MARKOV DECISION PROCESSES WITH TERMINAL STATES: OCCUPANCY MEASURE AND ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by text generation in large language models, where the policy determines both the response content and termination via an end-of-sequence token, we study Markov decision processes (MDPs) with terminal states without discounting. In contrast to the fixed-horizon and infinite-horizon MDPs widely studied in the literature, in our setting, both cumulative reward and trajectory length depend on the policy. While infinite-horizon MDPs typically use discounted rewards to ensure finite returns, the undiscounted cumulative reward objective is finite in this setting, provided that every stationary policy terminates almost surely. Assuming every stationary policy terminates almost surely, we establish that feasible state-action occupancy measures form a compact convex polytope, with expected return and termination time acting as linear functionals. Building on this geometry, we prove that any convex combination of these measures is realizable by an explicitly constructed stationary policy, and that the K-fold undiscounted Bellman operator is a strict contraction. Finally, we derive Trust-Region Policy Optimization (TRPO) and Natural Policy Gradient (NPG) algorithms for this setting, providing sufficient conditions for monotonic improvement and non-asymptotic convergence rates.

1 INTRODUCTION

Reinforcement learning (RL) is increasingly applied to domains where the policy determines both the trajectory content and its length. An important example is reward-based optimization for large language models (LLMs) (Ouyang et al., 2022; DeepSeek-AI et al., 2025), where the policy generates tokens autoregressively (Brown et al., 2020) until it selects a special end-of-sequence (EOS) token or reaches a length limit (Ramamurthy et al., 2022). This policy-dependent horizon differs fundamentally from standard RL formulations based on infinite-horizon discounted or fixed-horizon episodic Markov decision processes (MDPs) (Sutton et al., 1998; Bertsekas, 2019). Motivated by these features and the wide use of undiscounted rewards in LLM training (Ouyang et al., 2022; Shao et al., 2024), we study finite MDPs with terminal states under an undiscounted cumulative reward objective, where a key feature is that the horizon depends strictly on the policy.

Theoretical analysis of undiscounted MDPs with terminal states started within the stochastic shortest path (SSP) framework, where an agent seeks to reach a goal state while minimizing expected cumulative cost (Bertsekas & Tsitsiklis, 1991). Previous studies have developed online learning algorithms with regret guarantees for SSP problems, typically assuming the existence of at least one proper stationary policy, together with suitable conditions on the costs (Rosenberg et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021). A policy is proper if it reaches a terminal state almost surely from every initial state.

Given the wide use of policy gradient methods in settings with policy-dependent horizons (Ouyang et al., 2022; Shao et al., 2024), we focus on their improvement and convergence guarantees. In the undiscounted setting, the existence of a proper policy alone does not ensure that the objective is finite throughout the policy class: improper policies may generate nonterminating trajectories and divergent cumulative rewards. To ensure that the objective remains well defined throughout policy optimization, we adopt the stronger assumption that every stationary policy is proper. In finite MDPs with bounded rewards, this assumption guarantees uniformly bounded expected termination times and finite expected returns over all stationary policies.

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

This assumption has been studied in the SSP literature (Bertsekas & Tsitsiklis, 1991; Yu & Bertsekas, 2013; Guin & Bhatnagar, 2025) and is natural in many applications. For example, LLM generation ends when an EOS token is selected or a prescribed token limit is reached (Ramamurthy et al., 2022). By including the remaining token budget in the state and treating budget exhaustion as termination, every policy is guaranteed to terminate. This assumption is also satisfied in games whose rules prevent indefinite play. For example, in Go under the American Go Association (AGA) rules, the superko rule prohibits stone placements that recreate an earlier board position with the same player to move. Since there are finitely many such positions, this restriction, together with the rules for ending the game through consecutive passes, ensures termination under every legal policy (American Go Association, 1991). Chess also has similar rules.

To the best of our knowledge, previous studies (Bojun, 2020; Borrell et al., 2025) have only established formulas for occupancy measures and policy gradients in this setting. In addition, (Guin & Bhatnagar, 2025) propose a policy gradient algorithm with only asymptotic convergence guarantees. However, the geometry of occupancy measures and the non-asymptotic convergence properties of policy optimization in this setting remain unclear.

In this work, we close this theoretical gap. Our contributions can be summarized as follows:

- We characterize the geometry of occupancy measures for MDPs with terminal states. For a fixed initial distribution, the feasible measures form a compact convex polytope, and both expected return and expected termination time are linear functionals of these measures.
- We provide an explicit construction of a stationary policy realizing any convex combination of feasible occupancy measures.
- We establish that the K-fold composition of the undiscounted Bellman operator is a contraction for sufficiently large K.
- We formulate and analyze trust-region policy optimization (TRPO) in this setting and establish sufficient conditions for monotonic improvement of trust-region updates.
- We derive natural policy gradient (NPG) updates in this setting and establish their convergence rates.

We would like to highlight that although the results take a form similar to that for infinite-horizon discounted MDPs, the technical analysis differs substantially because the horizon is a random variable whose distribution depends on the policy.

2 Preliminaries

Notation. For a finite nonempty set $\mathcal{X}$, let $\Delta_{\mathcal{X}} := \{\mu \in \mathbb{R}^{\mathcal{X}}_{\geq 0} \mid \sum_{x\in\mathcal{X}} \mu(x) = 1\}$ denote the probability simplex over $\mathcal{X}$. We will write μ_x and $\mu(x)$ interchangeably in this paper. For finite nonempty sets $\mathcal{X}$ and $\mathcal{Y}$, let $\Delta^{\mathcal{X}}_{\mathcal{Y}}$ denote the set of Markov kernels from $\mathcal{X}$ to $\mathcal{Y}$. Throughout this paper, we identify functions and measures on finite sets with vectors of their values. Specifically, for a finite set $\mathcal{I}$, a function $f : \mathcal{I} \to \mathbb{R}$ is represented by the vector $f = (f(i))_{i\in\mathcal{I}}$, whose entry indexed by i is $f(i)$. For a function $g : \mathcal{I} \times \mathcal{J} \to \mathbb{R}$, we write $g = (g(i,j))_{(i,j)\in\mathcal{I}\times\mathcal{J}}, g(i) := (g(i,j))_{j\in\mathcal{J}}$, where $g(i)$ collects the values obtained by fixing the first index and varying the second. The same convention applies to measures: $\nu = (\nu(i))_{i\in\mathcal{I}}$ denotes the vector of masses assigned to the elements of $\mathcal{I}$.

Markov decision processes. A finite Markov decision process (MDP) is specified by a tuple $(\mathcal{S}, \mathcal{A}, \alpha, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite nonempty state and action spaces, respectively. The transition kernel $\alpha \in \Delta^{\mathcal{S}\times\mathcal{A}}_{\mathcal{S}}$ specifies the probability $\alpha(s' \mid s, a)$ of transitioning to state s' after taking action a in state s. The reward function $r \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$ assigns an instantaneous reward $r(s,a)$ to each state–action pair. We assume that there exists a constant $R_{\max} < \infty$ such that $|r(s,a)| \leq R_{\max}$ for all $(s,a) \in \mathcal{X} \times \mathcal{A}$. The set $\mathcal{X}$ represents all the non-terminal states as we will define later.

A stationary randomized policy is a Markov kernel $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$, where $\pi(a \mid s)$ denotes the probability of selecting action a in state s. Given a policy π, the induced transition kernel of the state process, $p_\pi \in \Delta^{\mathcal{S}}_{\mathcal{S}}$, is obtained by averaging over the action selected under π: $p_\pi(s' \mid s) := \sum_{a\in\mathcal{A}} \pi(a \mid s)\, \alpha(s' \mid s, a)$. We define $p^{\mathcal{X}}_\pi := \left[p_\pi(s' \mid s)\right]_{s,s'\in\mathcal{X}}$.

Given an initial state distribution $\mu_0 \in \Delta_{\mathcal{S}}$ and a stationary policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, let $\zeta_{\pi,\mu_0} = (s_t, a_t)_{t\geq 0}$ denote a trajectory generated according to $s_0 \sim \mu_0, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \alpha(\cdot \mid s_t, a_t), t \geq 0$. We further let $\zeta_{\pi,s}$ denote a trajectory generated according to $s_0 = s, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \alpha(\cdot \mid s_t, a_t), t \geq 0, s \in \mathcal{X}$, and $\zeta_{\pi,(s,a)}$ denote a trajectory generated according to $s_0 = s, a_0 = a, a_{t+1} \sim \pi(\cdot \mid s_{t+1}), s_{t+1} \sim \alpha(\cdot \mid s_t, a_t), t \geq 0, (s, a) \in \mathcal{X} \times \mathcal{A}$.

Markov decision processes with terminal states. Let $\mathcal{S}_\perp \subseteq \mathcal{S}$ denote the set of terminal states. Throughout this paper, we consider Markov decision processes whose trajectories terminate upon their first visit to $\mathcal{S}_\perp$. We assume that no rewards are collected upon reaching terminal states. Thus, all terminal states in $\mathcal{S}_\perp$ can be merged into a single absorbing state $s_\perp$. Let $\mathcal{X} := \mathcal{S} \setminus \mathcal{S}_\perp$ denote the set of non-terminal states. Under Assumption 1, every state $\mathcal{X}$ is transient because the process reaches $\mathcal{S}_\perp$ almost surely. Therefore, we will only consider initial distributions $\mu_0 \in \Delta_{\mathcal{X}}$ and stationary randomized policies $\pi \in \Delta_{\mathcal{A}}^{\mathcal{X}}$ in this paper.

For a trajectory ζ_{π,μ_0}, define the termination time as the first visit to $\mathcal{S}_\perp$ after time zero:

$$T(\zeta_{\pi,\mu_0}) := \inf\{t \geq 1 : s_t \in \mathcal{S}_\perp\}.$$

We define the expected cumulative reward under policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{X}}$ and initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$ as $V^\pi(\mu_0) := \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,\mu_0})-1} r(s_t, a_t)\right]$. For trajectories starting from a state $s \in \mathcal{X}$, the corresponding value is $V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,s})-1} r(s_t, a_t)\right]$. Therefore, for any initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$, $V^\pi(\mu_0) = \sum_{s\in\mathcal{X}} \mu_0(s) V^\pi(s)$.

As in (Bertsekas & Tsitsiklis, 1991; Yu & Bertsekas, 2013; Guin & Bhatnagar, 2025), we also assume that every stationary policy is proper, meaning that trajectories terminate almost surely from every initial state. As illustrated by the examples of LLM generation and Go in the introduction, this assumption is reasonable in many practical scenarios.

Assumption 1. *For every policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{X}}$ and every initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$, the induced trajectory ζ_{π,μ_0} terminates in finite time almost surely:*

$$\mathbb{P}\big(T(\zeta_{\pi,\mu_0}) < \infty\big) = 1.$$

Under Assumption 1, the following lemma establishes that the expected trajectory length is finite, which in turn ensures that the expected cumulative reward is finite.

Lemma 1. *Suppose that Assumption 1 holds. Then there exists a constant $T_{\max} < \infty$ such that, for every stationary policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{X}}$ and every initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$,*

$$\mathbb{E}[T(\zeta_{\pi,\mu_0})] \leq T_{\max}.$$

We provide the proof in Appendix B.

3 Policy evaluation and the K-fold Bellman operator

In this section, we first derive the Q-function and the Bellman operator for MDPs with terminal states, and then show that the K-fold composition of this operator is a contraction for sufficiently large K.

Definition 1. *Let $\mathcal{Q} := \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be the space of real-valued functions on $\mathcal{X} \times \mathcal{A}$. The action-value function $Q^\pi \in \mathcal{Q}$ is defined by*

$$Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,(s,a)})-1} r(s_t, a_t)\right], \qquad (s,a) \in \mathcal{X} \times \mathcal{A},$$

where the expectation is taken over trajectories initialized at $(s_0, a_0) = (s, a)$, with transitions governed by α and subsequent actions selected according to π.

The Bellman operator $\mathcal{F}^\pi : \mathcal{Q} \to \mathcal{Q}$ is defined, for every $Q \in \mathcal{Q}$, by

$$(\mathcal{F}^\pi Q)(s,a) := r(s,a) + \sum_{s'\in\mathcal{X}} \alpha(s' \mid s,a) \sum_{a'\in\mathcal{A}} \pi(a' \mid s') Q(s',a'), \qquad (s,a) \in \mathcal{X} \times \mathcal{A}.$$

Since r is bounded and the expected termination time is finite, Q^π is well defined. Conditioning on the first transition decomposes the return into the immediate reward and the expected remaining return:

$$Q^\pi(s,a) = r(s,a) + \sum_{s' \in \mathcal{X}} \alpha(s' \mid s,a) \sum_{a' \in \mathcal{A}} \pi(a' \mid s') Q^\pi(s',a') = (\mathcal{F}^\pi Q^\pi)(s,a).$$

Thus, Q^π is a fixed point of the Bellman operator: $Q^\pi = \mathcal{F}^\pi Q^\pi$.

(Bertsekas & Tsitsiklis, 1991) established a contraction property of $\mathcal{F}^\pi$ under a weighted ℓ_∞ norm rather than the standard ℓ_∞ norm. In the following theorem, we show that the K-fold composition of the Bellman operator for MDPs with terminal states is a contraction under the standard ℓ_∞ norm for sufficiently large K.

Theorem 1. *Suppose that Assumption 1 holds, and fix $\pi \in \Delta_{\mathcal{A}}^{\mathcal{X}}$. Define the supremum norm as $\|Q\|_\infty := \max_{(s,a)\in\mathcal{X}\times\mathcal{A}} |Q(s,a)|$. For every $Q \in \mathcal{Q}$, define the k-fold composition of the Bellman operator by*

$$(\mathcal{F}^\pi)^0 Q := Q, \qquad (\mathcal{F}^\pi)^{k+1} Q := \mathcal{F}^\pi\big((\mathcal{F}^\pi)^k Q\big), \quad k \geq 0.$$

Let $T^\pi_{max} := \max_{(s,a)\in\mathcal{X}\times\mathcal{A}} \mathbb{E}\left[T(\zeta_{\pi,(s,a)})\right] < \infty$. For every integer $K > T^\pi_{max}$, the operator $(\mathcal{F}^\pi)^K$ is a contraction on $(\mathcal{Q}, \|\cdot\|_\infty)$:

$$\left\|(\mathcal{F}^\pi)^K Q_1 - (\mathcal{F}^\pi)^K Q_2\right\|_\infty \leq \beta_{\pi,K} \|Q_1 - Q_2\|_\infty, \qquad Q_1, Q_2 \in \mathcal{Q},$$

where

$$\beta_{\pi,K} := \max_{(s,a)\in\mathcal{X}\times\mathcal{A}} \mathbb{P}\left(T(\zeta_{\pi,(s,a)}) > K\right) \leq \frac{T^\pi_{max}}{K} < 1.$$

The proof can be found in Appendix C.

4 Occupancy measure and geometry

Occupancy measures (Kakade & Langford, 2002; Altman, 2021; Paternain et al., 2019; Müller & Montúfar, 2024) are important tools for establishing the convergence of various policy gradient methods. Although previous studies (Bojun, 2020; Borrell et al., 2025) have derived their form for MDPs with terminal states, their geometry has not been studied. In this section, we first present the formulation of occupancy measures developed by (Borrell et al., 2025) and then study their geometry.

4.1 Occupancy measure of MDPs with terminal states.

For a trajectory $\zeta_{\pi,\mu_0} = (s_t, a_t)_{t\geq 0}$, define $\rho^t_{\pi,\mu_0}(s) := \mathbb{P}(s_t = s, t < T(\zeta_{\pi,\mu_0}))$, $s \in \mathcal{X}$. The corresponding unnormalized occupancy measure is

$$\eta_{\pi,\mu_0}(s) := \sum_{t=0}^{\infty} \rho^t_{\pi,\mu_0}(s) = \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,\mu_0})-1} \mathbb{1}\{s_t = s\}\right], \quad s \in \mathcal{X}.$$

Its normalizing constant is $Z_{\pi,\mu_0} := \sum_{s\in\mathcal{X}} \eta_{\pi,\mu_0}(s) = \mathbb{E}[T(\zeta_{\pi,\mu_0})]$. Thus, the normalized occupancy distribution is $\bar{\rho}_{\pi,\mu_0}(s) := \frac{\eta_{\pi,\mu_0}(s)}{Z_{\pi,\mu_0}}$.

For state-action pairs (s,a), define

$$\eta_{\pi,\mu_0}(s,a) := \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,\mu_0})-1} \mathbb{1}\{s_t = s, a_t = a\}\right] = \eta_{\pi,\mu_0}(s)\pi(a \mid s),$$

and

$$\bar{\rho}_{\pi,\mu_0}(s,a) := \frac{\eta_{\pi,\mu_0}(s,a)}{Z_{\pi,\mu_0}} = \bar{\rho}_{\pi,\mu_0}(s)\pi(a \mid s).$$

Therefore, for any state-action reward r, the expected random-horizon return can be written as

$$V^\pi(\mu_0) = \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,\mu_0})-1} r(s_t, a_t)\right] = \sum_{s\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta_{\pi,\mu_0}(s,a) r(s,a) = Z_{\pi,\mu_0} \sum_{s\in\mathcal{X}}\sum_{a\in\mathcal{A}} \bar{\rho}_{\pi,\mu_0}(s,a) r(s,a).$$

For a fixed reward function r, the expected episodic return is a linear functional of the unnormalized occupancy measure:

$$V^\pi(\mu_0) = \langle \eta_{\pi,\mu_0}, r\rangle.$$

Consequently, $|V^\pi(s)| \le R_{\max} T_{max}$.

Thus, optimizing the expected return can be formulated as maximizing a linear functional over the set of occupancy measures. In next subsection, we will characterize the geometry of this set.

4.2 Convexity of the set of occupancy measure

The occupancy measure is not arbitrary, we first show it must satisfy a set of linear flow balance equations in the following lemma.

Lemma 2. *Suppose that Assumption 1 holds. Fix an initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$. A vector $\nu \in \mathbb{R}_{\ge 0}^{\mathcal{X}\times\mathcal{A}}$ is the non-terminal occupancy measure induced by some stationary policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{X}}$, i.e.,*

$$\nu(s,a) = \mathbb{E}\left[\sum_{t=0}^{T(\zeta_{\pi,\mu_0})-1} \mathbb{1}\{s_t = s, a_t = a\}\right],$$

if and only if

$$\nu(s) := \sum_{a\in\mathcal{A}} \nu(s,a) = \mu_0(s) + \sum_{s'\in\mathcal{S}\setminus\mathcal{S}_\perp}\sum_{a'\in\mathcal{A}} \nu(s',a')\alpha(s \mid s',a'), \qquad s\in\mathcal{X}. \tag{1}$$

The proof is provided in Appendix D.1.

We now define the feasible set of all valid occupancy measures under the bounded hitting-time assumption.

Definition 2. *Fix an initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$, the feasible set of non-terminal occupancy measures is*

$$\mathcal{N}_{\mu_0} := \left\{ \nu \in \mathbb{R}_{\ge 0}^{\mathcal{X}\times\mathcal{A}} \;\middle|\; \nu \text{ satisfies Eq. 1}, \sum_{s\in\mathcal{X}}\sum_{a\in\mathcal{A}} \nu(s,a) \le T_{max} \right\}. \tag{2}$$

We now ready to characterize the geometry of the set of occupancy measure in next theorem.

Theorem 2. *For a fixed initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$ and a finite bound T_{max}, the feasible occupancy set $\mathcal{N}_{\mu_0}$ defined in Eq. 2 is a compact convex polytope.*

We provide the proof in Appendix D.2.

In the following corollary, we provide a method for reconstructing a policy from an occupancy measure.

Corollary 1. *Suppose that Assumption 1 holds, and fix $\mu_0 \in \Delta_{\mathcal{X}}$. Let $\nu_1, \nu_2 \in \mathcal{N}_{\mu_0}$ and $\lambda \in [0,1]$. Define*

$$\nu_\lambda := \lambda\nu_1 + (1-\lambda)\nu_2, \qquad \nu_\lambda(s) := \sum_{a\in\mathcal{A}} \nu_\lambda(s,a).$$

For arbitrary distributions $q_s \in \Delta_{\mathcal{A}}$, define a stationary policy on $\mathcal{X}$ by

$$\pi_{\nu_\lambda}(a \mid s) := \begin{cases} \dfrac{\nu_\lambda(s,a)}{\nu_\lambda(s)}, & \nu_\lambda(s) > 0, \\ q_s(a), & \nu_\lambda(s) = 0. \end{cases}$$

Then,

$$\eta_{\pi_{\nu_\lambda},\mu_0}(s,a) = \nu_\lambda(s,a), \qquad (s,a) \in \mathcal{X}\times\mathcal{A}.$$

The proof can be found in Appendix D.3.

Remark 1. *Mixing policies does not generally produce the same mixture of their occupancy measures. Specifically, for $\pi_\lambda = \lambda\pi_1 + (1-\lambda)\pi_2$, we generally have*

$$\eta_{\pi_\lambda,\mu_0} \neq \lambda\eta_{\pi_1,\mu_0} + (1-\lambda)\eta_{\pi_2,\mu_0}.$$

To realize the occupancy mixture in Corollary 1, the policy must account for how often each policy visits each state. Writing $\nu_i(s) := \eta_{\pi_i,\mu_0}(s)$, the required policy is

$$\pi_{\nu_\lambda}(a \mid s) = \frac{\lambda\nu_1(s)\pi_1(a \mid s) + (1-\lambda)\nu_2(s)\pi_2(a \mid s)}{\lambda\nu_1(s) + (1-\lambda)\nu_2(s)},$$

whenever the denominator is positive, and may be chosen arbitrarily otherwise. Thus, the mixture weights depend on the state, rather than being fixed at λ and $1-\lambda$.

We now establish the linearity property of the expected hitting time under these linear combination of occupancy measure.

Corollary 2. *Suppose that Assumption 1 holds, and fix $\mu_0 \in \Delta_\mathcal{X}$. Let $\nu_1, \nu_2 \in \mathcal{N}_{\mu_0}$ and $\lambda \in [0,1]$, and define*

$$\nu_\lambda := \lambda\nu_1 + (1-\lambda)\nu_2.$$

Let π_{ν_1}, π_{ν_2}, and π_{ν_λ} be the stationary policies obtained by conditioning the corresponding occupancy measures as in Corollary 1. Then

$$\mathbb{E}\left[T(\zeta_{\pi_{\nu_\lambda},\mu_0})\right] = \lambda\mathbb{E}\left[T(\zeta_{\pi_{\nu_1},\mu_0})\right] + (1-\lambda)\mathbb{E}\left[T(\zeta_{\pi_{\nu_2},\mu_0})\right]. \tag{3}$$

The proof is in Appendix D.4.

5 Trust-region policy optimization for MDPs with terminal states

Trust-region policy optimization (TRPO) is a useful method for improving expected return by maximizing a surrogate objective while constraining policy changes (Schulman et al., 2015). In this section, we formulate TRPO for MDPs with terminal states and establish sufficient conditions for monotonic improvement of trust-region updates.

The performance difference lemma (Schulman et al., 2015) is central to the derivation of TRPO. Here, we establish its form for MDPs with terminal states.

Lemma 3 (Performance difference lemma)**.** *Define the advantage function as $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s), (s,a) \in \mathcal{X}\times\mathcal{A}$.*

Suppose that Assumption 1 holds. For any two stationary policies $\pi, \pi' \in \Delta_\mathcal{A}^\mathcal{X}$ and any $s_0 \in \mathcal{X}$,

$$V^{\pi'}(s_0) - V^\pi(s_0) = \mathbb{E}_{\pi',s_0}\left[\sum_{t=0}^{T(\zeta_{\pi',s_0})-1} A^\pi(s_t,a_t)\right], \tag{4}$$

where $\mathbb{E}_{\pi',s_0}$ denotes expectation under the trajectory law induced by π' with initial state s_0.

If we define

$$\bar{A}_\pi^{\pi'}(s) := \sum_{a\in\mathcal{A}} \pi'(a \mid s)A^\pi(s,a), \qquad s \in \mathcal{X}.$$

Then, for any $\mu_0 \in \Delta_\mathcal{X}$,

$$V^{\pi'}(\mu_0) - V^\pi(\mu_0) = \sum_{s\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta_{\pi',\mu_0}(s,a)A^\pi(s,a) = \sum_{s\in\mathcal{X}} \eta_{\pi',\mu_0}(s)\bar{A}_\pi^{\pi'}(s). \tag{5}$$

The proof is given in Appendix E.

The performance difference lemma (Lemma 3) relates the return of a candidate policy π' to that of the current policy π. Trajectories generated under π can be used to estimate its value and advantage functions at the states and actions visited. However, evaluating the return of π' through this

identity also requires the occupancy measure η_{π',μ_0}, which is not directly observed from trajectories generated under π.

Trust-region policy optimization (TRPO) addresses this difficulty by replacing the candidate occupancy measure η_{π',μ_0} with the current occupancy measure η_{π,μ_0}, yielding the local surrogate

$$V^{\pi'}(\mu_0) \approx \mathcal{L}_{\pi,\mu_0}(\pi') := V^{\pi}(\mu_0) + \sum_{s\in\mathcal{X}} \eta_{\pi,\mu_0}(s)\bar{A}_{\pi}^{\pi'}(s). \tag{6}$$

The following lemma establishes that, for MDPs with terminal states, the error in approximating $V^{\pi'}(\mu_0)$ by $\mathcal{L}_{\pi,\mu_0}(\pi')$ is bounded by a term quadratic in the distance between π and π'.

Lemma 4. *Suppose that Assumption 1 holds. Define*

$$\epsilon_\pi := \max_{(s,a)\in\mathcal{X}\times\mathcal{A}} |A^{\pi}(s,a)|, \qquad B_{\pi,\mu_0} := 4\epsilon_\pi Z_{\pi,\mu_0} T_{max}.$$

Define the maximum statewise total variation distance by

$$D_{\mathrm{TV}}^{\max}(\pi,\pi') := \max_{s\in\mathcal{X}} \frac{1}{2}\sum_{a\in\mathcal{A}} |\pi'(a \mid s) - \pi(a \mid s)|,$$

Then, for any policy π',

$$\left|V^{\pi'}(\mu_0) - \mathcal{L}_{\pi,\mu_0}(\pi')\right| \le B_{\pi,\mu_0}\left[D_{\mathrm{TV}}^{\max}(\pi,\pi')\right]^2. \tag{7}$$

Define the maximum statewise KL divergence by

$$D_{\mathrm{KL}}^{\max}(\pi,\pi') := \max_{s\in\mathcal{X}} \sum_{a\in\mathcal{A}} \pi(a \mid s)\log\frac{\pi(a \mid s)}{\pi'(a \mid s)}.$$

When $D_{\mathrm{KL}}^{\max}(\pi,\pi') < \infty$, we can have

$$V^{\pi'}(\mu_0) \ge \mathcal{L}_{\pi,\mu_0}(\pi') - \frac{B_{\pi,\mu_0}}{2} D_{\mathrm{KL}}^{\max}(\pi,\pi'). \tag{8}$$

We provide the proof in Appendix F.1.

The preceding lemma establishes that the surrogate provides a first-order approximation to the true return at π. This motivates maximizing the surrogate subject to a trust-region constraint. At iteration k, given the current policy π_k and a trust-region radius $\delta_k > 0$, we consider the optimization problem

$$\begin{aligned} \widetilde{\pi}_k \in \arg\max_{\pi'\in\Delta_{\mathcal{A}}^{\mathcal{X}}} \quad & \mathcal{L}_{\pi_k,\mu_0}(\pi') \\ \text{subject to} \quad & D_{\mathrm{KL}}^{\max}(\pi_k,\pi') \le \delta_k. \end{aligned} \tag{9}$$

The feasible set is compact and nonempty, since it contains π_k. As the objective is continuous, a maximizer exists.

By Lemma 4, the solution to Eq. 9 satisfies

$$V^{\widetilde{\pi}_k}(\mu_0) - V^{\pi_k}(\mu_0) \ge \mathcal{L}_{\pi_k,\mu_0}(\widetilde{\pi}_k) - V^{\pi_k}(\mu_0) - B_{\pi_k,\mu_0}\left[D_{\mathrm{TV}}^{\max}(\pi_k,\widetilde{\pi}_k)\right]^2. \tag{10}$$

The right-hand side is not necessarily nonnegative. To ensure monotonic improvement, we therefore consider a combination of the candidate policy $\widetilde{\pi}_k$ and the current policy π_k in next theorem.

Theorem 3. *Suppose that Assumption 1 holds. Let $\widetilde{\pi}_k$ be any proposal satisfying*

$$D_{\mathrm{KL}}^{\max}(\pi_k,\widetilde{\pi}_k) \le \delta_k,$$

and define

$$G_k := \mathcal{L}_{\pi_k,\mu_0}(\widetilde{\pi}_k) - V^{\pi_k}(\mu_0), \qquad d_k := D_{\mathrm{TV}}^{\max}(\pi_k,\widetilde{\pi}_k).$$

If $G_k \le 0$, set $\pi_{k+1} := \pi_k$. If $G_k > 0$, define

$$\beta_k := \min\left\{1,\ \frac{G_k}{2B_{\pi_k,\mu_0} d_k^2}\right\}, \tag{11}$$

and update

$$\pi_{k+1}(a \mid s) := (1-\beta_k)\pi_k(a \mid s) + \beta_k \widetilde{\pi}_k(a \mid s), \quad (s,a) \in \mathcal{X} \times \mathcal{A}. \tag{12}$$

Then the updated policy remains in the trust region and $V^{\pi_{k+1}}(\mu_0) \ge V^{\pi_k}(\mu_0)$.

We provide the proof in Appendix F.2.

However, the monotonic improvement alone does not imply convergence to a globally optimal policy, we will provide an algorithm with global convergence in next section.

6 Natural policy gradient for MDPs with terminal states

The natural policy gradient (NPG) preconditions the policy gradient with the pseudoinverse Fisher information matrix to account for the geometry of the policy-space. In this section, we derive NPG for MDPs with terminal states under the undiscounted cumulative reward objective and establish its convergence rate under a softmax parameterization.

We consider a twice continuously differentiable policy family π_θ, with θ in an open subset of $\mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$, such that $\pi_\theta(a \mid s) > 0$ for every $(s,a) \in \mathcal{X} \times \mathcal{A}$.

We first derive the gradient of the value function with respect to parameter θ.

Theorem 4. *The gradient of the expected return satisfies*

$$\begin{aligned} \nabla_\theta V^{\pi_\theta}(\mu_0) &= \sum_{s\in\mathcal{X}} \sum_{a\in\mathcal{A}} \eta_{\pi_\theta,\mu_0}(s,a) A^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a \mid s) \\ &= Z_{\pi_\theta,\mu_0} \mathbb{E}_{(s,a)\sim\bar{\rho}_{\pi_\theta,\mu_0}} \left[A^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a \mid s)\right]. \end{aligned} \tag{13}$$

Similar results were also given by (Bojun, 2020) and (Borrell et al., 2025) , we provide a different proof in Appendix G.1.

We present the closed-form NPG update below; the full derivation is provided in Section G.2.

Consider the tabular softmax parameterization

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{b\in\mathcal{A}} \exp(\theta_{s,b})}.$$

Define the Fisher information matrix as

$$F(\theta) := \mathbb{E}_{(s,a)\sim\bar{\rho}_{\pi_\theta,\mu_0}} \left[\nabla_\theta \log \pi_\theta(a \mid s) \big(\nabla_\theta \log \pi_\theta(a \mid s)\big)^\top\right].$$

Writing $g_\theta := \nabla_\theta V^{\pi_\theta}(\mu_0)$, the NPG parameter update is

$$\theta_{k+1} = \theta_k + \beta_k F(\theta_k)^\dagger g_{\theta_k},$$

where † denotes the Moore–Penrose pseudoinverse.

Let $\pi_k := \pi_{\theta_k}$ and choose $\beta_k = h/Z_{\pi_k,\mu_0}$ for a fixed step size $h > 0$. At states reachable from μ_0, the induced policy update takes the form

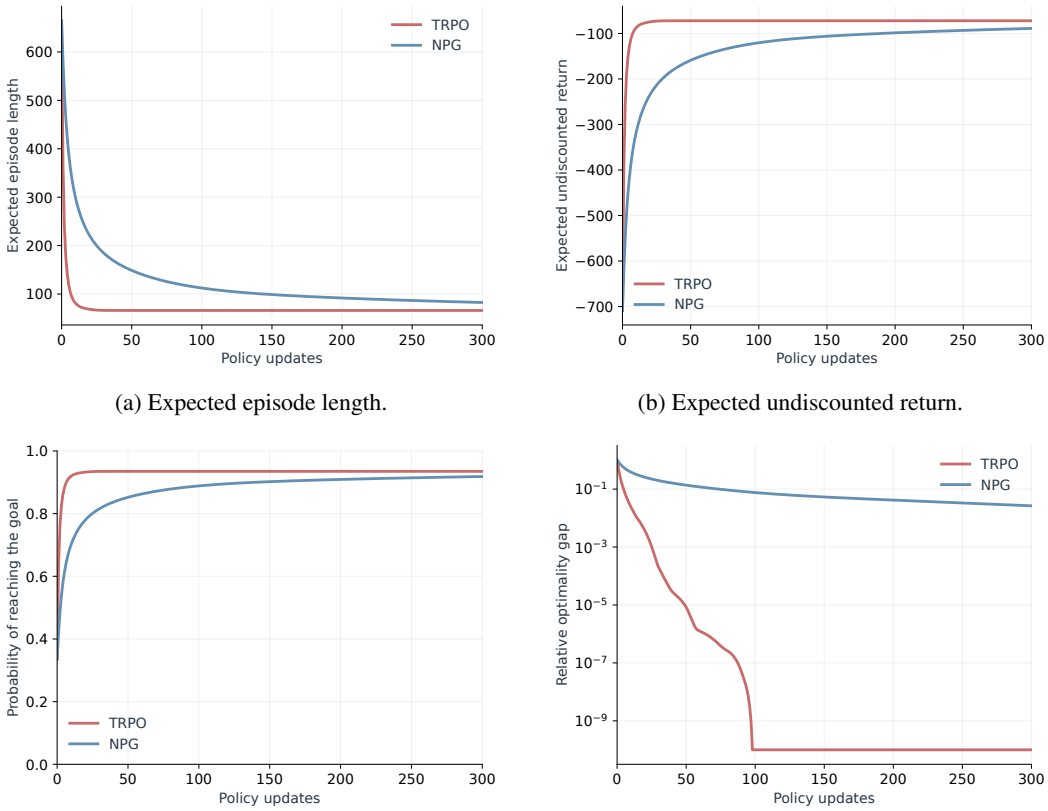
$$\pi_{k+1}(a \mid s) = \frac{\pi_k(a \mid s)\exp\big(hA^{\pi_k}(s,a)\big)}{\Lambda_k(s)}, \qquad \Lambda_k(s) := \sum_{b\in\mathcal{A}} \pi_k(b \mid s)\exp\big(hA^{\pi_k}(s,b)\big). \tag{14}$$

We adapt the discounted NPG analysis of (Agarwal et al., 2021, Theorem 16) using the unnormalized occupancy measure and the uniform termination-time bound from Lemma 1.

Theorem 5. *Suppose that Assumption 1 holds, and let π_k follow Eq. 14 with exact advantages, a fixed $h > 0$, and uniform initialization $\pi_0(a \mid s) = 1/|\mathcal{A}|$. For an optimal stationary policy $\pi^\star$ and every integer $K \geq 1$,*

$$V^{\pi^\star}(\mu_0) - V^{\pi_K}(\mu_0) \leq \frac{T_{max} \log |\mathcal{A}|}{hK} + \frac{2R_{\max} T_{max}^2}{K}. \tag{15}$$

Thus, the NPG for MDPs with terminal states achieves an $\mathcal{F}(K^{-1})$ convergence rate in expected return.



(a) Expected episode length.

(b) Expected undiscounted return.

(c) Probability of reaching the goal.

(d) Relative optimality gap.

Figure 1: Comparison of TRPO (red) and NPG (blue) on discrete mountain car with policy-dependent trajectory lengths. Subfigures 1a, 1b, 1c, and 1d show expected trajectory length, expected undiscounted return, goal-reaching probability, and relative optimality gap, respectively, as functions of the number of policy updates. In Figure 1d, the relative optimality gap is plotted on a logarithmic scale, with values below 10^{-10} clipped for visualization. Trajectory end upon reaching the goal or external termination, with external termination counted as failure.

7 Simulation

We evaluate TRPO and NPG in a discretized mountain car environment, where episodes end upon reaching the goal or through stochastic external termination. As shown in Figure 1, both methods improve expected return and goal-reaching probability while reducing expected episode length. TRPO achieves faster improvement per policy update than NPG in this experiment. The detail of the simulation environment can be found in section H.

AI USE STATEMENT

In this work, we used generative AI tools for literature retrieval and discovery, including finding related work, and to aid and polish writing by improving grammar, clarity, and readability. We have not used generative AI tools for research ideation or execution, drafting sections of the paper, or proving mathematical claims, and tasks involving synthetic dataset generation are not applicable to this work. We have reviewed all AI-assisted work. Our review covered AI-assisted language revisions and suggestions for related literature. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

American Go Association. AGA rules of Go, 1991. URL https://britgo.org/rules/agarules.html.

Dimitri P Bertsekas. *Reinforcement learning and optimal control*, volume 1. Athena Scientific Belmont, MA, 2019.

Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991. doi: 10.1287/moor.16.3.580.

Huang Bojun. Steady state analysis of episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9335–9345, 2020.

Enric Ribera Borrell, Lorenz Richter, and Christof Schütte. Reinforcement learning with random time horizons. *arXiv preprint arXiv:2506.00962*, 2025.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.14165.

Liyu Chen, Haipeng Luo, and Aviv Rosenberg. Policy optimization for stochastic shortest path. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 982–1046, 2022. URL https://proceedings.mlr.press/v178/chen22a.html.

Suei-Wen Chen, Keith Ross, and Pierre Youssef. Finite-sample analysis of the Monte Carlo Exploring Starts algorithm for reinforcement learning, 2024. URL https://arxiv.org/abs/2410.02994.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2103.13056.

Hugo Cruz-Suárez, Rocio Ilhuicatzi-Roldán, and Raúl Montes-de Oca. Markov decision processes on Borel spaces with total cost and random horizon. *Journal of Optimization Theory and Applications*, 162(1):329–346, 2014. doi: 10.1007/s10957-012-0262-8. URL https://doi.org/10.1007/s10957-012-0262-8.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Soumyajit Guin and Shalabh Bhatnagar. Convergent reinforcement learning algorithms for stochastic shortest path problem. *arXiv preprint arXiv:2508.13963*, 2025.

Tetsuo Iida and Masao Mori. Markov decision processes with random horizon. *Journal of the Operations Research Society of Japan*, 39(4):592–603, 1996. doi: 10.15807/jorsj.39.592. URL https://doi.org/10.15807/jorsj.39.592.

Rocio Ilhuicatzi-Roldán, Hugo Cruz-Suárez, and Selene Chávez-Rodríguez. Markov decision processes with time-varying discount factors and random horizon. *Kybernetika*, 53(1): 82–98, 2017. doi: 10.14736/kyb-2017-1-0082. URL https://doi.org/10.14736/kyb-2017-1-0082.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Jongmin Lee and Ernest K. Ryu. Why policy gradient algorithms work for undiscounted total-reward MDPs, 2025. URL https://arxiv.org/abs/2510.18340v2. Version 2, revised February 2026.

Debmalya Mandal, Goran Radanović, Jiarui Gan, Adish Singla, and Rupak Majumdar. Online reinforcement learning with uncertain episode lengths. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):9064–9071, 2023. doi: 10.1609/aaai.v37i7.26088. URL https://ojs.aaai.org/index.php/AAAI/article/view/26088.

Johannes Müller and Guido Montúfar. Geometry and convergence of natural policy gradient methods: J. müller, g. montúfar. *Information Geometry*, 7(Suppl 1):485–523, 2024.

Phalguni Nanda and Zaiwei Chen. Natural policy gradient as doubly smoothed policy iteration: A bellman-operator framework, 2026. URL https://arxiv.org/abs/2605.10671.

Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.02155.

Lina Palmborg. Convergence of SARSA with linear function approximation: The random horizon case. *Journal of Experimental & Theoretical Artificial Intelligence*, 38(1):137–153, 2026. doi: 10.1080/0952813X.2025.2541954. URL https://doi.org/10.1080/0952813X.2025.2541954.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8210–8219, 2020. URL https://proceedings.mlr.press/v119/rosenberg20a.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Jean Tarbouriech, Runlong Zhou, Simon S. Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2104.11186.

Huizhen Yu and Dimitri P Bertsekas. Q-learning and policy iteration algorithms for stochastic shortest path problems. *Annals of Operations Research*, 208(1):95–132, 2013.