PROVABLY VARIANCE-REDUCED FEDERATED CATEGORICAL DISTRIBUTIONAL TD LEARNING WITH LINEAR FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated reinforcement learning (FRL) has been successfully applied in various domains. To date, existing FRL methods focus on estimating the expected return. However, it is difficult to capture the uncertainty and risk, which are very important for handling risk-sensitive decision-making tasks. To address this issue, we propose a variance-reduced federated categorical distributional TD learning with linear function approximation, referred to as FedVR-DistTD, which uses a global control direction and same-sample TD differences to reduce variance. Furthermore, we establish a rigorous theoretical analysis of FedVR-DistTD. Concretely, we prove that, with a fixed per-round control batch, the final server iterate converges geometrically in mean square to a statistical neighborhood whose size is determined by the control-batch variance. Meanwhile, we also establish a finite-sample analysis of FedVR-DistTD that can converge to the optimal return distribution at rate of $\mathcal{O}(1/\sqrt{T})$, where T is the per-client local-iteration budget. Finally, we demonstrate the effectiveness of FedVR-DistTD on a four-state Markov reward process by experiments.

1 INTRODUCTION

Reinforcement learning (RL) has been widely applied in various domains, including game playing and robotic policy learning (Mnih et al., 2015; Levine et al., 2016). Despite the great success of RL, however, persistent obstacles are encountered in practical implementations of RL. In general, the training process needs massive data samples to achieve a desirable performance in large state and action spaces, whereas the training data samples are often scarcely available. Furthermore, the data samples are privacy-sensitive and cannot be shared directly in many applications. Therefore, it is difficult to acquire the sufficient number of data samples for effectively training RL algorithms. In addition, it is also impossible to explore all potential states in a large state space. For these reasons, a collaborative learning framework with RL, which includes a multiple RL agents, is imperative for reducing the risk of data privacy leakage and improving the performance of RL real-world scenarios.

Inspired by the superiority of federated learning (McMahan et al., 2017; Kairouz & McMahan, 2021), federated reinforcement learning (FRL) has been developed to address the above-mentioned challenges. In FRL, multiple agents learn a global policy guided by a centralized server. Concretely, each agent interacts with its local environment to learn a local policy, and then each agent sends its own local policy to the server. The server then aggregates the local policies to obtain a global policy. In this learning process, the raw trajectories remain undisclosed, thus preserving data privacy of agents. In this direction, various FRL algorithms have been proposed in recent years (Khodadadian et al., 2022; Zhu et al., 2026). Specifically, federated TD learning, as a cornerstone FRL framework, has also been developed recently (Wang et al., 2024; Mangold et al., 2024; Zhu et al., 2025). Despite the significant advancements in federated TD learning, however, existing federated TD learning only focuses on estimating the expected returns for guiding the behavior of agents. Consequently, it is difficult to capture the uncertainty and risk, which are very important for handling risk-sensitive decision-making tasks such as healthcare and automatic driving. Moreover, federated averaging of local value estimates targets the expected return and therefore discards information about return variability, tail risk, and multimodal outcomes(Bellemare et al., 2017; Lyle et al., 2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To tackle the above-mentioned issues, integrating distributional TD learning with federated learning framework can offer a promising solution. Specifically, distributional TD learning has been well-studied in recent years (Bellemare et al., 2017; 2023), which provides richer predictive information for capturing uncertainty, tail risk, and potentially multimodal return distributions (Lyle et al., 2019). To accommodate practical settings in risk-sensitive tasks, the non-asymptotic performance guarantees is also essential for distributional TD learning. In this direction, the theoretical analysis of distributional TD learning have been established recently (Peng et al., 2024; Rowland et al., 2024a), which only focus on the tabular settings. In real-world scenarios, however, function approximation is indispensable, especially for large or continuous state spaces (Bellemare et al., 2019). Very recently, the theoretical understanding of distributional TD learning with function approximation has been established in (Peng et al., 2025; Jin et al., 2025). Despite the great progress, however, the aforementioned theoretical results are still established for centralized settings, which is hard to be applied directly to distributed settings that include multiple RL agents. Although recent efforts have been made for federated distributional reinforcement learning (Millard et al., 2026), where the critic-side stability is analyzed theoretically, the non-asymptotic performance guarantees are still barely known. Consequently, there arises an open question.

Can we develop a federated distributional TD learning with function approximation that guarantees statistical efficiency as classic FRL?

Our objective is to provide an affirmative answer to the above-mentioned question. To this end, we consider the linear function approximation due to its simplicity and efficiency (Sutton & Barto, 2018). Despite the simplicity of linear function approximation, however, integrating it with federated distributional TD learning with theoretical guarantees is highly nontrivial due to the following major technical challenges.

(1) Existing federated TD learning methods mainly estimate the expected return and therefore cannot characterize the uncertainty, tail risk, or multimodality of return distributions. Extending federated TD learning to the categorical distributional setting is nontrivial, since the global update must simultaneously account for the distributional Bellman operator, the fixed-support categorical projection, and the linear function approximation architecture. Moreover, the server cannot access or aggregate the raw trajectories collected by different agents. This makes it challenging to construct a federated distributional TD update that preserves the structure of the return distribution while remaining statistically efficient.

(2) The main difficulty lies in characterizing the dependent errors induced by multiple local updates. In each communication round, all local steps reuse the same finite-batch control direction, so their errors share a common control-batch component. Meanwhile, the residual error of the same-sample TD difference depends on the local parameter displacement and therefore acts as parameter-dependent multiplicative noise. The resulting client drift further couples the local recursions. Consequently, standard stochastic approximation arguments based on independent additive noise are not directly applicable, necessitating new techniques for controlling the resulting errors.

(3) Even when the global control direction is conditionally unbiased, establishing a complete finite-sample convergence guarantee remains challenging. The projected linear categorical distributional TD operator is generally nonsymmetric due to the interaction among the Markov transition dynamics, the feature covariance matrix, discounting, and the categorical projection. Moreover, the finite control batch introduces a persistent statistical error, so the last server iterate contracts geometrically only up to a statistical neighborhood of the projected fixed point, whose size depends on the control-batch variance. The main difficulty is therefore to derive a unified two-stage error bound that captures both the geometric behavior of the last iterate and the statistical improvement from tail averaging, and then transfers these parameter-space guarantees to the stationary-weighted Cramér error.

To overcome the above-mentioned challenges, we develop a variance-reduced federated linear categorical distributional TD learning with theoretical guarantee for estimating the distribution of returns, referred to as FedVR-DistTD. To the best of our knowledge, the theoretical analysis of federated distributional linear categorical TD learning is first established by this work. In a nutshell, our main contributions are three-fold.

- We propose FedVR-DistTD, a variance-reduced federated linear categorical distributional TD learning method that combines same-sample TD differences with a globally aggregated

control direction. Specifically, this method reduces update variance without sharing raw trajectories and supports multiple local updates at each client.

- We also establish finite-time guarantees under Markovian sampling, reused control batches, and local client drift. A round-specific auxiliary target separates control-batch error from difference noise, yielding geometric convergence of the last server iterate at rate $\mathcal{O}(\rho^R)$ to a finite-batch neighborhood. Furthermore, FedVR-DistTD converges to an optimal distribution of returns at rate of $\mathcal{O}(\sqrt{E/(TNB)})$.
- We conduct various experiments to verify the theoretical results of FedVR-DistTD. The experimental results demonstrate the effectiveness of FedVR-DistTD versus tail-averaged performance. Moreover, the experimental results support the main convergence predictions.

2 Related Work

Federated policy evaluation for Markov reward processes. Policy evaluation has traditionally relied on temporal-difference learning to estimate expected returns. Linear TD methods have established the fundamental convergence and finite-sample theory for this problem (Sutton, 1988; Tsitsiklis & Van Roy, 1997; Bhandari et al., 2021). However, in many applications, interaction data are collected by multiple agents and cannot be directly centralized because of communication and privacy constraints. Distributed TD methods address this issue by coordinating local estimates through inter-agent communication (Doan et al., 2019), while federated learning allows clients to retain local samples and exchange only model information (McMahan et al., 2017; Kairouz & McMahan, 2021). Building on this framework, federated TD methods have been developed under Markovian sampling, local updates, environmental heterogeneity, and on-policy sampling (Khodadadian et al., 2022; Wang et al., 2024; Zhu et al., 2025; Zhang et al., 2024). Nevertheless, these methods estimate only the expected return. Consequently, they cannot adequately describe return variability, tail risk, or multimodal outcomes, which are important in risk-sensitive decision-making.

Distributional TD with linear categorical approximation. To overcome the limitation of expectation-based evaluation, distributional reinforcement learning estimates the full return distribution rather than only its mean (Bellemare et al., 2017; 2023). Different representations have been developed, including nonparametric, quantile, and implicit-quantile methods (Morimura et al., 2010; Dabney et al., 2018b;a). Categorical distributional TD represents the return distribution on fixed support points and enables analysis through the Cramér geometry (Rowland et al., 2018). Its combination with linear function approximation makes distributional policy evaluation applicable to large state spaces (Bellemare et al., 2019). However, obtaining reliable distributional estimates also requires finite-sample guarantees. Existing analyses cover centralized tabular or generative-model settings (Peng et al., 2024; Rowland et al., 2024b), as well as linear categorical policy evaluation (Peng et al., 2025), where samples and updates are centrally accessible. These assumptions are difficult to maintain when data are distributed across multiple clients.

Federated distributional TD and variance reduction. The combination of federated learning and distributional TD remains relatively underexplored. Millard et al. (Millard et al., 2026) proposed a federated distributional RL method with quantile critics and a Wasserstein trust region, but did not address linear categorical TD or finite-sample parameter guarantees. For variance reduction, local-update and proximal methods study the communication–heterogeneity tradeoff, while SCAFFOLD uses control variates to mitigate client drift (Stich, 2019; Li et al., 2020; Karimireddy et al., 2020); SCAFFLSA extends this idea to federated linear stochastic approximation and expectation-based TD (Mangold et al., 2024), and VrFLCTD improves the finite-sample efficiency of centralized linear categorical TD through reference-point recentering and same-batch direction differences (Jin et al., 2025). These techniques provide useful variance-reduction ingredients, but their interaction with categorical distributional Bellman updates under federated local computation remains unexplored. Consequently, no existing method simultaneously addresses federated data locality, categorical distributional representation, repeated local updates, and non-asymptotic convergence, motivating FedVR-DistTD.

Table 1: Comparison of policy-evaluation methods and convergence rates.

Method	Setting	Representation	VR	Rate
Khodadadian et al. (2022)	Federated	Exp./Lin.	✗	Finite-time; linear speedup
Wang et al. (2024)	Federated	Exp./Lin.	✗	$\widetilde{\mathcal{O}}((NKT)^{-1/2}) + \text{bias}$
Mangold et al. (2024)	Federated	Exp./Lin.	✓	$\widetilde{\mathcal{O}}((NS)^{-1/2})$
Bellemare et al. (2019)	Centralized	Dist./Cat.-Lin.	✗	Convergence guarantee
Peng et al. (2025)	Centralized	Dist./Cat.-Lin.	✗	$\widetilde{\mathcal{O}}(T^{-1/2})$
Jin et al. (2025)	Centralized	Dist./Cat.-Lin.	✓	$\mathcal{O}(2^{-n/2} + M^{-1/2})$
Millard et al. (2026)	Federated	Dist./Quantile	✗	Contractive update
FedVR-DistTD (ours)	**Federated**	**Dist./Cat.-Lin.**	✓	**Last:** $\rho^{R/2+\mathcal{O}((NB)^{-1/2})}$ **Tail:** $\mathcal{O}(\sqrt{E/(TNB)})$

Note: **Exp.**, **Dist.**, **Lin.**, **Cat.**, and **Quantile** denote expected-value, distributional, linear, categorical, and quantile representations, respectively. **Last** and **Tail** denote the last-server iterate θ^R and tail-averaged output $\bar{\theta}_T$, respectively. The **Rate** column reports root-error rates or convergence guarantees. $\boldsymbol{n}$ denotes algorithmic iterations or variance-reduction rounds, whereas N denotes the number of clients.

3 Preliminaries and Problem Setup

3.1 Markov Decision Process and Fixed Policy

We consider a discounted Markov decision process

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma),$$

where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R} = [0, 1]$ are the state, action, and reward spaces, respectively, $\gamma \in (0, 1)$ is the discount factor, and $P(\mathrm{d}\omega, \mathrm{d}s'|s, a)$ gives the joint distribution of the reward ω and next state s' given (s, a).

We fix a policy $\pi(\cdot|s)$ and study policy evaluation. It induces

$$P^\pi(\mathrm{d}\omega, \mathrm{d}s'|s) = \int_{\mathcal{A}} P(\mathrm{d}\omega, \mathrm{d}s'|s, a)\, \pi(\mathrm{d}a|s),$$

and let μ_π denote the corresponding stationary state distribution. The target $\eta^\pi(s)$ is the distribution of the discounted return from state s, supported on $[0, (1-\gamma)^{-1}]$. It satisfies

$$\eta^\pi = \mathcal{T}^\pi \eta^\pi,$$

where

$$(\mathcal{T}^\pi \eta)(s) = \mathbb{E}[(b_{\omega,\gamma})_{\#}\eta(S') \,|\, S = s], \qquad b_{\omega,\gamma}(z) = \omega + \gamma z.$$

Here, $(b_{\omega,\gamma})_{\#}\eta(S')$ is the distribution of $\omega + \gamma Z$ when $Z \sim \eta(S')$. Hence, the operator maps the return distribution at the next state to the distribution of the one-step reward plus the discounted future return.

3.2 Linear Categorical Distributional Representation

To make distributional policy evaluation tractable in large state spaces, we approximate return distributions with a linear categorical family and study its projected Bellman fixed point.

Let $\mathcal{P}^{\mathrm{sign}}$ be the set of unit-mass signed measures supported on $[0, (1-\gamma)^{-1}]$, and let $F_\nu(z) = \nu([0, z])$ denote the cumulative function of $\nu \in \mathcal{P}^{\mathrm{sign}}$. The stationary-weighted squared Cramér distance between $\eta_1, \eta_2 \in (\mathcal{P}^{\mathrm{sign}})^{\mathcal{S}}$ is

$$\ell^2_{2,\mu_\pi}(\eta_1, \eta_2) = \mathbb{E}_{S\sim\mu_\pi}\left[\int_0^{(1-\gamma)^{-1}} \left(F_{\eta_1(S)}(z) - F_{\eta_2(S)}(z)\right)^2 \mathrm{d}z\right].$$

To obtain a finite-dimensional representation of the return distribution, we discretize its support using $K + 1$ equally spaced atoms $x_k = k\iota_K$, where $\iota_K = [K(1-\gamma)]^{-1}$. We then parameterize the first K cumulative masses with a linear feature map. Specifically, let $\Theta = [\theta^{(0)}, \ldots, \theta^{(K-1)}] \in \mathbb{R}^{d\times K}$, $\theta = \mathrm{vec}(\Theta)$, and $\phi(s) \in \mathbb{R}^d$, and define

$$F_k(s; \Theta) = \phi(s)^\top \theta^{(k)} + \frac{k+1}{K+1}, \qquad k = 0, \ldots, K-1.$$

With $F_{-1} = 0$, $F_K = 1$, and $p_k(s;\Theta) = F_k(s;\Theta) - F_{k-1}(s;\Theta)$, the resulting categorical representation is

$$\eta_\theta(s) = \sum_{k=0}^{K} p_k(s;\Theta)\delta_{x_k}, \qquad \mathcal{P}^{\text{sign}}_{\phi,K} := \{\eta_\theta : \theta \in \mathbb{R}^{dK}\}.$$

The linear parameterization may produce signed masses, so the class is defined over finite signed measures.

We next project the distributional Bellman equation onto this linear categorical class. For $\eta \in (\mathcal{P}^{\text{sign}})^{\mathcal{S}}$, define

$$\Pi^{\mu_\pi}_{\phi,K}\eta := \arg\min_{\widetilde{\eta}\in\mathcal{P}^{\text{sign}}_{\phi,K}} \ell^2_{2,\mu_\pi}(\eta,\widetilde{\eta}). \tag{3.1}$$

The target of policy evaluation is the projected fixed point

$$\eta_{\theta^\star} = \Pi^{\mu_\pi}_{\phi,K}\mathcal{T}^\pi\eta_{\theta^\star}. \tag{3.2}$$

Finally, let $\Sigma_\phi = \mathbb{E}_{\mu_\pi}[\phi(S)\phi(S)^\top]$ and $J = I_K \otimes \Sigma_\phi$. The parameterization satisfies the exact isometry

$$\ell^2_{2,\mu_\pi}(\eta_{\theta_1},\eta_{\theta_2}) = \iota_K\|\theta_1 - \theta_2\|^2_J. \tag{3.3}$$

Therefore, convergence of the parameter vector is equivalent to convergence in the squared Cramér distance.

3.3 Federated Problem Setting

We consider a synchronous federated system with N statistically homogeneous clients that share π, P^π, ϕ, and the categorical support, while retaining local data and communicating only parameter updates.

Assumption 1 (Sampling and features). We assume that every client participates in each communication round and draws samples $\xi = (S,\omega,S')$ according to

$$S \sim \mu_\pi, \qquad (\omega,S') \sim P^\pi(\cdot,\cdot|S).$$

Conditional on the pre-round history, control samples are i.i.d., and local samples are fresh and independent across clients, local steps, and rounds. Control and local samples are independent, while each same-sample difference uses the same local sample. We further assume

$$\|\phi(s)\|_2 \le 1, \qquad 0 < \lambda_{\min} I_d \preceq \Sigma_\phi \preceq \lambda_{\max} I_d.$$

Under this setting, each client performs E local updates between consecutive communications. Hence, after R communication rounds, the per-client local-update budget is $T := RE$.

4 FedVR-DistTD

4.1 Stochastic Linear Categorical TD Direction

We begin by converting the projected Bellman equation into a finite-dimensional root equation that can be estimated from individual samples. To this end, let $\xi = (s,\omega,s')$ denote a sample. Vectorizing the linear categorical representation gives the affine stochastic Linear-CTD direction

$$h(\theta;\xi) = A(\xi)\theta - b(\xi), \tag{4.1}$$

where

$$A(\xi) = I_K \otimes [\phi(s)\phi(s)^\top] - M(\omega)\otimes[\phi(s)\phi(s')^\top], \qquad b(\xi) = \zeta(\omega)\otimes\phi(s).$$

Here, $M(\omega)$ and $\zeta(\omega)$ are the categorical projection matrix and affine offset, respectively; their explicit forms are given in Appendix A.1.

Taking expectations yields

$$\bar{h}(\theta) = \mathbb{E}[h(\theta;\xi)] = \bar{A}\theta - \bar{b}, \qquad \bar{A} := \mathbb{E}[A(\xi)], \quad \bar{b} := \mathbb{E}[b(\xi)]. \tag{4.2}$$

The projected Bellman equation (3.2) is equivalent to

$$\bar{h}(\theta^\star) = 0, \tag{4.3}$$

as shown in Appendix A.1. Thus, policy evaluation reduces to solving a finite-dimensional affine root equation, which is estimated in the federated update introduced next.

4.2 FEDERATED VARIANCE-REDUCED UPDATE

Because clients perform multiple local updates between communication rounds, directly using stochastic TD directions can accumulate estimation noise and client drift. We therefore construct a shared control variate from independent client batches to estimate the global mean TD direction.

At round r, the server broadcasts θ^r. Each client i draws a control batch $\{\widetilde{\xi}_{i,q}^r\}_{q=1}^B$ and computes

$$c_i^r = \frac{1}{B}\sum_{q=1}^{B} h(\theta^r; \widetilde{\xi}_{i,q}^r), \qquad c^r = \frac{1}{N}\sum_{i=1}^{N} c_i^r. \tag{4.4}$$

The clients upload c_i^r, and the server broadcasts c^r. Conditional on the pre-round history, c^r is unbiased for $\bar{h}(\theta^r)$ with variance $\mathcal{O}(1/\sqrt{NB})$.

Each client initializes $\theta_{i,0}^r = \theta^r$. For $t = 0, \ldots, E-1$, it draws a fresh sample $\xi_{i,t}^r$ and performs

$$g_{i,t}^r = h(\theta_{i,t}^r; \xi_{i,t}^r) - h(\theta^r; \xi_{i,t}^r) + c^r, \qquad \theta_{i,t+1}^r = \theta_{i,t}^r - \alpha g_{i,t}^r. \tag{4.5}$$

Using the same sample in both TD directions cancels the affine offset:

$$g_{i,t}^r = A(\xi_{i,t}^r)(\theta_{i,t}^r - \theta^r) + c^r. \tag{4.6}$$

Hence, the stepwise noise depends on the local displacement, whereas the control-batch error is shared throughout the round.

After E local steps, the server aggregates

$$\theta^{r+1} = \frac{1}{N}\sum_{i=1}^{N} \theta_{i,E}^r = \theta^r - \frac{\alpha}{N}\sum_{i=1}^{N}\sum_{t=0}^{E-1} g_{i,t}^r. \tag{4.7}$$

The algorithm outputs the final server iterate θ^R. For statistical estimation, the server additionally forms the tail average

$$\bar{\theta}_T = \frac{1}{m}\sum_{r=s}^{R-1} \theta^r, \qquad s = \lfloor R/2 \rfloor, \qquad m = R - s. \tag{4.8}$$

The last iterate is used to characterize round-wise contraction, while the tail average provides the statistical convergence guarantee. Algorithm 1 summarizes the update procedure.

Algorithm 1 FedVR-DistTD

Initialize N, R, E, B, α, and θ^0.
for $r = 0, 1, \ldots, R-1$ **do**
  Server broadcasts θ^r.
  for $i = 1, 2, \ldots, N$ **in parallel do**
    Draw $\{\widetilde{\xi}_{i,q}^r\}_{q=1}^B$ and compute $c_i^r \leftarrow B^{-1}\sum_{q=1}^B h(\theta^r; \widetilde{\xi}_{i,q}^r)$.
    Upload c_i^r to the server.
  end for
  Server computes $c^r \leftarrow N^{-1}\sum_{i=1}^N c_i^r$ and broadcasts c^r.
  for $i = 1, 2, \ldots, N$ **in parallel do**
    Set $\theta_{i,0}^r \leftarrow \theta^r$.
    for $t = 0, 1, \ldots, E-1$ **do**
      Draw $\xi_{i,t}^r$ and compute $g_{i,t}^r \leftarrow h(\theta_{i,t}^r; \xi_{i,t}^r) - h(\theta^r; \xi_{i,t}^r) + c^r$.
      Update $\theta_{i,t+1}^r \leftarrow \theta_{i,t}^r - \alpha g_{i,t}^r$.
    end for
    Upload $\theta_{i,E}^r$.
  end for
  Server updates $\theta^{r+1} \leftarrow N^{-1}\sum_{i=1}^N \theta_{i,E}^r$.
end for
Return θ^R.

5 Convergence Analysis

5.1 Variance-Reduction Ingredients

To establish contraction of the mean TD recursion, we first quantify the stability of its linear operator. Let $\kappa_\gamma = 1 - \sqrt{\gamma}$. For every $x \in \mathbb{R}^{dK}$,

$$x^\top \bar{A} x \geq \kappa_\gamma \|x\|_J^2. \tag{5.1}$$

Although $\bar{A}$ is generally nonsymmetric, this inequality implies strong monotonicity in the J-weighted norm and hence invertibility. Together with Equation (4.3), it guarantees the uniqueness of $\theta^\star$; the proof is given in Appendix A.5.

To quantify the finite-control error, condition on the pre-batch history $\mathcal{F}_r$. By Equation (4.4) and sample independence,

$$\mathbb{E}[c^r|\mathcal{F}_r] = \bar{h}(\theta^r), \qquad \mathbb{E}\big[\|c^r - \bar{h}(\theta^r)\|_2^2|\mathcal{F}_r\big] = \frac{1}{NB}\mathbb{E}\big[\|h(\theta^r;\xi) - \bar{h}(\theta^r)\|_2^2|\mathcal{F}_r\big]. \tag{5.2}$$

Thus, the global control direction is conditionally unbiased and its variance decreases as $1/(NB)$; the derivation is given in Appendix B.

For the same-sample local difference, define

$$\Delta_{i,t}^r := h(\theta_{i,t}^r;\xi_{i,t}^r) - h(\theta^r;\xi_{i,t}^r) - \big[\bar{h}(\theta_{i,t}^r) - \bar{h}(\theta^r)\big] = [A(\xi_{i,t}^r) - \bar{A}](\theta_{i,t}^r - \theta^r).$$

The random-operator bound in Appendix A.6 gives

$$\mathbb{E}[\Delta_{i,t}^r|\mathcal{G}_{r,t}] = 0, \qquad \mathbb{E}\big[\|\Delta_{i,t}^r\|_2^2|\mathcal{G}_{r,t}\big] \leq 2(1+\gamma)\|\theta_{i,t}^r - \theta^r\|_J^2. \tag{5.3}$$

Hence, the stepwise noise is conditionally mean-zero and controlled by the local displacement. These two bounds are the basis of the convergence analysis.

5.2 Global Parameter Convergence

Because the finite-batch control error is shared across all local steps, we analyze the last iterate and the tail average separately. Let

$$\varepsilon_c^r := c^r - \bar{h}(\theta^r), \qquad \vartheta^r := \theta^\star - \bar{A}^{-1}\varepsilon_c^r.$$

Equivalently,

$$\vartheta^r = \theta^\star - \bar{A}^{-1}\varepsilon_c^r. \tag{5.4}$$

Using $c^r = \bar{h}(\theta^r) + \varepsilon_c^r$ and $\bar{h}(\theta) = \bar{A}(\theta - \theta^\star)$, the conditional mean of each local direction is centered at ϑ^r:

$$\bar{A}(\theta_{i,t}^r - \theta^r) + c^r = \bar{A}(\theta_{i,t}^r - \vartheta^r). \tag{5.5}$$

Appendix B.3 shows that ϑ^r is conditionally centered at $\theta^\star$ and has mean-square displacement $\mathcal{O}((NB)^{-1})$. Let a, q_{loc}, $C_{\beta,\tau}$, and ρ denote the contraction constants defined in Appendix C.2; $\widetilde{\Sigma}_e$ and c_0 are defined in Lemma B.3. We assume $0 \leq a < 1$, $0 < q_{\mathrm{loc}} < 1$, and $0 < \rho < 1$.

Proposition 1 (Ideal-control geometric contraction) *Under Assumption 1, suppose that $0 \leq a < 1$, $0 < q_{\mathrm{loc}} < 1$, and $c^r = \bar{h}(\theta^r)$. Then*

$$\mathbb{E}\|\theta^R - \theta^\star\|_J^2 \leq q_{\mathrm{loc}}^R \mathbb{E}\|\theta^0 - \theta^\star\|_J^2. \tag{5.6}$$

Thus, exact global control yields geometric contraction of the last-server iterate to $\theta^\star$.

Proposition 2 (Finite-control last-iterate contraction) *Under Assumption 1, suppose that the above contraction conditions hold. Then*

$$\mathbb{E}\|\theta^R - \theta^\star\|_J^2 \leq \rho^R \mathbb{E}\|\theta^0 - \theta^\star\|_J^2 + \frac{2C_{\beta,\tau}}{NB}\frac{1-\rho^R}{1-\rho}\operatorname{Tr}(\widetilde{\Sigma}_e). \tag{5.7}$$

The bound combines geometric contraction via $\mathcal{O}(\rho^R)$ with a finite-control floor. Thus, the last-server iterate converges to an $\mathcal{O}((NB)^{-1})$ mean-square neighborhood; see Appendix C.4 for proof.

Theorem 1 (Tail-averaged mean-square convergence) *Under Assumption 1 and the conditions of Proposition 2, let $T = RE$ and suppose that*

$$\|I - \alpha\bar{A}\|_J < 1, \qquad \rho \le \bar{\rho} < 1.$$

Then there exists a constant $C_E < \infty$, independent of T, such that

$$\mathbb{E}\|\bar{\theta}_T - \theta^\star\|_J^2 \le C_E \left[\frac{E}{TNB}\,\mathrm{Tr}(\widetilde{\Sigma}_e) + \frac{E^2}{T^2}\rho^{\lfloor T/(2E)\rfloor}\mathbb{E}\|\theta^0 - \theta^\star\|_J^2\right]. \tag{5.8}$$

The tail average converges to the projected fixed point at root-error rate $\mathcal{O}(\sqrt{E/(TNB)})$, reducing to $\mathcal{O}(T^{-1/2})$ for fixed E, N, and B.

Corollary 1 (Tail-averaged squared Cramér convergence) *By the parameter–Cramér isometry, Theorem 1 implies*

$$\mathbb{E}\,\ell_{2,\mu_\pi}^2\big(\eta_{\bar{\theta}_T}, \eta_{\theta^\star}\big) \le \iota_K C_E \left[\frac{E}{TNB}\,\mathrm{Tr}(\widetilde{\Sigma}_e) + \frac{E^2}{T^2}\rho^{\lfloor T/(2E)\rfloor}\mathbb{E}\|\theta^0 - \theta^\star\|_J^2\right]. \tag{5.9}$$

For fixed N, B, and E, the root Cramér error is $\mathcal{O}(1/\sqrt{T})$. This bound concerns the projected fixed point $\eta_{\theta^\star}$; approximation and discretization errors relative to η^π are treated in Appendix E.2.

Remark 1 *Compared with existing centralized distributional TD methods, which derive finite-sample guarantees from pooled samples but require centralized access to interaction data, our analysis incorporates data locality, communication, and repeated client-side updates. FedVR-DistTD characterizes these effects through the geometric term $\mathcal{O}(\rho^R)$, the finite-control error $\mathcal{O}((NB)^{-1/2})$, and the tail-averaging rate $\mathcal{O}(\sqrt{E/(TNB)})$. In practical policy-evaluation settings where trajectories are distributed across agents and cannot be directly shared, this provides a theoretically grounded approach to estimating return distributions while reducing update variance and preserving uncertainty information.*

6 Experiments

6.1 Experimental Setup

We evaluate the methods on a four-state Markov reward process with homogeneous clients and generative-model sampling, using the projected fixed point $\theta^\star$ as reference. We report parameter, variance, disagreement, and distributional errors; method definitions, algorithm configurations, and full protocols are provided in Appendix F.

6.2 Overall Convergence and Ablation Study

We evaluate last-server and tail-averaged MSEs over communication rounds and $T = RE$, respectively, observing geometric decay to a finite-control floor and an $\mathcal{O}(T^{-1})$ tail-averaged mean-square error rate.

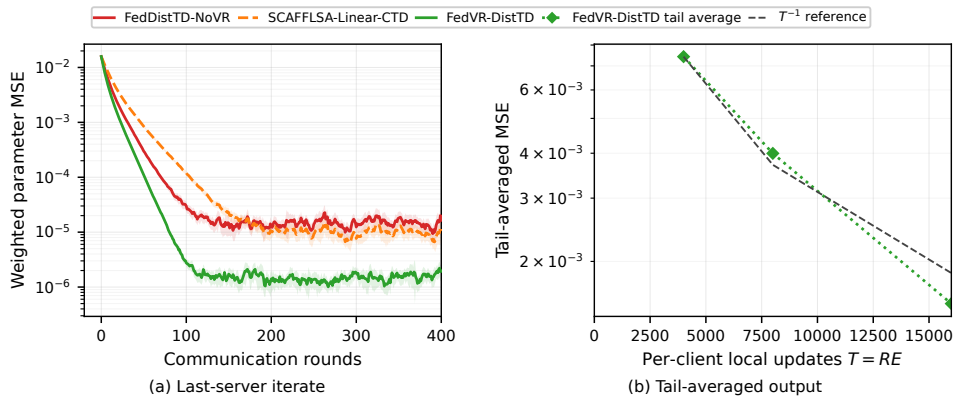


Figure 1: Weighted parameter MSE of (a) the last-server iterates over communication rounds and (b) the tail-averaged FedVR-DistTD output versus the per-client update budget $T = RE$.

6.3 Variance Reduction and Client Drift

We measure J-weighted update variance and pre-aggregation RMS client disagreement using fresh probes. NoVR, Diff, Control, and FedVR denote the four methods; details are provided in Appendix F

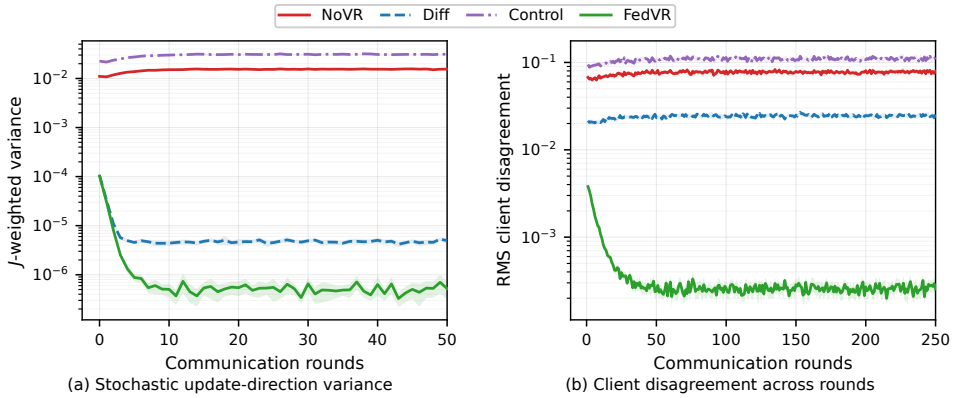


Figure 2: (a) J-weighted update variance and (b) pre-aggregation RMS client disagreement.

FedVR-DistTD achieves low variance and disagreement, while Diff and Control isolate one variance-reduction component and NoVR removes both.

6.4 Effect of the Number of Local Steps

We vary local updates under a fixed communication budget to assess accuracydisagreement tradeoff.

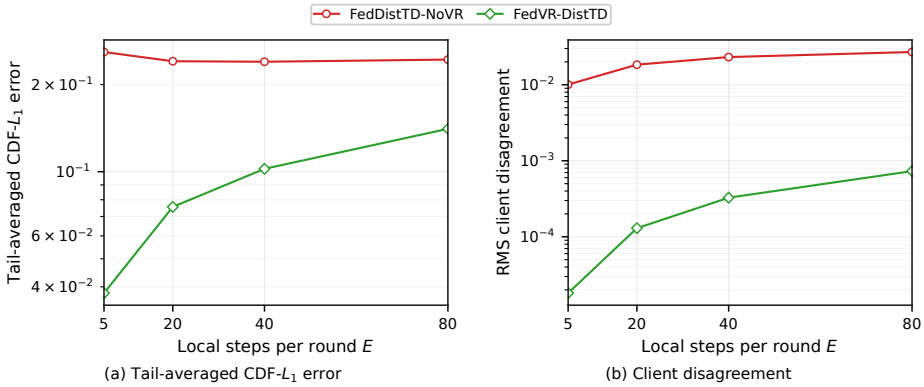


Figure 3: Effect of local steps. (a) Tail-averaged CDF-L_1 error. (b) Mean RMS client disagreement.

The results reveal a communication–accuracy tradeoff: increasing the number of local updates raises client disagreement and, for FedVR-DistTD, increases the tail-averaged CDF-L_1 error.

7 Conclusion

We proposed FedVR-DistTD, combining a global control direction with same-sample TD differences. The analysis distinguishes two outputs and two phases of the method. Exact mean control exposes an intrinsic geometric contraction of the last server iterate, while the implementable finite-control algorithm contracts geometrically to an explicit $\mathcal{O}(1/\sqrt{NB})$ statistical neighborhood. Tail averaging of the round-level server iterates then removes the residual round-wise fluctuation and yields $\mathcal{O}(E/(TNB))$ mean-square convergence to the projected fixed point; for fixed N, B, and E, this is $\mathcal{O}(1/\sqrt{T})$. The parameter–Cramér isometry yields the corresponding expected squared Cramér-error guarantee. Experiments illustrate the predicted variance-reduction mechanisms; heterogeneity, partial participation, and Markovian sampling remain open.

AI USE STATEMENT

In this work, Generative AI tools were used to improve English grammar and academic style, refine the organization and presentation of the manuscript based on the authors intended structure, and assist with LATEX formatting. The research ideas, algorithmic design, theoretical derivations, and initial experimental design were developed by the authors; AI-generated suggestions were used only for review and refinement. The tools were not used to generate experimental data or execute experiments. All AI-assisted suggestions and resulting revisions were independently reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 449–458. PMLR, 2017.

Marc G. Bellemare, Nicolas Le Roux, Pablo Samuel Castro, and Subhodeep Moitra. Distributional reinforcement learning with linear function approximation. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 2203–2211. PMLR, 2019.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, Cambridge, MA, 2023.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):2892–2901, 2018b.

Thinh T. Doan, Siva Theja Maguluri, and Justin Romberg. Finite-time analysis of distributed td(0) with linear function approximation on multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1626–1635. PMLR, 2019.

Kaicheng Jin, Yang Peng, Jiansheng Yang, and Zhihua Zhang. Accelerated distributional temporal difference learning with linear function approximation, 2025. URL https://arxiv.org/abs/2511.12688.

Peter Kairouz and H. Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 06 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5132–5143. PMLR, 2020.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 10997–11057. PMLR, 2022.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Clare Lyle, Marc G. Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):4504–4511, 2019.

Paul Mangold, Sergey Samsonov, Safwan Labbi, Ilya Levin, Reda Alami, Alexey Naumov, and Eric Moulines. SCAFFLSA: Taming heterogeneity in federated linear stochastic approximation and TD learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13927–13981, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

David Millard, Cecilia Alm, Rashid Ali, Pengcheng Shi, and Ali Baheri. Federated distributional reinforcement learning with distributional critic regularization, 2026. URL https://arxiv.org/abs/2603.17820.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshihiro Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 799–806, 2010.

Yang Peng, Liangyu Zhang, and Zhihua Zhang. Statistical efficiency of distributional temporal difference learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24724–24761, 2024.

Yang Peng, Kaicheng Jin, Liangyu Zhang, and Zhihua Zhang. A finite sample analysis of distributional TD learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170729–170777, 2025.

Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pp. 29–37. PMLR, 2018.

Mark Rowland, Rémi Munos Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G. Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25(163):1–47, 2024a. URL http://jmlr.org/papers/v25/23-0154.html.

Mark Rowland, Li Kevin Wenliang, Rémi Munos, Clare Lyle, Yunhao Tang, and Will Dabney. Near-minimax-optimal distributional reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 132774–132823, 2024b.

Sebastian U. Stich. Local sgd converges fast and communicates little. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3: 9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.

Han Wang, Aritra Mitra, Hamed Hassani, George J. Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity, 2024. URL https://arxiv.org/abs/2302.02212.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 501–545, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/026aff87942ce636ada884d934cde0ae-Paper-Conference.pdf.

Junlong Zhu, Haotong Dong, Mingchuan Zhang, Gaofeng Chen, Ruijuan Zheng, Quanbo Ge, and Qingtao Wu. Pfopg: A personalized federated first-order policy gradient algorithm and its nonasymptotic analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(1): 192–203, 2026. doi: 10.1109/TSMC.2025.3627478. URL https://doi.org/10.1109/TSMC.2025.3627478.

Ye Zhu, Xiaowen Gong, and Shiwen Mao. Finite-time analysis of heterogeneous federated temporal difference learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025. doi: 10.24963/ijcai.2025/808.