Average-Reward Q-Learning: From Single Agent to Federated Learning

Yuejie Chi

Yale Statistics & Data Science

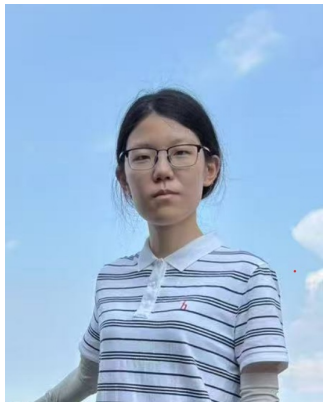
Joint Statistics Meeting

August 2026

In collaboration with

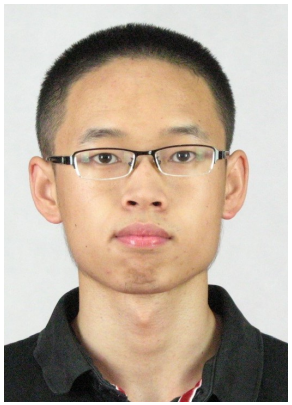
Jiin Woo

CMU



Yuchen Jiao

CUHK



Gen Li

CUHK

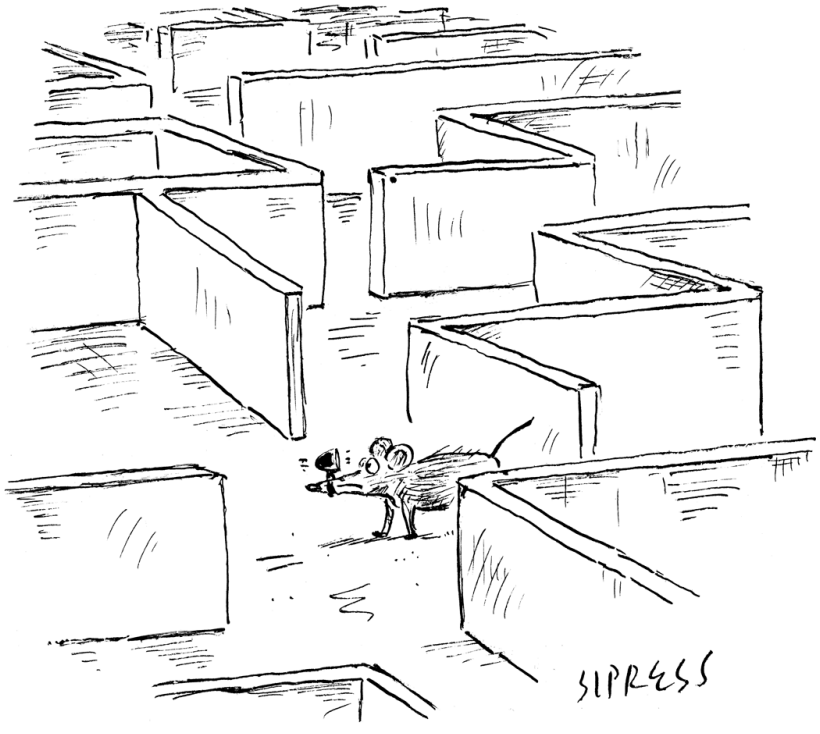
Gauri Joshi

CMU

"Sample Complexity of Average-Reward Q-Learning: From Single-agent to Federated Reinforcement Learning", arXiv:2601.13642.

Reinforcement learning (RL)

In RL, an agent learns by interacting with an *unknown* environment through trial-and-error to maximize long-term total reward.

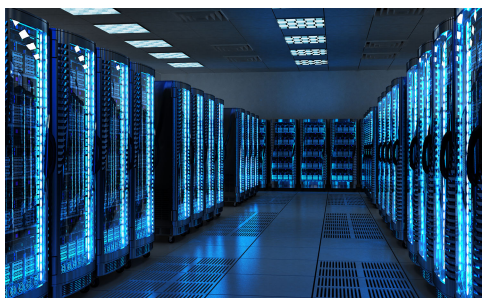


"Recalculating ... recalculating ..."

Average-reward RL

Many decision systems operate continuously—there is no natural terminal time at which performance resets.

Cloud resource allocation



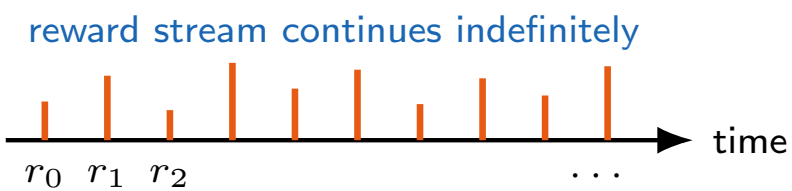
jobs served versus energy consumed

Autonomous driving

safe progress over sustained operation

Communication networks

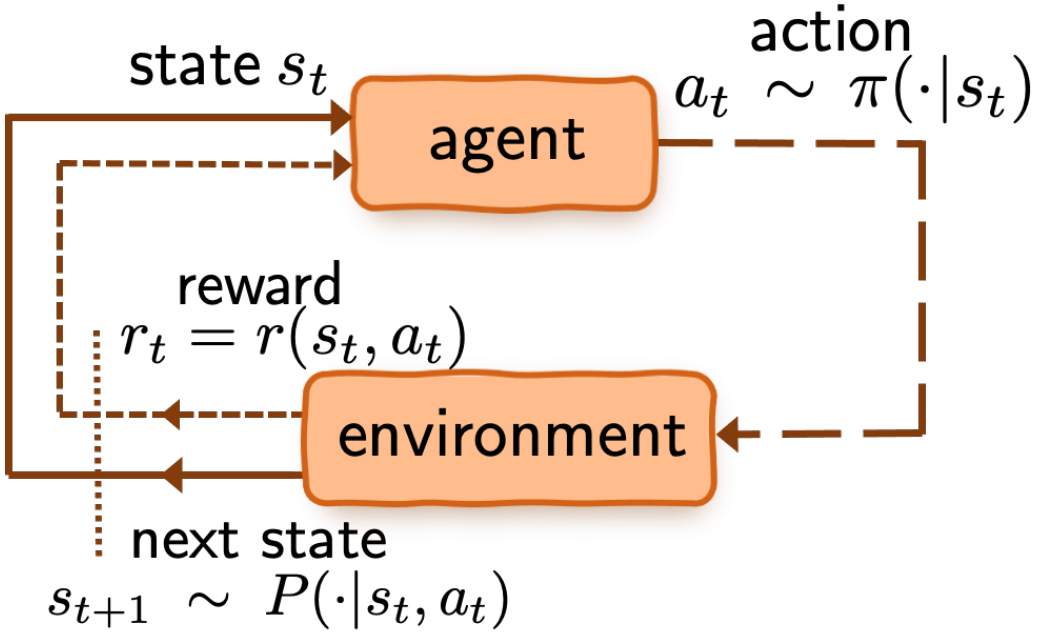
throughput and latency over time

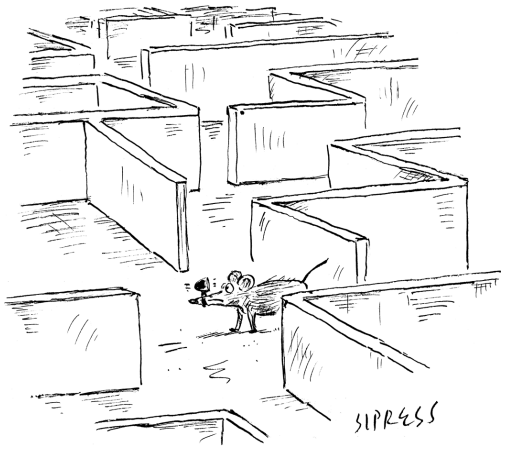


$$J^\pi = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\left[\sum_{t=0}^{T-1} r_t\right]$$

Optimize sustained performance per unit time.

Markov decision process (MDP)

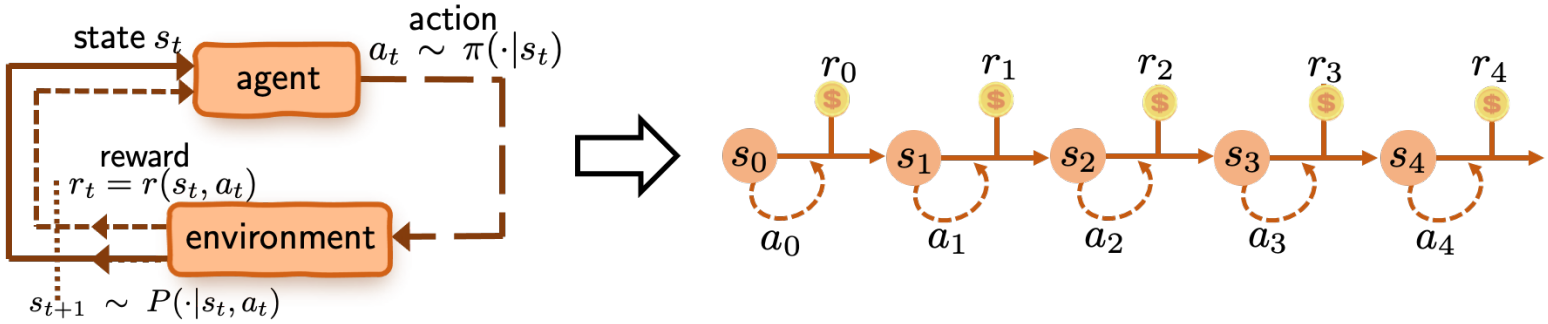




"Recalculating ... recalculating ..."

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot \mid s)$: policy (or action selection rule)
- $P(\cdot \mid s, a)$: transition probabilities

From discounted return to average reward

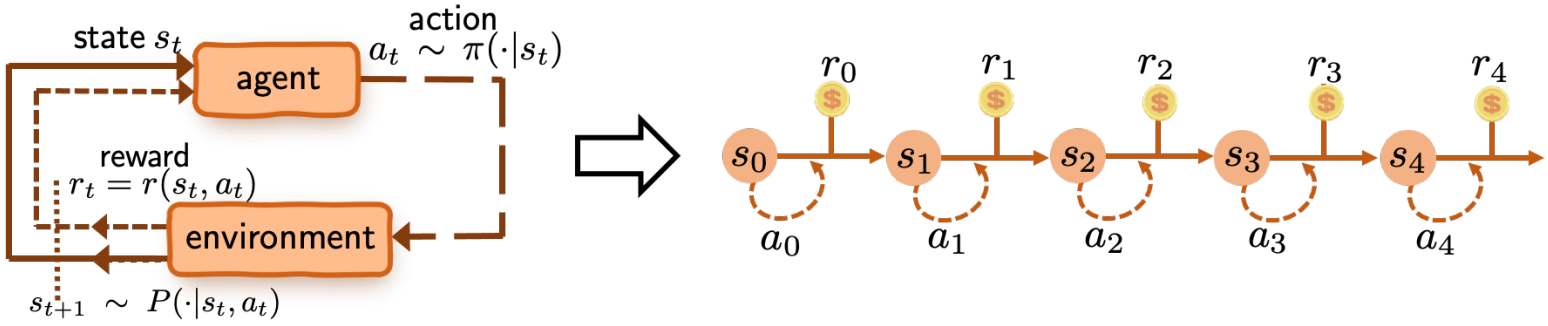


Discounted value function of policy π:

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

- The value function $0 \le V^\pi(s) \le 1$.
- Rewards t steps ahead receive weight $(1-\gamma)\gamma^t$.
- The effective planning horizon is $H_\gamma = (1-\gamma)^{-1}$.

From discounted return to average reward



Average-reward value function of policy π:

$$\forall s \in \mathcal{S}: \qquad J^\pi(s) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r_t \,\middle|\, s_0 = s\right].$$

- Every time step contributes equally to long-run performance.
- Measures steady-state reward per step, with no chosen horizon.

Goal: find a policy $\pi^\star$ maximizing $J^\pi(s)$, with $J^\star = J^{\pi^\star}$.

The bias function and optimality

The bias function h^π measures transient advantage relative to the long-run average:

$$h^\pi(s) = \mathrm{C}\text{–}\lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(r_t - J^\pi\right)\right],$$

where $\mathrm{C}\text{–}\lim$ is the Cesaro limit.

Bellman equation. For a **weakly communicating** MDP, $J^\star$ is independent of the initial state and satisfies

$$J^\star + h^\star(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \sum_{s'\in\mathcal{S}} P(s' \mid s,a) h^\star(s') \right\},$$

Challenge: no contraction property.

Approximation via discounted Q-functions

Discounted Q-function: for every $\gamma < 1$, the optimum Q-function $Q^\star_\gamma$ is the unique fixed point $Q^\star_\gamma = \mathcal{T}_\gamma(Q^\star_\gamma)$:

$$T_\gamma(Q) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_a Q(s',a)$$

- As $\gamma \to 1$, the normalized discounted optimum approaches the optimal average reward:

$$Q^\star_\gamma(s,a) \longrightarrow J^\star, \qquad \left\| Q^\star_\gamma - J^\star \mathbb{1} \right\|_\infty \leq 4(1-\gamma)\|h^\star\|_{\mathrm{sp}}.$$

- *A direct approach:* choose one sufficiently large γ so that the approximation error is at most ε:

$$1 - \gamma \lesssim \frac{\varepsilon}{\|h^\star\|_{\mathrm{sp}}}.$$

Approximation via discounted Q-functions

Discounted Q-function: for every $\gamma < 1$, the optimum Q-function $Q^\star_\gamma$ is the unique fixed point $Q^\star_\gamma = \mathcal{T}_\gamma(Q^\star_\gamma)$:

$$T_\gamma(Q) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_a Q(s',a)$$

- As $\gamma \to 1$, the normalized discounted optimum approaches the optimal average reward:

$$Q^\star_\gamma(s,a) \longrightarrow J^\star, \qquad \left\| Q^\star_\gamma - J^\star \mathbb{1} \right\|_\infty \leq 4(1-\gamma)\|h^\star\|_{\mathrm{sp}}.$$

- *A direct approach:* choose one sufficiently large γ so that the approximation error is at most ε:

$$1 - \gamma \lesssim \frac{\varepsilon}{\|h^\star\|_{\mathrm{sp}}}.$$

But what is the price of learning with such a long horizon?

Why not choose one discount factor close to one?

To estimate $J^\star$, decompose the error through a discounted optimum:

$$\|Q_t - J^\star \mathbb{1}\|_\infty \leq \underbrace{\|Q_t - Q^\star_\gamma\|_\infty}_{\text{discounted Q-learning error}} + \underbrace{\|Q^\star_\gamma - J^\star \mathbb{1}\|_\infty}_{\leq 4(1-\gamma)\|h^\star\|_{\mathrm{sp}}}.$$

- A smaller $1-\gamma$ reduces the approximation bias.
- But the effective horizon $H_\gamma = (1-\gamma)^{-1}$ grows, making the discounted problem harder to learn.
- Choosing γ extremely close to one from the start wastes samples on a poorly initialized long-horizon problem.

Why not choose one discount factor close to one?

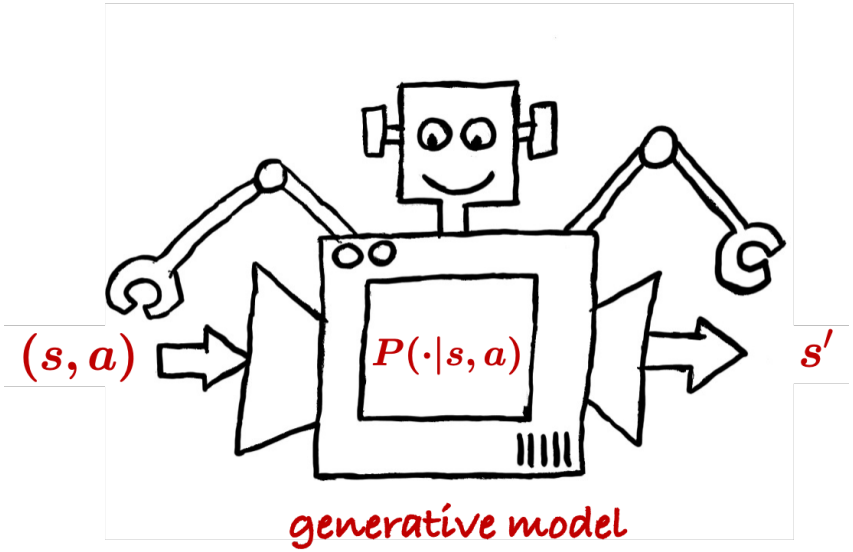
To estimate $J^\star$, decompose the error through a discounted optimum:

$$\|Q_t - J^\star \mathbb{1}\|_\infty \leq \underbrace{\|Q_t - Q^\star_\gamma\|_\infty}_{\text{discounted Q-learning error}} + \underbrace{\|Q^\star_\gamma - J^\star \mathbb{1}\|_\infty}_{\leq 4(1-\gamma)\|h^\star\|_{\mathrm{sp}}}.$$

- A smaller $1-\gamma$ reduces the approximation bias.
- But the effective horizon $H_\gamma = (1-\gamma)^{-1}$ grows, making the discounted problem harder to learn.
- Choosing γ extremely close to one from the start wastes samples on a poorly initialized long-horizon problem.

Our solution: increase γ_k stage by stage and warm-start each discounted problem.

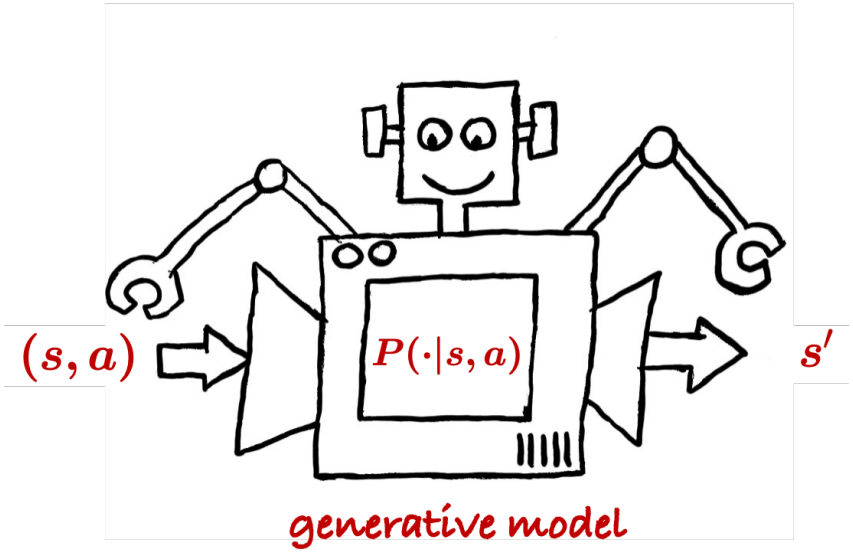
Discounted synchronous Q-learning



Stochastic approximation for solving Bellman equation $Q^\star = \mathcal{T}(Q^\star)$ using samples collected from the generative model:

$$\underbrace{Q_{t+1}(s,a) = (1-\eta)Q_t(s,a) + \eta \mathcal{T}_t(Q_t)(s,a),}_{\text{draw the transition } (s,a,s') \text{ for all } (s,a)} \quad t \geq 0$$

Discounted synchronous Q-learning



Stochastic approximation for solving Bellman equation $Q^\star = \mathcal{T}(Q^\star)$ using samples collected from the generative model:

$$\underbrace{Q_{t+1}(s,a) = (1-\eta)Q_t(s,a) + \eta \mathcal{T}_t(Q_t)(s,a),}_{\text{draw the transition } (s,a,s') \text{ for all } (s,a)} \quad t \geq 0$$

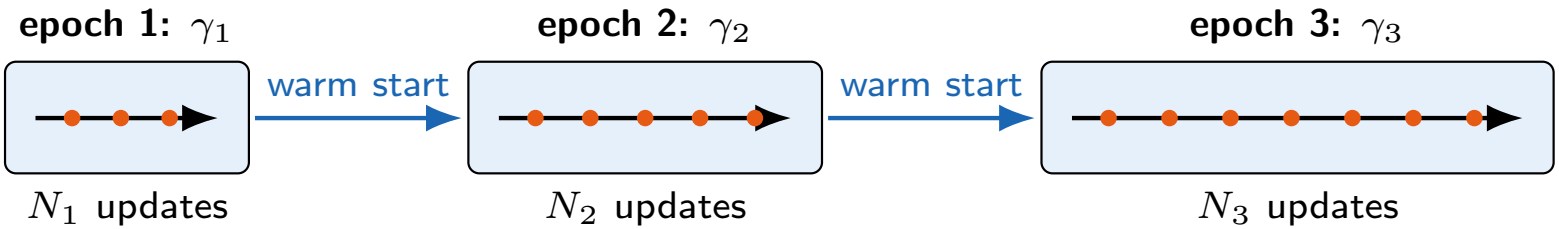
$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

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

Stage-wise average-reward Q-learning

Instead of committing to one very long horizon, solve a sequence of increasingly long discounted problems.

$$\gamma_1 < \gamma_2 < \gamma_3 \uparrow 1 \quad \Longleftrightarrow \quad H_k = (1-\gamma_k)^{-1} \nearrow$$



At epoch k:

1. initialize from the previous epoch, $Q_{k,0} = Q_{k-1,N_{k-1}}$;
2. run N_k synchronous Q-updates using discount factor γ_k:

$$\begin{aligned} Q_{k,t}(s,a) = (1-\eta_{k,t})Q_{k,t-1}(s,a) \\ + \eta_{k,t}\Big[(1-\gamma_k)r(s,a) + \gamma_k V_{\iota(k,t)}(s_{k,t}(s,a))\Big], \end{aligned}$$

where $\iota(k,t)$ is some proper historical index.

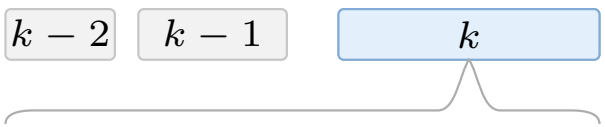
Two update strategies

same update rule $\quad Q_{k,t} \leftarrow (1-\eta_{k,t})Q_{k,t-1} + \eta_{k,t}\left[(1-\gamma_k)r + \gamma_k V_{\iota(k,t)}(s')\right]$

Group 1: polynomial rates

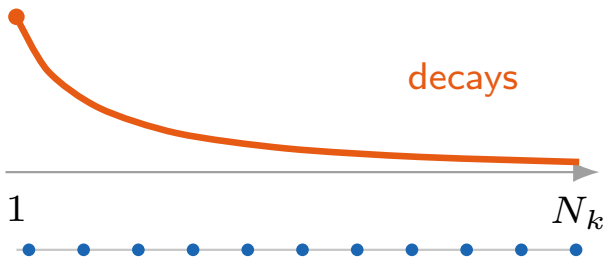
$N_k = \Theta(2^k)$

EPOCHS
not to scale



LEARNING RATE
$t = 1, \ldots, N_k$

$\eta_{k,t} = \widetilde{\Theta}(t^{-2/3})$



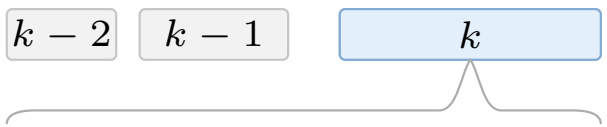
$V_{k,t-1}$: refresh every update

DISCOUNT

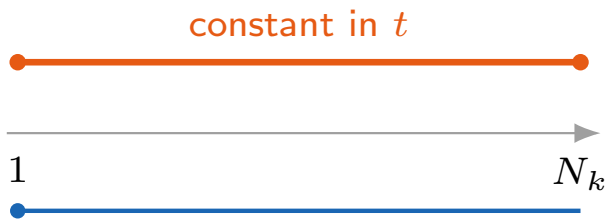
$\gamma_k \uparrow 1: \quad 1-\gamma_k = \dfrac{2\log(4N_k)}{N_k^{1/3}}$

Group 2: constant rates

$N_k = \widetilde{\Theta}(k^2)$



$\eta_{k,t} = \eta_k = \widetilde{\Theta}(N_k^{-1})$



$V_{k-1,N_{k-1}}$: held fixed

$\gamma_k \uparrow 1: \quad 1-\gamma_k = \dfrac{1}{k+1}$

Our theorem

Theorem (Single-agent average-reward Q-learning)

Assume $\|h^\star\|_{\mathrm{sp}} < \infty$. *For either parameter group, with probability at least* $1-\delta$,

$$\|Q_{K,N_K} - J^\star \mathbb{1}\|_\infty \le \frac{C\|h^\star\|_{\mathrm{sp}}}{T_K^{1/3}} \log^4\left(\frac{4|\mathcal{S}||\mathcal{A}|T_K}{\delta}\right), \qquad T_K = \sum_{k=1}^{K} N_k.$$

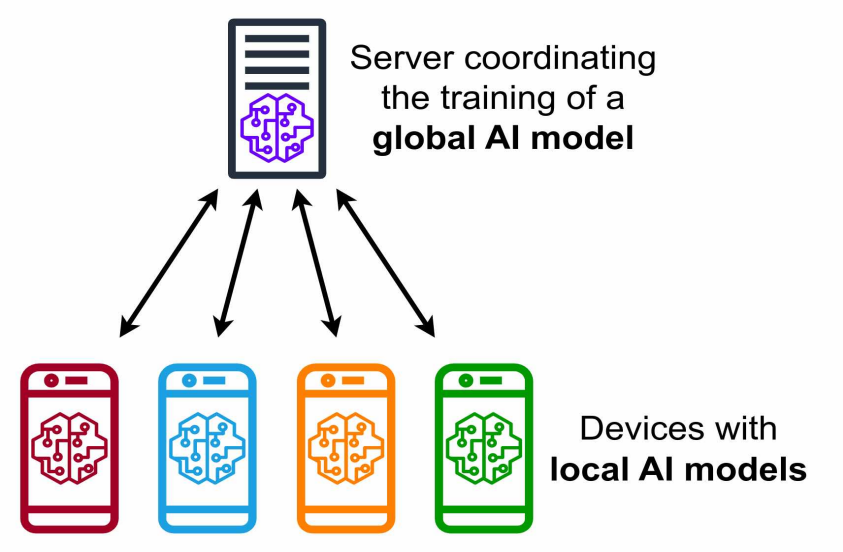
Consequently, to achieve ε*-accuracy, it needs at most*

$$|\mathcal{S}||\mathcal{A}|T_K = \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\|h^\star\|_{\mathrm{sp}}^3}{\varepsilon^3}\right) \qquad \textit{samples.}$$

- Improvement over prior average-reward Q-learning (Jin et al. 2024) by a factor $\widetilde{O}(\|h^\star\|_{\mathrm{sp}}^2/\varepsilon^2)$.
- Best known guarantee for *vanilla* Q-learning in weakly communicating average-reward MDPs.

Federated average-reward Q-learning

Can we harness the power of federated learning?



FORBES > INNOVATION > AI

IBM Federated Learning Research – Extracting Machine Learning Models From Multiple Data Pools

Kevin Krewell Contributor
Tirias Research Contributor Group

Follow

Oct 15, 2021, 02:51pm EDT

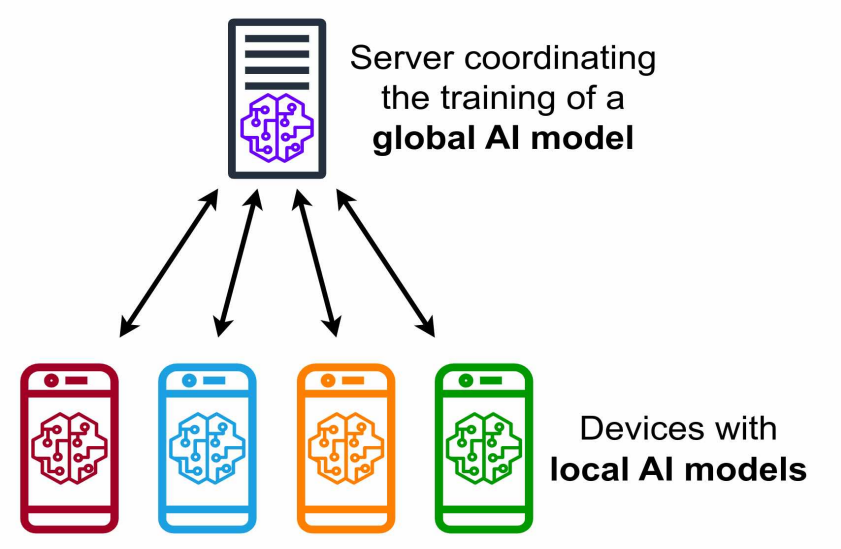
How Apple personalizes Siri without hoovering up your data

The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

By Karen Hao

December 11, 2019

Can we harness the power of federated learning?



FORBES > INNOVATION > AI

IBM Federated Learning Research – Extracting Machine Learning Models From Multiple Data Pools

Kevin Krewell Contributor
Tirias Research Contributor Group

Follow

Oct 15, 2021, 02:51pm EDT

How Apple personalizes Siri without hoovering up your data

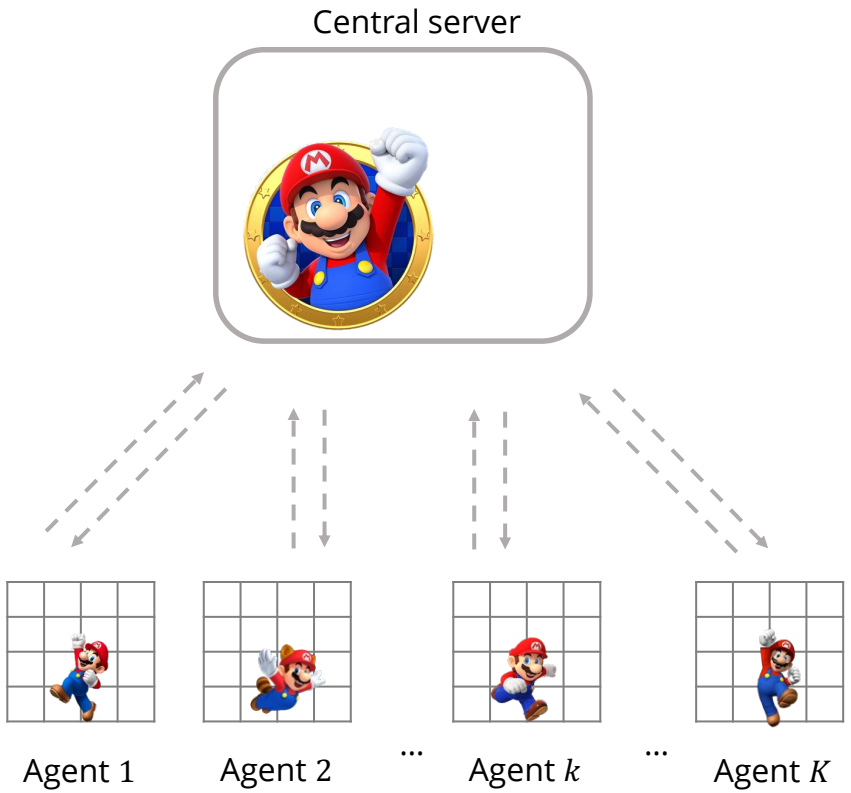
The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

By Karen Hao

December 11, 2019

Can we harness the power of federated learning for RL?

RL meets federated learning



Federated reinforcement learning: enables multiple agents to collaboratively learn a global policy without sharing datasets.

Federated synchronous Q-learning with local updates

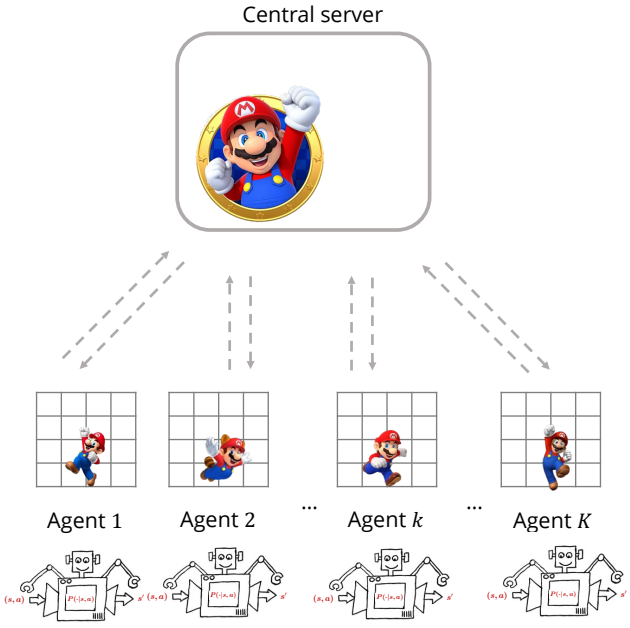
- **Each agent** m performs τ rounds of local Q-learning:

$$Q_{t+1}^m \leftarrow (1-\eta)Q_t^m + \eta \mathcal{T}_{\gamma,t}^m(Q_t^m),$$

then sends its Q-table to the server.

- **The server** averages and broadcasts the global estimate:

$$Q_t = \frac{1}{M}\sum_{m=1}^{M} Q_t^m.$$

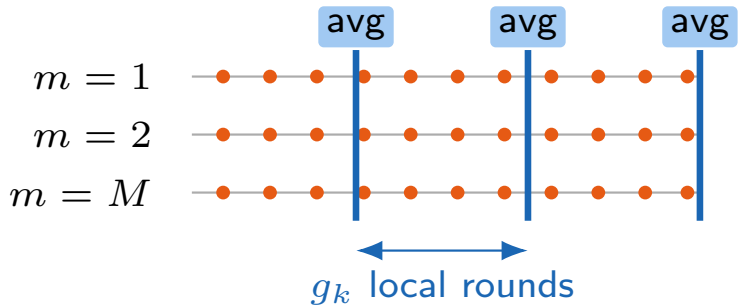


Can we achieve linear speedup with low communication complexity?

Federated update schedules

Group 1: polynomial rates

periodic averages within epoch k



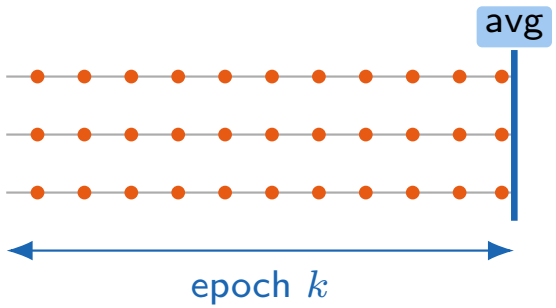
SYNC RULE $g_k = \widetilde{\Theta}\big((N_k^2/M)^{1/3}\big)$

LOCAL ROUNDS $N_k = \Theta(2^k)$

DISCOUNT $\gamma_k = 1 - \dfrac{2\log(4MN_k)}{(MN_k)^{1/3}}$

Group 2: constant rates

one average at the epoch boundary



$\mathcal{C}(k) = \{N_k\}$

$N_k = \widetilde{\Theta}\big(\max\{k^2/M, k\}\big)$

$\gamma_k = \dfrac{k}{k+1}$

Our theorem

Theorem (Federated average-reward Q-learning)

Assume $\|h^\star\|_{\mathrm{sp}} < \infty$. *For either federated schedule, with probability at least* $1-\delta$,

$$\|Q_{K,N_K} - J^\star \mathbb{1}\|_\infty \le \frac{C\|h^\star\|_{\mathrm{sp}}}{(MT_K)^{1/3}} \log^4\left(\frac{4M|\mathcal{S}||\mathcal{A}|T_K}{\delta}\right).$$

Hence the sample complexity per agent is

$$|\mathcal{S}||\mathcal{A}|T_K = \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\|h^\star\|_{\mathrm{sp}}^3}{M\varepsilon^3}\right).$$

The number of communication rounds is

$$\sum_{k=1}^{K} |\mathcal{C}(k)| = \widetilde{O}\left(\frac{\|h^\star\|_{\mathrm{sp}}}{\varepsilon}\right).$$

- Linear speedup with low communication cost independent of M.

Learning an optimal policy

Set $\widehat{\pi}(s) \in \arg\max_a Q_{K,N_K}(s,a)$ and use a policy-learning schedule with $1-\gamma_k \asymp (MN_k)^{-1/5}$.

Theorem (Federated policy learning)

With probability at least $1-\delta$, $J^\star - J^{\widehat{\pi}} \leq \varepsilon$ *using per-agent sample complexity*

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\|h^\star\|_{\mathrm{sp}}^5}{M\varepsilon^5}\right),$$

and communication complexity

$$\widetilde{O}\left(\frac{\|h^\star\|_{\mathrm{sp}}}{\varepsilon}\right).$$

The key conversion is

$$J^\star - J^{\widehat{\pi}} \leq 8(1-\gamma_K)\|h^\star\|_{\mathrm{sp}} + \frac{4\|Q^\star_{\gamma_K} - Q_{K,N_K}\|_\infty}{1-\gamma_K},$$

which explains the fifth-order dependence after balancing the two terms.

Summary: average-reward Q-learning

- **Single-agent Q-learning:**

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\|h^\star\|_{\mathrm{sp}}^3}{\varepsilon^3}\right) \quad \text{samples for value estimation.}$$

- **Federated Q-learning:**

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\|h^\star\|_{\mathrm{sp}}^3}{M\varepsilon^3}\right) \quad \text{samples per agent.}$$

$$\widetilde{O}\left(\frac{\|h^\star\|_{\mathrm{sp}}}{\varepsilon}\right) \quad \text{communication rounds, independent of } M.$$

Careful discount, learning-rate, and communication schedules are sufficient for sample-efficient average-reward Q-learning.

Thanks!

- "Sample Complexity of Average-Reward Q-Learning: From Single-agent to Federated Reinforcement Learning", arXiv:2601.13642.

https://yuejiechi.github.io/