align*}
\mu^{(t+1)} &\propto [\mu^{(t)}]^{1-\eta\tau} \exp \left( \frac{[A\nu^{(t+1)}]}{\tau} \right)^{\eta\tau} \\
\nu^{(t+1)} &\propto [\nu^{(t)}]^{1-\eta\tau} \exp \left( -\frac{[A^\top \mu^{(t+1)}]}{\tau} \right)^{\eta\tau}
\end{align*}
\]
Motivation: an implicit update method

Implicit update (IU) method

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

\[
\begin{align*}
\mu^{(t+1)} &\propto [\mu^{(t)}]^{1-\eta} \exp \left( \frac{[A \nu^{(t+1)}] / \tau}{\eta} \right) \\
\nu^{(t+1)} &\propto [\nu^{(t)}]^{1-\eta} \exp \left( -\frac{[A^\top \mu^{(t+1)}] / \tau}{\eta} \right)
\end{align*}
\]

Theorem (Cen, Wei, Chi, 2021)

Suppose that $0 < \eta \leq 1/\tau$, then for all $t \geq 0$,

\[
KL(\zeta_\tau^* \parallel \zeta^{(t)}) \leq (1 - \eta \tau)^t KL(\zeta_\tau^* \parallel \zeta^{(0)}),
\]

where $KL(\zeta_\tau^* \parallel \zeta^{(t)}) = KL(\mu_\tau^* \parallel \mu^{(t)}) + KL(\nu_\tau^* \parallel \nu^{(t)})$. 

Can we make this practical?
Motivation: an implicit update method

**Implicit update (IU) method**

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

$$\begin{cases} 
\mu^{(t+1)} \propto \left[ \mu^{(t)} \right]^{1-\eta}\exp\left(\left[A\nu^{(t+1)}\right]/\tau\right)^{\eta}\tau \\
\nu^{(t+1)} \propto \left[ \nu^{(t)} \right]^{1-\eta}\exp\left(-\left[A^\top \mu^{(t+1)}\right]/\tau\right)^{\eta}\tau 
\end{cases}$$

**Theorem (Cen, Wei, Chi, 2021)**

Suppose that $0 < \eta \leq 1/\tau$, then for all $t \geq 0$,

$$KL(\zeta_\tau^* \parallel \zeta^{(t)}) \leq (1 - \eta\tau)^t KL(\zeta_\tau^* \parallel \zeta^{(0)}),$$

where $KL(\zeta_\tau^* \parallel \zeta^{(t)}) = KL(\mu_\tau^* \parallel \mu^{(t)}) + KL(\nu_\tau^* \parallel \nu^{(t)})$.

Can we make this practical?
From implicit updates to policy extragradient methods

**Predictive update (PU) method**

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

1. **extrapolate/predict:**

\[
\bar{\mu}(t+1) \propto \mu(t)^{1-\eta\tau} \exp \left( \frac{A \tilde{\nu}(t+1)}{\tau} \right)^{\eta\tau}
\]

\[
\bar{\nu}(t+1) \propto \nu(t)^{1-\eta\tau} \exp \left( -\frac{A^\top \bar{\mu}(t+1)}{\tau} \right)^{\eta\tau}
\]

2. **update:**

\[
\begin{align*}
\mu(t+1) & \propto \mu(t)^{1-\eta\tau} \exp \left( \frac{A \bar{\nu}(t+1)}{\tau} \right)^{\eta\tau} \\
\nu(t+1) & \propto \nu(t)^{1-\eta\tau} \exp \left( -\frac{A^\top \bar{\mu}(t+1)}{\tau} \right)^{\eta\tau}
\end{align*}
\]
Predictive update (PU) method

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

1. extrapolate/predict:

\[
\begin{align*}
\bar{\mu}^{(t+1)} & \propto [\mu^{(t)}]^{1-\eta \tau} \exp \left( \frac{[A\nu^{(t)}]}{\tau} \right)^{\eta \tau} \\
\bar{\nu}^{(t+1)} & \propto [\nu^{(t)}]^{1-\eta \tau} \exp \left( -\frac{[A^\top \mu^{(t)}]}{\tau} \right)^{\eta \tau}
\end{align*}
\]

2. update:

\[
\begin{align*}
\mu^{(t+1)} & \propto [\mu^{(t)}]^{1-\eta \tau} \exp \left( \frac{[A\bar{\nu}^{(t+1)}]}{\tau} \right)^{\eta \tau} \\
\nu^{(t+1)} & \propto [\nu^{(t)}]^{1-\eta \tau} \exp \left( -\frac{[A^\top \bar{\mu}^{(t+1)}]}{\tau} \right)^{\eta \tau}
\end{align*}
\]
Optimistic multiplicative weights update (OMWU) method

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

1. **extrapolate/predict:**

\[
\begin{align*}
\bar{\mu}(t+1) &\propto [\mu(t)]^{1-\eta} \exp \left( \frac{[A\bar{\nu}(t)]}{\tau} \right)^{\eta \tau} \\
\bar{\nu}(t+1) &\propto [\nu(t)]^{1-\eta} \exp \left( -\frac{[A^\top \bar{\mu}(t)]}{\tau} \right)^{\eta \tau}
\end{align*}
\]

2. **update:**

\[
\begin{align*}
\mu(t+1) &\propto [\mu(t)]^{1-\eta} \exp \left( \frac{[A\bar{\nu}(t+1)]}{\tau} \right)^{\eta \tau} \\
\nu(t+1) &\propto [\nu(t)]^{1-\eta} \exp \left( -\frac{[A^\top \bar{\mu}(t+1)]}{\tau} \right)^{\eta \tau}
\end{align*}
\]
Last-iterate convergence

- **Entropy-regularized matrix game:** To get an $\epsilon$-optimal solution to the regularized problem ($\epsilon$-QRE), the iteration complexity is at most

  $$\tilde{O}\left(\left(1 + \frac{\|A\|_{\infty}}{\tau}\right) \log \frac{1}{\epsilon}\right).$$

- **Unregularized matrix game:** To get an $\epsilon$-optimal solution to the unregularized problem ($\epsilon$-NE), the iteration complexity is at most

  $$\tilde{O}\left(\frac{\|A\|_{\infty}}{\epsilon}\right).$$

No need to assume unique Nash equilibrium!
Concluding remarks

Future directions:

- function approximation
- sample complexities
- Markov games
- multi-agent RL

Policy Gradient

Natural Policy Gradient

fast global linear convergence of RL enabled by regularization + preconditioning
Thanks!


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