The Annals of Statistics
2025, Vol. 53, No. 5, 1987–2011
https://doi.org/10.1214/25-AOS2527

ESTIMATION AND INFERENCE IN DISTRIBUTIONAL REINFORCEMENT LEARNING

BY LIANGYU ZHANG[1,a], YANG PENG[2,b], JIADONG LIANG[2,c], WENHAO YANG[2,d] AND ZHIHUA ZHANG[3,e]

[1]*School of Statistics and Data Science, Shanghai University of Finance and Economics,* [a]*zhangliangyu@sufe.edu.cn*
[2]*School of Mathematical Sciences, Peking University,* [b]*pengyang@pku.edu.cn,* [c]*jdliang@pku.edu.cn,* [d]*yangwenhaosms@pku.edu.cn*
[3]*School of Mathematical Sciences, School of Computer Science, Peking University,* [e]*zhzhang@math.pku.edu.cn*

In this paper, we study distributional reinforcement learning from the perspective of statistical efficiency. We investigate distributional policy evaluation, aiming to estimate the complete return distribution (denoted η^π) attained by a given policy π. We use the certainty equivalence method to construct our estimator $\hat{\eta}_n^\pi$, based on a generative model. In this circumstance, we need a dataset of size $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\varepsilon^{-2p}(1-\gamma)^{-(2p+2)})$ to guarantee the supremum p-Wasserstein metric between $\hat{\eta}_n^\pi$ and η^π less than ε with high probability. This implies the distributional policy evaluation problem can be solved with sample efficiency. Also, we show that under different mild assumptions a dataset of size $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\varepsilon^{-2}(1-\gamma)^{-4})$ suffices to ensure the supremum Kolmogorov-Smirnov metric and supremum total variation metric between $\hat{\eta}_n^\pi$ and η^π is below ε with high probability. Furthermore, we investigate the asymptotic behavior of $\hat{\eta}_n^\pi$. We demonstrate that the "empirical process" $\sqrt{n}(\hat{\eta}_n^\pi - \eta^\pi)$ converges weakly to a Gaussian process in the space of bounded functionals on a Lipschitz function class $\ell^\infty(\mathcal{F}_{W_1})$, also in the space of bounded functionals on an indicator function class $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$ and a bounded measurable function class $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$ when some mild conditions hold. Our findings give rise to a unified approach to statistical inference of a wide class of statistical functionals of η^π.

1. Introduction.

Reinforcement learning has achieved remarkable advancements in various fields, including game-playing [44, 53], robotics systems [22], large language models [34, 35], among others. In classical reinforcement learning which relies on the reward hypothesis [48, 49], one evaluates the performance of a learning agent by the expected returns (i.e., the expected cumulative sum of a received reward). However, in many applications of reinforcement learning, it is not enough to simply consider the expected returns because other factors such as uncertainty or risks might also be crucial. For example, when we ask a large language model a question we not only expect a useful answer but want to know how reliable the answer is. An investor should consider the risk-return tradeoff when making investment decisions in financial markets, as high expected returns usually mean greater risks [15]. In the area of healthcare, we are not only interested in the expected performance of a dynamic treatment regime but care about its long-tail performance. Otherwise, it would have the potential to cause serious consequences for patients [23].

Distributional reinforcement learning [3, 32] goes beyond the notion of expected returns and proposes to model the complete distribution of the random returns. From the statistical perspective, the main difference between standard reinforcement learning and distributional reinforcement learning is that the former considers the expectation estimation of the return,

Received September 2024; revised March 2025.
MSC2020 subject classifications. Primary 62C05, 62G05, 62G15; secondary 90C40, 68Q32.
Key words and phrases. Distributional reinforcement learning, probability metrics, sample complexity bounds, weak convergence.

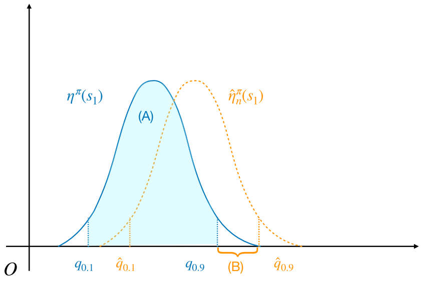

FIG. 1. *An illustration of two types of uncertainty in reinforcement learning. Blue distribution: ground-truth return distribution $\eta^\pi(s_1)$ with quantiles $q_{0.1}$ and $q_{0.9}$. Orange distribution: estimated return distribution $\hat{\eta}_n^\pi(s_1)$ with quantiles $\hat{q}_{0.1}$ and $\hat{q}_{0.9}$. Shaded area (A): intrinsic uncertainty in reinforcement learning. Error (B): error caused by statistical uncertainty in reinforcement learning.*

while the latter studies the estimation of the full return distribution. This distinction brings two advantages. First, the full return distribution contains richer information than its expectation alone. It offers a comprehensive depiction of the intrinsic uncertainty (also known as aleatoric uncertainty) in the performance of learning agents, which arises due to both the stochastic nature of environments and the actions taken by the agents. For example, once we have an estimator of the return distribution, we can directly evaluate any statistical functionals of interest such as variance, quantiles, and conditional value-at-risk (CVaR), and obtain a better understanding of the consequences of the agents' behaviors. We can also optimize a chosen statistical functional beyond the expectation, which will lead to robust or risk-sensitive policies [28, 46]. Second, estimating the full return distribution could be helpful even in standard reinforcement learning tasks (e.g., estimation or optimization of expected returns) under appropriate conditions [39, 54] because the estimation of the return distribution can provide richer and stabler learning signals, ultimately improving the efficiency and stability of the policy evaluation or policy learning process."

In the setting of reinforcement learning, we are usually not fully aware of the stochastic environment and must rely on a dataset to evaluate or train a learning agent. This induces another kind of uncertainty that we call statistical uncertainty (also known as epistemic uncertainty) [8, 29], which stems from limited data. In this paper, we seek to simultaneously address the two kinds of uncertainties aforementioned by developing statistical understandings for distributional reinforcement learning (see Figure 1 for an illustration of the two types of uncertainty). Specifically, we aim to answer the following two fundamental questions: (a) How many samples are required to learn the full distribution of random returns? (b) Is it possible to perform statistical inferences from the learned return distributions? We give affirmative answers to both questions with the benefit of a statistical analysis of distributional reinforcement learning presented in our paper. We believe that our work could shed light on uncertainty quantification in reinforcement learning.

1.1. *Our contributions.*

In this paper, we focus on the problem of distributional policy evaluation, which lies at the core of distributional reinforcement learning.[1] Consider a γ-discounted infinite-horizon Markov decision process (MDP). The MDP is assumed to be tabular, that is, it has finite state space $\mathcal{S}$ and action space $\mathcal{A}$. Let $\eta^\pi(s)$ denote the random return gained by running the policy π from the initial state s and define $\eta^\pi := (\eta^\pi(s_1), \dots, \eta^\pi(s_{|\mathcal{S}|}))$. When the underlying MDP is known, we may find η^π by solving the

[1] Indeed, the control problem in distributional reinforcement learning can be solved by a two-stage procedure. First, we estimate a near-optimal policy $\hat{\pi}$ using some policy learning subroutines. Then it remains to solve a distributional policy evaluation problem, that is, the return distribution of $\hat{\pi}$. See Section 7.3 of [4].

distributional Bellman equation $\eta^{\pi} = \mathcal{T}^{\pi}\eta^{\pi}$ with the distributional dynamic programming algorithm. Here, $\mathcal{T}^{\pi}$ is called the distributional Bellman operator.

Our goal is to estimate η^{π} when the underlying MDP is unknown. Following common practice in the literature on reinforcement learning, we assume that the distribution of the random reward is fully known and that the transition probability of that MDP is unknown. Our estimator $\hat{\eta}_n^{\pi}$ is constructed using the certainty equivalence approach [45, 50]. More specifically, we first build an explicit model of the underlying transition dynamics (denoted $\widehat{P}_n$) on an offline dataset of $n|\mathcal{S}||\mathcal{A}|$ entries obtained from a generative model.[2] Then we form an empirical MDP $\widehat{M}$ whose transition dynamic is $\widehat{P}_n$. An estimator of η^{π} is then formulated as the return distribution of $\widehat{M}$, which we denote as $\hat{\eta}_n^{\pi} := (\hat{\eta}_n^{\pi}(s_1), \dots, \hat{\eta}_n^{\pi}(s_{|\mathcal{S}|}))$. Note that we consider a fully nonparametric setting where $\hat{\eta}_n^{\pi}$ is not restricted in some parametric model and can be any probability distribution.

In this paper, we analyze the statistical performance of the estimated return distribution $\hat{\eta}_n^{\pi}$. Concretely, we would like to: (a) prove nonasymptotic bounds for the l_{∞}-type estimation error $\sup_{s\in\mathcal{S}} d(\eta^{\pi}(s), \hat{\eta}_n^{\pi}(s))$, where d is some probability metric; (b) study the asymptotics of $\hat{\eta}_n^{\pi}(s)$, particularly identifying the limit of the process $\sqrt{n}(\eta^{\pi}(s) - \hat{\eta}_n^{\pi}(s))$. Our analysis has close relationships with the works on the perturbation theory of Markov chains. Given the MDP M and policy π, we define a Markov chain on the state space of the MDP, and use P^{π} to denote its Markov kernel. Our problem can be restated as how the gap between η^{π} and $\hat{\eta}_n^{\pi}$ is related to the gap between the Markov kernels P^{π} and $\widehat{P}_n^{\pi}$. Or, more broadly, how the "characteristics" of a Markov chain would change when we perturb its Markov kernel. This is the core question in the field of perturbation theory of Markov chains. While prior works mostly focus on "characteristics" like invariant distributions and t-step distributions, the novelty of our work is that the object of interest is the distribution of the cumulative rewards. We will elaborate on the differences between our work and prior works on perturbation analysis of Markov chain in the related work section.

To the best of our knowledge, we are among the first to develop statistical theories for distributional reinforcement learning. Our main contributions are outlined below.

- We show that under mild conditions, the distributional dynamic programming algorithm converges to the fixed point when measured by the Kolmogorov-Smirnov (KS) metric and by the total variation (TV) metric. Interestingly, this convergence occurs despite the distributional Bellman operator no longer being (strictly) contractive. Our findings correct the misconception that distributional dynamic programming is not guaranteed to converge in the KS and TV metrics.
- We provide nonasymptotic bounds for the p-Wasserstein metric, the KS metric and the TV metric between η^{π} and $\hat{\eta}_n^{\pi}$. Specifically, we prove $\sup_{s\in\mathcal{S}} W_p(\eta^{\pi}(s), \hat{\eta}_n^{\pi}(s)) = \widetilde{O}(n^{-1/(2p)}(1-\gamma)^{-1-1/(2p)})$,[3] $\sup_{s\in\mathcal{S}} \mathrm{KS}(\eta^{\pi}(s), \hat{\eta}_n^{\pi}(s)) = \widetilde{O}(n^{-1/2}(1-\gamma)^{-2})$ and $\sup_{s\in\mathcal{S}} \mathrm{TV}(\eta^{\pi}(s), \hat{\eta}_n^{\pi}(s)) = \widetilde{O}(n^{-1/2}(1-\gamma)^{-2})$, with high probability. Our nonasymptotic results translate can be translated to an $\widetilde{O}(\varepsilon^{-2p}(1-\gamma)^{-2p-2})$ complexity bound for the case of the p-Wasserstein metric and $\widetilde{O}(\varepsilon^{-2}(1-\gamma)^{-4})$ complexity bound for the cases of the KS and TV metrics. We also generalize our results to more challenging settings with less exploratory datasets and unknown reward distributions.
- We give a characterization of the asymptotic behavior of $\hat{\eta}_n^{\pi}$. We show that for any $s \in \mathcal{S}$, $\sqrt{n}(\hat{\eta}_n^{\pi}(s) - \eta^{\pi}(s))$ converges weakly to a Gaussian process in $\ell^{\infty}(\mathcal{F}_{W_1})$. Under mild conditions, $\sqrt{n}(\hat{\eta}_n^{\pi}(s) - \eta^{\pi}(s))$ also converges weakly to a Gaussian process in both $\ell^{\infty}(\mathcal{F}_{\mathrm{KS}})$

[2] We also consider the more challenging settings where the reward is unknown either and the offline dataset is not perfectly exploratory, in the later part of the paper.

[3] $\widetilde{O}$ hides all terms of logarithmic order.

and $\ell^{\infty}(\mathcal{F}_{\mathrm{TV}})$ for each $s \in \mathcal{S}$. Here, $\mathcal{F}_{W_1}$, $\mathcal{F}_{\mathrm{KS}}$, and $\mathcal{F}_{\mathrm{TV}}$ represent the 1-Lipschitz function class, the indicator function class, and the bounded measurable function class, respectively. We generalize the asymptotic results to the settings with less exploratory datasets and unknown reward distributions. These asymptotic results enable us to perform statistical inference for η^{π}. Concretely, we construct asymptotically valid confidence regions for η^{π} in the forms of 1-Wasserstein, KS, and TV metric balls, and asymptotically valid confidence intervals for $\phi(\eta^{\pi}(s))$, where ϕ can be any Hadamard differentiable functional.

- At the technical level, our main challenge is that we must work in the infinite-dimensional space of probability measures. Therefore, most of the techniques developed for classical reinforcement learning theory are not valid anymore. We address the challenge through an analysis of the concentration behaviors as well as asymptotics of $(\mathcal{I} - \widehat{\mathcal{T}}_n^{\pi})^{-1}(\widehat{\mathcal{T}}_n^{\pi} - \mathcal{T}^{\pi})\eta^{\pi}$. Here $\widehat{\mathcal{T}}_n^{\pi}$ is the distributional Bellman operator associated with the empirical MDP $\widehat{M}$ and $(\mathcal{I} - \mathcal{T}^{\pi})^{-1} : = \sum_{i=0}^{\infty}(\mathcal{T}^{\pi})^i$ is defined on a subspace of interest. We achieve this by carefully examining the properties of the distributional Bellman operator $\mathcal{T}^{\pi}$ on the vector space of signed measures equipped with different metrics to decouple the dependencies between operators $(\mathcal{I} - \widehat{\mathcal{T}}_n^{\pi})^{-1}$ and $(\widehat{\mathcal{T}}_n^{\pi} - \mathcal{T}^{\pi})$.

1.2. *The related work.*

1.2.1. *Distributional reinforcement learning.* Distributional reinforcement learning has achieved remarkable success in fields such as communications [17], transportation systems [33], and algorithm discovery [12]. Notable distributional reinforcement learning algorithms include categorical temporal-difference learning [3], quantile temporal-difference learning [9, 10], GAN-based methods [11, 14], actor-critic methods [30], etc. For a comprehensive treatment of distributional reinforcement learning, readers could refer to a very recent book by Bellemare, Dabney and Rowland [4].

Despite its empirical success, there is a relative lack of theoretical understanding of distributional reinforcement learning. Rowland et al. [37] analyzed the convergence properties of categorical temporal-difference learning. However, their convergence analysis is asymptotic and does not consider sample complexities as well as asymptotic distributions. Recently, Rowland et al. [38] presented similar consistency results for quantile temporal-difference learning. Wu, Uehara and Sun [56] showed that the distribution of returns can be learned using an algorithm called Fitted Likelihood Estimation (FLE) given an offline dataset. They also proposed nonasymptotic bounds for the statistical distance between the learned distribution and the ground truth. The major difference between Wu, Uehara and Sun [56] and our work is that Wu, Uehara and Sun [56] focused on parametrized return distributions, while we study the nonparametric case. Both the FLE algorithm and the associated nonasymptotic statistical bounds would be invalid under the nonparametric scenario. Another line of work treats learning the distribution of returns as an auxiliary task and aims to understand how this auxiliary task can improve policy learning within the framework of classical reinforcement learning. Sun et al. [47] found that such auxiliary tasks can be viewed as a form of regularization and can make the optimization process more stable. Wang et al. [54] explored the statistical benefits of distributional reinforcement learning. They showed that distributional reinforcement learning can yield better nonasymptotic bounds than classical reinforcement learning in the “small loss” scenarios.

1.2.2. *Statistical inference in reinforcement learning.* Statistical inference in the context of reinforcement learning has drawn growing interest in the community. There are a number of works studying the statistical inference problems for expected returns (or value functions). Thomas, Theocharous and Ghavamzadeh [51] and Jiang and Li [19] proposed high-confidence bounds for value functions in the setting of off-policy evaluation. Hao et al. [16] devised a bootstrapping procedure to perform statistical inference in off-policy evaluation. Shi et al. [43] modeled the value function with the series/sieve methods and devised confidence intervals for value functions in both the settings of policy evaluation and policy learning. Zhu, Dong and Lam [59] also constructed asymptotically tight confidence intervals for learned (optimal) value functions. Li et al. [26] and Li, Liang and Zhang [25] considered online statistical inference for value functions in an online reinforcement learning setting.

At the same time, fewer works focus on statistical inferences for other statistical functionals of the return distribution. Yang, Zhang and Zhang [57] investigated the asymptotic behaviors of distributionally robust value functions and constructed asymptotically tight confidence bounds. Chandak et al. [6] and Huang et al. [18] proposed methods to estimate the cumulative distribution function (CDF) and confidence band for the ground truth CDF. And statistical inference for statistical functionals can be achieved by the plug-in approach. However, their estimator is based on importance sampling, causing the errors to grow exponentially w.r.t. the horizon length. Also, their confidence intervals are based on nonasymptotic bounds and might thus be too conservative.

1.2.3. *Perturbation theory of Markov chains.* Early works on the perturbation theory of Markov chains can date back to [21, 42]. Mitrophanov [31] showed perturbation bounds of the t-step distributions for uniformly ergodic Markov chains. Ferré, Hervé and Ledoux [13] analyzed behaviors of perturbed V-geometrically ergodic Markov chains. Rudolf and Schweizer [40] proved bounds of the Wasserstein distance between the t-step distributions of a Wasserstein ergodic Markov chain and its perturbed counterpart. The perturbation theory of Markov chains is widely used in Markov chain Monte-Carlo algorithms and Bayesian statistics [1, 2, 20]. One can refer to the very recent book chapter [41] for a more thorough review.

Compared with previous works, the novelty of our work is that we study perturbation bounds for a novel object, that is, the return distribution of an MDP (or MRP). Our results are not simple corollaries of prior perturbation bounds on t-step distributions because the rewards obtained at different timesteps are dependent. See Wiltzeret al. [55] for a more detailed discussion. To get the new results, we developed new analysis techniques. Our proof techniques are specifically developed based on a thorough understanding of the theoretical properties of the distributional Bellman operator, making them particularly suited for the analysis of return distributions.

The remainder of this paper is organized as follows. In Section 2, we introduce some basic concepts of distributional reinforcement learning. In Section 3, we present our statistical analysis of distributional reinforcement learning. In Section 4, we propose a series of inferential procedures for the return distribution. Section 5 verifies our theoretical findings and tests our inferential procedures through numerical simulations.

2. Preliminaries.

2.1. *Problem settings and the certainty equivalence estimator.* An Markov decision process (MDP) is represented by a 5-tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_R, P, \gamma \rangle$, where $\mathcal{S}$ represents a finite state space, $\mathcal{A}$ a finite action space, $\mathcal{P}_R : \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$ the distribution of rewards, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the transition dynamics, and $\gamma \in (0,1)$ a discounted factor. Here we use $\Delta(\cdot)$ to represent the set of probability distributions over some set. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and an initial state $S_0 = s \in \mathcal{S}$, a random trajectory $\{(S_t, A_t, R_t)_{t=0}^{\infty}\}$ can be sampled from the MDP using the following procedure:

$$A_t|S_t \sim \pi(\cdot|S_t),$$
$$R_t|(S_t, A_t) \sim \mathcal{P}_R(\cdot|S_t, A_t),$$
$$S_{t+1}|(S_t, A_t) \sim P(\cdot|S_t, A_t).$$

We define the return of such trajectories by

$$G^\pi(s) := \sum_{t=0}^{\infty} \gamma^t R_t.$$

Note that $G^\pi(s)$ is a random variable (see Proposition 6.1 in Section 6 in the Supplementary Material [58]). And we always have $G^\pi(s) \in [0, (1-\gamma)^{-1}]$. The expected return $\mathbb{E}G^\pi(s)$ is called the value function and denoted by $V^\pi(s)$. We also define $\eta^\pi(s)$ as the distribution of $G^\pi(s)$.

In this paper, we focus on the problem of learning η^π for some policy π when the underlying MDP is unknown and must be estimated from a prespecified dataset. This problem is called the distributional policy evaluation problem. We assume the dataset is generated by a generative model, which is able to return a value of the next state s' according to $P(\cdot|s, a)$ for any given pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. For each pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the generative model is called n times, producing an array of n samples $X_1^{(s,a)}, \dots, X_n^{(s,a)} \overset{\text{i.i.d.}}{\sim} P(\cdot|s, a)$.

Given the dataset, we obtain the estimate of the transition probability as

$$\widehat{P}_n(s'|s, a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i^{(s,a)} = s'\}.$$

Thus, $\widehat{P}_n$ defines an empirical MDP $\widehat{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_R, \widehat{P}_n, \gamma\rangle$ with the corresponding distribution of returns $\hat{\eta}_n^\pi$. We call $\hat{\eta}_n^\pi$ the certainty equivalence estimator, and aim to explore the statistical properties of $\hat{\eta}_n^\pi$.

REMARK 1. An alternative problem formulation is to transform the MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}_R, P, \gamma\rangle$ to an Markov reward process (MRP) $\langle \mathcal{S}, \mathcal{P}_R^\pi, P^\pi, \gamma\rangle$ induced by the original MDP and the policy π to be evaluated. Here $\mathcal{P}_R^\pi(\cdot|s) = \sum_{a\in\mathcal{A}} \pi(a|s)\mathcal{P}_R(\cdot|s, a)$ and $P^\pi(\cdot|s) = \sum_{a\in\mathcal{A}} \pi(a|s)P(\cdot|s, a)$. Then we may use samples from the MRP to form the certainty equivalence estimator. With such problem formulation, we no longer need to care about the action space $\mathcal{A}$ and the policy π explicitly. However, the main drawback of this formulation is that the data collection process depends on the policy π. On the other hand, in our problem formulation we can evaluate arbitrary policies with the same predefined dataset, which fits better with real-world applications. This formulation also allows our theoretical results to draw further implications on various fields of reinforcement learning (see Corollary 3.2 and Example 4.3).

2.2. *Metrics on the space of measures.* Suppose μ and ν are two probability distributions on $\mathbb{R}$ with finite pth moments ($p \geq 1$). The p-Wasserstein metric between μ and ν is defined as

$$W_p(\mu, \nu) = \left(\inf_{\kappa\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^p \kappa(dx, dy)\right)^{1/p}.$$

Elements $\kappa \in \Gamma(\mu, \nu)$ are called couplings of μ and ν, that is, joint distributions on $\mathbb{R}^2$ with prescribed marginals μ and ν on each "axis". Suppose μ and ν have cumulative distribution function F_μ and F_ν, respectively. In the case of $p = 1$, we have

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|\, dx.$$

The Kolmogorov–Smirnov (KS) metric is defined as

$$\mathrm{KS}(\mu, \nu) = \sup_{t \in \mathbb{R}} \bigl|\mu\bigl((-\infty, t]\bigr) - \nu\bigl((-\infty, t]\bigr)\bigr|.$$

We may bound $\mathrm{KS}(\mu, \nu)$ with $W_1(\mu, \nu)$ when either of μ, ν has bounded densities.

PROPOSITION 2.1 ([36], Proposition 1.2). *Assume that $\mu \in \Delta(\mathbb{R})$ has finite first moment and μ has a Lebesgue density p_μ that is bounded by C. Then for any $\nu \in \Delta(\mathbb{R})$ with finite first moment, we have* $\mathrm{KS}(\mu, \nu) \le \sqrt{2C W_1(\mu, \nu)}$.

The total variation (TV) metric is defined as

$$\mathrm{TV}(\mu, \nu) = \sup_{A \in \mathcal{B}(\mathbb{R})} \bigl|\mu(A) - \nu(A)\bigr|,$$

where $\mathcal{B}(\mathbb{R})$ denotes the collection of all Borel sets in $\mathbb{R}$. $\mathrm{TV}(\mu, \nu)$ can be also bounded by $W_1(\mu, \nu)$ when μ and ν have smooth densities.

PROPOSITION 2.2 ([5], Theorem 2.1). *Assume that $\mu, \nu \in \Delta(\mathbb{R})$ have Lebesgue densities $p_\mu, p_\nu \in H_1^1(\mathbb{R})$. Specifically, $H_1^1(\mathbb{R})$ represents the L^1 Sobolev space of order* 1 *defined as*

$$H_1^1(\mathbb{R}): = \bigl\{f \in L^1(\mathbb{R}): D^1 f \in L^1(\mathbb{R});\ \|f\|_{H_1^1} = \|f\|_1 + \bigl\|D^1 f\bigr\|_1 < \infty\bigr\}.$$

Here $L^1(\mathbb{R})$ is the space of Lebesgue integrable functions and $\|\cdot\|_1$ is the associated L^1 norm, $D^1 f$ represents the weak derivative of f. Then we have

$$\mathrm{TV}(\mu, \nu) \le \sqrt{K\bigl(\|p_\mu\|_{H_1^1} + \|p_\nu\|_{H_1^1}\bigr) W_1(\mu, \nu)}.$$

Here K is a universal constant.

The 1-Wasserstein metric, the KS metric, and the TV metric are all special cases of integral probability metrics. Specifically, we define

$$\|\mu - \nu\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mu f - \nu f|,$$

where $\mathcal{F}$ denotes some function class and μf represents $\mathbb{E}_{X \sim \mu}[f(X)]$. If we choose

- $\mathcal{F}_{W_1} := \{f | f \text{ is 1-Lipschitz}\}$, then $\|\mu - \nu\|_{\mathcal{F}_{W_1}} = W_1(\mu, \nu)$;
- $\mathcal{F}_{\mathrm{KS}} := \{\mathbb{1}\{\cdot \le z\} | z \in \mathbb{R}\}$, then $\|\mu - \nu\|_{\mathcal{F}_{\mathrm{KS}}} = \mathrm{KS}(\mu, \nu)$; and
- $\mathcal{F}_{\mathrm{TV}} := \{f | f \text{ is measuable and } \operatorname{ess\,sup}_{x \in \mathbb{R}} |f(x)| \le 1\}$, then $\|\mu - \nu\|_{\mathcal{F}_{\mathrm{TV}}} = \mathrm{TV}(\mu, \nu)$.

2.3. *Distributional Bellman operator and distributional dynamic programming.* It is well known that the expected returns (also called the value functions) satisfy the Bellman equation. In particular, letting V^π denote $(V^\pi(s_1), \ldots, V^\pi(s_{|\mathcal{S}|}))$, we have for any $s \in \mathcal{S}$,

$$\begin{aligned} V^\pi(s) &= \bigl[T^\pi\bigl(V^\pi\bigr)\bigr](s) \\ &:= \mathbb{E}_{A \sim \pi(\cdot|s), R \sim \mathcal{P}(\cdot|s,A)} R + \mathbb{E}_{A \sim \pi(\cdot|s), S' \sim P(\cdot|s,A)} V^\pi\bigl(S'\bigr) \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) \int_0^1 r \mathcal{P}_R(dr|s, a) + \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} \pi(a|s) P\bigl(s'|s, a\bigr) V^\pi\bigl(s'\bigr). \end{aligned} \tag{1}$$

We call the operator $T^\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ the Bellman operator. And the Bellman equation says that the value function V^π is a fixed point of T^π.

The distributional Bellman equation describes a similar relationship to Equation (1) for the return distributions. Letting η^π denote $(\eta^\pi(s_1), \dots, \eta^\pi(s_{|\mathcal{S}|}))$, we have for any $s \in \mathcal{S}$

$$\begin{aligned}\eta^\pi(s) &= [\mathcal{T}^\pi(\eta^\pi)](s) \\ &:= \mathbb{E}_{A\sim\pi(\cdot|s),R\sim\mathcal{P}_R(\cdot|s,A),S'\sim P(\cdot|s,A)}(b_{R,\gamma})_\#\eta^\pi(S') \\ &= \sum_{a\in\mathcal{A},s'\in\mathcal{S}} \pi(a|s)P(s'|s,a)\int_0^1 (b_{r,\gamma})_\#\eta^\pi(s')\mathcal{P}_R(dr|s,a).\end{aligned} \tag{2}$$

Here $b_{r,\gamma}: \mathbb{R} \to \mathbb{R}$ is an affine function defined by $b_{r,\gamma}(x) = r + \gamma x$, and $g_\#\mu$ is the push-forward measure of μ through function g so that $g_\#\mu(A) = \mu(g^{-1}(A))$ for any Borel set A. The integral $\int_0^1 (b_{r,\gamma})_\#\eta^\pi(s')\mathcal{P}_R(dr|s,a)$ is defined by

$$\left[\int_0^1 (b_{r,\gamma})_\#\eta^\pi(s')\mathcal{P}_R(dr|s,a)\right](B) = \int_0^1 [(b_{r,\gamma})_\#\eta^\pi(s')](B)\mathcal{P}_R(dr|s,a)$$

for any Borel set B in $[0, (1-\gamma)^{-1}]$. We call the operator $\mathcal{T}^\pi: \Delta([0,(1-\gamma)^{-1}])^{\mathcal{S}} \to \Delta([0,(1-\gamma)^{-1}])^{\mathcal{S}}$ the distributional Bellman operator, and the vector of return distributions η^π is a fixed point of $\mathcal{T}^\pi$.

Suppose the MDP M is already known. For a policy π, we compute the value function V^π by the dynamic programming algorithm. Specifically, assuming $V_{k+1} = T^\pi(V_k)$, we have T^π is a γ-contraction w.r.t. the supremum norm $\|\cdot\|_\infty$ on $\mathbb{R}^{\mathcal{S}}$, and thus $\lim_{k\to\infty}\|V_k - V^\pi\|_\infty = 0$. In analogy to dynamic programming, we also define distributional dynamic programming, that is, $\eta^{(k+1)} = \mathcal{T}^\pi\eta^{(k)}$. It can be shown that $\mathcal{T}^\pi$ is a γ-contraction in the supremum p-Wasserstein metric. Thus, distributional dynamic programming exhibits the geometric convergence in the supremum p-Wasserstein metric.

PROPOSITION 2.3 ([4], Propositions 4.15 and 4.16). *The distributional Bellman operator is a γ-contraction on $\Delta(\mathbb{R})^{\mathcal{S}}$ in the supremum p-Wasserstein metric, for any $p \geq 1$. More precisely, for $\eta, \eta' \in \Delta(\mathbb{R})^{\mathcal{S}}$, we have*

$$\sup_{s\in\mathcal{S}} W_p([\mathcal{T}^\pi\eta](s), [\mathcal{T}^\pi\eta'](s)) \leq \gamma \sup_{s\in\mathcal{S}} W_p(\eta(s), \eta'(s)).$$

Furthermore, we have

$$\sup_{s\in\mathcal{S}} W_p(\eta^{(k)}(s), \eta^\pi(s)) \leq \gamma^k \sup_{s\in\mathcal{S}} W_p(\eta^{(0)}(s), \eta^\pi(s))$$

and

$$\lim_{k\to\infty}\sup_{s\in\mathcal{S}} W_p(\eta^{(k)}(s), \eta^\pi(s)) = 0.$$

When measured by other commonly used probability metrics like the supremum KS metric or the supremum TV metric, the distributional Bellman operator might not be a (strict) contraction and distributional dynamic programming would not converge at all [4]. This is because, unlike the cases of dynamic programming, now we operate in an infinite-dimensional space and a metric in an infinite-dimensional space may not be equivalent to one another. However, we find that under mild conditions, distributional dynamic programming does converge in the supremum KS metric and the supremum TV metric, and the convergences are also geometrically fast. To the best of our knowledge, we are the first to examine the convergence property of distributional dynamic programming w.r.t. the supremum KS metric and the supremum TV metric.

ASSUMPTION 1. Assume that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $\mathcal{P}_R(dr|s,a)$ has a Lebesgue density $p_{s,a}^R$ upper-bounded by a constant C.

ASSUMPTION 2. Assume that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $\mathcal{P}_R(dr|s,a)$ has a Lebesgue density $p_{s,a}^R \in H_1^1(\mathbb{R})$ and $\|p_{s,a}^R\|_{H_1^1(\mathbb{R})} \le M$.

Assumption 2 is strictly stronger than Assumption 1 as a consequence of Sobolev's inequality.

PROPOSITION 2.4. *The distributional Bellman operator is nonexpansive on* $\Delta(\mathbb{R})^{\mathcal{S}}$ *in the supremum KS metric. Moreover, if Assumption* 1 *holds, we have*

$$\sup_{s\in\mathcal{S}} \mathrm{KS}\big(\eta^{(k)}(s), \eta^\pi(s)\big) \le (\sqrt{\gamma})^k \sup_{s\in\mathcal{S}} \sqrt{C W_1\big(\eta^{(0)}(s), \eta^\pi(s)\big)}.$$

PROPOSITION 2.5. *The distributional Bellman operator is nonexpansive on* $\Delta(\mathbb{R})^{\mathcal{S}}$ *in the supremum TV metrics. If Assumption* 2 *holds, we also have*

$$\sup_{s\in\mathcal{S}} \mathrm{TV}\big(\eta^{(k)}(s), \eta^\pi(s)\big) \le (\sqrt{\gamma})^k \sup_{s\in\mathcal{S}} \sqrt{2MKW_1\big(\eta^{0}(s), \eta^\pi(s)\big)}.$$

To prove Proposition 2.4, we first show that when Assumption 1 is true, the distribution of return $\eta^\pi(s)$ must have a bounded density. Then by Proposition 2.1 the KS metric can be controlled by the 1-Wasserstein metric. The proof strategy of Proposition 2.5 is similar to that of Proposition 2.4. We first show that when Assumption 2 holds, the $\eta^\pi(s)$ and $\eta^{(k)}(s)$ both have densities in $H_1^1(\mathbb{R})$. Then we can bound the TV metric with the 1-Wasserstein metric through Proposition 2.2. The full proof can be found in Section 6 in the Supplementary Material [58].

For the certainty equivalence estimator $\hat{\eta}_n^\pi$, we also have the following distributional Bellman equation

$$\begin{aligned}\hat{\eta}_n^\pi(s) &= \big[\widehat{\mathcal{T}}_n^\pi(\hat{\eta}_n^\pi)\big](s)\\ &:= \mathbb{E}_{A\sim\pi(\cdot|s), R\sim\mathcal{P}(\cdot|s,A), S'\sim\widehat{P}_n(\cdot|s,A)}(b_{R,\gamma})_\#\hat{\eta}_n^\pi(S')\\ &= \sum_{a\in\mathcal{A}, s'\in\mathcal{S}} \pi(a|s)\widehat{P}_n(s'|s,a)\int_0^1 (b_{r,\gamma})_\#\hat{\eta}_n^\pi(s')\mathcal{P}_R(dr|s,a),\end{aligned} \tag{3}$$

where $\widehat{\mathcal{T}}_n^\pi$ is called the empirical distributional Bellman operator. Thus, $\hat{\eta}_n^\pi$ can be computed via the empirical version of distributional dynamic programming, that is, $\hat{\eta}^{(k+1)} = \widehat{\mathcal{T}}_n^\pi(\hat{\eta}^{(k)})$.

3. Statistical analysis. In this section, we analyze distributional reinforcement learning from both the nonasymptotic and asymptotic viewpoints. We give the nonasymptotic convergence rates of $\sup_{s\in\mathcal{S}} W_1(\hat{\eta}_n^\pi(s), \eta^\pi(s))$, $\sup_{s\in\mathcal{S}} \mathrm{KS}(\hat{\eta}_n^\pi(s), \eta^\pi(s))$, and $\sup_{s\in\mathcal{S}} \mathrm{TV}(\hat{\eta}_n^\pi(s), \eta^\pi(s))$, which suggest distributional policy evaluation is sample-efficient under a generative model setting. We also study the asymptotics of $\sqrt{n}(\hat{\eta}_n^\pi(s) - \eta^\pi(s))$ for any $s \in \mathcal{S}$. Under mild conditions, we demonstrate that $\sqrt{n}(\hat{\eta}_n^\pi(s) - \eta^\pi(s))$ converges weakly to a Gaussian random element in the spaces $\ell^\infty(\mathcal{F}_{W_1})$, $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$ and $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$.

3.1. *Results on nonasymptotic analysis.* Our main results of nonasymptotic analysis is given in the following theorems.

THEOREM 3.1. *For any fixed policy π, we have that*

$$\mathbb{E}\sup_{s\in\mathcal{S}} W_1\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big) \le \sqrt{\frac{9\log|\mathcal{S}|}{n(1-\gamma)^4}},$$

and that for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\sup_{s\in\mathcal{S}} W_1\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big) \le \frac{\sqrt{9\log|\mathcal{S}|}+\sqrt{\log(1/\delta)/2}}{\sqrt{n(1-\gamma)^4}}.$$

To sum up, we show $n=\widetilde{O}(\varepsilon^{-2}(1-\gamma)^{-4})$ suffices to ensure $\mathbb{E}\sup_{s\in\mathcal{S}} W_1(\hat{\eta}_n^\pi(s), \eta^\pi(s))\le\varepsilon$ and $\sup_{s\in\mathcal{S}} W_1(\hat{\eta}_n^\pi(s),\eta^\pi(s))\le\varepsilon$ with high probability, which implies model-based distributional policy evaluation is sample-efficient. The key idea of our proof is that we first analyze the concentration behaviors of the infinite-dimensional operator $(\widehat{\mathcal{T}}_n^\pi-\mathcal{T}^\pi)$. Then we examine the properties of $\mathcal{T}^\pi$ in the vector space of signed measures equipped with the 1-Wasserstein metric and give a reasonable definition of $(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}:=\sum_{i=0}^\infty(\widehat{\mathcal{T}}_n^\pi)^i$ on a product of vector spaces consisting of signed measures μ such that $\mu([0,(1-\gamma)^{-1}])=0$. This allows us to write $\hat{\eta}_n-\eta^\pi=(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}(\widehat{\mathcal{T}}_n^\pi-\mathcal{T}^\pi)\eta^\pi$. We can draw the conclusion noting that the operator norm of $(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}$ w.r.t. the 1-Wasserstein metric is always bounded by $(1-\gamma)^{-1}$. The detailed proof is given in Section 2 in the Supplementary Material [58].

Compared with the minimax optimal $\widetilde{O}(\varepsilon^{-2}(1-\gamma)^{-3})$ sample complexity bound for the model-based policy evaluation [24], our sample complexity bound has an additional $(1-\gamma)^{-1}$ factor. In fact, learning the distribution of returns is harder than the classic policy evaluation problem, because we always have $|V(s)-\widehat{V}(s)|\le\varepsilon$ as long as $W_1(\eta^\pi(s),\hat{\eta}_n^\pi(s))\le\varepsilon$. However, we speculate that the additional $(1-\gamma)^{-1}$ factor can be eliminated with more refined analysis techniques specially developed for handling infinite-dimensional cases.

Combining Theorem 3.1 with the fact $[W_p(\hat{\eta}_n^\pi(s),\eta^\pi(s))]^p\le(1-\gamma)^{1-p}W_1(\hat{\eta}_n^\pi(s), \eta^\pi(s))$, we can derive the nonasymptotic results for the p-Wasserstein metric.

COROLLARY 3.1. *For any fixed policy π and $p>1$, we have that*

$$\mathbb{E}\sup_{s\in\mathcal{S}} W_p\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big)\le\left[\frac{9\log|\mathcal{S}|}{n(1-\gamma)^{2p+2}}\right]^{\frac{1}{2p}},$$

and that for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\sup_{s\in\mathcal{S}} W_p\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big)\le\left[\frac{\sqrt{9\log|\mathcal{S}|}+\sqrt{\log(1/\delta)/2}}{\sqrt{n(1-\gamma)^{2p+2}}}\right]^{\frac{1}{p}}.$$

Note that the slow rate $n^{-1/(2p)}$ for the p-Wasserstein metric is inevitable without assuming additional regularity conditions. Consider an MDP with $\mathcal{S}=\{s_1,s_2,s_3\}$ and $\mathcal{A}=\{a_1\}$. As there is only one single action a_1 available, the action variable can be omitted. We start from s_1 with deterministic reward $r(s_1)=0$ and $P(s_2|s_1)=P(s_3|s_1)=\frac{1}{2}$. And s_2 and s_3 are absorbing states with deterministic rewards $r(s_2)=1$ and $r(s_3)=0$. Suppose that after n calls of the generative model we have gained an estimator of $P(s_2|s_1)$ (denoted $\hat{p}$), then $\eta(s_1)=\frac{1}{2}\delta_0+\frac{1}{2}\delta_{(1-\gamma)^{-1}}$ and $\hat{\eta}_n(s_1)=(1-\hat{p})\delta_0+\hat{p}\delta_{(1-\gamma)^{-1}}$. We have

$$W_p\big(\eta(s_1),\hat{\eta}_n(s_1)\big)=\frac{1}{1-\gamma}\left|\hat{p}-\frac{1}{2}\right|^{\frac{1}{p}}.$$

Since $\hat{p}\sim\mathsf{Binomial}(n,\frac{1}{2})$, $|\hat{p}-\frac{1}{2}|$ is of the order $n^{-1/2}$ by CLT. Thus, $W_p(\eta(s_1),\hat{\eta}_n(s_1))$ is of order $n^{-1/(2p)}$.

Under Assumption 1, we also have the following bounds on the KS metric.

THEOREM 3.2. *Suppose Assumption 1 holds. For any fixed policy π,*

$$\mathbb{E}\sup_{s\in\mathcal{S}}\mathrm{KS}\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big)\le C'\sqrt{\frac{\log|\mathcal{S}|}{n(1-\gamma)^4}}.$$

And for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\sup_{s\in\mathcal{S}}\mathrm{KS}\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big)\le\frac{C''(\sqrt{\log|\mathcal{S}|}+\sqrt{\log(1/\delta)})}{\sqrt{n(1-\gamma)^4}}.$$

Here C' and C'' are constants only depending on C in Assumption 1.

The upper bound of the supremum KS metric between $\hat{\eta}_n^\pi$ and η^π is of the same order as that of the supremum 1-Wasserstein metric, which indicates that under mild conditions learning a near-optimal return distribution in the sense of the KS metric is not more difficult than learning a near-optimal return distribution in the sense of the 1-Wasserstein metric. This is somewhat a surprise because the distributional Bellman operator exhibits benign behaviors only when measured by Wasserstein metrics, and for μ, ν with bounded support $W_1(\mu,\nu)$ can be always bounded by $\mathrm{KS}(\mu,\nu)$ multiplying a constant factor.

Simply combining the results in Theorem 3.1 and Proposition 2.1 can only yield a suboptimal $n^{-1/4}$ convergence rate for $\sup_{s\in\mathcal{S}}\mathrm{KS}(\hat{\eta}_n^\pi(s),\eta^\pi(s))$. Instead, we obtain a $n^{-1/2}$ rate using a quite different proof strategy from that of Theorem 3.1. The first challenge is that the operator $(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}$ may be unbounded on its domain measured with the norm induced by the KS metric. Therefore, although we can still write $\hat{\eta}_n^\pi-\eta^\pi=(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}(\widehat{\mathcal{T}}_n^\pi-\mathcal{T}^\pi)\eta^\pi$, bounds of $(\widehat{\mathcal{T}}_n^\pi-\mathcal{T}^\pi)\eta^\pi$ cannot be directly translated to bounds of $\hat{\eta}_n^\pi-\eta^\pi$. We handle this challenge by an “expansion trick”, which raises yet another technical challenge: we need a stronger notion of concentration of $(\widehat{\mathcal{T}}_n^\pi-\mathcal{T}^\pi)\eta^\pi$. Specifically, unlike in the proof of Theorem 3.1 where it suffices to bound $W_1(\widehat{\mathcal{T}}_n^\pi\eta^\pi,\mathcal{T}^\pi\eta^\pi)$, here we further need to bound $\mathrm{TV}(\widehat{\mathcal{T}}_n^\pi\eta^\pi,\mathcal{T}^\pi\eta^\pi)$. We achieve this with an analysis through the lens of density functions. The detailed proof can be found in Section 3 in the Supplementary Material [58].

THEOREM 3.3. *Suppose Assumption 2 holds. For any fixed policy π,*

$$\mathbb{E}\sup_{s\in\mathcal{S}}\mathrm{TV}\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big)\le K'\sqrt{\frac{\log|\mathcal{S}|}{n(1-\gamma)^4}}.$$

And for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\sup_{s\in\mathcal{S}}\mathrm{TV}\big(\hat{\eta}_n^\pi(s),\eta^\pi(s)\big)\le\frac{K''(\sqrt{\log|\mathcal{S}|}+\sqrt{\log(1/\delta)})}{\sqrt{n(1-\gamma)^4}}.$$

Here K' and K'' are absolute constants that depend only on M in Assumption 2.

Based on the above theorem, we observe that our upper bounds for the supremum TV metric are of the same order as those for the supremum 1-Wasserstein metric or the supremum KS metric. Directly applying Theorem 3.1 and Proposition 2.2 only attains a slow $n^{-1/4}$ rate. We rather employ a similar analytical approach as in the case of the KS metric to establish a standard convergence rate of $n^{-1/2}$. The detailed proof is given in Section 4 in the Supplementary Saterial [58].

Extension I: *Less Exploratory Offline Dataset*. In our analysis, we assume that the offline dataset is obtained using a generative model. We can relax such an assumption to that the dataset is generated from some probability measure ξ and the transition dynamic P [7]. Specifically, we first sample a batch of state-action pairs $\{(s_i, a_i)\}_{i=1}^m$ with ξ. For a state-action pair (s_i, a_i), we sample the next-state s_i' according to $P(\cdot|s_i, a_i)$. Then we have the dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^m$. The empirical estimate of the transition probability is constructed as follows:

$$\widehat{P}_m(s'|s, a) = \frac{\sum_{i=1}^m \mathbb{1}\{(s_i, a_i, s_i') = (s, a, s')\}}{1 \vee \sum_{i=1}^m \mathbb{1}\{(s_i, a_i) = (s, a)\}}.$$

We define the certainty equivalence estimator $\hat{\eta}_m^\pi$ with $\widehat{P}_m$ in the same way as before.

THEOREM 3.4. *Let* $\xi_{\min} = \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \xi(s, a)$. *For any fixed policy* π *and for any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *as long as* $m \geq 8\log(2|\mathcal{S}||\mathcal{A}|/\delta)/\xi_{\min}$, *we have*:

(a) $\sup_{s\in\mathcal{S}} W_1(\hat{\eta}_m^\pi(s), \eta^\pi(s)) \leq \frac{\sqrt{18\log|\mathcal{S}|}+\sqrt{\log(2/\delta)}}{\sqrt{\xi_{\min} m(1-\gamma)^4}}$.

(b) $\sup_{s\in\mathcal{S}} \mathrm{KS}(\hat{\eta}_m^\pi(s), \eta^\pi(s)) \leq \frac{C'(\sqrt{\log|\mathcal{S}|}+\sqrt{\log(1/\delta)})}{\sqrt{\xi_{\min} m(1-\gamma)^4}}$ *when Assumption* 1 *is true. Here* C' *is a constant only depending on* C *in Assumption* 1.

(c) $\sup_{s\in\mathcal{S}} \mathrm{TV}(\hat{\eta}_m^\pi(s), \eta^\pi(s)) \leq \frac{K'(\sqrt{\log|\mathcal{S}|}+\sqrt{\log(1/\delta)})}{\sqrt{\xi_{\min} m(1-\gamma)^4}}$ *when Assumption* 2 *is true. Here* K' *is a constant only depending on* M *in Assumption* 2.

One may refer to Section 7 in the Supplementary Material [58] for the detailed proof. The dependence on $\xi_{\min}$ is inevitable because our aim is to bound the l_∞-type estimation error (e.g., $\sup_{s\in\mathcal{S}} W_1(\hat{\eta}_m^\pi(s), \eta^\pi(s))$). See Example 7.1 in the Supplementary Material [58] for more concrete discussions. And we also observe such a phenomenon in the asymptotic results (see Theorem 3.7). If ξ is the uniform distribution and $\xi_{\min} = 1/(|\mathcal{S}||\mathcal{A}|)$, then the bounds here are equivalent to the bounds in Theorems 3.1, 3.2, 3.3 up to constant factors.

Extension II: *Unknown Reward Distributions*. In our previous analysis, we assume that the reward distribution $\mathcal{P}_R$ is known. Here we remove this assumption and extend our analysis to the scenario where the reward distribution is estimated on a set of finite samples. Specifically, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, in addition to the n next-state samples $X_1^{(s,a)}, \dots, X_n^{(s,a)} \overset{\text{i.i.d.}}{\sim} P(\cdot|s, a)$ used to obtain the estimator $\widehat{P}_n(\cdot|s, a)$, we also sample n rewards $R_1^{(s,a)}, \dots, R_n^{(s,a)} \overset{\text{i.i.d.}}{\sim} \mathcal{P}_R(\cdot|s, a)$ to estimate $\mathcal{P}_R(\cdot|s, a)$. We denote the estimated reward distribution as $\widehat{\mathcal{P}}_{R,n}$, and the empirical distributional Bellman operator as $\widetilde{\mathcal{T}}_n^\pi$. The explicit forms of $\widehat{\mathcal{P}}_{R,n}$ and $\widetilde{\mathcal{T}}_n^\pi$ will be given later, depending on the metric we choose. Now, we define the estimator $\tilde{\eta}_n^\pi$ as the solution to the fixed point equation $\eta = \widetilde{\mathcal{T}}_n^\pi \eta$.

THEOREM 3.5. *For any fixed policy* π *and for any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *we have*:

(a) $\sup_{s\in\mathcal{S}} W_1(\tilde{\eta}_n^\pi(s), \eta^\pi(s)) \leq \frac{\sqrt{9\log|\mathcal{S}|}+\sqrt{\log(1/\delta)/2}}{\sqrt{n(1-\gamma)^4}}$. *In this case, we directly use the empirical reward distributions as an estimator of* $\mathcal{P}_R$, *and the empirical distributional Bellman operator is given by*

$$[\widetilde{\mathcal{T}}_n^\pi(\eta)](s) = \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}} \pi(a|s)(b_{R_i^{(s,a)},\gamma})_\# \eta(X_i^{(s,a)}).$$

(b) $\sup_{s\in\mathcal{S}} \mathrm{KS}(\tilde{\eta}_n^\pi(s), \eta^\pi(s)) \le \frac{C'(\sqrt{\log|\mathcal{S}|}+\sqrt{\log(1/\delta)})}{\sqrt{n(1-\gamma)^4}} + \frac{C'}{1-\gamma}\sup_{s\in\mathcal{S},a\in\mathcal{A}}\|\widehat{\mathcal{P}}_{R,n}(\cdot|s,a) - \mathcal{P}_R(\cdot|s,a)\|_{\mathcal{F}_{\mathrm{TV}}}$ *when Assumption* 1 *is true. Here* C' *is a constant only depending on* C *in Assumption* 1. *In this case,* $\widehat{\mathcal{P}}_{R,n}(\cdot|s,a)$ *can be any density estimator as long as it has a Lebesgue density upper-bounded by* $2C$ *for each* $(s,a)\in\mathcal{S}\times\mathcal{A}$, *and* $\widetilde{\mathcal{T}}_n^\pi$ *is the distributional Bellman operator of the empirical MDP* $\widetilde{M} = \langle\mathcal{S},\mathcal{A},\widehat{\mathcal{P}}_{R,n},\widehat{P}_n,\gamma\rangle$.

(c) $\sup_{s\in\mathcal{S}} \mathrm{TV}(\tilde{\eta}_n^\pi(s), \eta^\pi(s)) \le \frac{K'(\sqrt{\log|\mathcal{S}|}+\sqrt{\log(1/\delta)})}{\sqrt{n(1-\gamma)^4}} + \frac{K'}{1-\gamma}\sup_{s\in\mathcal{S},a\in\mathcal{A}}\|\widehat{\mathcal{P}}_{R,n}(\cdot|s,a) - \mathcal{P}_R(\cdot|s,a)\|_{\mathcal{F}_{\mathrm{TV}}}$ *when Assumption* 2 *is true. Here* K' *is a constant only depending on* M *in Assumption* 2. *In this case,* $\widehat{\mathcal{P}}_{R,n}(\cdot|s,a)$ *can be any density estimator as long as it has a Lebesgue density* $\hat{p}_{s,a}^R \in H_1^1(\mathbb{R})$ *with* $\|\hat{p}_{s,a}^R\|_{H_1^1(\mathbb{R})} \le 2M$, *and* $\widetilde{\mathcal{T}}_n^\pi$ *is the distributional Bellman operator of the empirical MDP* $\widetilde{M} = \langle\mathcal{S},\mathcal{A},\widehat{\mathcal{P}}_{R,n},\widehat{P}_n,\gamma\rangle$.

See Section 7 in the Supplementary Material [58] for a detailed proof. Our nonasymptotic bounds here depend on specific choices of the reward estimator $\widehat{\mathcal{P}}_{R,n}$. One can refer to Remark 1 in the Supplementary Material [58] for some choices of density estimators and the corresponding nonasymptotic bounds of total variation $\|\widehat{\mathcal{P}}_{R,n}(\cdot|s,a) - \mathcal{P}_R(\cdot|s,a)\|_{\mathcal{F}_{\mathrm{TV}}}$.

Implications in Risk-sensitive RL. We conclude this section with a brief discussion that highlights the implications of our results in the field of risk-sensitive RL. The main goal of risk-sensitive RL is to find a policy $\pi^\star$ minimizing a risk functional of the return distribution. Concretely, we define $L_\rho(\pi) := \rho(\eta^\pi(s))$ and $\pi^\star := \arg\min_\pi L_\rho(\pi)$. Here $\rho(\cdot)$ is some risk functional such as value-at-risk or conditional value-at-risk, and s the initial state. When the underlying MDP is not explicitly known, π^* is estimated by $\hat{\pi} := \arg\min_\pi \widehat{L}_\rho(\pi)$, where $\widehat{L}_\rho(\pi) := \rho(\hat{\eta}_n^\pi(s))$. If $\rho(\cdot)$ is Lipschitz continuous w.r.t. some probability metric, then we derive the following sample complexity bounds with our above results.

COROLLARY 3.2. *It is sufficient for* $n = \widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4})$ *to guarantee* $L_\rho(\hat{\pi}) - L_\rho(\pi^\star) \le \varepsilon$ *with probability at least* $1-\delta$ *as long as one of the following conditions is true*:

(a) $\rho(\cdot)$ *is Lipschitz continuous w.r.t. the* 1*-Wasserstein metric*;
(b) $\rho(\cdot)$ *is Lipschitz continuous w.r.t. the KS metric and Assumption* 1 *holds*;
(c) $\rho(\cdot)$ *is Lipschitz continuous w.r.t. the TV metric and Assumption* 2 *holds.*

The Lipschitz condition is common because it covers a wide range of risk measures, including the class of distortion risk measures, convex and coherent measures, etc. [27].

The idea behind Corollary 3.2 is straightforward. As long as we have nonasymptotic bounds for $d(\eta^\pi(s), \hat{\eta}_n^\pi(s))$ that hold for arbitrary fixed π (Theorems 3.1, 3.2, 3.3), then we use the covering argument to bound $\sup_\pi d(\eta^\pi(s), \hat{\eta}_n^\pi(s))$. And such uniform convergence results can further lead to the bounds $L_\rho(\hat{\pi}) - L_\rho(\hat{\pi}) \le \varepsilon$.

3.2. *Results on asymptotic analysis.* We first give our main results of the asymptotic analysis in Theorem 3.6.

THEOREM 3.6. *For any fixed policy* π, *we have for any* $s\in\mathcal{S}$:

(a) $\sqrt{n}(\hat{\eta}_n^\pi(s) - \eta^\pi(s))$ *converges weakly to the process* $f \mapsto [(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f$ *in* $\ell^\infty(\mathcal{F}_{W_1})$, *where* $\mathcal{F}_{W_1} := \{f | f$ *is supported on* $[0,(1-\gamma)^{-1}]$ *and* 1*-Lipschitz*$\}$;

(b) *If Assumption* 1 *is true, then* $\sqrt{n}(\hat{\eta}_n^\pi(s) - \eta^\pi(s))$ *converges weakly to the process* $f \mapsto [(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f$ *in* $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$, *where* $\mathcal{F}_{\mathrm{KS}} := \{\mathbb{1}_{(-\infty,z]} | z\in[0,(1-\gamma)^{-1}]\}$;

(c) *If Assumption* 2 *is true, then* $\sqrt{n}(\hat{\eta}_n^\pi(s) - \eta^\pi(s))$ *converges weakly to the process* $f \mapsto [(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f$ *in* $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$, *where* $\mathcal{F}_{\mathrm{TV}} := \{\mathbb{1}_A | A\subseteq[0,(1-\gamma)^{-1}]$ *is Borel*$\}$.

Here the random element $\widetilde{\mathbb{G}}^{\pi}$ *is defined as*

$$\widetilde{\mathbb{G}}^{\pi}(s) := \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} Z_{s,a,s'} \int_0^1 (b_{r,\gamma})_{\#}\eta^{\pi}(s')\mathcal{P}_R(dr|s,a) \quad \forall s \in \mathcal{S},$$

where $(Z_{s,a,s'})_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ *are mean-zero Gaussians with*

$$\mathsf{Cov}(Z_{s_1,a_1,s_1'}, Z_{s_2,a_2,s_2'}) = \mathbb{1}\{(s_1,a_1)=(s_2,a_2)\}P(s_1'|s_1,a_1)(\mathbb{1}\{s_1'=s_2'\} - P(s_2'|s_1,a_1)).$$

The operator $(\mathcal{I}-\mathcal{T}^{\pi})^{-1}$ *is defined as* $(\mathcal{I}-\mathcal{T}^{\pi})^{-1} := \sum_{i=0}^{\infty}(\mathcal{T}^{\pi})^{i}$.

At a high level, we depict the asymptotic behavior of $\sqrt{n}(\hat{\eta}_n^{\pi} - \eta^{\pi})$ by showing that the "empirical processes" induced by $\sqrt{n}(\hat{\eta}_n^{\pi} - \eta^{\pi})$ converge to a Gaussian random element. Moreover, the limiting random element has a simple structure in the sense that it is a linear transformation applied to a finite mixture of probability distributions with Gaussian coefficients. Our asymptotic results are general in the sense that the conclusions are valid in different spaces under different regularity conditions: $\ell^{\infty}(\mathcal{F}_{W_1})$, $\ell^{\infty}(\mathcal{F}_{\mathrm{KS}})$, and $\ell^{\infty}(\mathcal{F}_{\mathrm{TV}})$. Therefore, our findings have the potential to yield numerous valuable inferential procedures for the field of distributional reinforcement learning. Our proof of Theorem 3.6 is built on the foundation of our nonasymptotic analysis in Section 3. The detailed proof is given in Section 5 in the Supplementary Material [58].

Extension I: *Less Exploratory Offline Dataset.* We also present asymptotic results in the setting of less exploratory dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^{m}$.

THEOREM 3.7. *For any fixed policy* π, *we have for any* $s \in \mathcal{S}$:

(a) $\sqrt{m}(\hat{\eta}_m^{\pi}(s) - \eta^{\pi}(s))$ *converges weakly to the process* $f \mapsto [(\mathcal{I}-\mathcal{T}^{\pi})^{-1}\mathring{\mathbb{G}}^{\pi}](s)f$ *in* $\ell^{\infty}(\mathcal{F}_{W_1})$;

(b) *If Assumption* 1 *is true, then* $\sqrt{m}(\hat{\eta}_m^{\pi}(s) - \eta^{\pi}(s))$ *converges weakly to the process* $f \mapsto [(\mathcal{I}-\mathcal{T}^{\pi})^{-1}\mathring{\mathbb{G}}^{\pi}](s)f$ *in* $\ell^{\infty}(\mathcal{F}_{\mathrm{KS}})$;

(c) *If Assumption* 2 *is true, then* $\sqrt{m}(\hat{\eta}_m^{\pi}(s) - \eta^{\pi}(s))$ *converges weakly to the process* $f \mapsto [(\mathcal{I}-\mathcal{T}^{\pi})^{-1}\mathring{\mathbb{G}}^{\pi}](s)f$ *in* $\ell^{\infty}(\mathcal{F}_{\mathrm{TV}})$.

Here the random element $\mathring{\mathbb{G}}^{\pi}$ *is defined as*

$$\mathring{\mathbb{G}}^{\pi}(s) := \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \frac{\mathring{Z}_{s,a,s'}}{\sqrt{\xi(s,a)}} \int_0^1 (b_{r,\gamma})_{\#}\eta^{\pi}(s')\mathcal{P}_R(dr|s,a) \quad \forall s \in \mathcal{S},$$

where $(\mathring{Z}_{s,a,s'})_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ *are mean-zero Gaussians with*

$$\begin{aligned}&\mathsf{Cov}(\mathring{Z}_{s_1,a_1,s_1'}, \mathring{Z}_{s_2,a_2,s_2'}) \\ &\quad = \mathbb{1}\{(s_1,a_1)=(s_2,a_2)\}P(s_1'|s_1,a_1)(\mathbb{1}\{s_1'=s_2'\} - \xi(s_1,a_1)P(s_2'|s_1,a_1)).\end{aligned}$$

The main difference between the limiting random elements $\mathring{\mathbb{G}}^{\pi}$ and $\widetilde{\mathbb{G}}^{\pi}$ is that $\mathring{\mathbb{G}}^{\pi}$ has larger "variance." This is because we introduce additional randomness in the process of sampling the state-action pairs from distribution ξ. We also observe the $1/\sqrt{\xi(s,a)}$ factor in $\mathring{\mathcal{B}}^{\pi}(s)$. This suggests that the l_{∞}-type error bound would inevitably depend on the factor $1/\sqrt{\xi_{\min}}$ as in Theorem 3.4.

Extension II: *Unknown Reward Distributions.* We give the asymptotic results in the setting of unknown reward distributions.

THEOREM 3.8. *For any fixed policy π, we have for any $s \in \mathcal{S}$:*

(a) *$\sqrt{n}(\tilde{\eta}_n^\pi(s) - \eta^\pi(s))$ converges weakly to the process $f \mapsto [(\mathcal{I} - \mathcal{T}^\pi)^{-1}\bar{\mathbb{G}}^\pi](s)f$ in $\ell^\infty(\mathcal{F}_{\mathrm{W}_1})$. Here $\widetilde{\mathcal{T}}_n^\pi$ is defined as in part (a) of Theorem 3.5, and the random element $\bar{\mathbb{G}}^\pi(s)$ is a zero-mean Gaussian process in $\ell^\infty(\mathcal{F}_{\mathrm{W}_1})$ with covariance function: $\forall f, g \in \mathcal{F}_{\mathrm{W}_1}$,*

$$
\begin{aligned}
&\mathsf{Cov}(\bar{\mathbb{G}}^\pi(s)f, \bar{\mathbb{G}}^\pi(s)g) \\
&\quad= \sum_{a\in\mathcal{A}} \pi(a|s)^2 \biggl\{ \sum_{s'\in\mathcal{S}} P(s'|s,a) \int_0^1 [(b_{r,\gamma})_\# \eta^\pi(s') f][(b_{r,\gamma})_\# \eta^\pi(s') g] \mathcal{P}_R(dr|s,a) \\
&\qquad - \biggl[\sum_{s'\in\mathcal{S}} P(s'|s,a) \int_0^1 (b_{r,\gamma})_\# \eta^\pi(s') f \mathcal{P}_R(dr|s,a) \biggr] \\
&\qquad \times \biggl[\sum_{s'\in\mathcal{S}} P(s'|s,a) \int_0^1 (b_{r,\gamma})_\# \eta^\pi(s') g \mathcal{P}_R(dr|s,a) \biggr] \biggr\}.
\end{aligned}
$$

(b) *If Assumption 1 is true and $\sqrt{n}(\widehat{\mathcal{P}}_{R,n}(\cdot|s,a) - \mathcal{P}_R(\cdot|s,a))$ converges weakly to a tight random element $\mathbb{G}_{s,a}^R$ in $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then $\sqrt{n}(\tilde{\eta}_n^\pi(s) - \eta^\pi(s))$ converges weakly to the process $f \mapsto [(\mathcal{I} - \mathcal{T}^\pi)^{-1}(\widetilde{\mathbb{G}}^\pi + \mathbb{G}_R^\pi)](s)f$ in $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$, where $\widetilde{\mathbb{G}}^\pi$ is defined in Theorem 3.6 and $\mathbb{G}_R^\pi(s)$ is independent of $\widetilde{\mathbb{G}}^\pi$ and given by*

$$
\mathbb{G}_R^\pi(s) = \sum_{a\in\mathcal{A}, s'\in\mathcal{S}} \pi(a|s) P(s'|s,a) \int_0^1 (b_{r,\gamma})_\# \eta^\pi(s') \mathbb{G}_{s,a}^R(dr).
$$

(c) *If Assumption 2 is true and $\sqrt{n}(\widehat{\mathcal{P}}_{R,n}(\cdot|s,a) - \mathcal{P}_R(\cdot|s,a))$ converges weakly to a tight random element $\mathbb{G}_{s,a}^R$ in $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then $\sqrt{n}(\tilde{\eta}_n^\pi(s) - \eta^\pi(s))$ converges weakly to the process $f \mapsto [(\mathcal{I} - \mathcal{T}^\pi)^{-1}(\widetilde{\mathbb{G}}^\pi + \mathbb{G}_R^\pi)](s)f$ in $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$, where $\widetilde{\mathbb{G}}^\pi$ is given in Theorem 3.6 and $\mathbb{G}_R^\pi(s)$ is independent of $\widetilde{\mathbb{G}}^\pi$ and given by*

$$
\mathbb{G}_R^\pi(s) = \sum_{a\in\mathcal{A}, s'\in\mathcal{S}} \pi(a|s) P(s'|s,a) \int_0^1 (b_{r,\gamma})_\# \eta^\pi(s') \mathbb{G}_{s,a}^R(dr).
$$

For the weak convergence in $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$ (or $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$) we require the reward estimator to satisfy that $\sqrt{n}(\widehat{\mathcal{P}}_{R,n}(\cdot|s,a) - \mathcal{P}_R(\cdot|s,a))$ converges weakly to a tight random element in $\ell^\infty(\mathcal{F}_{\mathrm{KS}})$ (or $\ell^\infty(\mathcal{F}_{\mathrm{TV}})$). Such an assertion is invalid for most nonparametric density estimators such as the histogram estimators and kernel density estimators, because the typical convergence rates of most density estimators are slower than $O(n^{-1/2})$. However, such an assertion holds if we confine $\mathcal{P}_R(\cdot|s,a)$ to some parametric families, for example, the truncated normal family or the Beta family. Note that we observe the phenomenon of inflated variance in the limiting distribution as before, because estimating the reward distribution induces new randomness.

4. Statistical inference.

In this section, we consider the statistical inference of distributional reinforcement learning. First, we present nonparametric confidence sets for $\eta^\pi(s)$ in the forms of 1-Wasserstein, KS, and TV metric balls. Second, we study inference on Hadamard differentiable functionals, with moments, quantiles, and uniform advantage of policy as examples.

4.1. *Inferences for $\eta^\pi(s)$.* Our theoretical findings in Theorems 3.6 allow us to construct confidence sets in the space $\Delta([0,(1-\gamma)^{-1}])$ for the true return distribution $\eta^\pi(s)$, given any initial state $s \in \mathcal{S}$. Specifically, we can construct three types of confidence sets for $\eta^\pi(s)$: 1-Wasserstein, KS and TV metric balls.

THEOREM 4.1. *For some fixed policy π and initial state $s \in \mathcal{S}$, define $\rho_1(\alpha) := \frac{z_1(1-\alpha)}{\sqrt{n}}$, where $z_1(p)$ is defined as the p-quantile of $\sup_{f\in\mathcal{F}_{W_1}}[(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f$. Define the confidence set as*

$$C_1(\alpha) := \{\eta \in \Delta([0,(1-\gamma)^{-1}]) | W_1(\eta, \hat{\eta}_n^\pi(s)) \le \rho_1(\alpha)\}.$$

Then $\lim_{n\to\infty} \mathbb{P}(\eta^\pi(s) \in C_1(\alpha)) = 1-\alpha$.

Furthermore, if Assumption 1 holds, we have

$$\sup_{f\in\mathcal{F}_{W_1}} [(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f = \int_0^{\frac{1}{1-\gamma}} |[(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)\mathbb{1}\{\cdot \le x\}| \, dx.$$

The proof is given in Section 8 of the Supplementary Material [58]. Recall that for two probability distributions μ_1, μ_2 supported on $[0,(1-\gamma)^{-1}]$ we have

$$W_1(\mu_1,\mu_2) = \sup_{f\in\mathcal{F}_{W_1}} |\mu_1 f - \mu_2 f| = \int_0^{\frac{1}{1-\gamma}} |F_1(x) - F_2(x)| \, dx,$$

where F_1 and F_2 are the cumulative distribution functions of μ_1 and μ_2, respectively. Hence, the asymptotic distribution of $W_1(\hat{\eta}_n^\pi(s), \eta^\pi(s))$ can be described in two different ways using asymptotic results in Theorem 3.6 and the continuous mapping theorem. And $\rho_1(\alpha)$ can be determined accordingly.

The confidence set $C_1(\alpha)$ is asymptotically valid, but it relies on the quantile, $z_1(1-\alpha)$, of the unknown limiting distributions that depend on $\mathcal{T}^\pi$ and η^π. We can obtain a consistent estimate of $z_1(1-\alpha)$ using the plug-in approach.

PROPOSITION 4.1. *For any fixed policy π and initial state $s \in \mathcal{S}$, define*

$$\widehat{\mathbb{G}}^\pi(s) := \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \widehat{Z}_{s,a,s'} \int_0^1 (b_{r,\gamma})_\# \hat{\eta}_n^\pi(s') \, d\mathcal{P}_R(dr|s,a) \quad \forall s \in \mathcal{S},$$

where $(\widehat{Z}_{s,a,s'})_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ are mean-zero Gaussians with

$$\mathsf{Cov}(\widehat{Z}_{s_1,a_1,s_1'}, \widehat{Z}_{s_2,a_2,s_2'}) = \mathbb{1}_{\{(s_1,a_1)=(s_2,a_2)\}} \widehat{P}_n(s_1'|s_1,a_1)(\mathbb{1}_{\{s_1'=s_2'\}} - \widehat{P}_n(s_2'|s_1,a_1)),$$

and

$$\hat{z}_1(p) := \inf\Big\{t \Big| \mathbb{P}\Big(\sup_{f\in\mathcal{F}_{W_1}} [(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}\widehat{\mathbb{G}}^\pi](s)f \le t\Big) \ge p\Big\}.$$

Then $\hat{z}_1(p) \xrightarrow{p} z_1(p)$ if $z_1(\cdot)$ is continuous at p.

Furthermore, if Assumption 1 holds, we have

$$\sup_{f\in\mathcal{F}_{W_1}} [(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}\widehat{\mathbb{G}}^\pi](s)f = \int_0^{\frac{1}{1-\gamma}} |[(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}\widehat{\mathbb{G}}^\pi](s)\mathbb{1}_{(-\infty,x]}| \, dx,$$

which can be computed efficiently, and $z_1(\cdot)$ is continuous at any $p \in (0,1)$.

The proof is given in Section 8 of the Supplementary Material [58]. We can also construct confidence sets in the form of KS balls and TV balls for $\eta^\pi(s)$ when Assumption 1 or Assumption 2 holds.

THEOREM 4.2. *For some fixed policy π and $s \in \mathcal{S}$, define*:

$\rho_2(\alpha) := \frac{z_2(1-\alpha)}{\sqrt{n}}$, *where $z_2(p)$ is defined as the p-quantile of* $\sup_{f\in\mathcal{F}_{\mathrm{KS}}}[(\mathcal{I} - \mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f$,

$\rho_3(\alpha) := \frac{z_3(1-\alpha)}{\sqrt{n}}$, *where $z_3(p)$ is defined as the p-quantile of* $\sup_{f\in\mathcal{F}_{\mathrm{TV}}}[(\mathcal{I} - \mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)f$.

Let

$$C_2(\alpha) := \{\eta \in \Delta([0,(1-\gamma)^{-1}]) | \mathrm{KS}(\eta, \hat{\eta}_n^\pi(s)) \le \rho_2(\alpha)\},$$
$$C_3(\alpha) := \{\eta \in \Delta([0,(1-\gamma)^{-1}]) | \mathrm{TV}(\eta, \hat{\eta}_n^\pi(s)) \le \rho_3(\alpha)\}.$$

Then we have:

(a) $\lim_{n\to\infty} \mathbb{P}(\eta^\pi(s) \in C_2(\alpha)) = 1-\alpha$ *under Assumption* 1;
(b) $\lim_{n\to\infty} \mathbb{P}(\eta^\pi(s) \in C_3(\alpha)) = 1-\alpha$ *under Assumption* 2.

Here $\rho_2(\alpha)$ and $\rho_3(\alpha)$ asymptotically describe the quantiles of $\mathrm{KS}(\hat{\eta}_n^\pi(s), \eta^\pi(s))$, $\mathrm{TV}(\hat{\eta}_n^\pi(s), \eta^\pi(s))$, respectively. They are determined using results in Theorem 3.6 and the continuous mapping theorem. Note that $C_2(\alpha)$ and $C_3(\alpha)$ are asymptotically valid confidence sets. Although they rely on the unknown quantile function $z_2(1-\alpha)$ and $z_3(1-\alpha)$, they can be consistently estimated using a plug-in approach as in the case of $z_1(1-\alpha)$.

PROPOSITION 4.2. *For any fixed π and $s \in \mathcal{S}$, define*

$$\hat{z}_2(p) := \inf\Big\{t \Big| \mathbb{P}\Big(\sup_{f\in\mathcal{F}_{\mathrm{KS}}} [(\mathcal{I} - \widehat{\mathcal{T}}_n^\pi)^{-1}\widehat{\mathbb{G}}^\pi](s)f \le t\Big) \ge p\Big\}.$$

$$\hat{z}_3(p) := \inf\Big\{t \Big| \mathbb{P}\Big(\sup_{f\in\mathcal{F}_{\mathrm{TV}}} [(\mathcal{I} - \widehat{\mathcal{T}}_n^\pi)^{-1}\widehat{\mathbb{G}}^\pi](s)f \le t\Big) \ge p\Big\}.$$

Then $\hat{z}_2(p) \xrightarrow{p} z_2(p)$ if Assumption 1 *holds, and $\hat{z}_3(p) \xrightarrow{p} z_3(p)$ if Assumption* 2 *holds.*

One can refer to the proof in Section 8 of the Supplementary Material [58].

4.2. *Inference for Hadamard differentiable functionals.* We consider the problem of statistical inference for $\phi(\eta^\pi(s))$, where $\phi: \ell^\infty(\mathcal{F}_{W_1}) \to \mathbb{R}$ represents a statistical functional. When ϕ is Hadamard differentiable, we can determine the limiting distribution of $\sqrt{n}(\phi(\hat{\eta}_n^\pi(s)) - \phi(\eta^\pi(s)))$ using the functional delta method. Subsequently, we can construct asymptotic confidence sets for $\phi(\eta^\pi(s))$ based on this result.

THEOREM 4.3. *For a fixed policy π and $s \in \mathcal{S}$, let $\phi: \ell^\infty(\mathcal{F}_{W_1}) \to \mathbb{R}$ be Hadamard differentiable at $\eta^\pi(s)$ tangentially to $\mathbb{D}_0 \subset \ell^\infty(\mathcal{F}_{W_1})$. Suppose $[(\mathcal{I} - \mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s) \in \mathbb{D}_0$ and define*

$$C_\phi(\alpha) := \left[\phi(\hat{\eta}_n^\pi(s)) + \frac{z_\phi(\alpha/2)}{\sqrt{n}}, \phi(\hat{\eta}_n^\pi(s)) + \frac{z_\phi(1-\alpha/2)}{\sqrt{n}}\right],$$

where z_ϕ is the quantile function of $\phi'_{\eta^\pi(s)}([(\mathcal{I} - \mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s))$, which is indeed a one-dimensional Gaussian variable with zero means. We have

$$\lim_{n\to\infty} \mathbb{P}(\phi(\eta^\pi(s)) \in C_\phi(\alpha)) = 1-\alpha.$$

Under Assumption 1 *or Assumption* 2, *we have similar results for $\phi: \ell^\infty(\mathcal{F}_{\mathrm{KS}}) \to \mathbb{R}$ or $\ell^\infty(\mathcal{F}_{\mathrm{TV}}) \to \mathbb{R}$ that is Hadamard differentiable at $\eta^\pi(s)$.*

Theorem 4.3 follows directly from our asymptotic results described in Theorem 3.6 and the functional delta method (Theorem 20.8 in [52]). Since the derivative ϕ' is continuous, the plug-in approach is still valid for estimating z_ϕ.

PROPOSITION 4.3. *Whenever the Hadamard derivative ϕ' is properly defined, let*

$$\hat{z}_\phi(p) := \inf\{t \mid \mathbb{P}(\phi'_{\hat{\eta}_n^\pi(s)}([(\mathcal{I}-\widehat{\mathcal{T}}_n^\pi)^{-1}\mathbb{G}^\pi](s)) \le t) \ge p\}.$$

Then $\hat{z}_\phi(p) \xrightarrow{p} z_\phi(p)$ if $z_\phi(\cdot)$ is continuous at p.

The proof of the proposition is nearly identical to those of Proposition 4.1 and Proposition 4.2. We demonstrate the use of Theorem 4.3 with three concrete examples.

EXAMPLE 4.1 (The rth moment of returns). We first consider a simple example of statistical inference for the rth moments of returns. Let $\phi_r(\mu) := \mathbb{E}_{X\sim\mu}(X^r)$, where μ is a signed measure supported on $[0, (1-\gamma)^{-1}]$. It can be easily verified that $\phi\colon \ell^\infty(\mathcal{F}_{W_1}) \to \mathbb{R}$ is Hadamard differentiable with the derivative $\phi'_r(h) = \phi_r(h) = \mathbb{E}_{X\sim h}(X^r)$ for any signed measure h supported on $[0, (1-\gamma)^{-1}]$. Then by Theorem 4.3, we have

$$\sqrt{n}(\phi_r(\hat{\eta}_n^\pi(s)) - \phi_r(\eta^\pi(s))) \rightsquigarrow \phi_r([(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)),$$

according to which we perform statistical inference for $\phi(\eta^\pi(s))$. When $r=1$, we have

$$\sqrt{n}(\widehat{V}^\pi(s) - V^\pi(s)) \rightsquigarrow [(I-\gamma P^\pi)^{-1}\widetilde{G}^\pi](s).$$

Here $P^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ is the transition matrix under policy π, and $\widehat{V}^\pi$ and V^π are the estimated value function and ground-truth value function. $\widetilde{G}^\pi$ is defined as

$$\widetilde{G}^\pi(s) := \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} Z_{s,a,s'} V^\pi(s'),$$

where Z is a Gaussian vector as defined in Theorem 3.6. This recovers the results of limiting distributions of errors of model-based policy evaluations in the generative model setting. Another simple corollary is the limiting distribution of the variance of returns. Concretely, we have

$$\begin{aligned}&\sqrt{n}(\mathsf{Var}_{X\sim\hat{\eta}_n^\pi(s)}(X) - \mathsf{Var}_{X\sim\eta^\pi(s)}(X))\\ &\quad\rightsquigarrow \phi_2([(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)) - 2\phi_1([(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s))\phi_1(\eta^\pi(s)).\end{aligned}$$

EXAMPLE 4.2 (Quantiles of returns). We next consider statistical inference for quantiles of returns when Assumption 1 holds,. Let $\phi_p(\mu) := \inf\{t \mid \mu\mathbb{1}_{(-\infty,t]} \ge p\}$ be the p-quantile of probability distribution μ. Lemma 6.2 in the proof of Proposition 2.4 in the supplement indicates that $\eta^\pi(s)$ must have a bounded density. Hence we have ϕ_p is Hadamard differentiable tangentially to $C[0, (1-\gamma)^{-1}]$ by Lemma 21.4 in [52]. And the derivative $\phi'_p(\eta^\pi(s))$ is the map $h \mapsto -\frac{h(\phi_p(\eta^\pi(s)))}{g(\phi_p(\eta^\pi(s)))}$, where g is the density of $\eta^\pi(s)$. Therefore, the cumulative distribution function of $[(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)$ is in $C[0, (1-\gamma)^{-1}]$ almost surely, we have

$$\sqrt{n}(\phi_p(\hat{\eta}_n^\pi(s)) - \phi_p(\eta^\pi(s))) \rightsquigarrow -\frac{[(\mathcal{I}-\mathcal{T}^\pi)^{-1}\widetilde{\mathbb{G}}^\pi](s)\mathbb{1}_{(-\infty,\phi_p(\eta^\pi(s))]}}{g(\phi_p(\eta^\pi(s)))},$$

which leads to asymptotically valid inferential procedures for quantiles of returns.

EXAMPLE 4.3 (Uniform advantage). Policy improvement [48] is a key ingredient in many reinforcement learning algorithms. The goal is to find a new policy π such that the advantage function $V^{\pi}(s_0) - V^{\pi_0}(s_0) \geq 0$ where π_0 is a given baseline policy. We now propose a new notion of policy improvement called (near)-uniform policy improvement. Specifically, the aim is to find a new policy π such that the uniform advantage $\phi(\eta^{\pi}(s_0), \eta^{\pi_0}(s_0)) := \mathbb{P}(G^{\pi}(s_0) \geq G^{\pi_0}(s_0))$ is above some threshold.

A natural estimator of $\phi(\eta^{\pi}(s_0), \eta^{\pi_0}(s_0))$ is $\phi(\hat{\eta}_n^{\pi}(s_0), \hat{\eta}_n^{\pi_0}(s_0))$. For technical convenience, assume that $\hat{\eta}_n^{\pi}(s_0)$ and $\hat{\eta}_n^{\pi_0}(s_0)$ are estimated by data splitting techniques. From Lemma 20.10 in [52], ϕ is Hadamard differentiable tangentially to $C[0, (1-\gamma)^{-1}]$. For $h_1, h_2 \in C[0, (1-\gamma)^{-1}]$, the derivative $\phi'(\eta^{\pi}(s_0), \eta^{\pi_0}(s_0))$ is $(h_1, h_2) \mapsto -\eta^{\pi}(s_0)h_2 + \eta^{\pi_0}(s_0)h_1$. When Assumption 1 is true,

$$\sqrt{n}\big[\big(\hat{\eta}_n^{\pi}(s_0), \hat{\eta}_n^{\pi_0}(s_0)\big) - \big(\eta^{\pi}(s_0), \eta^{\pi_0}(s_0)\big)\big] \rightsquigarrow \big(\big[(\mathcal{I}-\mathcal{T}^{\pi})^{-1}\widetilde{\mathbb{G}}^{\pi}\big](s_0), \big[(\mathcal{I}-\mathcal{T}^{\pi})^{-1}\widetilde{\mathbb{G}}^{\pi_0}\big](s)\big)$$

and the cumulative distribution functions of $[(\mathcal{I}-\mathcal{T}^{\pi})^{-1}\widetilde{\mathbb{G}}^{\pi}](s_0)$ and $[(\mathcal{I}-\mathcal{T}^{\pi_0})^{-1}\widetilde{\mathbb{G}}^{\pi_0}](s_0)$ (denoted F and F_0) are in $C[0, (1-\gamma)^{-1}]$ almost surely. Thus, we have

$$\sqrt{n}\big[\phi\big(\hat{\eta}_n^{\pi}(s_0), \hat{\eta}_n^{\pi_0}(s_0)\big) - \phi\big(\eta^{\pi}(s_0), \eta^{\pi_0}(s_0)\big)\big] \rightsquigarrow \eta^{\pi}(s_0)F_0 - \eta^{\pi_0}(s_0)F.$$

5. Numerical simulations.

In this section, we conduct numerical simulations to validate our theoretical findings as well as the proposed inferential procedures. All of the numerical simulations are conducted on a desktop computer with a single TITAN RTX GPU. The code is available in the Supplementary Material [58].

5.1. *Implementations.*

To make computations tractable, we confine the return distributions to the class of categorical distributions. The categorical distributions are expressed as a vector $\eta = (\eta(s_1), \ldots, \eta(s_{|\mathcal{S}|}))$, where $\eta(s) := \sum_{k=0}^{K} w_k \delta_{x_k}$, with weights $\sum_{k=0}^{K} w_k = 1$ and particles $x_k := k/[(K+1)(1-\gamma)]$. We set $K = 1000$, which is large enough to make the categorical class rich enough and good approximations of continuous return distributions.

The categorical distributions can be updated with a categorical version of distributional dynamical programming [3], which is also a good approximation of the original version of distributional dynamic programming considered in our paper when K is large. Throughout our simulation studies, the ground-truth return distributions η^{π} are obtained via a sufficiently large number of iterations of distributional dynamic programming with the ground-truth distributional Bellman operator $\mathcal{T}^{\pi}$. The estimated return distributions $\hat{\eta}_n^{\pi}$ are obtained by the same procedure while the ground-truth distributional Bellman operator $\mathcal{T}^{\pi}$ is replaced by the estimated distributional Bellman operator $\widehat{\mathcal{T}}_n^{\pi}$.

Recall that in our inferential procedures, we must explicitly form the operator $(\mathcal{I}-\mathcal{T}^{\pi})^{-1}$, which might cause computational intractability. We instead use a Neumann expansion to approximate $(\mathcal{I}-\mathcal{T}^{\pi})^{-1}$. That is, $(\mathcal{I}-\mathcal{T}^{\pi})^{-1} \approx \sum_{j=0}^{J}(\mathcal{T}^{\pi})^{j}$ where J takes a sufficiently large value. In summary, by these approximation techniques, our implementations achieve computational tractability while ensuring that the approximation error is negligible compared with the statistical error that is of primary interest.

5.2. *Linear convergence of distributional dynamic programming.*

We first verify the results in Propositions 2.3, 2.4 and 2.5. We perform the simulations in randomly generated tabular MDPs with $|\mathcal{S}| = 5$, $|\mathcal{A}| = 2$ and $\gamma = 0.9$. Without loss of generality, we always use the first state s_1 as the initial state. The reward distribution is chosen as truncated Gaussians. Specifically, $\mathcal{P}_R(\cdot|s,a)$ is set as $\mathsf{N}(l_{s,a}, 0.1)$ truncated to $[0,1]$, with the location parameter

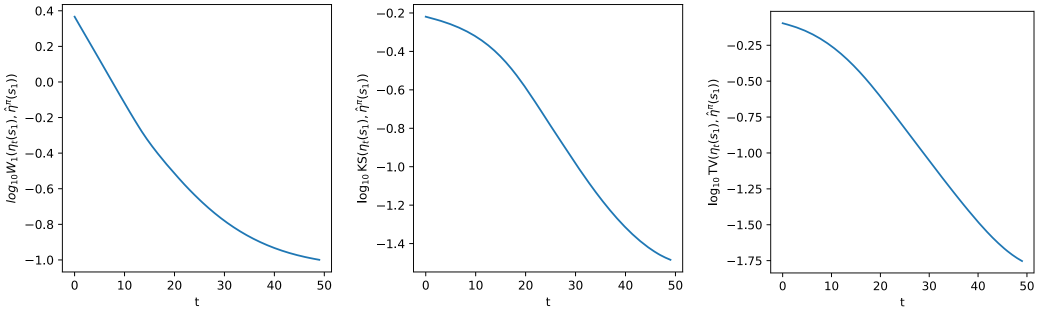

FIG. 2. *Convergence of* $\log W_1(\eta^{(t)}(s_1), \hat{\eta}_n^\pi(s_1))$, $\log \mathrm{KS}(\eta^{(t)}(s_1), \hat{\eta}_n^\pi(s_1))$, *and* $\log \mathrm{TV}(\eta^{(t)}(s_1), \hat{\eta}_n^\pi(s_1))$ *with sample size* $n = 10{,}000$ *and* $\gamma = 0.9$. t *is the iteration number.*

$l_{s,a}$ randomly determined. The dataset we use to form the estimator $\hat{\eta}_n^\pi$ is obtained from a generative model with $n = 10{,}000$. We first perform distributional dynamical programming for N iterations, with N sufficiently large, and use $\eta^{(N)}$ as a proxy of the estimator $\hat{\eta}_n^\pi$. Figure 2 depicts how $d(\eta^{(t)}, \hat{\eta}_n^\pi)$ changes as t increases from 0 to $N/2$. The probability metric d is set as the 1-Wasserstein metric, the KS metric, and the TV metric. We see that distributional dynamical programming does exhibit linear convergence when measured by each of these three metrics.

5.3. *Finite-sample convergence performance.* We investigate the finite-sample convergence performances of empirical distributional dynamic programming and verify our nonasymptotic results. The simulation environments remain the same as in the preceding subsection. And we consider more choices of γ. Specifically, we set $\gamma \in \{0.7, 0.8, 0.9, 0.97\}$. We also consider more choices of n. Specifically, we set $n \in \{10, 100, 1000, 10{,}000\}$. We repeat the estimation process for 100 times and report the averaged errors.

Figures 3–5 show the convergence performance of empirical distributional dynamic programming measured by the 1-Wasserstein metric, KS metric and TV metric, respectively. We see that in all cases, the convergence consists of two phases. In the first phase, the dynamic programming algorithm does not converge, and we observe a linear convergence rate. In the second phase, the error terms are dominated by the statistical error: $W_1(\hat{\eta}_n^\pi(s_1), \eta^\pi(s_1))$, $\mathrm{KS}(\hat{\eta}_n^\pi(s_1), \eta^\pi(s_1))$ or $\mathrm{TV}(\hat{\eta}_n^\pi(s_1), \eta^\pi(s_1))$, which exhibits strong correlations with n and $(1-\gamma)^{-1}$.

We also examine how the error terms $W_1(\hat{\eta}_n^\pi(s_1), \eta^\pi(s_1))$, $\mathrm{KS}(\hat{\eta}_n^\pi(s_1), \eta^\pi(s_1))$ or $\mathrm{TV}(\hat{\eta}_n^\pi(s_1), \eta^\pi(s_1))$ change as n increases. From Figures 6–8, we can verify that for all cases, the convergence rates are indeed of the typical $n^{-1/2}$ order as described in Theorems 3.1, 3.2 and 3.3.

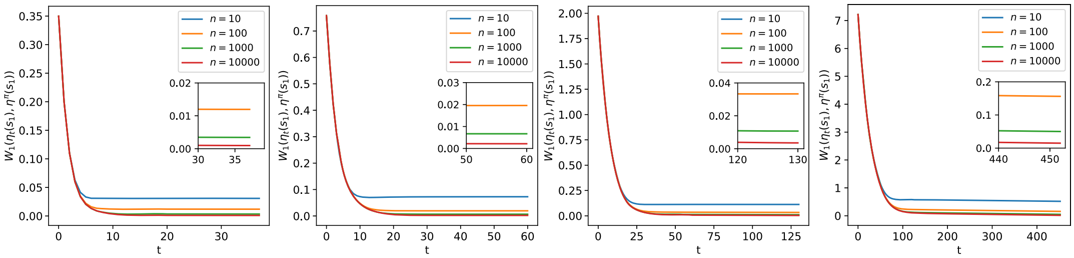

FIG. 3. *Two-phase convergence of* $W_1(\eta^{(t)}(s_1), \eta^\pi(s_1))$ *with different sample sizes.* t *is the iteration number. From left to right*: $\gamma = 0.7$; $\gamma = 0.8$; $\gamma = 0.9$; $\gamma = 0.97$.

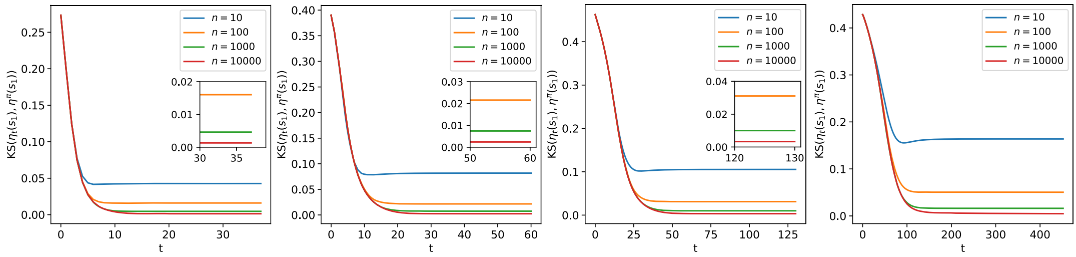

FIG. 4. *Two-phase convergence of* $\mathrm{KS}(\eta^{(t)}(s_1), \eta^\pi(s_1))$ *with different sample sizes. t is the iteration number. From left to right*: $\gamma = 0.7$; $\gamma = 0.8$; $\gamma = 0.9$; $\gamma = 0.97$.

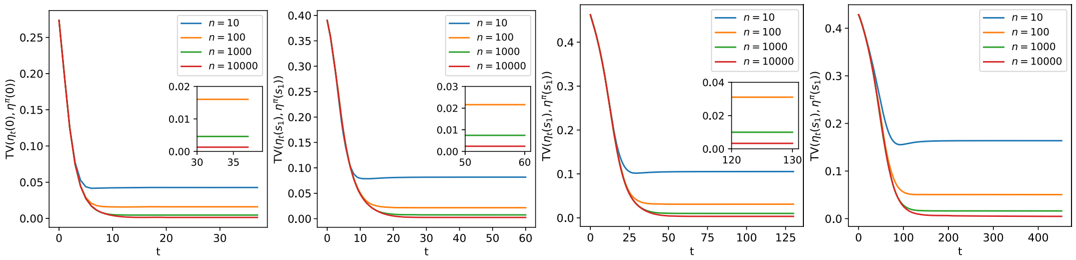

FIG. 5. *Two-phase convergence of* $\mathrm{TV}(\eta^{(t)}(s_1), \eta^\pi(s_1))$ *with different sample sizes. t is the iteration number. From left to right*: $\gamma = 0.7$; $\gamma = 0.8$; $\gamma = 0.9$; $\gamma = 0.97$.

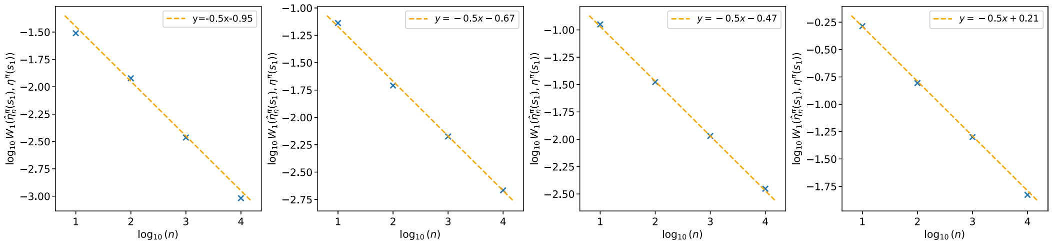

FIG. 6. *The statistical error* $W_1(\hat{\eta}^\pi_n(s_1), \eta^\pi(s_1))$ *with different sample sizes. From left to right*: $\gamma = 0.7$; $\gamma = 0.8$; $\gamma = 0.9$; $\gamma = 0.97$.

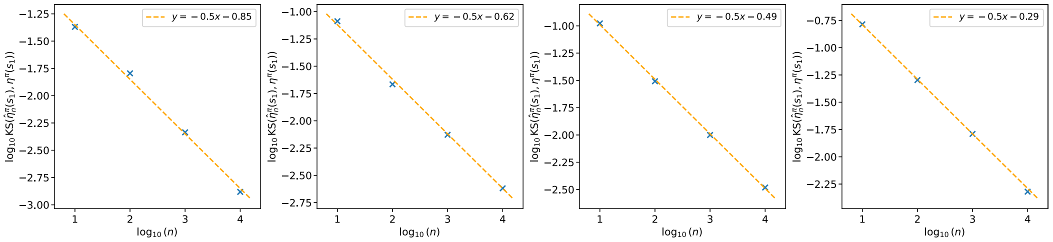

FIG. 7. *The statistical error* $\mathrm{KS}(\hat{\eta}^\pi_n(s_1), \eta^\pi(s_1))$ *with different sample sizes. From left to right*: $\gamma = 0.7$; $\gamma = 0.8$; $\gamma = 0.9$; $\gamma = 0.97$.

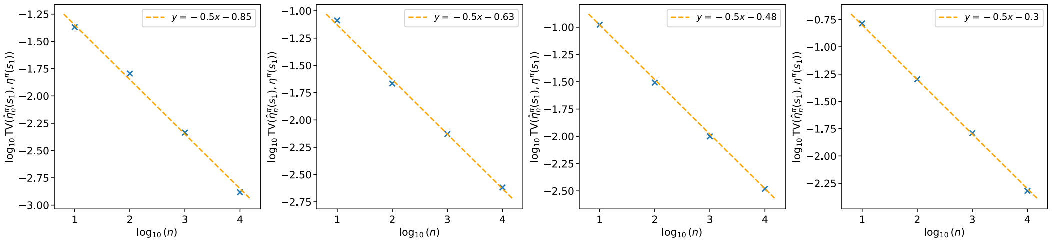

FIG. 8. *The statistical error* $\mathrm{TV}(\hat{\eta}^\pi_n(s_1), \eta^\pi(s_1))$ *with different sample sizes. From left to right*: $\gamma = 0.7$; $\gamma = 0.8$; $\gamma = 0.9$; $\gamma = 0.97$.

TABLE 1
Coverage rate (CR) and confidence set radius (CSR) of our proposed nonparametric confidence sets for $\eta^{\pi}(s_1)$ under different choices of n

Type of confidence sets	W_1 ball		KS ball		TV ball	
n	CR	CSR ± std	CR	CSR ± std	CR	CSR ± std
5	0.918	0.3467 ± 0.0719	0.939	0.3302 ± 0.0538	0.934	0.3310 ± 0.0530
10	0.945	0.2528 ± 0.0366	0.934	0.2364 ± 0.0250	0.956	0.2367 ± 0.0246
100	0.941	0.0804 ± 0.0041	0.956	0.0759 ± 0.0032	0.944	0.0761 ± 0.0030
1000	0.945	0.0255 ± 0.0008	0.951	0.0241 ± 0.0007	0.950	0.0241 ± 0.0007

5.4. *Validity of inferential procedures.* We also perform numerical simulations to validate the inferential procedures proposed in Section 4. The environment is exactly the same as that of the previous subsection with γ fixed as 0.9. The confidence sets are constructed using plug-in approaches. The nominal coverage probability is set as 0.95, and the quantiles of the estimated limiting distributions are computed using Monte Carlo methods. Here our Monte Carlo implementations are fully vectorized to further improve computational efficiency. We repeat our inferential procedures for 1000 times, and report the empirical coverage rates and the averaged radius of confidence sets. The results are presented in Tables 1 and 2. We observe that the empirical coverage rates approach the nominal confidence level and the radius of confidence sets decreases as n increases in all cases.

6. Discussions. In this paper, we have analyzed the statistical performance of distributional reinforcement learning from both nonasymptotic and asymptotic perspectives. We have presented nonasymptotic rates for $\sup_{s\in\mathcal{S}} W_p(\hat{\eta}_n^{\pi}(s), \eta^{\pi}(s))$, $\sup_{s\in\mathcal{S}} \mathrm{KS}(\hat{\eta}_n^{\pi}(s), \eta^{\pi}(s))$ and $\sup_{s\in\mathcal{S}} \mathrm{TV}(\hat{\eta}_n^{\pi}(s), \eta^{\pi}(s))$. We have also derived that given an initial state s, the "empirical process" $\sqrt{n}(\hat{\eta}_n^{\pi}(s) - \eta^{\pi}(s))$ converges weakly to a Gaussian random element. Based on our theoretical findings, we have devised inferential procedures for a wide class of statistical functionals of the return distribution. We hope our work can spur further research in the uncertainty quantification of reinforcement learning.

One future direction is whether we can close the gap between our sample complexity bound $\widetilde{O}(\varepsilon^{-2}(1-\gamma)^{-4})$ and the lower bound $\widetilde{O}(\varepsilon^{-2}(1-\gamma)^{-3})$. We speculate that the minimax optimal sample complexity is indeed $\widetilde{O}(\varepsilon^{-2}(1-\gamma)^{-3})$, and could be attained through more sophisticated analysis techniques. Another interesting future direction is to develop nonasymptotic bounds as well as asymptotic results that are uniform for $\pi \in \Pi$, where Π denotes a policy class of interest. This might give rise to a wider range of inferential applications in reinforcement learning.

TABLE 2
Coverage rate (CR) and confidence set radius (CSR) of our proposed confidence intervals for different Hardamard differentiable statistical functionals of $\eta^{\pi}(s_1)$ under different choices of n

Functionals of interest	Variance		0.1 quantile		0.9 quantile	
n	CR	CSR ± std	CR	CSR ± std	CR	CSR ± std
5	0.928	0.0627 ± 0.0202	0.917	0.4020 ± 0.0929	0.916	0.3210 ± 0.0660
10	0.939	0.0442 ± 0.0101	0.945	0.2949 ± 0.0436	0.930	0.2288 ± 0.0308
100	0.959	0.0137 ± 0.0010	0.949	0.0912 ± 0.0050	0.946	0.0733 ± 0.0036
1000	0.946	0.0043 ± 0.0002	0.935	0.0289 ± 0.0010	0.952	0.0232 ± 0.0008

Acknowledgments. The authors would like to thank the Editor, the Associate Editor and the referees for the constructive feedback, which helped us improve the quality of our work.

Yang Peng is the corresponding author.

Funding. This work has been supported by the National Key Research and Development Project of China (No. 2022YFA1004002), the National Natural Science Foundation of China (No. 12271011 and No. 12350001), and the MOE Project of Key Research Institute of Humanities and Social Sciences (No.22JJD110001).

SUPPLEMENTARY MATERIAL

Supplement: Technical proofs (DOI: 10.1214/25-AOS2527SUPPA; .pdf). This supplementary material contains proof outlines of the theoretical results and omitted proofs of technical lemmas.

Supplement: Code (DOI: 10.1214/25-AOS2527SUPPB; .zip). This supplementary material contains all code used in the numerical simulations.

REFERENCES

[1] ALQUIER, P., FRIEL, N., EVERITT, R. and BOLAND, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Stat. Comput.* **26** 29–47. MR3439357 https://doi.org/10.1007/s11222-014-9521-x

[2] BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *International Conference on Machine Learning* 405–413. PMLR.

[3] BELLEMARE, M. G., DABNEY, W. and MUNOS, R. (2017). A distributional perspective on reinforcement learning. In *International Conference on Machine Learning* 449–458. PMLR.

[4] BELLEMARE, M. G., DABNEY, W. and ROWLAND, M. (2023). *Distributional Reinforcement Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. http://www.distributional-rl.org. MR4649766

[5] CHAE, M. and WALKER, S. G. (2020). Wasserstein upper bounds of the total variation for smooth densities. *Statist. Probab. Lett.* **163** 108771, 6. MR4083852 https://doi.org/10.1016/j.spl.2020.108771

[6] CHANDAK, Y., NIEKUM, S., DA SILVA, B., LEARNED-MILLER, E., BRUNSKILL, E. and THOMAS, P. S. (2021). Universal off-policy evaluation. *Adv. Neural Inf. Process. Syst.* **34** 27475–27490.

[7] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning* 1042–1051. PMLR.

[8] CLEMENTS, W. R., VAN DELFT, B., ROBAGLIA, B.-M., SLAOUI, R. B. and TOTH, S. (2019). Estimating risk and uncertainty in deep reinforcement learning. arXiv preprint. Available at arXiv:1905.09638.

[9] DABNEY, W., OSTROVSKI, G., SILVER, D. and MUNOS, R. (2018). Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning* 1096–1105. PMLR.

[10] DABNEY, W., ROWLAND, M., BELLEMARE, M. and MUNOS, R. (2018). Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[11] DOAN, T., MAZOURE, B. and LYLE, C. (2018). Gan q-learning. arXiv preprint. Available at arXiv:1805.04874.

[12] FAWZI, A., BALOG, M., HUANG, A., HUBERT, T., ROMERA-PAREDES, B., BAREKATAIN, M., NOVIKOV, A., RUIZ, F. J., SCHRITTWIESER, J. et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610** 47–53.

[13] FERRÉ, D., HERVÉ, L. and LEDOUX, J. (2013). Regular perturbation of V-geometrically ergodic Markov chains. *J. Appl. Probab.* **50** 184–194. MR3076780 https://doi.org/10.1239/jap/1363784432

[14] FREIRICH, D., SHIMKIN, T., MEIR, R. and TAMAR, A. (2019). Distributional multivariate policy evaluation and exploration with the Bellman gan. In *International Conference on Machine Learning* 1983–1992. PMLR.

[15] GHYSELS, E., SANTA-CLARA, P. and VALKANOV, R. (2005). There is a risk-return trade-off after all. *J. Financ. Econ.* **76** 509–548.

[16] HAO, B., JI, X., DUAN, Y., LU, H., SZEPESVARI, C. and WANG, M. (2021). Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning* 4074–4084. PMLR.

[17] HUA, Y., LI, R., ZHAO, Z., CHEN, X. and ZHANG, H. (2019). GAN-powered deep distributional reinforcement learning for resource management in network slicing. *IEEE J. Sel. Areas Commun.* **38** 334–349.

[18] HUANG, A., LEQI, L., LIPTON, Z. and AZIZZADENESHELI, K. (2022). Off-policy risk assessment for Markov decision processes. In *International Conference on Artificial Intelligence and Statistics* 5022–5050. PMLR.

[19] JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* 652–661. PMLR.

[20] JOHNDROW, J. E. and MATTINGLY, J. C. (2017). Error bounds for approximations of Markov chains used in Bayesian sampling. arXiv preprint. Available at arXiv:1711.05382.

[21] KARTASHOV, N. V. (1986). Inequalities in theorems of ergodicity and stability for Markov chains with common phase space. I. *Theory Probab. Appl.* **30** 247–259.

[22] KOBER, J., BAGNELL, J. A. and PETERS, J. (2013). Reinforcement learning in robotics: A survey. *Int. J. Robot. Res.* **32** 1238–1274.

[23] LAVORI, P. W. and DAWSON, R. (2004). Dynamic treatment regimes: Practical design considerations. *Clin. Trials* **1** 9–20.

[24] LI, G., WEI, Y., CHI, Y. and CHEN, Y. (2024). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Oper. Res.* **72** 203–221. MR4705834

[25] LI, X., LIANG, J. and ZHANG, Z. (2023). Online statistical inference for nonlinear stochastic approximation with Markovian data. arXiv preprint. Available at arXiv:2302.07690.

[26] LI, X., YANG, W., LIANG, J., ZHANG, Z. and JORDAN, M. I. (2023). A statistical analysis of Polyak-ruppert averaged q-learning. In *International Conference on Artificial Intelligence and Statistics* 2207–2261. PMLR.

[27] LIANG, H. and LUO, Z. (2024). Regret bounds for risk-sensitive reinforcement learning with Lipschitz dynamic risk measures. In *Proceedings of the* 27*th International Conference on Artificial Intelligence and Statistics* (S. Dasgupta, S. Mandt and Y. Li, eds.). *Proceedings of Machine Learning Research* **238** 1774–1782. PMLR.

[28] LIM, S. H. and MALIK, I. (2022). Distributional reinforcement learning for risk-sensitive policies. In *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh, eds.) **35** 30977–30989. Curran Associates, Red Hook.

[29] LOCKWOOD, O. and SI, M. (2022). A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* **18** 155–162.

[30] MA, X., XIA, L., ZHOU, Z., YANG, J. and ZHAO, Q. (2020). Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. arXiv preprint. Available at arXiv:2004.14547.

[31] MITROPHANOV, A. Y. (2005). Sensitivity and convergence of uniformly ergodic Markov chains. *J. Appl. Probab.* **42** 1003–1014. MR2203818 https://doi.org/10.1239/jap/1134587812

[32] MORIMURA, T., SUGIYAMA, M., KASHIMA, H., HACHIYA, H. and TANAKA, T. (2010). Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML*-10) 799–806.

[33] NAEEM, F., SEIFOLLAHI, S., ZHOU, Z. and TARIQ, M. (2020). A generative adversarial network enabled deep distributional reinforcement learning for transmission scheduling in Internet of vehicles. *IEEE Trans. Intell. Transp. Syst.* **22** 4550–4559.

[34] OPENAI (2023). GPT-4 Technical Report.

[35] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K. et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35** 27730–27744.

[36] ROSS, N. (2011). Fundamentals of Stein's method. *Probab. Surv.* **8** 210–293. MR2861132 https://doi.org/10.1214/11-PS182

[37] ROWLAND, M., BELLEMARE, M., DABNEY, W., MUNOS, R. and TEH, Y. W. (2018). An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* 29–37. PMLR.

[38] ROWLAND, M., MUNOS, R., AZAR, M. G., TANG, Y., OSTROVSKI, G., HARUTYUNYAN, A., TUYLS, K., BELLEMARE, M. G. and DABNEY, W. (2024). An analysis of quantile temporal-difference learning. *J. Mach. Learn. Res.* **25** Paper No. [163], 47. MR4758134

[39] ROWLAND, M., TANG, Y., LYLE, C., MUNOS, R., BELLEMARE, M. G. and DABNEY, W. (2023). The statistical benefits of quantile temporal-difference learning for value estimation. In *International Conference on Machine Learning* 29210–29231. PMLR.

[40] RUDOLF, D. and SCHWEIZER, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **24** 2610–2639. MR3779696 https://doi.org/10.3150/17-BEJ938

[41] RUDOLF, D., SMITH, A. and QUIROZ, M. (2024). Perturbations of Markov chains. arXiv preprint. Available at arXiv:2404.10251.

[42] SCHWEITZER, P. J. (1968). Perturbation theory and finite Markov chains. *J. Appl. Probab.* **5** 401–413. MR0234527 https://doi.org/10.2307/3212261

[43] SHI, C., ZHANG, S., LU, W. and SONG, R. (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 765–793. MR4460575

[44] SILVER, D., HUBERT, T., SCHRITTWIESER, J., ANTONOGLOU, I., LAI, M., GUEZ, A., LANCTOT, M., SIFRE, L., KUMARAN, D. et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362** 1140–1144. MR3888768 https://doi.org/10.1126/science.aar6404

[45] SIMON, H. A. (1956). Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica* **24** 74–81. MR0077847 https://doi.org/10.2307/1905261

[46] SINGH, R., ZHANG, Q. and CHEN, Y. (2020). Improving robustness via risk averse distributional reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control* (A. M. Bayen, A. Jadbabaie, G. Pappas, P. A. Parrilo, B. Recht, C. Tomlin and M. Zeilinger, eds.). *Proceedings of Machine Learning Research* **120** 958–968. PMLR.

[47] SUN, K., ZHAO, Y., LIU, Y., SHI, E., WANG, Y., YAN, X., JIANG, B. and KONG, L. (2022). *Interpreting Distributional Reinforcement Learning: A Regularization Perspective*.

[48] SUTTON, R. S. (2004). The reward hypothesis.

[49] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[50] THEIL, H. (1957). A note on certainty equivalence in dynamic planning. *Econometrica* **25** 346–349. MR0091227 https://doi.org/10.2307/1910260

[51] THOMAS, P., THEOCHAROUS, G. and GHAVAMZADEH, M. (2015). High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[52] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[53] VINYALS, O., BABUSCHKIN, I., CZARNECKI, W. M., MATHIEU, M., DUDZIK, A., CHUNG, J., CHOI, D. H., POWELL, R., EWALDS, T. et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575** 350–354.

[54] WANG, K., ZHOU, K., WU, R., KALLUS, N. and SUN, W. (2023). The benefits of being distributional: Small-loss bounds for reinforcement learning. *Adv. Neural Inf. Process. Syst.* **36** 2275–2312.

[55] WILTZER, H., FAREBROTHER, J., GRETTON, A., TANG, Y., BARRETO, A., DABNEY, W., BELLEMARE, M. G. and ROWLAND, M. (2024). A distributional analogue to the successor representation. In *Proceedings of the 41st International Conference on Machine Learning* 52994–53016.

[56] WU, R., UEHARA, M. and SUN, W. (2023). Distributional offline policy evaluation with predictive error guarantees. In *International Conference on Machine Learning* 37685–37712. PMLR.

[57] YANG, W., ZHANG, L. and ZHANG, Z. (2022). Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *Ann. Statist.* **50** 3223–3248. MR4524495 https://doi.org/10.1214/22-aos2225

[58] ZHANG, L., PENG, Y., LIANG, J., YANG, W. and ZHANG, Z. (2025). Supplement to "Estimation and Inference in Distributional Reinforcement Learning." https://doi.org/10.1214/25-AOS2527SUPPA, https://doi.org/10.1214/25-AOS2527SUPPB

[59] ZHU, Y., DONG, J. and LAM, H. (2024). Uncertainty quantification and exploration for reinforcement learning. *Oper. Res.* **72** 1689–1709. MR4812023