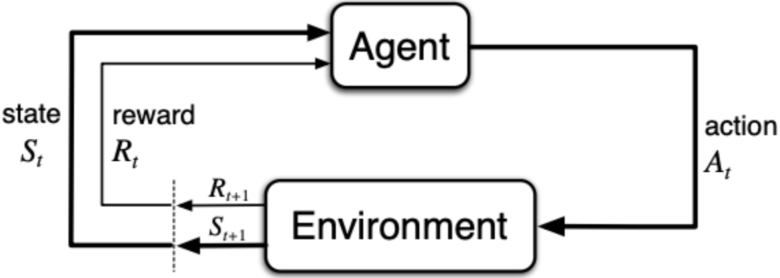
Sample Efficient Reinforcement Learning with REINFORCE

Junzi Zhang[1], Jongho Kim[1], Brendan O'Donoghue[2], Stephen Boyd[1]

[1]EE & ICME Departments, Stanford University
[2]Google DeepMind

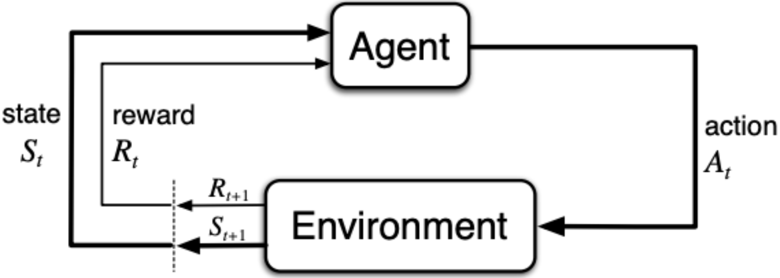
AAAI 2021 Virtual Presentation

Overview



- **RL**: algorithms for solving MDPs with incomplete information of $\mathcal{M}$ (*e.g.*, p, r accessible by interacting with the environment) as input.

Reinforcement Learning (RL)

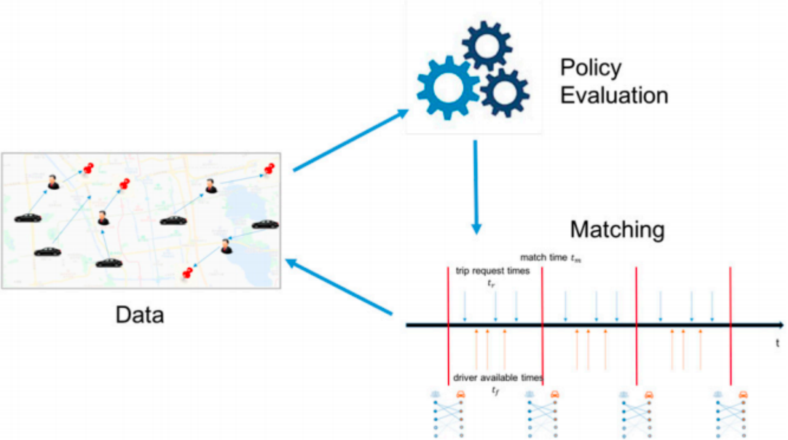


- **RL**: algorithms for solving MDPs with incomplete information of $\mathcal{M}$ (*e.g.*, p, r accessible by interacting with the environment) as input.
- **Today**: episodic (allow restart in the trajectory) and model-free (no storage of transition & reward models).

Success of RL


14:32
AlphaStar
177 /200
945
758
64
113
940
SUPPLY
MINERALS
GAS
WORKERS
ARMY
APM
PRODUCTION
LiquidTLO
147 /172
335
442
61
86
1377

Success of RL



Why Policy Gradient?

Heroes Behind the Success: RL algorithms

- Value function learning (global convergence ✓)
  - Q-learning, SARSA, Bellman Residue Minimization, etc.

Why Policy Gradient?

Heroes Behind the Success: RL algorithms

- Value function learning (global convergence ✓)
    - Q-learning, SARSA, Bellman Residue Minimization, etc.
- Monte Carlo Tree Search (global convergence ✓):
    - ϵ-greedy tree search, UCT, BRUE, etc.

Why Policy Gradient?

Heroes Behind the Success: RL algorithms

- Value function learning (global convergence ✓)
  - Q-learning, SARSA, Bellman Residue Minimization, etc.
- Monte Carlo Tree Search (global convergence ✓):
  - ϵ-greedy tree search, UCT, BRUE, etc.
- Policy optimization (global convergence ✓✗)
  - Policy gradient, random search, actor-critic, etc.

Why Policy Gradient?

Heroes Behind the Success: RL algorithms

- Value function learning (global convergence ✓)
  - Q-learning, SARSA, Bellman Residue Minimization, etc.
- Monte Carlo Tree Search (global convergence ✓):
  - ϵ-greedy tree search, UCT, BRUE, etc.
- Policy optimization (global convergence ✓✗)
  - Policy gradient, random search, actor-critic, etc.

Today: global convergence & sample efficiency of practical versions of policy gradient methods such as REINFORCE

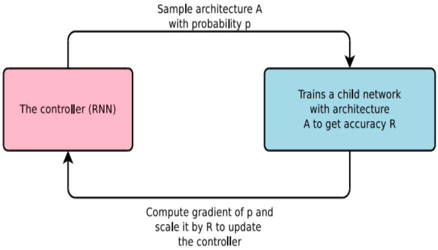
Why REINFORCE?

REINFORCE: balance between **good empirical performance** & **implementation simplicity**

Why REINFORCE?

REINFORCE: balance between **good empirical performance** & **implementation simplicity**

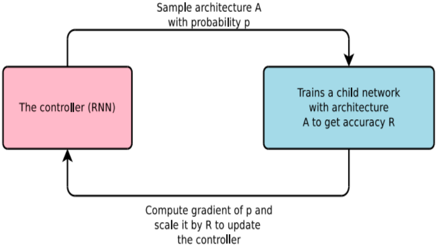
- Neural Architecture Search
- Semantic Program Parser
- Visual Question Answering
- Dialogue generation
- Coreference resolution
- ...



Why REINFORCE?

REINFORCE: balance between **good empirical performance** & **implementation simplicity**

- Neural Architecture Search
- Semantic Program Parser
- Visual Question Answering
- Dialogue generation
- Coreference resolution
- ...



A good baseline and starting point!

Underlying MDP

MDP (stationary, discounted): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, \rho)$, $\gamma \in [0, 1)$.

Underlying MDP

MDP (stationary, discounted): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, \rho)$, $\gamma \in [0, 1)$.

- $\rho > 0$, $S = |\mathcal{S}| < \infty$, $A = |\mathcal{A}| < \infty$. W.l.o.g., $r(s, a) \in [0, 1]$.
- Goal: maximize $\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, where $s_0 \sim \rho$, $a_t \sim \pi(s_t, \cdot)$, $s_{t+1} \sim p(\cdot|s_t, a_t)$, and $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is called policy.

Policy Optimization

- Policy optimization reformulation:

$$\text{maximize}_{\pi \in \Pi}\, F(\pi),$$

where

$$F(\pi) = \mathbf{E} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t),$$

$s_0 \sim \rho,\ a_t \sim \pi(s_t, \cdot),\ s_{t+1} \sim p(\cdot | s_t, a_t),\ \forall t \geq 0$, and

$$\Pi = \left\{ \pi \in \mathbf{R}^{SA} \,\middle|\, \sum_{a=1}^{A} \pi_{s,a} = 1\,(\forall s \in \mathcal{S}),\ \pi_{s,a} \geq 0\,(\forall s \in \mathcal{S},\ a \in \mathcal{A}) \right\}.$$

Policy Optimization

- Policy optimization reformulation:

$$\text{maximize}_{\pi \in \Pi} F(\pi),$$

- $F(\pi)$ is also written as $V^{\pi}(\rho)$ in the value function learning literature.

Policy Optimization

- Policy optimization reformulation:

$$\text{maximize}_{\pi \in \Pi} F(\pi),$$

- $F(\pi)$ is also written as $V^\pi(\rho)$ in the value function learning literature.
- Policy parametrization: $\pi_\theta : \Theta \to \Pi$.
- New problem:

$$\text{maximize}_{\theta \in \Theta} F(\pi_\theta).$$

Policy Optimization

- Policy optimization reformulation:

$$\text{maximize}_{\pi\in\Pi}\ F(\pi),$$

- $F(\pi)$ is also written as $V^\pi(\rho)$ in the value function learning literature.
- Policy parametrization: $\pi_\theta : \Theta \to \Pi$.
- New problem:

$$\text{maximize}_{\theta\in\Theta}\ F(\pi_\theta).$$

- **Today** – energy-based policies: $\pi_\theta(s,a) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})}$, $\Theta = \mathbf{R}^{SA}$.
- Practical choice in reality, common basis for more advanced (*e.g.*, neural) parametrization.

Policy Gradient Existence

- Question: Is $F(\pi_\theta)$ differentiable?

Policy Gradient Existence

- Question: Is $F(\pi_\theta)$ differentiable?
- Answer: yes!
  - Indeed, $F(\pi_\theta)$ is at least C^2 and $\nabla_\theta F(\pi_\theta)$ is $8/(1-\gamma)^3$-Lipschitz.

Policy Gradient Methods

- (Vanilla) policy gradient method:

$$\theta^{k+1} = \theta^k + \alpha^k \nabla_\theta L_{\lambda^k}(\theta^k),$$

where $L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta)$: *e.g.*, entropy reg R.

- Some other variants: NPG, TRPO/PPO, DPG etc.

Policy Gradient Methods

- (Vanilla) policy gradient method:

$$\theta^{k+1} = \theta^k + \alpha^k \nabla_\theta L_{\lambda^k}(\theta^k),$$

  where $L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta)$: *e.g.*, entropy reg R.
- Some other variants: NPG, TRPO/PPO, DPG etc.
- What does the policy gradient look like?

Policy Gradient Methods

- (Vanilla) policy gradient method:

$$\theta^{k+1} = \theta^k + \alpha^k \nabla_\theta L_{\lambda^k}(\theta^k),$$

  where $L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta)$: *e.g.*, entropy reg R.
- Some other variants: NPG, TRPO/PPO, DPG etc.
- What does the policy gradient look like?
  - **Policy gradient theorems** (PGT): hold for general C^1-smooth π_θ.
  - **Policy gradient estimators** (PGE): Monte Carlo approx of PGT.

Policy Gradient Methods

- (Vanilla) policy gradient method:

$$\theta^{k+1} = \theta^k + \alpha^k \nabla_\theta L_{\lambda^k}(\theta^k),$$

  where $L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta)$: *e.g.*, entropy reg R.
- Some other variants: NPG, TRPO/PPO, DPG etc.
- What does the policy gradient look like?
  - **Policy gradient theorems** (PGT): hold for general C^1-smooth π_θ.
  - **Policy gradient estimators** (PGE): Monte Carlo approx of PGT.
- How to reduce variance caused by Monte Carlo approximation?

Policy Gradient Methods

- (Vanilla) policy gradient method:

$$\theta^{k+1} = \theta^k + \alpha^k \nabla_\theta L_{\lambda^k}(\theta^k),$$

where $L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta)$: *e.g.*, entropy reg R.

- Some other variants: NPG, TRPO/PPO, DPG etc.
- What does the policy gradient look like?
  - **Policy gradient theorems** (PGT): hold for general C^1-smooth π_θ.
  - **Policy gradient estimators** (PGE): Monte Carlo approx of PGT.
- How to reduce variance caused by Monte Carlo approximation?
  - **Mini-batch updates**.

PGE in Theoretical Analysis

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{s \sim d_\rho^{\pi_\theta}} \mathbf{E}_{a \sim \pi_\theta(s,\cdot)} \left[Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(s,a) \right].$$

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{s\sim d_\rho^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(s,\cdot)} \left[Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(s,a) \right].$$

Here $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ denotes a trajectory, and

$$Q^\pi(s,a) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a, a_t \sim \pi(s_t, \cdot), s_{t+1} \sim p(\cdot|s_t, a_t), \forall t > 0 \right],$$

$$d_\rho^\pi = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbf{Prob}_\pi(s_t = s | s_0 \sim \rho).$$

PGE in Theoretical Analysis

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{s \sim d_\rho^{\pi_\theta}} \mathbf{E}_{a \sim \pi_\theta(s,\cdot)} \left[Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(s,a) \right].$$

- Vistation measure based PGE **(used in theory)**:

$$\hat{\nabla}_\theta F(\pi_{\theta^k}) = \tfrac{1}{1-\gamma} (\hat{Q}^k(s,a) - b(s)) \nabla_\theta \log \pi_\theta(s,a),$$

where $s \sim d_\rho^{\pi_{\theta^k}}$, $a \sim \pi_{\theta^k}(s,\cdot)$, $\hat{Q}^k(s,a) \approx Q^{\pi_{\theta^k}}(s,a)$, b is baseline:

PGE in Theoretical Analysis

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{s \sim d_\rho^{\pi_\theta}} \mathbf{E}_{a \sim \pi_\theta(s,\cdot)} \left[Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(s,a) \right].$$

- Vistation measure based PGE **(used in theory)**:

$$\hat{\nabla}_\theta F(\pi_{\theta^k}) = \tfrac{1}{1-\gamma} (\hat{Q}^k(s,a) - b(s)) \nabla_\theta \log \pi_\theta(s,a),$$

where $s \sim d_\rho^{\pi_{\theta^k}}$, $a \sim \pi_{\theta^k}(s,\cdot)$, $\hat{Q}^k(s,a) \approx Q^{\pi_{\theta^k}}(s,a)$, b is baseline:

  - Trajectory for sampling s is wasted, rarely used in practice.

PGE in Theoretical Analysis

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{s \sim d_\rho^{\pi_\theta}} \mathbf{E}_{a \sim \pi_\theta(s,\cdot)} \left[Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(s,a) \right].$$

- Vistation measure based PGE **(used in theory)**:

$$\hat{\nabla}_\theta F(\pi_{\theta^k}) = \tfrac{1}{1-\gamma} (\hat{Q}^k(s,a) - b(s)) \nabla_\theta \log \pi_\theta(s,a),$$

where $s \sim d_\rho^{\pi_{\theta^k}}$, $a \sim \pi_{\theta^k}(s,\cdot)$, $\hat{Q}^k(s,a) \approx Q^{\pi_{\theta^k}}(s,a)$, b is baseline:
  - Trajectory for sampling s is wasted, rarely used in practice.
  - Example $\hat{Q}$: $\hat{Q}^k(s,a) = \sum_{t'=t}^{H^k} \gamma^{t'-t} r_{t'}^k$, H^k is a truncation horizon, $\tau^k = (s, a, r_0^k, \dots, s_{H^k}^k, a_{H^k}^k, r_{H^k}^k) \sim \mathbf{Prob}_{s,a}^{\pi_{\theta^k}}$.

PGE in Practice

- Trajectory-based PGT:

$$\nabla_\theta F(\pi_\theta) = \mathbf{E}_{\tau \sim \mathbf{Prob}_\rho^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(s_t, a_t) \right]$$

- Trajectory-based PGT:

$$\nabla_\theta F(\pi_\theta) = \mathbf{E}_{\tau \sim \mathbf{Prob}_\rho^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(s_t, a_t) \right]$$

- REINFORCE PGE **(used in practice)**:

$$\hat{\nabla}_\theta F(\pi_{\theta^k}) = \sum_{t=0}^{\lfloor \beta H^k \rfloor} \gamma^t (\widehat{Q}^k(s_t^k, a_t^k) - b(s_t^k)) \nabla_\theta \log \pi_{\theta^k}(a_t^k | s_t^k),$$

where $\beta \in (0,1)$, $\widehat{Q}^k(s,a) \approx Q^{\pi_{\theta^k}}(s,a)$, b is baseline, H^k is the truncation horizon, $\tau^k = (s_0^k, a_0^k, r_0^k, \ldots, s_{H^k}^k, a_{H^k}^k, r_{H^k}^k) \sim \mathbf{Prob}_\rho^{\pi_{\theta^k}}$.

  - Example $\widehat{Q}$: $\widehat{Q}^k(s_t^k, a_t^k) = \sum_{t'=t}^{H^k} \gamma^{t'-t} r_{t'}^k$.

PGE in Practice (Continued)

- Another trajectory-based PGT:

$$\nabla_\theta F(\pi_\theta) = \mathbf{E}_{\tau \sim \mathbf{Prob}_\rho^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \sum_{h=0}^{t} \nabla_\theta \log \pi_\theta(s_h, a_h) \right]$$

PGE in Practice (Continued)

- Another trajectory-based PGT:

$$\nabla_\theta F(\pi_\theta) = \mathbf{E}_{\tau \sim \mathbf{Prob}_\rho^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \sum_{h=0}^{t} \nabla_\theta \log \pi_\theta(s_h, a_h) \right]$$

- GPOMDP PGE **(used in practice)**:

$$\hat{\nabla}_\theta F(\pi_{\theta^k}) = \sum_{t=0}^{H^k} \gamma^t (r_t^k - b_t) \sum_{h=0}^{t} \nabla_\theta \log \pi_{\theta^k}(a_h^k | s_h^k),$$

where b is baseline, H^k is the truncation horizon.

Additional (Practical) PGE

- Actor-Critic PGE: Q-functions estimated using TD algorithms.

Additional (Practical) PGE

- Actor-Critic PGE: Q-functions estimated using TD algorithms.
- Zeroth-Order/Random Search PGE:
  - Corresponding to a random perturbation/smoothing type "policy gradient theorem", widely used in PG + LQR literature.

Additional (Practical) PGE

- Actor-Critic PGE: Q-functions estimated using TD algorithms.
- Zeroth-Order/Random Search PGE:
  - Corresponding to a random perturbation/smoothing type "policy gradient theorem", widely used in PG + LQR literature.
- Question 1: Can we deal with all kinds of (practical) estimators (*e.g.*, REINFORCE)?

Mini-batch Updates

- Sample M independent trajectories $\tau_1^k, \ldots, \tau_M^k$ from $\mathcal{M}$ following policy π_{θ^k} and then compute an approximate gradient $\hat{\nabla}_\theta^{(i)} L_{\lambda^k}(\theta^k)$ $(i = 1, \ldots, M)$ using each of these M trajectories.

Mini-batch Updates

- Sample M independent trajectories $\tau_1^k, \ldots, \tau_M^k$ from $\mathcal{M}$ following policy π_{θ^k} and then compute an approximate gradient $\hat{\nabla}_\theta^{(i)} L_{\lambda^k}(\theta^k)$ $(i = 1, \ldots, M)$ using each of these M trajectories.
- Then update as follows:

$$\theta^{k+1} = \theta^k + \alpha^k \frac{1}{M} \sum_{i=1}^{M} \hat{\nabla}_\theta^{(i)} L_{\lambda^k}(\theta^k).$$

Mini-batch Updates

- Sample M independent trajectories $\tau_1^k, \ldots, \tau_M^k$ from $\mathcal{M}$ following policy π_{θ^k} and then compute an approximate gradient $\hat{\nabla}_\theta^{(i)} L_{\lambda^k}(\theta^k)$ $(i = 1, \ldots, M)$ using each of these M trajectories.
- Then update as follows:

$$\theta^{k+1} = \theta^k + \alpha^k \frac{1}{M} \sum_{i=1}^{M} \hat{\nabla}_\theta^{(i)} L_{\lambda^k}(\theta^k).$$

- Question 2: Can we accurately characterize the effect of M?

Theory vs. Practice: What was Missing?

	Global?	Practical PGE?	Finite MB?	High-Prob Rate?
Long Ago	No	Yes	Yes	No (a.s. Asymp)
$\sim$ 10 years	No	Yes	Yes	No (Rate in Expect.)
$\sim$ 2 years	Yes	No	No: $\Omega(\frac{1}{M^p})$	No (Rate in Expect.)
Our Work	Yes	Yes	Yes	Yes (High-Prob + a.s.)

Table: PGE: policy gradient estimators; MB: mini-batch

Theory vs. Practice: What was Missing?

	Global?	Practical PGE?	Finite MB?	High-Prob Rate?
Long Ago	No	Yes	Yes	No (a.s. Asymp)
$\sim$ 10 years	No	Yes	Yes	No (Rate in Expect.)
$\sim$ 2 years	Yes	No	No: $\Omega(\frac{1}{M^p})$	No (Rate in Expect.)
Our Work	Yes	Yes	Yes	Yes (High-Prob + a.s.)

Table: PGE: policy gradient estimators; MB: mini-batch

- Exceptions:
  - LQR [JSW20] (our work: general MDPs);
  - NPG [AYBB+19, CYJW19, ESRM20] (our work: vanilla PG).

Algorithm Specification & PGE Assumptions

1. Choose regularization $R(\theta) = \frac{1}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \log \pi_\theta(s,a)$ ($\frac{\lambda}{S}$-smooth);

Algorithm Specification & PGE Assumptions

1. Choose regularization $R(\theta) = \frac{1}{SA} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \log \pi_\theta(s, a)$ ($\frac{\lambda}{S}$-smooth);
2. Decrease λ^k in doubling phases – indexing: $k \to (l, k)$;

Algorithm Specification & PGE Assumptions

1. Choose regularization $R(\theta) = \frac{1}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \log \pi_\theta(s,a)$ ($\frac{\lambda}{S}$-smooth);
2. Decrease λ^k in doubling phases – indexing: $k \to (l,k)$;
3. Add simple truncation after each phase (to bound log).

Assumption (PGE: nearly unbiased & bounded variance)

There exist constants C, C_1, C_2, M_1, $M_2 > 0$, such that for all l, $k \geq 0$, we have $\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})\|_2 \leq C_1$ almost surely and that

$$\nabla_\theta L_{\lambda^l}(\theta^{l,k})^T \mathbf{E}_{l,k}\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) \geq C_2\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 - \delta_{l,k}, \tag{1}$$

$$\mathbf{E}_{l,k}\|\widehat{\nabla}_\theta L_{\lambda^k}(\theta^{l,k})\|_2^2 \leq M_1 + M_2\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2, \tag{2}$$

where $\sum_{k=0}^{T_l-1}\delta_{l,k}^2 \leq C$, $\forall\ l \geq 0$. Also, $H^{l,k} \geq \log_{1/\gamma}(k+1)$, $\forall\ l,\ k \geq 0$.

Main Results

Then we obtain (N is the number of episodes):

Main Results

Then we obtain (N is the number of episodes):

1. any-time sub-linear high-prob regret bound

$$O((M^{\frac{1}{6}} + M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}} + M(\log N)^2) = \tilde{O}(N^{\frac{5}{6}}).$$

Main Results

Then we obtain (N is the number of episodes):

1. any-time sub-linear high-prob regret bound

$$O((M^{\frac{1}{6}} + M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}} + M(\log N)^2) = \tilde{O}(N^{\frac{5}{6}}).$$

2. a.s. convergence of average regret with asymptotic rate

$$O\left((M^{\frac{1}{6}} + M^{-\frac{5}{6}})N^{-\frac{1}{6}}\left(1+\frac{M}{N}\right)^{\frac{5}{6}}(\log N)^{\frac{5}{2}} + \frac{M(\log N)^2}{N}\right) = \tilde{O}(N^{-\frac{1}{6}}).$$

Main Results (Continued)

For REINFORCE & GPOMDP PGEs:

- PGE assumptions easily verified with $\Theta(\log k)$ truncated horizon H^k.

Main Results (Continued)

For REINFORCE & GPOMDP PGEs:

- PGE assumptions easily verified with $\Theta(\log k)$ truncated horizon H^k.

1. any-time sub-linear high-prob regret bound (w.p. at least $1-\delta$)

$$O\left(\left(\frac{S^2A^2}{(1-\gamma)^7}+\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty\right)(M^{\frac{1}{6}}+M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}}+M(\log N)^2\right).$$

Main Results (Continued)

For REINFORCE & GPOMDP PGEs:

- PGE assumptions easily verified with $\Theta(\log k)$ truncated horizon H^k.

1. any-time sub-linear high-prob regret bound (w.p. at least $1-\delta$)

$$O\left(\left(\frac{S^2A^2}{(1-\gamma)^7}+\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty\right)(M^{\frac{1}{6}}+M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}}+M(\log N)^2\right).$$

2. a.s. convergence of average regret with asymptotic rate

$$O\left(\left(\frac{S^2A^2}{(1-\gamma)^7}+\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty\right)(M^{\frac{1}{6}}+M^{-\frac{5}{6}})N^{-\frac{1}{6}}\left(1+\frac{M}{N}\right)^{\frac{5}{6}}(\log N)^{\frac{5}{2}}+\frac{M(\log N)^2}{N}\right).$$

Proof Sketch

- **Phase analysis**: bound regret in each phase (with λ^k fixed)

Proof Sketch

- **Phase analysis**: bound regret in each phase (with λ^k fixed)
  - **Control of "bad" episodes**: sub-linear upper bound on $\#$ episodes with large gradient norms $\|\nabla_\theta L_\lambda(\theta^k)\|_2$.

Proof Sketch

- **Phase analysis**: bound regret in each phase (with λ^k fixed)
  - **Control of "bad" episodes**: sub-linear upper bound on # episodes with large gradient norms $\|\nabla_\theta L_\lambda(\theta^k)\|_2$.
  - **Gradient domination condition** [AKLM19]: from gradient norm $\|\nabla_\theta L_\lambda(\theta^k)\|_2$ to sub-optimality gap $F^\star - F(\pi_{\theta^k})$.

Proof Sketch

- **Phase analysis**: bound regret in each phase (with λ^k fixed)
  - **Control of "bad" episodes**: sub-linear upper bound on $\#$ episodes with large gradient norms $\|\nabla_\theta L_\lambda(\theta^k)\|_2$.
  - **Gradient domination condition** [AKLM19]: from gradient norm $\|\nabla_\theta L_\lambda(\theta^k)\|_2$ to sub-optimality gap $F^\star - F(\pi_{\theta^k})$.
- **Doubling trick**:
  - stitch together phase regrets with $\log N$ additional terms.

Proof Sketch

- **Phase analysis**: bound regret in each phase (with λ^k fixed)
  - **Control of "bad" episodes**: sub-linear upper bound on # episodes with large gradient norms $\|\nabla_\theta L_\lambda(\theta^k)\|_2$.
  - **Gradient domination condition** [AKLM19]: from gradient norm $\|\nabla_\theta L_\lambda(\theta^k)\|_2$ to sub-optimality gap $F^\star - F(\pi_{\theta^k})$.
- **Doubling trick**:
  - stitch together phase regrets with $\log N$ additional terms.
- **From high prob (with** $\log(1/\delta)$ **dependency) to a.s.**:
  - Borel-Cantelli.

Open Problems

Extended version of this work (posting soon, check
`https://stanford.edu/~boyd/papers/conv_reinforce.html`):

- Episodic finite horizon MDPs.
- Additional PGEs.

Open Problems

Extended version of this work (posting soon, check `https://stanford.edu/~boyd/papers/conv_reinforce.html`):

- Episodic finite horizon MDPs.
- Additional PGEs.

Some future directions:

- Practically widely used (relative) entropy regularization, and empirical tests of the log-barrier one adopted in our work and [AKLM19].
- Remove the necessity of the positivity assumption ($\rho > 0$).
- Function approximation.

Any Questions?

Thank you all for listening! Any questions?