Joint MDPs and Reinforcement Learning in Coupled-Dynamics Environments

Ege C. Kaya[1] **Mahsa Ghasemi**[1] **Abolfazl Hashemi**[1]

[1] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

arXiv:2603.06946v2 [cs.LG] 16 Aug 2026

Abstract

Many distributional quantities in reinforcement learning are intrinsically joint across actions, including distributions of gaps and probabilities of superiority. However, the classical Markov decision process (MDP) formalism specifies only marginal laws and leaves the joint law of counterfactual one-step outcomes across multiple possible actions at a state unspecified. We study *coupled-dynamics environments* with a multi-action generative interface which can sample counterfactual one-step outcomes for multiple actions under shared exogenous randomness. We propose *joint MDPs* (JMDPs) as a formalism for such environments by augmenting an MDP with a multi-action sample transition model which specifies a coupling of one-step counterfactual outcomes, while preserving standard MDP interaction as marginal observations. We adopt a local one-step coupling regime in which each queried state admits a shared-randomness outcome table and future tables are freshly sampled recursively. In this regime, we derive Bellman operators for nth-order return moments, providing dynamic programming and incremental algorithms with convergence guarantees.

1 INTRODUCTION

Distributional reinforcement learning (DRL) studies the discounted return random variable (RV) $Z^\pi(s,a)$ belonging to a policy π, and seeks to represent or estimate its distributional properties beyond its expectation. Formally,

$$Z^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t R_t,\ S_0 = s, A_0 = a,\ A_t \sim \pi(\cdot \mid S_t)$$
$$R_t \sim P_R(\cdot \mid S_t, A_t),\ S_{t+1} \sim P_S(\cdot \mid S_t, A_t). \tag{1}$$

DRL methods customarily aim to learn per-action marginal return laws at state s, i.e., $\{\mathrm{Law}(Z^\pi(s,a))\}_{a\in\mathcal{A}}$ [Bellemare et al., 2017, 2023]. However, a range of distributional quantities that are of interest in decision-making [Artzner et al., 1999, Rockafellar et al., 2000] are not functions of per-action marginals alone. For actions $a, \tilde{a} \in \mathcal{A}$, examples include:

1. the *gap* RV $G^\pi(s;a,\tilde{a}) := Z^\pi(s,a) - Z^\pi(s,\tilde{a})$ and its distribution $\mathrm{Law}(G^\pi(s;a,\tilde{a}))$,
2. any tail functional of $G^\pi(s;a,\tilde{a})$, such as the quantile $q_\alpha(G^\pi(s;a,\tilde{a}))$ or the $\mathrm{CVaR}_\alpha(G^\pi(s;a,\tilde{a}))$,
3. the *probability of superiority* $\mathbb{P}(Z^\pi(s,a) > Z^\pi(s,\tilde{a}))$.

These quantities depend on the joint law of $(Z^\pi(s,a))$, $Z^\pi(s,\tilde{a}))$ and thus require knowledge of coupling structure.

A limitation of the Markov decision process (MDP) formalism. MDPs specify the marginal distributions of the reward and successor state under each action. However, they make no specification about the joint distribution of *counterfactual, one-step outcomes across multiple actions at a state*. Therefore, joint objects such as $\mathrm{Law}(Z^\pi(s,a) - Z^\pi(s,\tilde{a}))$ are not well-defined without adopting additional coupling conventions [Wiltzer et al., 2024a].

Coupled-dynamics environments. The preceding limitation is substantive in environments where counterfactual one-step outcomes under multiple actions are operationally meaningful and sampleable. A canonical setting is scenario-based simulation, where an exogenous disturbance U is realized at each decision point and one-step outcomes are functions of (s,a,U). Evaluating multiple candidate actions under the same realized disturbance is standard in simulation optimization [Glasserman, 2004, Shapiro et al., 2021] and yields a multi-action generative interface. Examples include logistics, dispatch, inventory, energy, robotics, and network simulators where actions are compared under the same demand, traffic, load, cloned-state or latency shock.

Contribution and scope.

- We propose *joint MDPs* (JMDPs) as a formalism for

coupled-dynamics environments. A JMDP augments an MDP with a multi-action sample transition model that specifies a coupling of one-step counterfactual outcomes.

- We focus exclusively on *policy evaluation*: For a fixed policy π, we derive moment Bellman operators up to arbitrary finite order and provide both dynamic programming (DP) and incremental estimation algorithms with convergence guarantees, along with computable Bellman residual certificates [Scherrer, 2010].
- **One-step coupling regime.** We adopt a local one-step coupling regime. Branches that query the same state share the current outcome table, while distinct successor states use independent fresh tables. The Bellman recursion then propagates mixed moments over successor pairs without constructing fully coupled counterfactual trajectory trees.

2 RELATED WORKS

MDP formalism and generative interfaces. In the standard MDP formalism, the environment is specified by a per-action reward law and a per-action transition kernel [Puterman, 1994, Sutton and Barto, 2018, Bellemare et al., 2023]. Interaction reveals only the reward and next state of the executed action. This is sufficient for objectives depending on *marginal* state-action return laws (e.g., expected value or a single-action return distribution), but it does not specify a joint law over counterfactual one-step outcomes across actions. We formalize a setting in which the environment exposes a *multi-action generative interface* returning counterfactual one-step outcomes for multiple queried actions under shared exogenous randomness. Such shared-randomness interfaces are standard in common-random-number Monte Carlo (MC) comparison and scenario-based decision models [Glasserman, 2004, Shapiro et al., 2021].

Distributional and multivariate RL. DRL studies the law of the scalar return and Bellman operators on spaces of probability measures, beginning with the distributional perspective of Bellemare et al. [2017] and subsequent practical approximations [Bellemare et al., 2017, Dabney et al., 2018b,a, Yang et al., 2019, Nguyen-Tang et al., 2021]. These methods learn per-state action *marginal* return distributions and do not define joint RVs $(Z^\pi(s,a))_{a\in\mathcal{A}}$. A complementary line studies multivariate return distributions induced by vector-valued rewards and cumulants, including Bellman-GAN [Freirich et al., 2019], MD3QN [Zhang et al., 2021], and the recent foundations and convergent DP/TD methods of Wiltzer et al. [2024b]. Our setting is orthogonal: even with scalar rewards, we study joint structure across counterfactual actions induced by coupled-environment randomness, which is not captured by standard multivariate DRL formulations.

Moment-based evaluation, risk sensitivity, and gap/advantage distributions. Return-moment policy evaluation has a long history [Sobel, 1982], including DP characterizations of variance and TD estimators for variance and related quantities [Mannor and Tsitsiklis, 2011, Tamar et al., 2013, 2016]. Coherent risk measures such as CVaR have likewise been incorporated via risk-aware Bellman operators [Chow and Ghavamzadeh, 2014]. These works remain within a single action's return law. By contrast, our target quantities are intrinsically joint across actions and require mixed moments encoding cross-action dependence. This distinction is especially visible for gap and advantage distributions: if only marginals are given, the law of $Z^\pi(s,a) - Z^\pi(s,\tilde{a})$ depends on an unspecified coupling, motivating canonical-coupling axioms [Wiltzer et al., 2024a]. Our approach instead models environments in which the coupling is part of the environment itself. Under our one-step coupling regime, gap and advantage RVs are well-defined by construction, and their moments are computable from learned mixed moments.

Counterfactual modeling perspectives. The multi-action generative interface admits a counterfactual interpretation. At a state, the environment reveals multiple potential one-step outcomes under shared exogenous randomness. This connects to potential-outcome and structural causal model (SCM) perspectives in RL [Lu et al., 2020, Amitai et al., 2024]. We do not address identifiability or learning such an SCM from data, rather, assuming simulator access to counterfactual one-step queries, we ask what formalism and policy-evaluation theory are appropriate. A POMDP or SCM is natural when latent variables must be inferred from ordinary trajectories. Here, the shared-randomness counterfactual table is queryable, and even if a latent U renders outcomes conditionally deterministic, the unconditional gap under a shared U remains a joint object rather than a marginal one.

Overall, prior DRL and risk-sensitive RL methods either model marginal per-action return laws or multivariate returns arising from vector-valued rewards along a realized trajectory. Our work isolates a different axis of jointness: across actions at a fixed state, induced by coupled-environment randomness. We propose joint MDPs as a formalism and algorithmic basis for policy evaluation of both marginal and joint return moments, enabling principled treatment of intrinsically joint distributional quantities.

3 PRELIMINARIES

Definition 3.1 (MDP)**.** An MDP is a quintuple $(\mathcal{S}, \mathcal{A}, P_R, P_S, \gamma)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space with $N = |\mathcal{A}|$, $\gamma \in (0,1)$ is the discount factor, $P_R(\cdot \mid s,a)$ is a reward kernel over $[0,1]$ and $P_S(\cdot \mid s,a)$ is a transition kernel over $\mathcal{S}$.

A policy is a Markov kernel $\pi(\cdot \mid s) \in \Delta(\mathcal{A})$. For measurable spaces $(\Omega_i, \mathcal{F}_i)$, let $\mathcal{P}(\Omega_i)$ denote the set of

probability measures over Ω_i. An m-variate distribution $\nu \in \mathcal{P}(\Omega_1 \times \dots \Omega_m)$ is a coupling of its marginals $\{\nu_i\}_{i=1}^m$, where ν_i is the ith coordinate marginal. We write $[m]$ for $\{1, \dots, m\}$. To conclude the section, let us recall a standard convention, used throughout RL, for describing the interaction with an environment via a generative model.

Definition 3.2 (Sample transition model (STM) [Bellemare et al., 2023])**.** Let $\xi \in \Delta(\mathcal{S})$ and π be a policy. The STM is the joint distribution of (S, A, R, S') generated by

$$\begin{aligned} S \sim \xi, \quad A \mid S \sim \pi(\cdot \mid S), \\ R \mid S, A \sim P_R(\cdot \mid S, A), \quad S' \mid S, A \sim P_S(\cdot \mid S, A). \end{aligned} \tag{2}$$

4 COUPLED-DYNAMICS ENVIRONMENTS AND JOINT MDPS

Many distributional quantities in RL are intrinsically joint across actions at a fixed state, including gaps and advantages, probabilities of superiority, and tail functionals of such quantities. Formalizing such quantities requires a joint law for *counterfactual* one-step outcomes that would occur under different actions taken from the same state, under the same realization of exogenous randomness. The MDP formalism does not account for this, since it only specifies marginal one-step laws for each action and leaves the joint counterfactual structure unspecified. Consequently, environments that agree on all one-step marginals when modeled using an MDP may still disagree on joint-action quantities.

We now present a multi-action variant of the STM, the m-joint STM (m-JSTM). Informally, an m-JSTM allows the learner to query m distinct actions at the same state, and obtain m coupled counterfactual one-step outcomes.

Definition 4.1 (m-JSTM)**.** For an integer $m \geq 1$, an m-JSTM is a kernel

$$\begin{aligned} \mathcal{J}_m(\cdot \mid s, a_{1:m}) \in \mathcal{P}\big(([0,1] \times \mathcal{S})^m\big), \\ s \in \mathcal{S}, a_{1:m} \in \mathcal{A}^m \text{ distinct}, \end{aligned} \tag{3}$$

such that each coordinate marginal matches the classical one-step marginals. Concretely, if $((R_i, S_i'))_{i=1}^m \sim \mathcal{J}_m(\cdot \mid s, a_{1:m})$, then for each coordinate $i \in [m]$,

$$(R_i, S_i') \sim \big(P_R(\cdot \mid s, a_i), P_S(\cdot \mid s, a_i)\big). \tag{4}$$

Specifically, when $m = N$ and $a_{1:N}$ is a fixed ordering of $\mathcal{A}$, $\mathcal{J}_m(\cdot \mid s, a_{1:N})$ specifies a joint coupling across all action at s. The additional expressiveness provided by an m-JSTM is the *dependence* it specifies across counterfactual outcomes for different actions. When this dependence is nontrivial, the MDP formalism omits information that is directly relevant for joint distributional questions. This motivates treating the coupling as part of the model.

Definition 4.2 (Coupled-dynamics environment)**.** An environment with marginal kernels (P_R, P_S) is a *coupled-dynamics* environment if it admits an m-JSTM $\mathcal{J}_m$ for some $m \geq 2$ such that for some query $(s, a_{1:m})$, $\mathcal{J}_m(\cdot \mid s, a_{1:m})$ is **not** the product coupling of its marginals. Equivalently, the environment contains joint one-step counterfactual structure that is not determined by (P_R, P_S) alone, and is therefore lost under a purely marginal MDP description.

The following examples illustrate the issue: Identical MDP marginals can conceal different joint structures.

Example 1. Let $\mathcal{A} = \{1, 2\}$ and fix $s \in \mathcal{S}$. Let $U \sim \mathrm{Bernoulli}(1/2)$, and let the counterfactual rewards be $R^{(1)} = U$, $R^{(2)} = 1 - U$. Each action's marginal reward law is $\mathrm{Bernoulli}(1/2)$, while their joint law is perfectly anti-correlated. $\mathbb{P}(R^{(1)} > R^{(2)}) = 1/2$ and $\mathbb{E}[R^{(1)} R^{(2)}] = 0$, neither fact determined by the marginals alone.

Example 2. Let $\mathcal{S} = \{0, 1\}$, $\mathcal{A} = \{1, 2\}$, and fix $s \in \mathcal{S}$. Let $U \sim \mathrm{Bernoulli}(1/2)$, and the counterfactual next states be

(2.1) $S'^{(1)} = S'^{(2)} = U$. Each action's marginal transition law is uniform on $\mathcal{S}$, while $S'^{(1)} = S'^{(2)}$.

(2.2) $S'^{(1)} = U$ and $S'^{(2)} = 1 - U$. Each action's marginal transition law is uniform on $\mathcal{S}$, while $S'^{(1)} = 1 - S'^{(2)}$.

We now formalize a joint extension of the MDP model, appropriate for modeling coupled-dynamics environments.

Definition 4.3 (JMDP)**.** A *JMDP* is a quadruple $(\mathcal{S}, \mathcal{A}, \gamma, \mathcal{J})$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space with $N = |\mathcal{A}|$, $\gamma \in (0, 1)$ is the discount factor, and $\mathcal{J}(\cdot \mid s) \in \mathcal{P}(([0,1] \times \mathcal{S})^N)$ is a Markov kernel over counterfactual one-step outcome tables $((R^{(a)}, S'^{(a)}))_{a \in \mathcal{A}}$.

At time t, conditional on $S_t = s$, the environment samples a one-step outcome table $((R^{(a)}, S'^{(a)}))_{a \in \mathcal{A}} \sim \mathcal{J}(\cdot \mid s)$. The agent selects action A_t to execute, the realized transition is given by the executed action: $(R_t, S_{t+1}) = (R_t^{(A_t)}, S_{t+1}^{(A_t)})$. The remaining coordinates of the sampled outcome table are counterfactual and do not affect the realized trajectory.

Remark 4.4. Every JMDP induces a marginal MDP description of the environment when we take coordinate marginals of the kernel $\mathcal{J}$. To see this, consider the collection of 1-JSTMs $\mathcal{J}_1(\cdot \mid s, a)$ for each (s, a) such that

$$\begin{aligned} \mathcal{J}_1(\cdot \mid s, a) := \mathrm{Law}\big((R^{(a)}, S'^{(a)})\big), \\ \text{where } \big((R^{(b)}, S'^{(b)})\big)_{b \in \mathcal{A}} \sim \mathcal{J}(\cdot \mid s). \end{aligned} \tag{5}$$

Then, the reward and transition marginals $P_R(\cdot \mid s, a)$ and $P_S(\cdot \mid s, a)$ of $\mathcal{J}_1$ will constitute the reward and transition kernels of a classical MDP description of the environment.

In a similar vein, fixing $m \geq 2$, for any state s and any m-tuple of distinct actions $a_{1:m}$, we can induce an m-JSTM $\mathcal{J}_m$ as the joint law of the m queried coordinates:

$$\mathcal{J}_m(\cdot \mid s, a_{1:m}) := \mathrm{Law}\big((R^{(a)}, S'^{(a)})_{a\in a_{1:m}}\big), \quad \text{where } \big((R^{(b)}, S'^{(b)})\big)_{b\in\mathcal{A}} \sim \mathcal{J}(\cdot \mid s). \tag{6}$$

To summarize, a coupled-dynamics environment is one where the marginal description of an MDP formalism does not determine the joint law of counterfactual one-step outcomes across actions. JMDPs model the missing structure via a kernel $\mathcal{J}(\cdot \mid s)$ over counterfactual outcome tables, from which the standard MDP interaction model STM and the multi-action m-JSTM interface are obtained by marginalization. The JMDP formalism retains the dependence information discarded by an MDP description of the environment, essential for joint distributional questions.

The JMDP definition specifies how a counterfactual one-step outcome table is sampled at each visited state, and how a single executed action generates the realized transition. Before developing policy evaluation algorithms for joint quantities, it remains to formalize what we have referred to as the *one-step coupling* regime. The coupling is local in time, i.e., the current table shares exogenous randomness across actions at the queried state, and the next step samples fresh tables. If two coordinates later query the same successor state, they are coupled again through that local table. If they are at distinct states, their one-step randomness is independent. This matches multi-action simulators while also avoiding fully-coupled, exponentially-expanding counterfactual trajectory trees.

Assumption 4.5 (Exogenous noise representation [Bertsekas and Shreve, 1996, Ng and Jordan, 2000])**.** We assume that there exist an i.i.d. exogenous sequence $(U_t)_{t\geq 0}$ over a measurable space $\mathcal{U}$, and measurable maps

$$g : \mathcal{S}\times\mathcal{A}\times\mathcal{U} \to [0,1], \quad h : \mathcal{S}\times\mathcal{A}\times\mathcal{U}\to\mathcal{S}, \tag{7}$$

such that at any time t and state s, the counterfactual one-step outcomes for any action $a\in\mathcal{A}$ are related by

$$R_t^{(a)} = g(s,a,U_t), \quad S_{t+1}^{(a)} = h(s,a,U_t). \tag{8}$$

Thus, the full outcome table sampled by the JMDP kernel can alternatively be expressed as

$$\big((R_t^{(a)}, S_{t+1}^{(a)})\big)_{a\in\mathcal{A}} = \big((g(S_t,a,U_t), h(S_t,a,U_t))\big)_{a\in\mathcal{A}}. \tag{9}$$

Assumption 4.5 means that after the transition to successor states after time t, subsequent one-step outcome tables are generated using fresh exogenous variables, independent of U_t. Dependence is therefore mediated by local table draws and not by a single latent variable which is shared for an entire counterfactual tree.

We now define the main object of the rest of our inquiry.

Definition 4.6 (Joint return vector)**.** Let π be a stationary policy. For any state s, we define the *joint return vector* belonging to π as $Z^\pi(s) := (Z^\pi(s,a))_{a\in\mathcal{A}}$.

Under the one-step coupling regime, the components of $Z^\pi(s)$ are coupled through the shared outcome table at the initial state s, and then recursively through the pair process over successor states.

5 JOINT ITERATIVE POLICY EVALUATION

We now develop DP and incremental policy evaluation schemes for JMDPs. Throughout this section, we fix a policy π, and the goal is to estimate joint moments of $Z^\pi(s)$, including mixed moments. We start with the case of moments up to 2nd order, displaying an approachable and interpretable instance of the general case of nth-order moments. Proofs are deferred to Appendix A.

Let $\mathcal{X} := \mathcal{S}\times\mathcal{A}$ denote the state-action space. For each $(s,a)\in\mathcal{X}$, we define the first moments

$$\mu^\pi(s,a) := \mathbb{E}\left[Z^\pi(s,a)\right], \tag{10}$$

and for each $(s,a,\tilde{s},\tilde{a})\in\mathcal{X}^2$, the second moments

$$\Sigma^\pi(s,a,\tilde{s},\tilde{a}) := \mathbb{E}\left[Z^\pi(s,a)\, Z^\pi(\tilde{s},\tilde{a})\right]. \tag{11}$$

The algorithms defined below will operate on candidate moment collections $M = (M_\mu, M_\Sigma)$, where

$$M_\mu \in \{f : \mathcal{X}\to\mathbb{R}\}, \quad M_\Sigma \in \{f(s,a,\tilde{s},\tilde{a}) : \mathcal{X}^2\to\mathbb{R}\}. \tag{12}$$

Thus, $M_\mu(s,a)$ will be a candidate value for $\mu^\pi(s,a)$, and $M_\Sigma(s,a,\tilde{s},\tilde{a})$ for $\Sigma^\pi(s,a,\tilde{s},\tilde{a})$. The introduced algorithms will map any such dimensionally consistent moment collection M to the true collection of moments $M_2^\pi := (\mu^\pi, \Sigma^\pi)$.

We now define the 2nd-order joint Bellman operator.

Definition 5.1 (2nd-order joint Bellman operator)**.** The 2nd-order joint Bellman operator T_2^π maps a moment collection M to $T_2^\pi M$ coordinatewise through the equations

$$(T_2^\pi M)_\mu(s,a) := \mathbb{E}\big[R^{(a)} + \gamma M_\mu\big(S'^{(a)}, A'^{(a)}\big) \mid S = s\big] \tag{13}$$

for each (s,a). Furthermore, for each $(s,a,\tilde{s},\tilde{a})$,

$$\begin{aligned}(T_2^\pi M)_\Sigma(s,a,\tilde{s},\tilde{a}) := \mathbb{E}\big[&R^{(a)}\tilde{R}^{(\tilde{a})} \\ &+\gamma R^{(a)} M_\mu\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big) + \gamma\tilde{R}^{(\tilde{a})} M_\mu\big(S'^{(a)}, A'^{(a)}\big) \\ &+\gamma^2 M_\Sigma\big(S'^{(a)}, A'^{(a)}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big) \mid (S,\tilde{S}) = (s,\tilde{s})\big].\end{aligned} \tag{14}$$

It is instructional to consider the joint laws with respect to which we take the expectation in these equations. In (13), this is the joint law of $(R^{(a)}, S'^{(a)}, A'^{(a)})$, where

$$(R^{(a)}, S'^{(a)}) \sim J_1(\cdot \mid s,a), \quad A'^{(a)} \sim \pi(\cdot \mid S'^{(a)}), \tag{15}$$

where $\mathcal{J}_1$ is a 1-JSTM induced from the JMDP kernel $\mathcal{J}$.

This is also the case in the "diagonal" scenario of (14) ($s = \tilde{s}$ and $a = \tilde{a}$), where the expression simplifies to

$$\begin{aligned}&(T_2^\pi M)_\Sigma(s,a,s,a)\\ &= \mathbb{E}\big[\big(R^{(a)}\big)^2 + 2\gamma R^{(a)} M_\mu\big(S'^{(a)}, A'^{(a)}\big)\\ &+\gamma^2 M_\Sigma\big(S'^{(a)}, A'^{(a)}, S'^{(a)}, A'^{(a)}\big) \mid S = s\big],\end{aligned} \tag{16}$$

When $a \neq \tilde{a}$, the joint law is of $(R^{(a)}, S'^{(a)}, A'^{(a)}, \tilde{R}^{(\tilde{a})}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})})$. When we have $s \neq \tilde{s}$,

$$\begin{aligned}(R^{(a)}, S'^{(a)}) \sim \mathcal{J}_1(\cdot \mid s,a), \quad (\tilde{R}^{(\tilde{a})}, \tilde{S}'^{(\tilde{a})}) \sim \mathcal{J}_1(\cdot \mid \tilde{s},\tilde{a}),\\ A'^{(a)} \sim \pi(\cdot \mid S'^{(a)}), \quad \tilde{A}'^{(\tilde{a})} \sim \pi(\cdot \mid \tilde{S}'^{(\tilde{a})}).\end{aligned} \tag{17}$$

$\mathcal{J}_1$ is a 1-JSTM induced from $\mathcal{J}$, the draws are independent.

Crucially, in the "same-state" case where $s = \tilde{s}$, we have

$$\begin{aligned}\big(R^{(a)}, S'^{(a)}, \tilde{R}^{(\tilde{a})}, \tilde{S}'^{(\tilde{a})}\big) \sim \mathcal{J}_2\big(\cdot \mid s, (a,\tilde{a})\big),\\ A'^{(a)} \sim \pi(\cdot \mid S'^{(a)}), \quad \tilde{A}'^{(\tilde{a})} \sim \pi(\cdot \mid \tilde{S}'^{(\tilde{a})}),\end{aligned} \tag{18}$$

where $\mathcal{J}_2$ is a 2-JSTM induced from $\mathcal{J}$.

Note that, by construction, M_2^π is a fixed point of T_2^π, since

$$\begin{aligned}\mu^\pi(s,a) &= \mathbb{E}\big[R^{(a)} + \gamma Z^\pi\big(S'^{(a)}, A'^{(a)}\big) \mid S = s\big],\\ \Sigma^\pi(s,a,\tilde{s},\tilde{a}) &= \mathbb{E}\big[\big(R^{(a)} + \gamma Z^\pi\big(S'^{(a)}, A'^{(a)}\big)\big)\\ &\cdot\big(\tilde{R}^{(\tilde{a})} + \gamma Z^\pi\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big)\big) \mid (S,\tilde{S}) = (s,\tilde{s})\big].\end{aligned} \tag{19}$$

We now prove that T_2^π is a contraction mapping.

Lemma 5.2. *Let $\lambda := 2/(1-\gamma)$. For any moment collection M, define the norm*

$$\|M\|_\lambda := \max\left\{\|M_\mu\|_\infty, \frac{1}{\lambda}\|M_\Sigma\|_\infty\right\}. \tag{20}$$

Then, T_2^π is a γ-contraction in $\|\cdot\|_\lambda$.

We now consider the iterative process expressed by

$$M_{k+1} = T_2^\pi M_k. \tag{21}$$

Theorem 5.3. *T_2^π admits a unique fixed point M_2^π. Moreover, for any initialization M_0, the iteration (21) has geometric convergence with rate γ in $\|\cdot\|_\lambda$:*

$$\|M_k - M_2^\pi\|_\lambda \leq \gamma^k \|M_0 - M_2^\pi\|_\lambda. \tag{22}$$

Furthermore, for any moment collection M, define the 2nd-order Bellman residual as $\|M - T_2^\pi M\|_\lambda$. Then,

$$\|M - M_2^\pi\|_\lambda \leq \frac{1}{1-\gamma}\|M - T_2^\pi M\|_\lambda. \tag{23}$$

This enables the interpretation of (21) as an algorithm with a computable stopping criterion for ϵ accuracy in $\|\cdot\|_\lambda$, which we call 2nd-order joint iterative policy evaluation (JIPE-2).

5.1 JIPE-2 WITH FUNCTION APPROXIMATION

In high-dimensional or continuous state spaces, exact tabular JIPE-2 becomes impractical, so it is preferable to move to a function-approximation setting in which the moments are restricted to a parameterized $\hat{M}(\theta)$, with a PSD-structured parameterization for M_Σ to preserve valid 2nd moment geometry. The resulting objective is a projected fixed point equation of the form $\hat{M}(\theta^*) = \Pi T_2^\pi \hat{M}(\theta^*)$. This yields a projected JIPE-2 iteration with a standard projection for the mean term and a PSD-constrained projection for the 2nd-moment term. Under standard regularity conditions, Appendix B shows that the projected operator is contractive in a suitable weighted norm, giving existence and uniqueness of the projected fixed point, convergence of the iteration, and a standard approximation-error bound.

5.2 INCREMENTAL JIPE-2

We next give an incremental variant of JIPE-2, suitable when $\mathcal{J}$ is accessed only through samples from its induced 1- and 2-JSTMs. The method is an instance of stochastic approximation for the operator T_2^π. Let us define an index set to simplify indexing a moment collection $M = (M_\mu, M_\Sigma)$:

$$\mathcal{I}_2 := \mathcal{X} \times \{\mu\} \cup \mathcal{X}^2 \times \{\Sigma\}. \tag{24}$$

For $i = ((s,a),\mu)$, we write $M(i) = M_\mu(s,a)$, and for $i = ((s,a),(\tilde{s},\tilde{a}),\Sigma)$, we write $M(i) = M_\Sigma(s,a,\tilde{s},\tilde{a})$.

We define the random one-sample backup $\hat{T}_2^\pi(M,i)$:

(i) If $i = ((s,a),\mu)$, we draw $(R^{(a)}, S'^{(a)}) \sim \mathcal{J}_1(\cdot \mid s,a)$ from the induced 1-JSTM, $A'^{(a)} \sim \pi(\cdot \mid S'^{(a)})$. We set

$$\hat{T}_2^\pi(M,i) := R^{(a)} + \gamma M_\mu\big(S'^{(a)}, A'^{(a)}\big). \tag{25}$$

(ii) If $i = ((s,a),(s,a)),\Sigma)$, we draw $(R^{(a)}, S'^{(a)}) \sim \mathcal{J}_1(\cdot \mid s,a)$ from the induced 1-JSTM and $A'^{(a)} \sim \pi(\cdot \mid S'^{(a)})$. We set

$$\begin{aligned}\hat{T}_2^\pi(M,i) := \big(R^{(a)}\big)^2 + 2\gamma R^{(a)} M_\mu\big(S'^{(a)}, A'^{(a)}\big)\\ +\gamma^2 M_\Sigma\big(S'^{(a)}, A'^{(a)}, S'^{(a)}, A'^{(a)}\big).\end{aligned} \tag{26}$$

(iii) If $i = ((s,a),(\tilde{s},\tilde{a}),\Sigma)$, with $a \neq \tilde{a}$ but $s = \tilde{s}$, we draw $(R^{(a)}, S'^{(a)}, \tilde{R}^{(\tilde{a})}, \tilde{S}'^{(\tilde{a})}) \sim \mathcal{J}_2(\cdot \mid s,(a,\tilde{a}))$ from the induced 2-JSTM, $A'^{(a)} \sim \pi(\cdot \mid S'^{(a)})$, $\tilde{A}'^{(\tilde{a})} \sim \pi(\cdot \mid \tilde{S}'^{(\tilde{a})})$. If $s \neq \tilde{s}$, we independently draw $(R^{(a)}, S'^{(a)}) \sim \mathcal{J}_1(\cdot \mid s,a)$, $(\tilde{R}^{(\tilde{a})}, \tilde{S}'^{(\tilde{a})}) \sim \mathcal{J}_1(\cdot \mid \tilde{s},\tilde{a})$ and $A'^{(a)} \sim \pi(\cdot \mid S'^{(a)})$, $\tilde{A}'^{(\tilde{a})} \sim \pi(\cdot \mid \tilde{S}'^{(\tilde{a})})$. In either case, we set

$$\begin{aligned}\hat{T}_2^\pi(M,i) := R^{(a)}\tilde{R}^{(\tilde{a})}\\ +\gamma R^{(a)} M_\mu\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big) + \gamma\tilde{R}^{(\tilde{a})} M_\mu\big(S'^{(a)}, A'^{(a)}\big)\\ +\gamma^2 M_\Sigma\big(S'^{(a)}, A'^{(a)}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big).\end{aligned} \tag{27}$$

For $\alpha_k \in (0,1)$, we define the incremental JIPE-2 recursion:

$$M_{k+1}(i) = \begin{cases} (1-\alpha_k(i))M_k(i) + \alpha_k(i)\hat{T}_2^\pi(M_k, i), & i = I_k, \\ M_k(i), & i \neq I_k, \end{cases} \tag{28}$$

where I_k designates the kth (random) index updated. The following theorem establishes almost-sure convergence.

Theorem 5.4. *Assume each coordinate $i \in \mathcal{I}_2$ is selected infinitely often by I_k. Let the step size $\alpha_k(i)$ satisfy the conditions of Robbins and Monro [1951] for every index i:*

$$\sum_k \alpha_k(i) = \infty, \quad \sum_k \alpha_k(i)^2 < \infty. \tag{29}$$

Then, the incremental JIPE-2 iteration converges almost surely in $\|\cdot\|_\lambda$, i.e.,

$$\|M_k - M_2^\pi\|_\lambda \to 0 \ \text{ almost surely as } k \to \infty. \tag{30}$$

5.3 GENERALIZATION TO HIGHER MOMENTS

It is logically straightforward to extend the construction to the case of nth-order moments, with heavier notation. For each $k \in [n]$, we consider a kth-order moment collection

$$M^{(k)} := \mathcal{X}^k \to \mathbb{R}, \quad M^{(k)}(x_{1:k}) = \mathbb{E}\left[\prod_{i=1}^k Z^\pi(x_i)\right]. \tag{31}$$

The collection of all moments of orders 1 through n is $M := (M^{(1)}, \dots, M^{(n)})$.

Let $k \le n$ and $x_{1:k} \in \mathcal{X}^k$. Let $\{G_l\}_{l=1}^L$ be the partition of $[k]$ into groups that share the same current state, i.e., $i, j \in G_l$ for some l if and only if $s_i = s_j$. For each group G_l with common state $s^{(l)}$ and action subtuple $a_{G_l} := (a_i)_{i \in G_l}$, let

$$\big((R_i, S_i')\big)_{i \in G_l} \sim \mathcal{J}_{|G_l|}(\cdot \mid s^{(l)}, a_{G_l}) \tag{32}$$

under the induced $|G_l|$-JSTM. Assume these groupwise draws are independent across l. Conditioned on $(S_i')_{i=1}^k$, let actions be drawn independently as $A_i' \sim \pi(\cdot \mid S_i')$ for $i \in [k]$. We write $X_I' := (X_i')_{i \in I}$ for any index set $I \subseteq [k]$. This construction is the direct k-variate analogue of the joint laws used in the 2nd-order scenario. Within a state, the relevant coordinates share one coupled draw via induced JSTMs, whereas across distinct states, the one-step randomness is independent under the one-step coupling regime.

We define the operator T_n^π through the system of equations

$$(T_n^\pi M)_k(x_{1:k}) := \mathbb{E}\left[\sum_{I \subseteq [k]} \gamma^{|I|}\Big(\prod_{i \notin I} R_i\Big) M_{|I|}(X_I') \mid x_{1:k}\right], \tag{33}$$

for each $k \in [n]$ and each $x_{1:k} \in \mathcal{X}^k$. We also define

$$\|M\|_\lambda := \max_{k \in [n]} \frac{1}{\lambda_k} \left\|M^{(k)}\right\|_\infty, \tag{34}$$

for $\lambda_k := (2/(1-\gamma))^{k-1}$. The operator T_n^π is a γ-contraction with respect to $\|\cdot\|_\lambda$, and admits as unique fixed point the true collection of moments up to nth order, M_n^π. The JIPE-n algorithm and its incremental variant can be defined analogously to the previous section, and admit similar results to those of Theorems 5.3 and 5.4. These higher-order moments may be of use when mean and variance miss skewed or heavy-tailed action gap risks, analogous to high-order portfolio objective which extend mean-variance criteria with higher moments [Zhou and Palomar, 2021].

5.4 CONNECTION TO THE GAP RV

We now return to the discussion of the gap RV $G^\pi(s; a, \tilde{a})$. Quantities that only involve the first moments require no joint modeling, for instance, $\mathbb{E}[G^\pi(s; a, \tilde{a})] = \mu^\pi(s, a) - \mu^\pi(s, \tilde{a})$ is already determined by the marginal dynamics. However, joint structure becomes relevant when we consider comparative or risk-sensitive questions about gaps. JMDPs specify a concrete coupling of counterfactual outcomes for $(a, \tilde{a})$ at s, inducing a well-defined joint return law. The moment recursions of the previous section provide mixed moments $\mathbb{E}[Z^\pi(s, a)^i Z^\pi(s, \tilde{a})^j]$, which determine the moments of $G^\pi(s; a, \tilde{a})$ and enable moment-based bounds or approximations for comparative risk criteria. For example, 2nd-order mixed moments yield the variance of the gap:

$$\begin{aligned} \operatorname{var}\big(G^\pi(s; a, \tilde{a})\big) &= \Sigma^\pi(s, a, s, a) - \Sigma^\pi(s, \tilde{a}, s, \tilde{a}) \\ &\quad -2\Sigma^\pi(s, a, s, \tilde{a}) - \big(\mu^\pi(s, a) - \mu^\pi(s, \tilde{a})\big)^2, \end{aligned} \tag{35}$$

where all moments are computable through JIPE-2.

Similarly, if $\mu_G := \mathbb{E}[G^\pi(s; a, \tilde{a}) > 0]$, by the Chebyshev inequality [Boucheron et al., 2013] $\mathbb{P}(G^\pi(s; a, \tilde{a}) \le 0) \le \sigma_G^2/(\sigma_G^2 + \mu_G^2)$, where $\sigma_G^2 := \operatorname{var}(G^\pi(s; a, \tilde{a}))$.

6 EXPERIMENTS

We empirically validate the policy-evaluation theory developed in the previous sections in four complementary ways. First, in tabular coupled-dynamics environments we run JIPE-2 and track the Bellman residual $\|M - T_2^\pi M\|_\lambda$, which provides a certificate of accuracy. Second, we visualize the learned joint structure via return correlation matrices across actions, computed from the first and second moments. Third, we demonstrate that the learned mixed moments enable meaningful estimates and bounds for the gap RV, and validate these estimates against MC simulation. Finally, we show that the incremental JIPE-2 with neural function approximation scales beyond tabular settings in coupled ALE environments Bellemare et al. [2012].

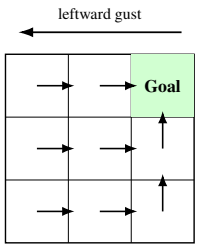


Figure 1: A 3 × 3 WGW environment and the evaluated policy.

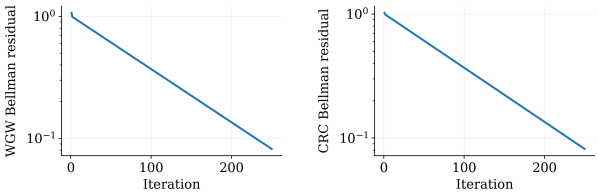


Figure 2: **Bellman residual convergence in tabular coupled-dynamics environments. Left:** 5×5 WGW. **Right:** CRC with $|\mathcal{S}| = 25$. We plot the Bellman residual $\|M_k - T_2^\pi M_k\|_\lambda$ on logarithmic scale.

All experiments are policy evaluation: We fix π and estimate the first moments $\mu^\pi(s,a)$ and second moments $\Sigma^\pi(s,a,\tilde{s},\tilde{a})$. In tabular experiments, we apply the exact operator T_2^π and report the Bellman residual. In large-scale experiments, we implement incremental JIPE-2 with neural approximation, and report moving-average TD errors.

Windy gridworld (WGW). We consider a WGW environment with actions $\{\mathsf{U},\mathsf{R},\mathsf{D},\mathsf{L}\}$ and a leftward "wind gust" that couples counterfactual next states under shared exogenous randomness (Figure 1). Figure 2 (left) shows linear decay of the Bellman residual on a log scale, consistent with the γ-contraction and geometric convergence results.

Coupled-reward chain (CRC). We consider a finite chain of states with two actions and anti-correlated rewards (Figure 3). Figure 2 (right) again displays linear decay of the Bellman residual on a log scale. To visualize the cross-action dependence, we form, for each state s, the covariance and correlation matrices from the estimated uncentered second moments. Figure 4 shows, for a 3×3 WGW under a fixed goal-directed policy, the 4×4 correlation matrix ρ_s^π at each state. The figure indicates that the coupled-dynamics induce a structured, state-dependent joint law across actions which is invisible to an MDP marginal description. Figure 5 similarly reports the learned return correlation matrix $\rho_{s_0}^\pi$ at the initial chain state s_0 of a CRC with $|\mathcal{S}| = 25$.

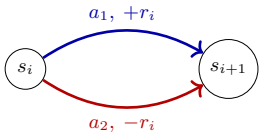


Figure 3: One CRC transition. Both actions reach the same next state but receive anti-correlated rewards.

We then validate the use of mixed moments for gap statistics. Figure 6 (left, middle) compare JIPE-2-derived predictions of $\mathbb{E}[G^\pi]$ and $\mathrm{var}(G^\pi)$ to MC estimates. The agreement indicates that the learned mixed moments meaningfully capture joint-action return structure. Furthermore, using the Chebyshev inequality, we derive an upper bound on the inferiority probability $\mathbb{P}(G^\pi \leq 0)$. We summarize the bound's tightness by the empirical cdf (ECDF) of the ratio $\hat{\mathbb{P}}(G^\pi \leq 0)/\mathrm{Chebyshev}(G^\pi)$ (where $\hat{\mathbb{P}}$ is an MC estimate). Figure 6 (right) shows ratios bounded by 1, indicating empirical non-violation of the upper-bound.

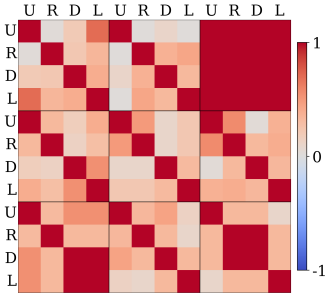


Figure 4: **Per-state action correlation matrices in WGW.** Each tile corresponds to a state in a 3×3 WGW, and displays the 4×4 correlation matrix ρ_s^π computed through JIPE-2.

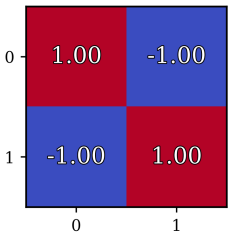


Figure 5: Correlation matrix $\rho_{s_0}^\pi$ computed from the converged JIPE-2 moments in CRC.

To demonstrate scalability beyond tabular settings, we implement incremental JIPE-2 with neural function approximation in four ALE environments with a coupled-dynamics interface. Concretely, we wrap the simulator with a multi-action one-step query mechanism that returns counterfactual one-step outcomes for multiple actions under shared emulator randomness at the queried state. We parameterize $M_\mu(s,a)$ and $M_\Sigma(s,a,\tilde{s},\tilde{a})$ with neural networks and perform SGD on TD errors. Figure 7 reports the moving-average TD errors for the μ coordinate and multiple Σ coordinate blocks. Across all cases, TD errors decrease by several orders of magnitude, providing evidence that incremental JIPE-2 can be combined with function approximation to mitigate the $|\mathcal{S}|^2|\mathcal{A}|^2$ complexity of tabular 2nd moment evaluation. Appendix C provides implementation and validation details.

7 CONCLUSION

We introduced JMDPs for coupled-dynamics environments that explicitly model the joint law of counterfactual outcomes across actions. Under a one-step coupling regime, we derived Bellman operators for return moments and obtained convergence guarantees for DP algorithms and their incremental variants, along with a projected function-approximation formulation. This makes intrinsically joint quantities such as gap moments and inferiority-probability bounds computable from multi-action simulator access. Experiments in coupled tabular and ALE environments show that learned mixed moments recover meaningful cross-action dependence absent from marginal MDP descriptions. A natural next step is control: extending joint moment evaluation to policy improvement under joint distributional objectives in JMDPs.

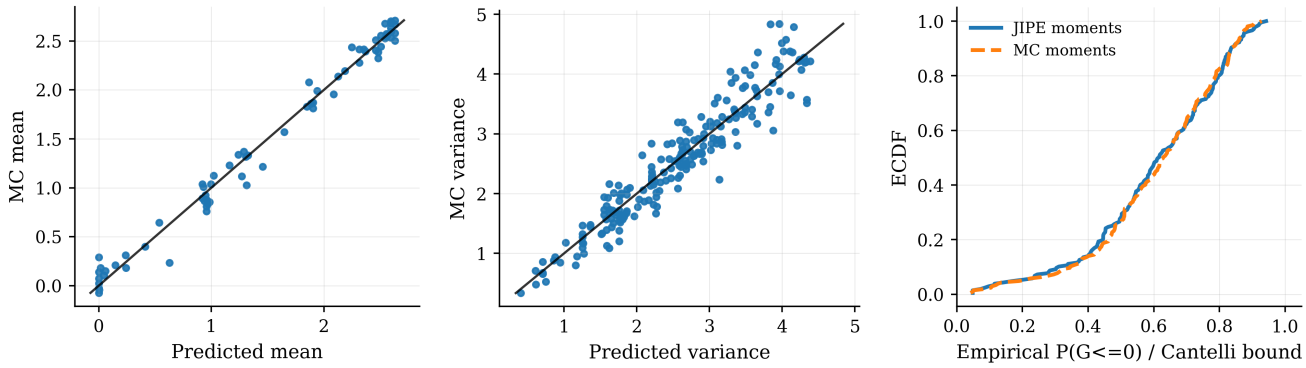


Figure 6: **Gap validation in WGW.** We evaluate a fixed goal-directed policy and consider gaps between the policy's action and alternatives. **Left:** Predicted vs. MC gap means $\mathbb{E}[G^\pi]$. **Middle:** Predicted vs. MC gap variances $\mathrm{var}(G^\pi)$. **Right:** Empirical cdf of the ratio $\hat{\mathbb{P}}(G^\pi \leq 0)/\mathrm{Chebyshev}(G^\pi)$. Ratios closer to 1 indicate tighter upper bounds, while small ratios indicate looseness. The **blue** curve computes the denominator using JIPE-2-estimated moments, while the **orange** curve uses MC-estimated moments (a near-ground-truth proxy), separating intrinsic bound looseness from moment-estimation error.

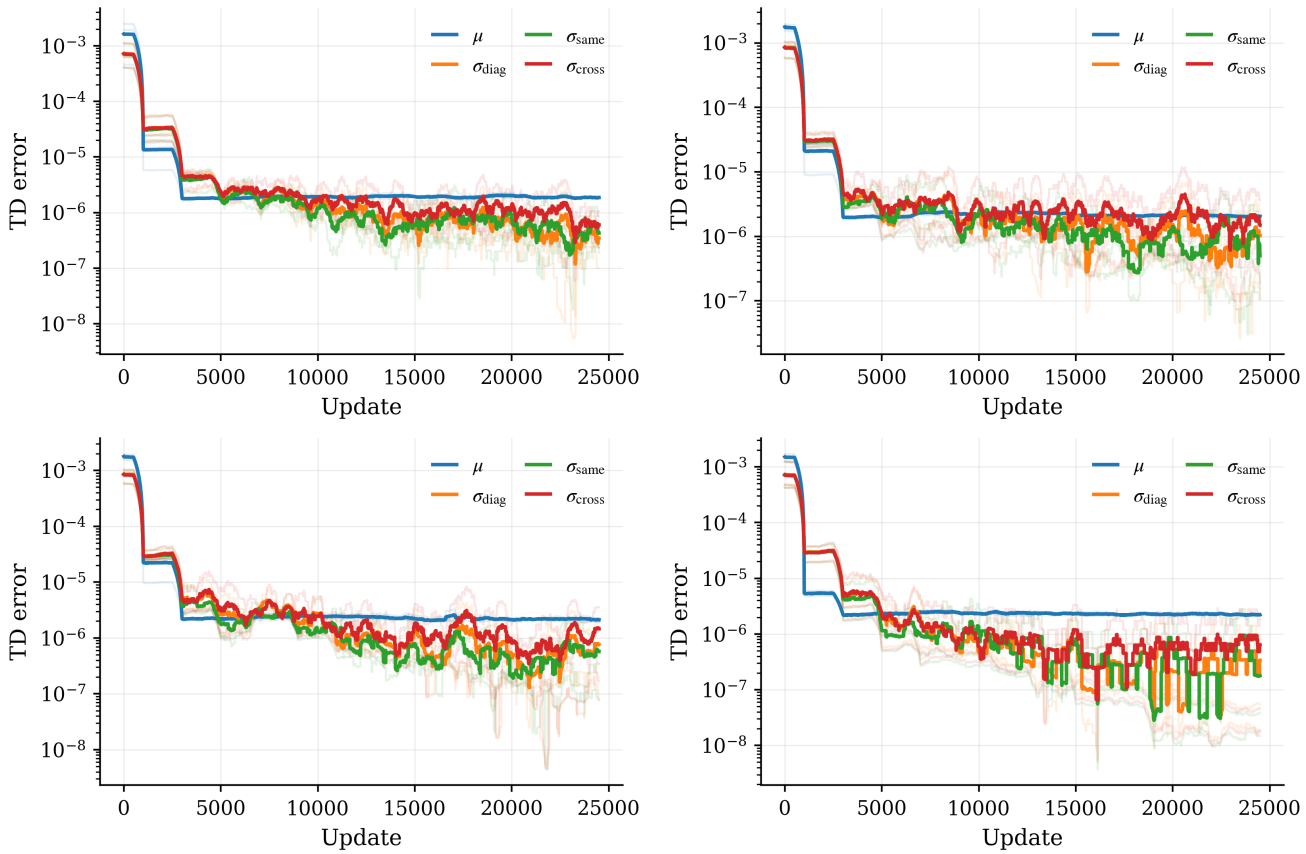


Figure 7: **Incremental JIPE-2 with neural functional approximation on coupled ALE environments.** Moving-average TD errors (log scale) during SGD-based incremental JIPE-2 with neural function approximation, shown for Pong, BattleZone, Boxing, and Atlantis (left-to-right, top-to-bottom). The plotted coordinates corresponds to the JIPE-2 blocks: μ is the 1st-moment (mean-return) component, σ_{diag} is the block where $s = \tilde{s}$ and $a = \tilde{a}$. σ_{same} is the block where $s = \tilde{s}$, $a \neq \tilde{a}$. σ_{cross} is the block where $s \neq \tilde{s}$, $a \neq \tilde{a}$. Bold curves show the across-seed mean TD error, and faint curves show the individual trajectories from **three random seeds**. Across all four ALE games, each block exhibits stable descent over training, supporting the practicality of the incremental JIPE-2 updates under nonlinear function approximation. Gap moments are evaluated separately against recursive-pair MC in Appendix C.

Acknowledgements

This work was supported in part by NSF under grants DMS-2502560 and CNS-2313109.

References

Yotam Amitai, Yael Septon, and Ofra Amir. Explaining reinforcement learning agents through counterfactual action outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10003–10011, 2024.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Marc Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47, 07 2012. doi: 10.1613/jair.3912.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. The MIT Press, 05 2023. ISBN 9780262374026. doi: 10.7551/mitpress/14207.001.0001. URL https://doi.org/10.7551/mitpress/14207.001.0001.

Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, volume 1, pages 560–564. IEEE, 1995.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3509–3517, Cambridge, MA, USA, 2014. MIT Press.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1096–1105. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018b. ISBN 978-1-57735-800-8.

Dror Freirich, Tzahi Shimkin, Ron Meir, and Aviv Tamar. Distributional multivariate policy evaluation and exploration with the Bellman GAN. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1983–1992. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/freirich19a.html.

Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer, 2004.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*, 2020.

Shie Mannor and John N. Tsitsiklis. Mean-variance optimization in markov decision processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 177–184, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Andrew Y. Ng and Michael Jordan. Pegasus: a policy search method for large mdps and pomdps. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, page 406–415, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607099.

Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment match-

ing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9144–9152, 2021.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Bruno Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In *27th International Conference on Machine Learning-ICML 2010*, 2010.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Matthew J. Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19 (4):794–802, 1982. ISSN 00219002. URL http://www.jstor.org/stable/3213832.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Temporal difference methods for the variance of the reward to go. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 495–503, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/tamar13.html.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36, 2016. URL http://jmlr.org/papers/v17/14-335.html.

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

Harley Wiltzer, Marc Bellemare, David Meger, Patrick Shafto, and Yash Jhaveri. Action gaps and advantages in continuous-time distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 47815–47848, 2024a.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37:101297–101336, 2024b.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34:1519–1529, 2021.

Rui Zhou and Daniel P. Palomar. Solving high-order portfolios via successive convex approximation algorithms. *IEEE Transactions on Signal Processing*, 69:892–904, 2021. doi: 10.1109/TSP.2021.3051369.

Joint MDPs and Distributional Reinforcement Learning in Coupled-Dynamics Environments (Supplementary Material)

Ege C. Kaya[1] **Mahsa Ghasemi**[1] **Abolfazl Hashemi**[1]

[1] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

A PROOFS

Lemma A.1. *Let $\lambda := 2/(1-\gamma)$. For any moment collection M, define the norm*

$$\|M\|_\lambda := \max\left\{\|M_\mu\|_\infty, \frac{1}{\lambda}\|M_\Sigma\|_\infty\right\}. \tag{36}$$

Then, T_2^π is a γ-contraction in $\|\cdot\|_\lambda$.

Proof. Firstly, $\|\cdot\|_\lambda$ is clearly a norm, since it is a maximum over two norms in their respective coordinate spaces. Now, we have, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$|(T_2^\pi M)_\mu(s,a) - (T_2^\pi M')_\mu(s,a)| = \gamma\left|\mathbb{E}\big[(M-M')_\mu\big(S'^{(a)}, A'^{(a)}\big) \mid S=s\big]\right| \le \gamma\|(M-M')_\mu\|_\infty. \tag{37}$$

Then, taking the supremum over (s,a) on the left-hand side directly yields

$$\begin{aligned}\|(T_2^\pi M - T_2^\pi M')_\mu\|_\infty \le \gamma\|(M-M')_\mu\|_\infty &\le \gamma\max\left\{\|(M-M')_\mu\|_\infty, \frac{1}{\lambda}\|(M-M')_\Sigma\|_\infty\right\}\\ &= \gamma\|M-M'\|_\lambda\end{aligned} \tag{38}$$

Similarly for the second-moment coordinate space, for any $(s,a,\tilde{s},\tilde{a}) \in (\mathcal{S}\times\mathcal{A})^2$,

$$\begin{aligned}|(T_2^\pi M)_\Sigma(s,a,\tilde{s},\tilde{a}) - (T_2^\pi M')_\Sigma(s,a,\tilde{s},\tilde{a})| &= \big|\mathbb{E}\big[\gamma R^{(a)}(M-M')_\mu\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big) + \gamma\tilde{R}^{(\tilde{a})}(M-M')_\mu\big(S'^{(a)}, A'^{(a)}\big)\\ &\quad +\gamma^2(M-M')_\Sigma\big(S'^{(a)}, A'^{(a)}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big) \mid (S,\tilde{S}) = (s,\tilde{s})\big]\big|\\ &\le \mathbb{E}\big[\gamma\left|R^{(a)}(M-M')_\mu\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big)\right| + \gamma\left|\tilde{R}^{(\tilde{a})}(M-M')_\mu\big(S'^{(a)}, A'^{(a)}\big)\right|\\ &\quad +\gamma^2\left|(M-M')_\Sigma\big(S'^{(a)}, A'^{(a)}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big)\right| \mid (S,\tilde{S}) = (s,\tilde{s})\big]\\ &\le 2\gamma\|(M-M')_\mu\|_\infty + \gamma^2\|(M-M')_\Sigma\|_\infty.\end{aligned} \tag{39}$$

Once again, taking the supremum over $(s,a,\tilde{s},\tilde{a})$ on the left-hand side and dividing by λ yields

$$\begin{aligned}\frac{1}{\lambda}\|(T_2^\pi M - T_2^\pi M')_\Sigma)\|_\infty &\le \frac{2\gamma}{\lambda}\|(M-M')_\mu\|_\infty + \frac{\gamma^2}{\lambda}\|(M-M')_\Sigma\|_\infty\\ &\le \big(\frac{2\gamma}{\lambda} + \gamma^2\big)\max\left\{\|(M-M')_\mu\|_\infty, \frac{1}{\lambda}\|(M-M')_\Sigma\|_\infty\right\}\\ &= \gamma\max\left\{\|(M-M')_\mu\|_\infty, \frac{1}{\lambda}\|(M-M')_\Sigma\|_\infty\right\}\\ &= \gamma\|M-M'\|_\lambda.\end{aligned} \tag{40}$$

Using (38) and (40), we conclude

$$\|T_2^\pi M - T_2^\pi M'\|_\lambda := \max\left\{\|(T_2^\pi M - T_2^\pi M')_\mu\|_\infty, \frac{1}{\lambda}\|(T_2^\pi M - T_2^\pi M')_\Sigma)\|_\infty\right\} \le \gamma\|M - M'\|_\lambda. \tag{41}$$

□

Theorem A.2. *T_2^π admits a unique fixed point M_2^π. Moreover, for $\lambda = 2/(1-\gamma)$ and initialization M_0, JIPE-2 has geometric convergence in γ with respect to $\|\cdot\|_\lambda$:*

$$\|M_k - M_2^\pi\|_\lambda \le \gamma^k \|M_0 - M_2^\pi\|_\lambda. \tag{42}$$

Furthermore, for any moment collection M, define the 2nd-order Bellman residual as $\|M - T_2^\pi M\|_\lambda$. Then,

$$\|M - M_2^\pi\|_\lambda \le \frac{1}{1-\gamma}\|M - T_2^\pi M\|_\lambda. \tag{43}$$

Proof. The proof of the first statement is a simple application of the Banach fixed-point theorem to Lemma 5.2. For the second statement, since $M_2^\pi = T_2^\pi M_2^\pi$,

$$\begin{aligned}\|M - M_2^\pi\|_\lambda &\le \|M - T_2^\pi M\|_\lambda + \|T_2^\pi M - T_2^\pi M_2^\pi\|_\lambda \\ &\le \|M - T_2^\pi M\|_\lambda + \gamma\|M - M_2^\pi\|_\lambda.\end{aligned} \tag{44}$$

Rearranging furnishes the result. □

Lemma A.3. *Let $\mathcal{F}_k$ be a filtration such that each update index $I_k \in \mathcal{I}_2$ is $\mathcal{F}_k$-measurable, and the randomness used to form $\hat{T}_2^\pi(M_k, I_k)$ is $\mathcal{F}_{k+1}$-measurable and conditionally independent of the past given $\mathcal{F}_k$. Let the coordinatewise noise term be*

$$\omega_{k+1}(i) := \hat{T}_2^\pi(M_k, i) - (T_2^\pi M_k)(i). \tag{45}$$

Then, for every index $i \in \mathcal{I}_2$ updated at time k

$$\mathbb{E}\big[\omega_{k+1}(i) \mid \mathcal{F}_k\big] = 0, \tag{46}$$

and there exist finite constants $C_0, C_1 > 0$ such that

$$\mathbb{E}\big[\omega_{k+1}(i)^2 \mid \mathcal{F}_k\big] \le C_0 + C_1\|M_k\|_\lambda^2. \tag{47}$$

Proof. Fix a time k and let $i \in \mathcal{I}_2$ be an index updated at time k. Equation (46) holds by construction, since $\hat{T}_2^\pi(M_k, i)$ is a sample-based estimator of the defining conditional expectations for $(T_2^\pi M_k)(i)$. In other words, $(T_2^\pi M_k)(i) = \mathbb{E}[\hat{T}_2^\pi(M_k, i) \mid \mathcal{F}_k]$. It remains to prove (47). Firstly,

$$\begin{aligned}\omega_{k+1}(i)^2 &= \big(\hat{T}_2^\pi(M_k, i) - (T_2^\pi M_k)(i)\big)^2 \\ &\le 2\hat{T}_2^\pi(M_k, i)^2 + 2(T_2^\pi M_k)(i)^2.\end{aligned} \tag{48}$$

Taking expectations conditional to $\mathcal{F}_k$ on both sides,

$$\begin{aligned}\mathbb{E}\big[\omega_{k+1}(i)^2 \mid \mathcal{F}_k\big] &\le 2\mathbb{E}\big[\hat{T}_2^\pi(M_k, i)^2 \mid \mathcal{F}_k\big] + 2(T_2^\pi M_k)(i)^2 \\ &\le 2\mathbb{E}\big[\hat{T}_2^\pi(M_k, i)^2 \mid \mathcal{F}_k\big] + 2\mathbb{E}\big[\hat{T}_2^\pi(M_k, i) \mid \mathcal{F}_k\big] \\ &= 4\mathbb{E}\big[\hat{T}_2^\pi(M_k, i)^2 \mid \mathcal{F}_k\big],\end{aligned} \tag{49}$$

using Jensen's inequality. Thus, it suffices to upper bound $\mathbb{E}[\hat{T}_2^\pi(M_k, i) \mid \mathcal{F}_k]$ by an affine function of $\|M_k\|_\lambda^2$ for the two cases of i being a μ-index or a Σ-index.

Case 1: If $i = ((s, a), \mu)$, then by definition of the sample backup,

$$\hat{T}_2^\pi(M_k, i) = R^{(a)} + \gamma(M_k)_\mu\big(S'^{(a)}, A'^{(a)}\big), \tag{50}$$

with $R^{(a)} \in [0,1]$ almost surely. Hence,

$$\left|\hat{T}_2^\pi(M_k, i)\right| \le 1 + \gamma \left\|(M_k)_\mu\right\|_\infty \le 1 + \gamma \left\|M_k\right\|_\lambda , \tag{51}$$

or alternatively,

$$\hat{T}_2^\pi(M_k, i)^2 \le 2 + 2\gamma^2 \left\|M_k\right\|_\lambda^2 . \tag{52}$$

Taking conditional expectations with respect to $\mathcal{F}_k$, and using (49), we have

$$\mathbb{E}\left[\omega_{k+1}(i)^2 \mid \mathcal{F}_k\right] \le 8 + 8\gamma^2 \left\|M_k\right\|_\lambda^2 . \tag{53}$$

Case 2: If $i = ((s,a),(\tilde{s},\tilde{a}),\Sigma)$, we have the sample backup

$$\hat{T}_2^\pi(M, i) := R^{(a)}\tilde{R}^{(\tilde{a})} + \gamma R^{(a)} M_\mu\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big) + \gamma \tilde{R}^{(\tilde{a})} M_\mu\big(S'^{(a)}, A'^{(a)}\big) + \gamma^2 M_\Sigma\big(S'^{(a)}, A'^{(a)}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big). \tag{54}$$

Once again, $R^{(a)}, \tilde{R}^{(\tilde{a})} \in [0,1]$ almost surely. Then,

$$\begin{aligned}\left|\hat{T}_2^\pi(M, i)\right| &\le 1 + \gamma \left|(M_k)_\mu\big(\tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big)\right| + \gamma \left|(M_k)_\mu\big(S'^{(a)}, A'^{(a)}\big)\right| + \gamma^2 \left|(M_k)_\Sigma\big(S'^{(a)}, A'^{(a)}, \tilde{S}'^{(\tilde{a})}, \tilde{A}'^{(\tilde{a})}\big)\right| \\ &\le 1 + 2\gamma \left\|(M_k)_\mu\right\|_\infty + \gamma^2 \left\|(M_k)_\Sigma\right\|_\infty .\end{aligned} \tag{55}$$

Furthermore, we have $\left\|(M_k)_\mu\right\|_\infty \le \left\|M_k\right\|_\lambda$ and $\left\|(M_k)_\Sigma\right\|_\infty \le \lambda \left\|M_k\right\|_\lambda$. Then,

$$\left|\hat{T}_2^\pi(M_k, i)\right| \le 1 + (2\gamma + \gamma^2\lambda) \left\|M_k\right\|_\lambda , \tag{56}$$

or alternatively,

$$\hat{T}_2^\pi(M_k, i)^2 \le 2 + 2(2\gamma + \gamma^2\lambda)^2 \left\|M_k\right\|_\lambda^2 . \tag{57}$$

Taking conditional expectations with respect to $\mathcal{F}_k$, and using (49), we have

$$\mathbb{E}\left[\omega_{k+1}(i)^2 \mid \mathcal{F}_k\right] \le 8 + 8(2\gamma + \gamma^2\lambda)^2 \left\|M_k\right\|_\lambda^2 . \tag{58}$$

Combining (53) and (58) to define

$$C_0 := 8, \qquad C_1 := 8 \max\left\{\gamma^2, (2\gamma + \gamma^2\lambda)^2\right\} \tag{59}$$

finishes the proof. □

Theorem A.4. *Assume each coordinate $i \in \mathcal{I}_2$ is selected infinitely often by I_k. Let the step size $\alpha_k(i)$ satisfy the conditions of Robbins and Monro [1951] for every index i:*

$$\sum_k \alpha_k(i) = \infty, \quad \sum_k \alpha_k(i)^2 < \infty. \tag{60}$$

Then, the incremental JIPE-2 iteration converges almost surely in $\left\|\cdot\right\|_\lambda$, i.e.,

$$\left\|M_k - M_2^\pi\right\|_\lambda \to 0 \ \text{ almost surely as } k \to \infty. \tag{61}$$

Proof. By Lemma 5.2, T_2^π is a contraction map on the space of moment collections under $\left\|\cdot\right\|_\lambda$, with unique fixed point M_2^π (Theorem 5.3). The incremental JIPE-2 recursion is an asynchronous stochastic approximation scheme [Jaakkola et al., 1993, Tsitsiklis, 1994, Bertsekas and Tsitsiklis, 1995] for T_2^π, with martingale-difference noise and controlled conditional second moment (Lemma A.3). The step size and visitation conditions are those required by the contraction-map stochastic approximation theorem by Bellemare et al. [2023], yielding almost-sure convergence to the fixed point. □

B JIPE-2 WITH FUNCTION APPROXIMATION

The exact DP algorithm, namely JIPE-2, is restricted to environments with discrete and appropriately small state-action spaces $\mathcal{X}$. In high-dimensional or continuous state spaces, representing the true moment collection $M = (M_\mu, M_\Sigma)$ exactly may become computationally intractable. We therefore study a projected fixed-point formulation in which the candidate moment collection is restricted to a parameterized family

$$\hat{M}(\theta) = \big(\hat{M}_\mu(\theta_\mu), \hat{M}_\Sigma(\theta_\Sigma)\big). \tag{62}$$

Function classes. For the first moment, let $\phi_\mu : \mathcal{X} \to \mathbb{R}^{d_\mu}$ be a feature map and define

$$\hat{M}_\mu(x; \theta_\mu) := \phi_\mu(x)^\top \theta_\mu, \qquad x = (s, a) \in \mathcal{X}. \tag{63}$$

For the second moment, we use a PSD-preserving quadratic form

$$\hat{M}_\Sigma(x, y; \Theta_\Sigma) := \phi_\mu(x)^\top \Theta_\Sigma \phi_\mu(y), \qquad \Theta_\Sigma \succeq 0. \tag{64}$$

Let $\mathcal{K}_\Sigma$ denote the closed convex cone of second-moment functions representable in this form. This parameterization ensures that every finite Gram matrix produced by $\hat{M}_\Sigma$ is positive semidefinite, so the approximation respects the second-moment geometry of return random variables.

Assumption B.1. The feature matrix $\Phi_\mu \in \mathbb{R}^{|\mathcal{X}| \times d_\mu}$ has full column rank.

Assumption B.2. The Markov chain over $\mathcal{X}$ induced by the stationary policy π is ergodic, with unique stationary state-action distribution ν satisfying $\nu(x) > 0$ for all $x \in \mathcal{X}$.

For first moments, we use the standard on-policy geometry

$$\langle f, g \rangle_\nu := \sum_{x \in \mathcal{X}} \nu(x) f(x) g(x), \qquad \|f\|_\nu := \sqrt{\langle f, f \rangle_\nu}. \tag{65}$$

For second moments, the natural transition operator is the pair kernel P_2 on $\mathcal{X}^2$ induced by the JMDP coupling and the fixed policy π. Since $\mathcal{X}^2$ is finite, P_2 has at least one stationary distribution. Fix any such distribution and call it μ_2, so that

$$\mu_2 P_2 = \mu_2. \tag{66}$$

The marginal dynamics of each coordinate of the pair process agree with the policy-induced marginal kernel P^π. By Assumption B.2, the marginals of any stationary μ_2 must therefore equal ν. The distribution μ_2 need not be unique and need not have full support. Let

$$\mathrm{supp}(\mu_2) := \{(x, y) \in \mathcal{X}^2 : \mu_2(x, y) > 0\}. \tag{67}$$

All $L_2(\mu_2)$ statements below are therefore support-relative, equivalently, stated modulo functions that agree on $\mathrm{supp}(\mu_2)$. Define

$$\langle h, \ell \rangle_{\mu_2} := \sum_{x, y \in \mathcal{X}} \mu_2(x, y) h(x, y) \ell(x, y), \qquad \|h\|_{\mu_2} := \sqrt{\langle h, h \rangle_{\mu_2}}. \tag{68}$$

Projection. We use a block-diagonal projection $\Pi M = (\Pi_\mu M_\mu, \Pi_\Sigma M_\Sigma)$, where

$$\Pi_\mu M_\mu := \arg \min_{\hat{M}_\mu \in \mathrm{span}(\Phi_\mu)} \|M_\mu - \hat{M}_\mu\|_\nu \tag{69}$$

and

$$\Pi_\Sigma M_\Sigma := \arg \min_{\hat{M}_\Sigma \in \mathcal{K}_\Sigma} \|M_\Sigma - \hat{M}_\Sigma\|_{\mu_2}. \tag{70}$$

The first projection is the usual orthogonal projection. The second is projection onto a closed convex cone in the Hilbert space $L_2(\mu_2)$, hence is non-expansive. If μ_2 is not full support, the projection does not identify values outside $\mathrm{supp}(\mu_2)$. The projected fixed-point equation is

$$\hat{M}^* = \Pi T_2^\pi \hat{M}^*. \tag{71}$$

B.1 PROJECTED JIPE-2

Projected JIPE-2 iterates

$$M_{k+1} = \Pi T_2^\pi M_k. \tag{72}$$

In parameters, the first-moment update has the usual least-squares form

$$\theta_{\mu,k+1} = \left(\Phi_\mu^\top D_\nu \Phi_\mu\right)^{-1} \Phi_\mu^\top D_\nu T_{2,\mu}^\pi \hat{M}(\theta_k), \tag{73}$$

where D_ν is diagonal with entries $\nu(x)$. The second-moment update is the PSD-constrained projection

$$\Theta_{\Sigma,k+1} = \arg\min_{\Theta \succeq 0} \left\| \Phi_\mu \Theta \Phi_\mu^\top - T_{2,\Sigma}^\pi \hat{M}(\theta_k) \right\|_{D_{\mu_2}}^2, \tag{74}$$

where D_{μ_2} is diagonal over $\mathcal{X}^2$ with entries $\mu_2(x,y)$, and $T_{2,\Sigma}^\pi \hat{M}(\theta_k)$ is viewed as a $|\mathcal{X}| \times |\mathcal{X}|$ matrix indexed by pair coordinates.

B.2 ANALYSIS OF PROJECTED JIPE-2

In the pair-stationary geometry, the coupled pair transition is automatically non-expansive in the norm used for the second-moment projection. The next lemma formalizes this.

Lemma B.3. *The marginal transition operator P^π is non-expansive in $L_2(\nu)$, and the pair transition operator P_2 is non-expansive in $L_2(\mu_2)$.*

Proof. For P^π, by Jensen's inequality and stationarity of ν,

$$\begin{aligned} \|P^\pi f\|_\nu^2 &= \sum_x \nu(x) \left(\sum_{x'} P^\pi(x' \mid x) f(x') \right)^2 \\ &\le \sum_x \nu(x) \sum_{x'} P^\pi(x' \mid x) f(x')^2 = \sum_{x'} \nu(x') f(x')^2 = \|f\|_\nu^2. \end{aligned} \tag{75}$$

The proof for P_2 is identical, using stationarity of μ_2:

$$\begin{aligned} \|P_2 h\|_{\mu_2}^2 &= \sum_z \mu_2(z) \left(\sum_{z'} P_2(z' \mid z) h(z') \right)^2 \\ &\le \sum_z \mu_2(z) \sum_{z'} P_2(z' \mid z) h(z')^2 = \sum_{z'} \mu_2(z') h(z')^2 = \|h\|_{\mu_2}^2, \end{aligned} \tag{76}$$

where $z = (x,y) \in \mathcal{X}^2$. □

Theorem B.4. *Under Assumptions B.1 and B.2, for any stationary distribution μ_2 of P_2, there exists $\beta > 0$ such that the projected joint Bellman operator ΠT_2^π is a κ-contraction in the weighted norm*

$$\|M\|_{\nu,\mu_2,\beta} := \max\left\{ \|M_\mu\|_\nu, \beta \|M_\Sigma\|_{\mu_2} \right\}, \tag{77}$$

with some $\kappa < 1$. Consequently, the projected iteration $M_{k+1} = \Pi T_2^\pi M_k$ converges to a unique projected fixed point $\hat{M}^*$ in this norm, equivalently a unique fixed point on $\mathrm{supp}(\mu_2)$ for the second-moment coordinate, and

$$\|\hat{M}^* - M_2^\pi\|_{\nu,\mu_2,\beta} \le \frac{1}{1-\kappa} \|\Pi M_2^\pi - M_2^\pi\|_{\nu,\mu_2,\beta}. \tag{78}$$

Moreover, for any pair (x,y) with $\mu_2(x,y) > 0$,

$$\left| \hat{M}_\Sigma^*(x,y) - M_{2,\Sigma}^\pi(x,y) \right| \le \frac{1}{\beta \sqrt{\mu_2(x,y)}} \|\hat{M}^* - M_2^\pi\|_{\nu,\mu_2,\beta}. \tag{79}$$

Thus, the theorem provides a pointwise guarantee for every pair state in the support of the chosen stationary pair distribution. If μ_2 has full support, this becomes a guarantee over all pair states.

Proof. Let M and $\bar{M}$ be arbitrary moment collections and write $\Delta = M - \bar{M}$. By non-expansiveness of Π_μ and Lemma B.3,

$$\|\Pi_\mu(T^\pi_{2,\mu}M - T^\pi_{2,\mu}\bar{M})\|_\nu \le \gamma\|\Delta_\mu\|_\nu. \tag{80}$$

For the second moment, for $X = (s, a)$ and $Y = (\tilde{s}, \tilde{a})$,

$$(T^\pi_{2,\Sigma}\Delta)(X, Y) = \gamma\mathbb{E}[R^{(a)}\Delta_\mu(Y') \mid X, Y] + \gamma\mathbb{E}[\tilde{R}^{(\tilde{a})}\Delta_\mu(X') \mid X, Y] + \gamma^2\mathbb{E}[\Delta_\Sigma(X', Y') \mid X, Y]. \tag{81}$$

The rewards are in $[0, 1]$, so Jensen's inequality and the fact that the Y' marginal under stationary pair dynamics is ν imply

$$\begin{aligned}\|\mathbb{E}[|\Delta_\mu(Y')| \mid X, Y]\|^2_{\mu_2} &\le \mathbb{E}_{\mu_2}\left[\mathbb{E}[\Delta_\mu(Y')^2 \mid X, Y]\right] \\ &= \mathbb{E}_{\mu_2}[\Delta_\mu(Y')^2] = \|\Delta_\mu\|^2_\nu.\end{aligned} \tag{82}$$

The same bound holds for the symmetric term involving X'. For the final term, we use Lemma B.3 to get

$$\|P_2\Delta_\Sigma\|_{\mu_2} \le \|\Delta_\Sigma\|_{\mu_2}. \tag{83}$$

Since Π_Σ is non-expansive in $L_2(\mu_2)$,

$$\|\Pi_\Sigma(T^\pi_{2,\Sigma}M - T^\pi_{2,\Sigma}\bar{M})\|_{\mu_2} \le 2\gamma\|\Delta_\mu\|_\nu + \gamma^2\|\Delta_\Sigma\|_{\mu_2}. \tag{84}$$

Therefore

$$\|\Pi T^\pi_2 M - \Pi T^\pi_2 \bar{M}\|_{\nu,\mu_2,\beta} \le \max\left\{\gamma, 2\beta\gamma + \gamma^2\right\}\|\Delta\|_{\nu,\mu_2,\beta}. \tag{85}$$

Choose any

$$0 < \beta < \frac{1 - \gamma^2}{2\gamma}. \tag{86}$$

Then $\kappa := \max\{\gamma, 2\beta\gamma + \gamma^2\} < 1$. Banach's fixed-point theorem guarantees convergence to a unique fixed point in the weighted norm. The approximation bound follows from the standard projected Bellman argument:

$$\begin{aligned}\|\hat{M}^* - M^\pi_2\| &= \|\Pi T^\pi_2\hat{M}^* - T^\pi_2 M^\pi_2\| \\ &\le \|\Pi T^\pi_2\hat{M}^* - \Pi T^\pi_2 M^\pi_2\| + \|\Pi M^\pi_2 - M^\pi_2\| \\ &\le \kappa\|\hat{M}^* - M^\pi_2\| + \|\Pi M^\pi_2 - M^\pi_2\|,\end{aligned} \tag{87}$$

where all norms are $\|\cdot\|_{\nu,\mu_2,\beta}$. Rearranging yields the claim. □

C ALE IMPLEMENTATION AND GAP DIAGNOSTICS

The ALE experiments use the same multi-action query interface as the tabular experiments, implemented by cloning the emulator state, restoring it for each queried action, applying shared exogenous shocks across the queried actions, and advancing only the executed branch. In the reward-coupled diagnostic below, the shared exogenous shock is the reward perturbation used to construct a nontrivial same-scenario gap law while preserving a controlled evaluation setting.

For neural second-moment approximation, we parameterize the covariance component through a Gram representation,

$$M_\Sigma(x, y) = M_\mu(x)M_\mu(y) + \langle v(x), v(y)\rangle, \tag{88}$$

which ensures that the estimated covariance matrix is positive semidefinite. Training also minimizes a TD residual for the implied gap second moment,

$$E[(Z_a - Z_b)^2] = M_\Sigma(x_a, x_a) + M_\Sigma(x_b, x_b) - 2M_\Sigma(x_a, x_b). \tag{89}$$

thereby directly supervising the combination of second-moment entries that determines action-gap variance.

Table 1: **Reward-coupled ALE gap-variance validation.** For each game, we sample emulator states, estimate the action-gap variance from recursive-pair MC, and report the ratios of the learned coupled prediction and an independent-marginals prediction to the same MC estimate. Values near 1.0 represent better calibrated estimates to the coupled MC gap variance.

Game	Coupled / MC	Independent / MC
Atlantis	1.03	83.3
BattleZone	3.87	115.7
Boxing	3.97	111.1
Pong	1.30	135.4

Across the four games, the coupled estimates are within a factor of 1.03–3.97 of the recursive-pair MC gap variance, while the independent-marginals estimates severely overestimate it by factors of 83.3–135.4.