Federated Reinforcement Learning: Statistical and Communication Trade-offs

Yuejie Chi

Carnegie Mellon University

L4DC
June 5, 2025

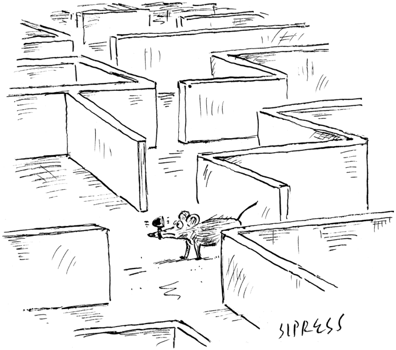
Jiin Woo
CMU

Sudeep Salgia
CMU

Gauri Joshi
CMU

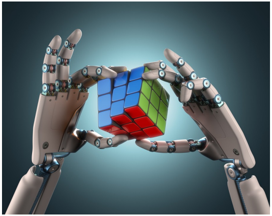
Reinforcement learning (RL)

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

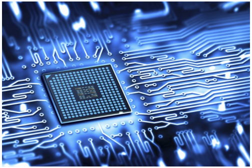


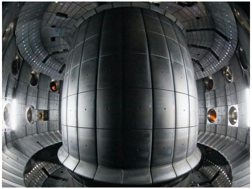
“Recalculating ... recalculating ...”

More successes of RL since AlphaGo


robotics


strategic games

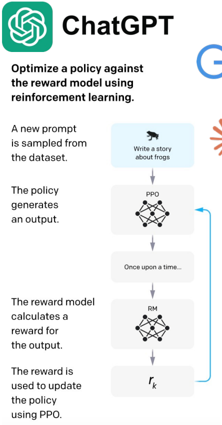

chip designs


nuclear plant control


resource management


UAV and drones

One more: RL for foundation models



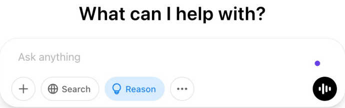
Gemini

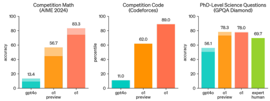
Claude

Meta AI

•
•
•

deepseek





Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\ldots$
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$
We started with the equation:
$\sqrt{a - \sqrt{a+x}} = x$
First, let's square both sides:
$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$
Next, I could square both sides again, treating the equation: $\cdots$
$\ldots$

Alignment: safety, human value..

Reasoning: math, coding…

Turing Award Goes to 2 Pioneers of Artificial Intelligence

Andrew Barto and Richard Sutton developed reinforcement learning, a technique vital to chatbots like ChatGPT.

RL holds great promise in accelerating scientific, engineering and societal discoveries.

Sample efficiency

However, collecting data samples might be expensive or time-consuming.

clinical trials

Prompt:
Should I add chorizo
to my paella?

Response 1: Absolutely! ...
Response 2: In Valencian...

Feedback (ranking):
Response 1 is better than 2

LLM alignment

autonomous driving

Sample efficiency

However, collecting data samples might be expensive or time-consuming.

clinical trials

Prompt:
Should I add chorizo
to my paella?

Response 1: Absolutely! ...
Response 2: In Valencian...

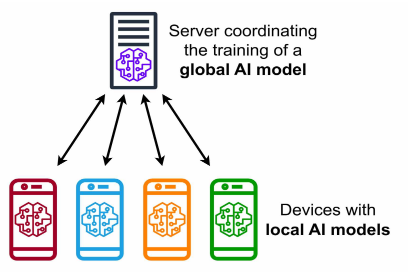
Feedback (ranking):
Response 1 is better than 2

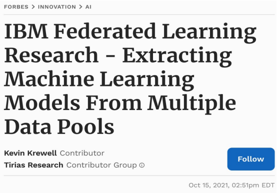
LLM alignment

autonomous driving

Calls for design of sample-efficient RL algorithms!

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

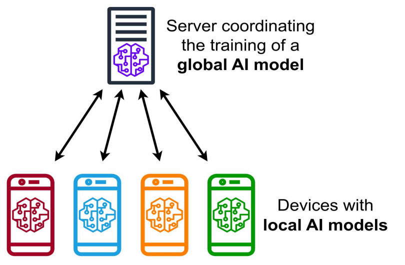
How Apple personalizes Siri without hoovering up your data

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

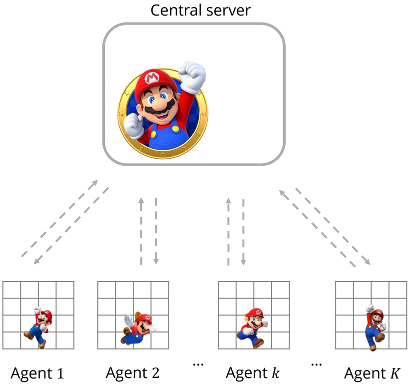
The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

By Karen Hao

December 11, 2019

Can we harness the power of federated learning for RL?

RL meets federated learning

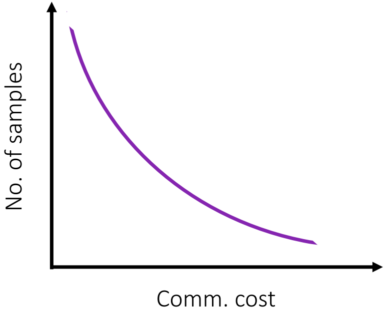


Federated reinforcement learning: enables multiple agents to collaboratively learn a global policy without sharing datasets.

Statistical-communication trade-offs

Statistical benefits

Communication overhead



Is linear speedup possible? What is the price in communication?

This talk: federated RL

<u>Statistical benefits</u>

<u>Communication overhead</u>

Linear speedup:

Can we achieve linear speedup when learning with multiple agents?

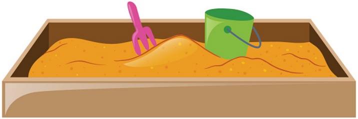
Communication efficiency:

What is the minimum amount of communication to achieve speedup?

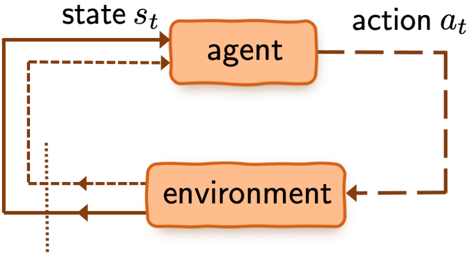
Taming heterogeneity:

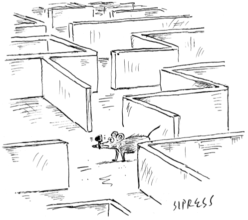
What if the agents are heterogeneous?

Backgrounds:
Markov decision processes and Q-learning



Markov decision process (MDP)

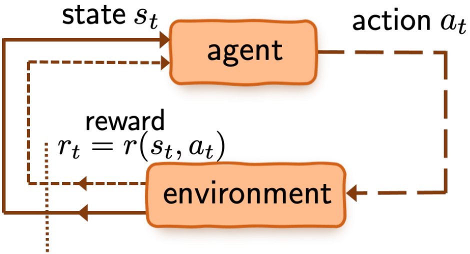


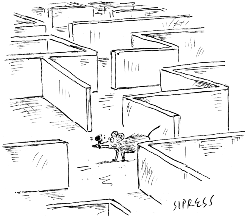


"Recalculating ... recalculating ..."

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

Markov decision process (MDP)

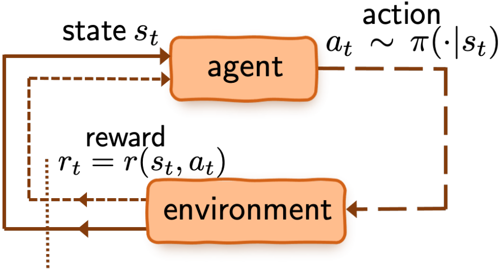


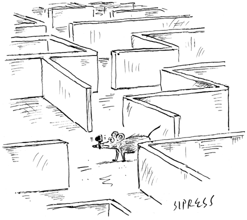


"Recalculating ... recalculating ..."

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

Markov decision process (MDP)

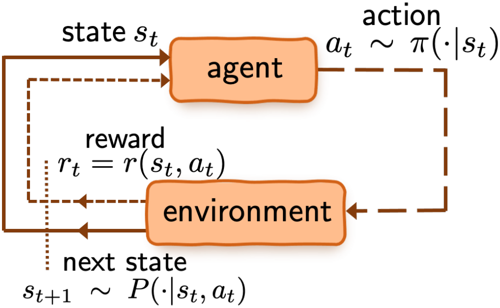


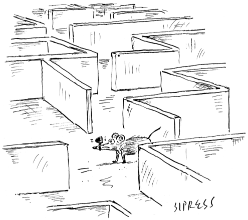


"Recalculating ... recalculating ..."

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

Markov decision process (MDP)

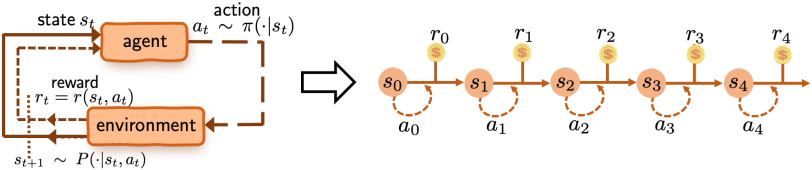




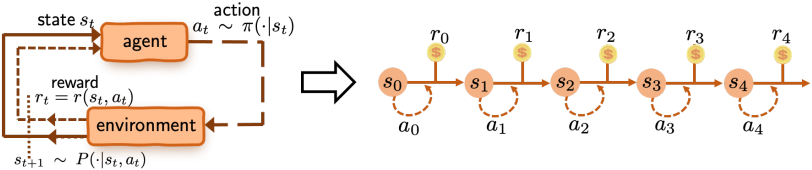
"Recalculating ... recalculating ..."

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



Value function



Value function of policy π:

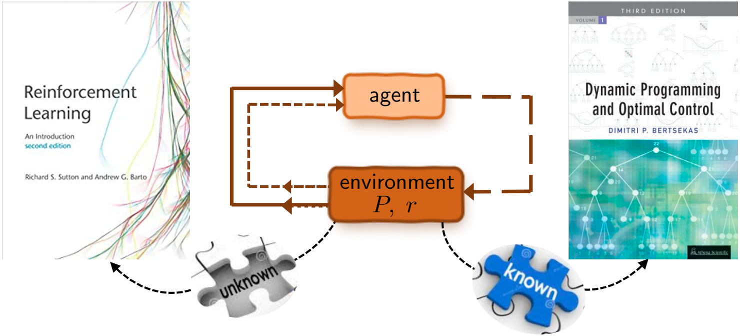
$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

Q-function of policy π:

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

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

Searching for the optimal policy



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

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$
- optimal policy $\pi^\star(s) = \mathrm{argmax}_{a \in \mathcal{A}} Q^\star(s, a)$

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro, 1951}}$ for solving the **Bellman equation**

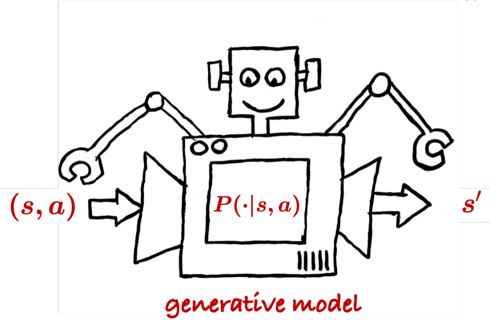
$$Q^\star = \mathcal{T}(Q^\star)$$

where

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\underbrace{\max_{a' \in \mathcal{A}} Q(s',a')}_{\text{next state's value}} \Big].$$

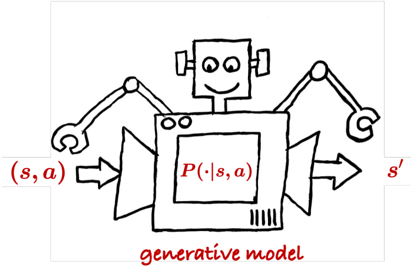
A generative model / simulator

— *Kearns & Singh, 1999*



Each iteration, draw an independent sample (s, a, s') for given (s, a)

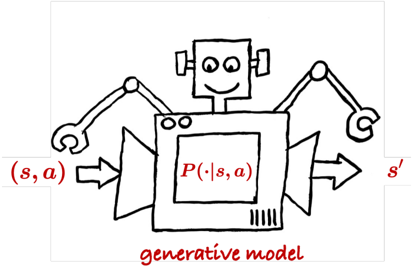
Q-learning with a generative model



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

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

Q-learning with a generative model



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

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

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

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

A sharp sample complexity of Q-learning

Question: How many samples are needed for $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$?

A sharp sample complexity of Q-learning

Question: How many samples are needed for $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$?

Theorem (Li, Cai, Chen, Wei, Chi, OR 2024)

For any $0 < \varepsilon \leq 1$, Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with sample complexity at most

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

Furthermore, this bound is tight for Q-learning.

- This is a factor of $\frac{1}{1-\gamma}$ away from the minimax lower bound, which is $\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$.
- The lower bound is based on analyzing the dynamic of Q-learning on a specific worst-case instance.

Federated Q-learning: towards linear speedup

Jiin Woo
CMU

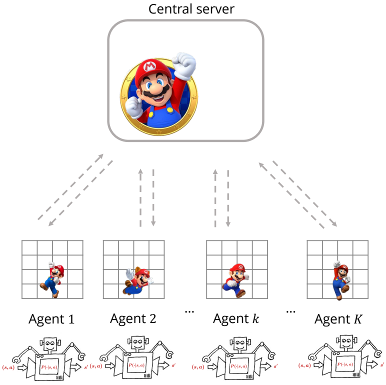
Gauri Joshi
CMU

Federated Q-learning with local updates

- **Local updates:** the agent k performs τ rounds of local Q-learning updates:

$$Q_{t+1}^{k} \leftarrow (1-\eta)Q_t^k + \eta \mathcal{T}_t(Q_t^k)$$

  and sends it to the server.



Federated Q-learning with local updates

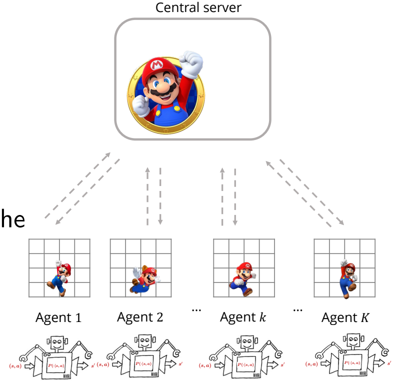
- **Local updates:** the agent k performs τ rounds of local Q-learning updates:

$$Q_{t+1}^k \leftarrow (1-\eta)Q_t^k + \eta \mathcal{T}_t(Q_t^k)$$

and sends it to the server.

- **Periodic averaging:** the server averages the local updates and sends it back to agents:

$$Q_t = \frac{1}{K}\sum_{k=1}^{K} Q_t^k$$



Federated Q-learning with local updates

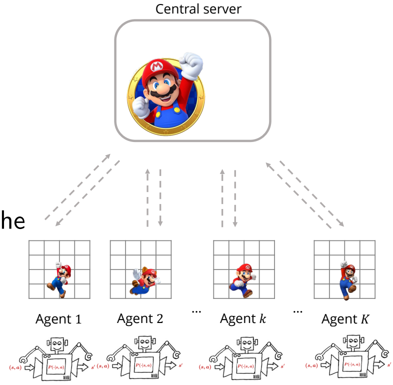
- **Local updates:** the agent k performs τ rounds of local Q-learning updates:

$$Q_{t+1}^k \leftarrow (1-\eta)Q_t^k + \eta \mathcal{T}_t(Q_t^k)$$

and sends it to the server.

- **Periodic averaging:** the server averages the local updates and sends it back to agents:

$$Q_t = \frac{1}{K}\sum_{k=1}^{K} Q_t^k$$



Can we achieve faster convergence, i.e. linear speedup, with low communication complexity?

Federated Q-learning with local updates

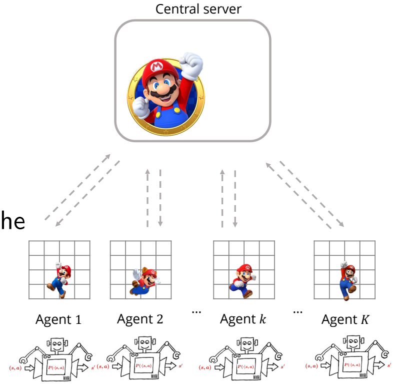
- **Local updates:** the agent k performs τ rounds of local Q-learning updates:

$$Q_{t+1}^k \leftarrow (1-\eta)Q_t^k + \eta\mathcal{T}_t(Q_t^k)$$

and sends it to the server.

- **Periodic averaging:** the server averages the local updates and sends it back to agents:

$$Q_t = \frac{1}{K}\sum_{k=1}^{K} Q_t^k$$



Can we achieve faster convergence, i.e. linear speedup, with low communication complexity?

Yes!!

Linear speedup of federated Q-learning

Theorem (Jiin, Joshi, Chi, ICML 2023)

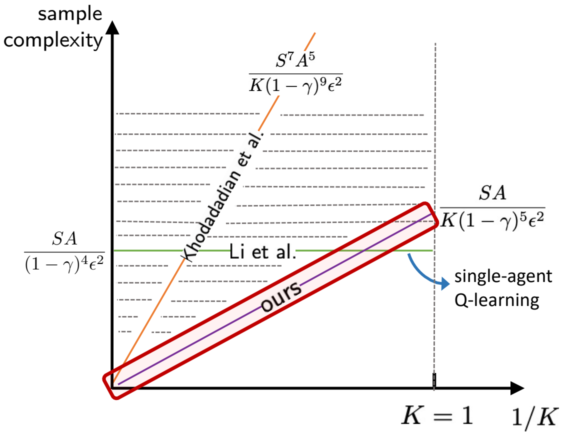
For any $0 < \varepsilon \le \frac{1}{1-\gamma}$, *federated synchronous Q-learning yields* $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ *with sample complexity at most*

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^5\varepsilon^2}\right)$$

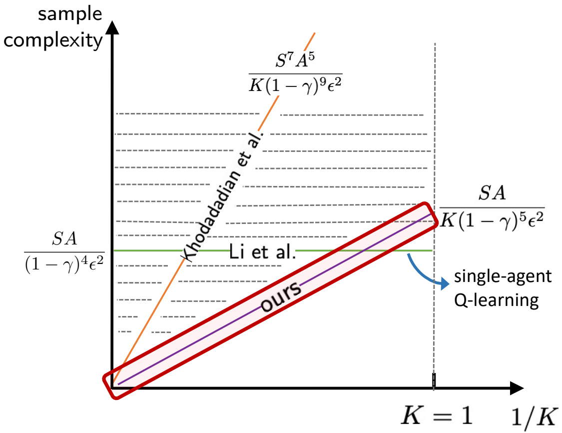
as long as $\tau - 1 \le \frac{1}{\eta}\min\left\{\frac{1-\gamma}{8\gamma}, \frac{1}{K}\right\}$ *and* $\eta = \widetilde{O}(K(1-\gamma)^4\varepsilon^2)$.

- **Linear speedup** compared with the single-agent sample complexity $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right)$.
- **Communication complexity:** ε-independent $T/\tau = \widetilde{O}\left(\frac{K}{1-\gamma}\right)$ for sufficiently small ε.

Comparison with prior art



Comparison with prior art

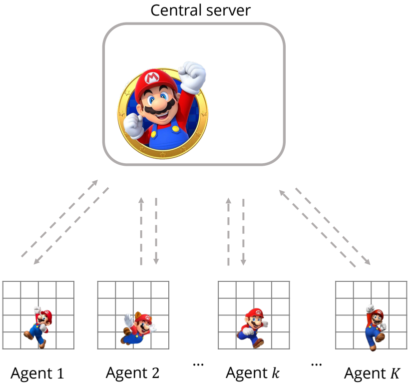


Linear speedup with near-optimal parameter dependencies!

The statistical-communication complexity trade-off in federated Q-Learning

Sudeep Salgia
CMU

Communication bottleneck



The price of communication: how much communication do we need to pay to achieve the linear speedup?

A communication lower bound

Theorem (Salgia and Chi, NeurIPS 2024; informal)

For a wide family of federated Q-learning algorithm with intermittent communication, *regardless of the choice of synchronization schedules, the number of communication rounds needs to be at least*

$$\widetilde{\Omega}\left(\frac{1}{1-\gamma}\right)$$

in order to achieve any speedup *with respect to the number of agents.*

- A similar lower bound holds for the number of communication *bits*.
- This is the first communication complexity barrier established for federated RL algorithms.

Key idea

$$\mathbb{E}[(\widehat{Q} - Q^\star)^2] = \underbrace{\mathbb{E}[(\mathbb{E}[\widehat{Q}] - Q^\star)^2]}_{\text{Bias}} + \underbrace{\mathbb{E}[(\widehat{Q} - \mathbb{E}[\widehat{Q}])^2]}_{\text{Variance}}$$

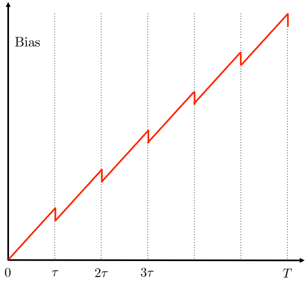
- Variance exhibits linear speedup on average.
- Bias increases between two communication rounds. Averaging has a small compensating effect, but the overall bias is independent of K.

Key idea

$$\mathbb{E}[(\widehat{Q} - Q^\star)^2] = \underbrace{\mathbb{E}[(\mathbb{E}[\widehat{Q}] - Q^\star)^2]}_{\text{Bias}} + \underbrace{\mathbb{E}[(\widehat{Q} - \mathbb{E}[\widehat{Q}])^2]}_{\text{Variance}}$$

- Variance exhibits linear speedup on average.
- Bias increases between two communication rounds. Averaging has a small compensating effect, but the overall bias is independent of K.

$$\text{Bias} \propto \tau = \Omega(T(1-\gamma))$$
$$\Downarrow$$
bias dominates the variance
$$\Downarrow$$
no collaboration gain

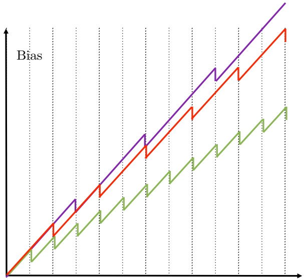


Key idea

$$\mathbb{E}[(\widehat{Q} - Q^\star)^2] = \underbrace{\mathbb{E}[(\mathbb{E}[\widehat{Q}] - Q^\star)^2]}_{\text{Bias}} + \underbrace{\mathbb{E}[(\widehat{Q} - \mathbb{E}[\widehat{Q}])^2]}_{\text{Variance}}$$

- Variance exhibits linear speedup on average.
- Bias increases between two communication rounds. Averaging has a small compensating effect, but the overall bias is independent of K.

$$\text{Bias} \propto \tau = \Omega(T(1-\gamma))$$
$$\Downarrow$$
bias dominates the variance
$$\Downarrow$$
no collaboration gain



Near-optimal algorithm design

Can one design a federated Q-learning algorithm that simultaneously offers optimal-order sample and communication complexities?

Near-optimal algorithm design

Can one design a federated Q-learning algorithm that simultaneously offers optimal-order sample and communication complexities?

Yes!!

Near-optimal algorithm design

Can one design a federated Q-learning algorithm that simultaneously offers optimal-order sample and communication complexities?

Yes!!

- **Fed-DVR-Q (Salgia and Chi, NeurIPS 2024):** achieves near-optimal statistical and communication complexities with communication compression and variance reduction:

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^3\varepsilon^2}\right) \text{ samples,} \qquad \widetilde{O}\left(\frac{1}{1-\gamma}\right) \text{ rounds.}$$

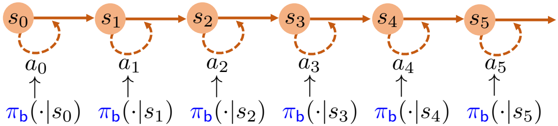
See our paper for details!

Dealing with heterogeneity in federated RL

Jiin Woo
CMU

Gauri Joshi
CMU

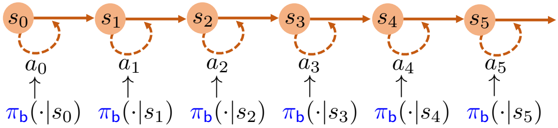
Q-learning following a behavior policy



Stochastic approximation for solving Bellman equation $Q^\star = \mathcal{T}(Q^\star)$ using samples collected from a behavior policy π_b:

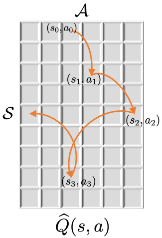
$$\underbrace{Q_{t+1}(s_t, a_t) = (1-\eta)Q_t(s_t, a_t) + \eta \mathcal{T}_t(Q_t)(s_t, a_t)}_{\textit{only}\text{ update } (s_t, a_t)\text{-th entry}}, \quad t \geq 0$$

Q-learning following a behavior policy



Stochastic approximation for solving Bellman equation $Q^\star = \mathcal{T}(Q^\star)$ using samples collected from a behavior policy π_b:

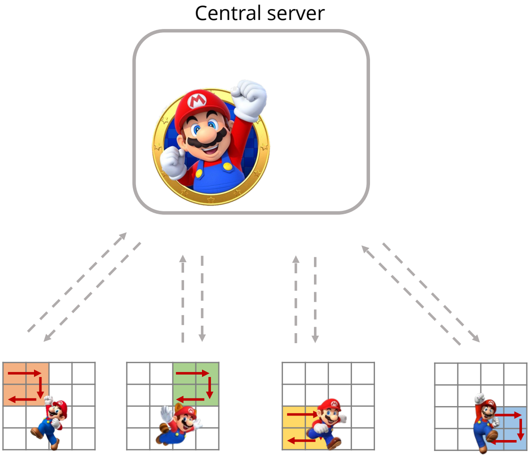
$$\underbrace{Q_{t+1}(s_t, a_t) = (1-\eta)Q_t(s_t, a_t) + \eta \mathcal{T}_t(Q_t)(s_t, a_t),}_{\textit{only}\text{ update } (s_t, a_t)\text{-th entry}} \quad t \ge 0$$



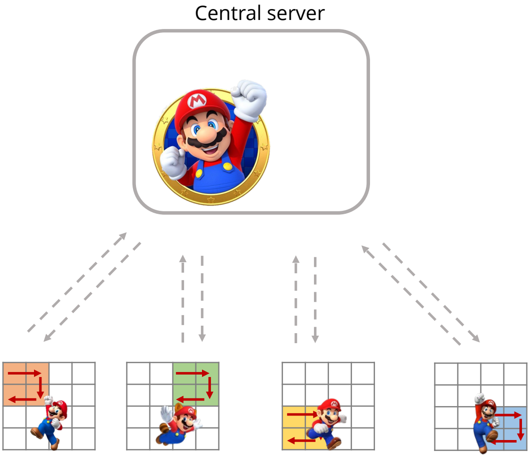
$$\mathcal{T}_t(Q)(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')$$

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

Tackling data heterogeneity



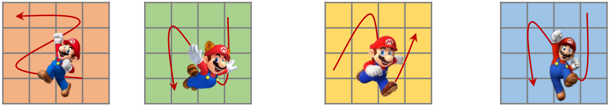
Tackling data heterogeneity



Can we achieve faster convergence with heterogeneous local behavior policies with low communication complexity?

The benefit of collaboration?

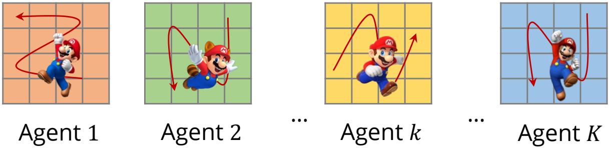
Prior art requires **full coverage** of every agent over the entire state-action space...



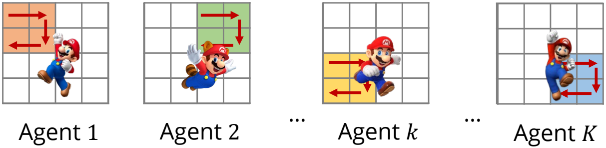
Agent 1 Agent 2 ... Agent k ... Agent K

The benefit of collaboration?

Prior art requires **full coverage** of every agent over the entire state-action space...

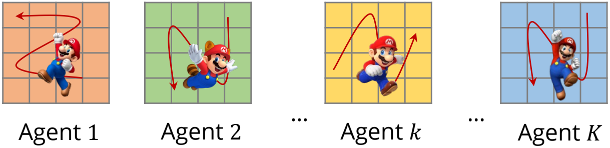


However, the power of collaboration really shines if we only need...

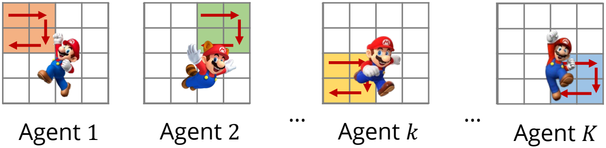


The benefit of collaboration?

Prior art requires **full coverage** of every agent over the entire state-action space...



However, the power of collaboration really shines if we only need...



Is collaborative coverage enough for federated Q-learning?

Key metrics

Collaborative coverage: minimum entry of the average stationary distribution

$$\mu_{\mathsf{avg}} = \min_{s,a} \frac{1}{K} \sum_{k=1}^{K} \mu_{\mathsf{b}}^{k}(s,a).$$

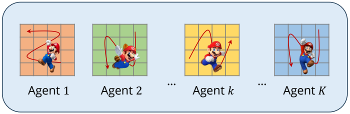
Key metrics

Collaborative coverage: minimum entry of the average stationary distribution

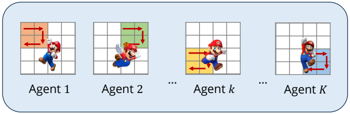
$$\mu_{\mathsf{avg}} = \min_{s,a} \frac{1}{K} \sum_{k=1}^{K} \mu_{\mathsf{b}}^{k}(s,a).$$

Heterogeneity of local behavior policies: density ratio of individual / average behavior policies

$$C_{\mathsf{het}} = K \max_{k,s,a} \frac{\mu_{\mathsf{b}}^{k}(s,a)}{\sum_{k=1}^{K} \mu_{\mathsf{b}}^{k}(s,a)} = \max_{k,s,a} \frac{\mu_{\mathsf{b}}^{k}(s,a)}{\mu_{\mathsf{avg}}(s,a)}.$$



$C_{\mathsf{het}} = 1$



$C_{\mathsf{het}} = K$

Our theorem

Theorem (Woo, Joshi, Chi, JMLR 2025)

For sufficiently small $\varepsilon > 0$, *federated asynchronous Q-learning yields* $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ *with sample complexity at most*

$$\widetilde{O}\left(\frac{C_{\mathsf{het}}}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

ignoring the burn-in cost that depends on the mixing times.

Our theorem

Theorem (Woo, Joshi, Chi, JMLR 2025)

For sufficiently small $\varepsilon > 0$, *federated asynchronous Q-learning yields* $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ *with sample complexity at most*

$$\widetilde{O}\left(\frac{C_{\mathsf{het}}}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

ignoring the burn-in cost that depends on the mixing times.

- Near-optimal linear speedup when the local behavior policies are similar, $C_{\mathsf{het}} \approx 1$.
- Key idea: leave-one-out arguments to decouple statistical dependencies due to Markovian sampling and local updates.

Our theorem

Theorem (Woo, Joshi, Chi, JMLR 2025)

For sufficiently small $\varepsilon > 0$, federated asynchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ with sample complexity at most

$$\widetilde{O}\left(\frac{C_{\mathsf{het}}}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

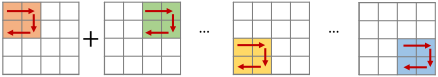
ignoring the burn-in cost that depends on the mixing times.

- Near-optimal linear speedup when the local behavior policies are similar, $C_{\mathsf{het}} \approx 1$.
- Key idea: leave-one-out arguments to decouple statistical dependencies due to Markovian sampling and local updates.

Curse of heterogeneity? Performance degenerates when local behavior policies are heterogeneous (i.e. $1 \ll C_{\mathsf{het}}$). ☹

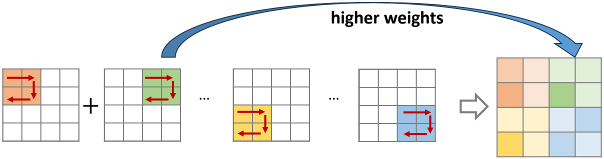
Importance averaging

Key observation: not all updates are of same quality due to limited visits induced by the behavior policy.



Importance averaging

Key observation: not all updates are of same quality due to limited visits induced by the behavior policy.



Importance averaging: the server averages the local updates based on importance via

$$Q_t(s,a) = \frac{1}{K}\sum_{k=1}^{K} \alpha_t^k(s,a) Q_t^k(s,a),$$

where

$$\alpha_t^k = \frac{(1-\eta)^{-N_{t-\tau,t}^k(s,a)}}{\sum_{k=1}^K (1-\eta)^{-N_{t-\tau,t}^k(s,a)}}, \quad N_{t-\tau,t}^k(s,a) = \text{number of visits in the sync period}.$$

Our theorem

Theorem (Woo, Joshi, Chi, JMLR 2025)

For sufficiently small $\varepsilon > 0$, federated asynchronous Q-learning with importance averaging yields $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ with at most

$$\widetilde{O}\left(\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

samples, ignoring the burn-in cost that depends on the mixing times.

Our theorem

Theorem (Woo, Joshi, Chi, JMLR 2025)

For sufficiently small $\varepsilon > 0$, federated asynchronous Q-learning with importance averaging yields $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ with at most

$$\widetilde{O}\left(\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

samples, ignoring the burn-in cost that depends on the mixing times.

- Similar results can be developed for the offline setting, too.

Our theorem

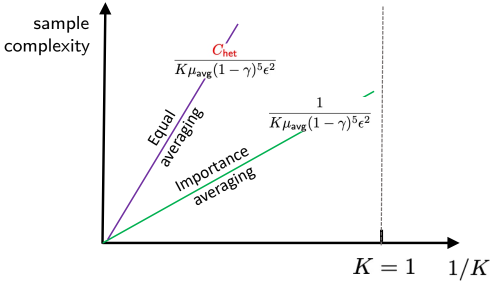
Theorem (Woo, Joshi, Chi, JMLR 2025)

For sufficiently small $\varepsilon > 0$, federated asynchronous Q-learning with importance averaging yields $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ with at most

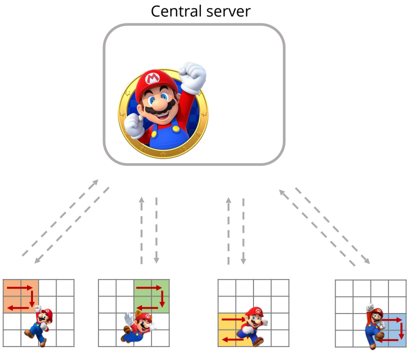
$$\widetilde{O}\left(\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

samples, ignoring the burn-in cost that depends on the mixing times.

- Similar results can be developed for the offline setting, too.

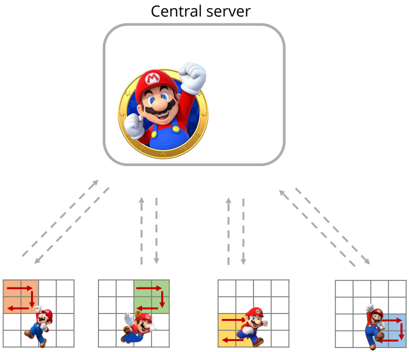


Summary



Synergy of statistics and RL: federated RL unleashes the collaborative power of agents even under heterogeneity!

Summary



Synergy of statistics and RL: federated RL unleashes the collaborative power of agents even under heterogeneity!

Future work:

- Multi-environment and personalized RL.
- Other MDP settings.

Thanks!

- The Blessing of Heterogeneity in Federated Q-Learning: Linear Speedup and Beyond, *JMLR 2025*. Preliminary version at ICML 2023.
- The Sample-Communication Complexity Trade-off in Federated Q-Learning, *NeurIPS 2024*, **oral**.
- Federated Offline Reinforcement Learning: Collaborative Single-Policy Coverage Suffices, *ICML 2024*.

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