Interpreting Primal-Dual Algorithms for Constrained Multiagent Reinforcement Learning

Preprint

Daniel Tabas,[1] Ahmed S. Zamzam,[2] and Baosen Zhang[1]

1 University of Washington
2 National Renewable Energy Laboratory

Presented at the 5th Annual Learning for Dynamics and Control Conference
Philadelphia, Pennsylvania
June 14–16, 2023

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-84649
September 2023

Interpreting Primal-Dual Algorithms for Constrained Multiagent Reinforcement Learning

Preprint

Daniel Tabas,[1] Ahmed S. Zamzam,[2] and Baosen Zhang[1]

1 University of Washington
2 National Renewable Energy Laboratory

Suggested Citation
Tabas, Daniel, Ahmed S. Zamzam, and Baosen Zhang. 2023. *Interpreting Primal-Dual Algorithms for Constrained Multiagent Reinforcement Learning: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5D00-84649. https://www.nrel.gov/docs/fy23osti/84649.pdf.

Conference Paper
NREL/CP-5D00-84649
September 2023

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported by the Laboratory Directed Research and Development (LDRD) Program at NREL. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

Interpreting Primal-Dual Algorithms for Constrained MARL

Daniel Tabas DTABAS@UW.EDU
Baosen Zhang ZHANGBAO@UW.EDU
University of Washington Department of Electrical Engineering, Seattle, WA, USA

Ahmed Zamzam AHMED.ZAMZAM@NREL.GOV
National Renewable Energy Laboratory, Golden, CO, USA

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

We study multiagent reinforcement learning (MARL) with constraints. This setting is gaining importance as MARL algorithms find new applications in real-world systems ranging from power grids to drone swarms. Most constrained MARL (C-MARL) algorithms use a primal-dual approach to enforce constraints through a penalty function added to the reward. In this paper, we study the structural effects of the primal-dual approach on the constraints and value function. First, we show that using the constraint evaluation as the penalty leads to a weak notion of safety, but by making simple modifications to the penalty function, we can enforce meaningful probabilistic safety constraints. Second, we show that the penalty term changes the value function in a way that is easy to model, and demonstrate the consequences of not doing so. We conclude with simulations in a simple constrained multiagent environment to back up the theoretical results.

Keywords: Multi-agent reinforcement learning, primal-dual methods, chance constraints, conditional value-at-risk

1. Introduction

As reinforcement learning (RL) algorithms progress from virtual to cyber-physical applications, it will be necessary to address the challenges of safety especially when systems are controlled by multiple agents. Examples of multi-agent safety-critical systems include power grids Cui et al. (2022), building energy management systems Biagioni et al. (2022), autonomous vehicle navigation Zhou et al. (2022), and drone swarms Chen et al. (2020). In each of these applications, agents must learn to operate in a complicated environment, while satisfying various local and system-wide constraints. Such constraints, derived from domain-specific k nowledge, a re d esigned t o prevent damage to equipment, humans, or infrastructure, or to preclude failure to complete some task or objective.

Constrained multiagent reinforcement learning (C-MARL) poses many challenges above and beyond the single-agent problem (C-RL) because the interaction between agents can influence both the satisfaction of constraints and the convergence to optimal policies. The potential scale of C-MARL problems eliminates the possibility of directly using common model-based methods for C-RL such as e.g. Chen et al. (2018); Ma et al. (2021); Tabas and Zhang (2022). The main strategy for tackling C-MARL problems found in the literature is the Lagrangian or primal-dual approach Lu et al. (2021); Li et al. (2020); Lee et al. (2018); Parnika et al. (2021). In this paper, we seek to expose and to mitigate some of the flaws in this class of algorithms.

In the primal-dual approach to C-MARL, each agent receives a reward signal that is augmented with a penalty term designed to incentivize constraint satisfaction. The magnitude of the penalty term is tuned to steer policies towards constraint satisfaction while keeping the penalty term from unnecessarily overshadowing the original reward. Although this approach has been shown to converge to a safe joint policy under certain assumptions and a specific notion of safety Lu et al. (2021), the primal-dual approach changes the structure of the learning task in a way that is not well understood.

In this paper, we study two challenges encountered when using the basic primal-dual algorithm for C-MARL. First, the primal-dual algorithm only enforces *discounted sum constraints* (DSCs) derived from the original safety constraints of the system. We show that DSCs guarantee safety only in expectation, providing bounds on neither the probability nor the severity of future constraint violations. We propose simple modifications to the penalty term that enable the enforcement of meaningful probabilistic constraints, namely chance constraints Mesbah (2016) and conditional value-at-risk constraints Rockafellar and Uryasev (2000). Although several single-agent RL algorithms deal directly with risk sensitivity García and Fernández (2015); Chow et al. (2018), the multi-agent context is less well-studied, and our contribution is to provide a novel understanding of the safety guarantees provided by existing C-MARL algorithms.

The second challenge encountered in the basic primal-dual algorithm is the fact that the reward function is constantly changing as the dual variables are updated. Every time the reward function changes, the accuracy of any value estimate diminishes. We quantify this loss of accuracy, and propose a new value estimation procedure that avoids it. Our proposal builds on results in Tessler et al. (2019) showing the linearity of the value function in the dual variables. The implications of this observation have not been isolated and studied, especially in the multi-agent setting. We develop a novel class of temporal difference algorithms for value function estimation that directly exploits this observation, giving rise to a value estimate that maintains an accurate derivative with respect to the dual variables. This first-order accuracy makes the estimate robust to dual variable updates.

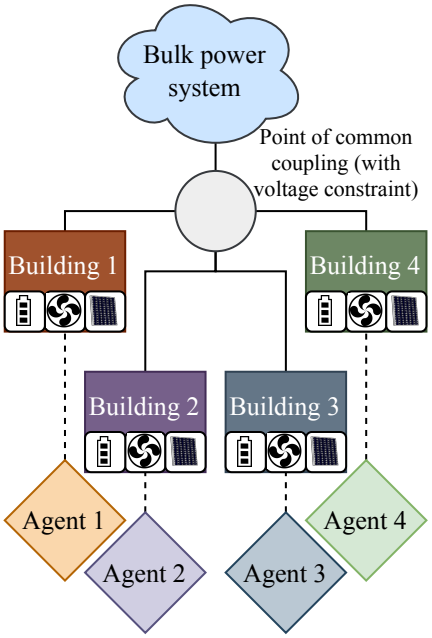

Figure 1: Building energy management with a voltage constraint at the point of common coupling.

The multi-agent RL (MARL) literature includes a wide array of problem formulations and solution techniques depending upon the extent to which states, rewards, and information are shared among agents. In this paper, we study a specific yet fairly general formulation inspired by the problem of building energy management (BEM) Biagioni et al. (2022); Molina-Solana et al. (2017), illustrated in Figure 1. The main objective of BEM is to control a building's resources to minimize the cost of energy consumption, while affording a degree of comfort and convenience to the occupants. However, when BEMs are deployed in multiple buildings, it is critical to ensure that the power network connecting them is safely operated since unco-

ordinated control of buildings can cause network voltage constraints to be violated. This mandates a level of coordination between agents in the learning stage. Thus, we adopt the commonly-studied centralized training/decentralized execution (CTDE) framework Lowe et al. (2017); Foerster et al. (2018) in which a simulator or coordinator provides global state information, constraint evaluations, and Lagrange multipliers (dual variables) to each agent during training.

The rest of the paper is organized as follows. In Section 2, we formulate the problem under consideration. In Section 3, we provide an overview of our main interpretive tool, the occupation measure Borkar and Bhatt (1996). In Section 4, we use the occupation measure to reformulate DSCs as probabilistic constraints. In Section 5, we study the value structure of the primal-dual problem and use the results to propose a new value estimation algorithm. In Section 6, we provide some simulation results affirming the theoretical observations and demonstrating the effectiveness of the proposed changes to the basic primal-dual algorithm for C-MARL.

1.1. Notation

The natural numbers and the nonnegative reals are denoted $\mathbb{N}$ and $\mathbb{R}_+$, respectively. Given a measurable set $\mathcal{S}$, the set of all possible probability densities over $\mathcal{S}$ is denoted $\Delta_{\mathcal{S}}$. For any discount factor $\gamma \in (0,1)$ and any sequence $\{y_t\}_{t=0}^T$, the discounted sum operator is $\prod_{t=0}^T[y_t \mid \gamma] = (1-\gamma)\sum_{t=0}^T \gamma^t y_t$, and $\prod_{t=0}^\infty[y_t \mid \gamma] = \lim_{T\to\infty} \prod_{t=0}^T[y_t \mid \gamma]$ if the limit exists. We will often drop the second argument γ for brevity. The positive component operator is $[y]_+ = \max\{y, 0\}$ and the logical indicator function $I[\cdot]$ maps $\{\text{True}, \text{False}\}$ to $\{1, 0\}$.

2. Problem formulation

2.1. Constrained MARL

We consider a noncooperative setting in which n agents pursue individual objectives, while being subjected to global constraints (e.g. a limited shared resource constraint). We assume there is no real-time communication and each agent's action is based only on their local observations. However, policy updates can use global information under the centralized training/decentralized execution (CTDE) framework Lowe et al. (2017); Foerster et al. (2018). In this paper, we consider the case of continuous state and action spaces.

The setting is described by the tuple $(\{\mathcal{X}_i\}_{i\in\mathcal{N}}, \{\mathcal{U}_i\}_{i\in\mathcal{N}}, \{R_i\}_{i\in\mathcal{N}}, f, C, p_0, \gamma)$ where $\mathcal{N}$ is the index set of agents, $\mathcal{X}_i$ and $\mathcal{U}_i$ are the state and action spaces of agent i, and $R_i : \mathcal{X}_i \times \mathcal{U}_i \to \mathbb{R}$ is the reward function of agent i. We assume the sets $\mathcal{X}_i$ and $\mathcal{U}_i$ are compact for all i. Let $\mathcal{X} = \prod_{i\in\mathcal{N}} \mathcal{X}_i$ and $\mathcal{U} = \prod_{i\in\mathcal{N}} \mathcal{U}_i$ be the joint state and action spaces of the system, respectively. Then $f : \mathcal{X}\times\mathcal{U} \to \Delta_{\mathcal{X}}$ describes the state transition probabilities, i.e., $f(\cdot \mid x, u) \in \Delta_{\mathcal{X}}$ for any $x \in \mathcal{X}$ and $u \in \mathcal{U}$. The function $C : \mathcal{X} \to \mathbb{R}^m$ is used to describe a set of safe states, $\mathcal{S} = \{x \in \mathcal{X} \mid C(x) \leq 0\}$.

Let $p_0 \in \Delta_{\mathcal{X}}$ denote the initial state probability density and $\gamma \in (0,1)$ be a discount factor. At time t, the state, action, and reward of agent i are x_t^i, u_t^i, and r_t^i, respectively, and constraint j evaluates to c_t^j. Using a quantity without a superscript to represent a stacked vector ranging over all $i \in \mathbb{N}$ or all $j \in \{1, \dots, m\}$, a system trajectory is denoted $\tau = \{(x_t, u_t, r_t, c_t)\}_{t=0}^\infty$.

In the noncooperative C-MARL framework, each agent seeks to learn a policy $\pi_i : \mathcal{X}_i \to \Delta_{\mathcal{U}_i}$ that maximizes the expected discounted accumulation of individual rewards. We let $\pi : \mathcal{X} \to \Delta_{\mathcal{U}}$ denote the joint policy, and $f^\pi : \mathcal{X} \to \Delta_{\mathcal{X}}$ is the state transition probability induced by a joint policy π, i.e. $f^\pi(\,\cdot \mid x) \in \Delta_{\mathcal{X}}$ for any $x \in \mathcal{X}$. The tuple (p_0, f, π) induces a state visitation probability

density at each time step, $p_t(x) = \int_{\mathcal{X}^t} p_0(x_0) \cdot \prod_{k=1}^{t} f^{\pi}(x_k \mid x_{k-1})\, dx_0 \ \ldots \ dx_{k-1}$, and we say $p_\infty(x) = \lim_{t\to\infty} p_t(x)$ for each $x \in \mathcal{X}$ if the limit exists. The collection of visitation probabilities $\{p_t\}_{t=0}^{\infty}$ gives rise to a probability density of trajectories τ, denoted $\mathcal{M} \in \Delta_{\prod_{t=0}^{\infty}(\mathcal{X}\times\mathcal{A}\times\mathbb{R}^n\times\mathbb{R}^m)}$. Thus, the objective of each agent can be stated precisely as maximizing $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} r_t^i]$.

However, the agents must settle on a joint policy that keeps the system in the safe set $\mathcal{S}$. Due to the stochastic nature of the system, satisfying this constraint for all time is too difficult and in some cases too conservative. A common relaxation procedure is to formulate an augmented reward $\tilde{r}_t^i = r_t^i - \lambda^T c_t$ where $\lambda \in \mathbb{R}_+^m$, the *Lagrange multiplier* or *dual variable*, is adjusted in order to incentivize constraint satisfaction. This leads to the primal-dual algorithm for C-MARL, discussed in the next section. The following mild assumption facilitates the analysis.

Assumption 1 *R^i, C^j, and p_t are bounded on $\mathcal{X}$ for all $i \in \mathcal{N}$, all $j \in \{1, \ldots, m\}$, and all $t \in \mathbb{N}$.*

The boundedness of R^i and C^j is a common assumption Lu et al. (2021); Tessler et al. (2019); Paternain et al. (2019) that we will use in order to exchange the order of limits, sums and integrals using the Dominated Convergence Theorem (DCT). The assumption of bounded p_t is not strictly necessary, but we use it throughout the paper in order to apply DCT with the standard Lebesgue measure.

2.2. Primal-dual algorithms

The augmented reward function leads to the following min-max optimization problem for agent i:

$$\min_{\lambda\geq 0} \max_{\pi_i} \mathbb{E}_{\tau\sim\mathcal{M}}\Big[\sum_{t=0}^{\infty}[r_t^i - \lambda^T c_t]\Big] \tag{1}$$

$$= \min_{\lambda\geq 0} \max_{\pi_i} \left(\mathbb{E}_{\tau\sim\mathcal{M}}\Big[\sum_{t=0}^{\infty}[r_t^i]\Big] - \lambda^T \mathbb{E}_{\tau\sim\mathcal{M}}\Big[\sum_{t=0}^{\infty}[c_t]\Big]\right) \tag{2}$$

where (2) uses absolute convergence (stemming from Assumption 1) to rearrange the terms of the infinite sum. Note that the minimization over λ is coupled across agents. Any fixed point of (2) will satisfy $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} c_t] \leq 0$ because if $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} c_t^j] \neq 0$ then the objective value can be reduced by increasing or decreasing λ_j, unless $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} c_t^j] < 0$ and $\lambda_j = 0$. In other words, the primal-dual method enforces a *discounted sum constraint* (DSC) derived from the safety set $\mathcal{S}$. Although DSCs are convenient, it is not obvious what satisfying a discounted sum constraint implies about safety guarantees with respect to the original constraints. We begin our investigation of DSCs by taking a closer look at a state visitation probability density known as the occupation measure.

3. Occupation measure

Definition 2 (Occupation measure Paternain et al. (2019)) *The occupation measure $\mu_\gamma \in \Delta_{\mathcal{X}}$ induced by a joint policy π is defined for any $x \in \mathcal{X}$ as $\mu_\gamma(x) = \Gamma_{t=0}^{\infty} p_t(x)$.*

Below, we provide some interpretations for the occupation measure itself before using it to ascribe meaning to discounted sum constraints. The natural first question to ask is whether μ_γ is itself a pdf. It is of course nonnegative, and the following proposition shows it integrates to unity under mild conditions.

Proposition 3 *Under Assumption 1,* $\int_{\mathcal{X}} \mu_\gamma(x)dx = 1$.

The proof for Proposition 3 is in Appendix A. What does μ_γ tell us about the behavior of a system under a given policy? It describes the probability of visiting a certain state but with more weight placed on states that are likely to be visited *earlier* in time. In fact, μ_γ describes the near-term behavior in the following sense.

Proposition 4 *Under Assumption 1, for any* $x \in \mathcal{X}$ *the following statements hold.*

1. $\lim_{\gamma \to 0^+} \mu_\gamma(x) = p_0(x)$.
2. $\lim_{\gamma \to 1^-} \mu_\gamma(x) = \lim_{t \to \infty} p_t$ *if the latter limit exists.*

The proof for Proposition 4 is in Appendix A. Figure 2 provides an illustration of the result in Proposition 4 when when p_t evolves as a normal distribution with mean α^t and constant variance, for some $\alpha \in (0, 1)$. The point at which μ_γ more closely resembles p_∞ than p_0 is exactly at $\gamma = \alpha$.

According to Proposition 4 the occupation measure describes a state distribution that lies between the initial and long-term behavior of the system. Next, we qualify this statement using a concept known as the *effective horizon*. The effective horizon of a discounted planning problem is often set to $T_1(\gamma) = \frac{1}{1-\gamma}$, which is the expected termination time if the probability of an episode terminating at any given time step is $(1-\gamma)$ Paternain et al. (2022). However, the concept of a random stopping time may not be sensible in all applications. Another way to define the effective horizon is to study the geometric distribution of weights. In this case, the effective horizon can be measured as $T_2(\gamma, \varepsilon) = \min\{K \in \mathbb{N} : \prod_{t=0}^{K-1}[1] \geq 1 - \varepsilon\}$ where $\varepsilon \in (0, 1)$ is a tolerance. Using either of the two definitions above, the occupation measure can be said to describe the behavior of the system up to the effective horizon.

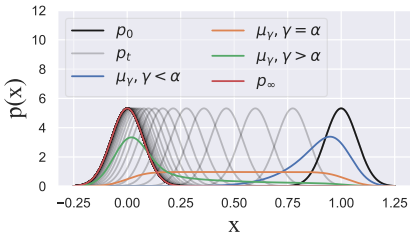

Figure 2: Example of the occupation measure for various levels of γ.

Depending on the application, either T_1 or T_2 can provide a more sensible connection between discounted and finite-horizon problems. But are these two definitions related? The next proposition answers this affirmatively, by showing that T_1 is actually a special case of T_2.

Proposition 5 $T_1(\gamma) = T_2(\gamma, \varepsilon)$ *when* ε *is set to* $\gamma^{\frac{1}{1-\gamma}} \approx \frac{1}{e}$.

The proof for Proposition 5 is in Appendix A. Proposition 5 is illustrated in Figure 3, where the effective horizon is plotted as a function of γ for three different values of ε. With an understanding of the occupation measure, we can begin to derive meaningful risk-related interpretations of DSCs. These interpretations lead directly to sensible recommendations for the design of C-MARL algorithms.

4. Discounted risk metrics

The discounted sum constraint can naturally be reinterpreted as a certain type of average constraint. In particular, Assumption 1 implies that $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} C(x_t)] = \mathbb{E}_{x\sim\mu_\gamma}[C(x)]$ Paternain et al. (2019). These near-term averages do not relate to any well-known risk metrics, and hence, do not provide practical safety guarantees. In general, information about the mean of a distribution cannot be used to infer information about its tails. Controlling the expected value of $C(x)$ leaves open the possibility of an infinite number of constraint violations of arbitrary severity. However, a simple change to the penalty function can yield information about the *probability* of incurring a constraint violation.

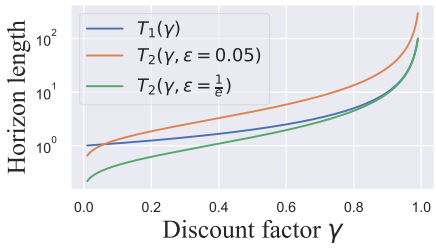

Figure 3: Effective horizon length as a function of γ.

Proposition 6 (Near-term probability of constraint violations) *Suppose that* $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} I[C^j(x_t) \geq \delta_j]] \leq \beta_j$ *for some* $\beta_j \in [0,1]$ *and* $\delta_j \in \mathbb{R}$. *Then under Assumption 1,* $\Pr\{C^j(x) \geq \delta_j \mid x \sim \mu_\gamma\} \leq \beta_j$.

Proof $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} I[C^j(x_t) \geq \delta_j]] = \mathbb{E}_{x\sim\mu_\gamma}[I[C^j(x) \geq \delta_j]] = \Pr\{C^j(x) \geq \delta_j \mid x \sim \mu_\gamma\}$. The first equality stems from Assumption 1 Paternain et al. (2019) and using DCT to exchange the order of the integral $\mathbb{E}$ and sum $\prod$, while the second follows from the definition of expectation. ■

Proposition 6 makes it easy to enforce chance constraints using primal-dual methods. When the penalty term $C^j(x)$ is replaced by the quantity $I[C^j(x) \geq \delta_j] - \beta_j$, the primal-dual algorithm enforces $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} I[C^j(x_t) \geq \delta_j]] - \beta_j \leq 0$. By Proposition 6, this guarantees that $\Pr\{C^j(x) \geq \delta_j \mid x \sim \mu_\gamma\} \leq \beta_j$. Since the probability of constraint violations is defined with x varying over μ_γ, we call the resulting guarantee a *near-term* or *discounted chance constraint.* This can be repeated for each $j \in \{1, \ldots, m\}$, providing a set of bounds on the probability of violating *each* constraint by more than its tolerance δ_j. On the other hand, we can control the probability of violating *any* constraint as follows. Define the statement $C(x) \geq \delta$ to be true if $C^j(x) \geq \delta_j \; \forall \; j \in \{1, \ldots, m\}$, and false otherwise. Then, applying Proposition 6 to the test condition $C(x) \geq \delta$ will result in a bound on $\Pr\{C(x) \geq \delta \mid x \sim \mu_\gamma\}$.

While chance constraints enable one to control the *probability* of extreme events in the near future, conditional value-at-risk (CVaR) constraints Rockafellar and Uryasev (2000) afford control over the *severity* of such events.

Definition 7 (Rockafellar and Uryasev (2000)) *Given a risk level* $\beta \in (0,1)$, *a cost* $h : \mathcal{X} \to \mathbb{R}$, *and a probability density* μ *on* $\mathcal{X}$, *the value at risk (VaR) and conditional value-at-risk (CVaR) are defined as*

$$\mathrm{VaR}(\beta, h, \mu) = \min\{\alpha \in \mathbb{R} : \Pr\{h(x) \leq \alpha \mid x \sim \mu\} \geq \beta\},$$
$$\mathrm{CVaR}(\beta, h, \mu) = \frac{1}{1-\beta} \int_{h(x)\geq \mathrm{VaR}(\beta,h,\mu)} h(x)\mu(x)dx.$$

In other words, $\mathrm{VaR}(\beta, h, \mu)$ is the least upper bound on h that can be satisfied with probability β. On the other hand, the conditional value-at-risk (CVaR) describes the expected value in the

VaR-tail of the distribution of h, thus characterizing the expected severity of extreme events. Such "extreme events" can be defined precisely as the $(1-\beta)$ fraction of events x with the worst outcomes as ranked by the cost incurred, $h(x)$. The VaR and CVaR for $h(x) = x,\ x \sim \mathcal{N}(0,1)$ are illustrated in Figure 4, where the shaded region has an area of $(1-\beta)$. For the rest of the paper, we assume that the cdf of $h(x)$ is continuous when $x \sim \mu$. For cases in which this assumption does not hold, we refer the reader to Rockafellar and Uryasev (2002).

Proposition 8 (Near-term CVaR) *For any $\alpha_j \geq 0$, suppose that $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty}[[C^j(x_t) - \alpha_j]_+]] \leq \eta_j$. Then* $\text{CVaR}(\beta, C^j, \mu_\gamma) \leq \alpha_j + (1-\beta)^{-1}\eta_j$.

Proof Under Assumption 1, the identity $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty}[C^j(x_t) - \alpha_j]_+] = \mathbb{E}_{x\sim\mu_\gamma}[C^j(x) - \alpha_j]_+$ holds Paternain et al. (2019). Next, we use the fact that the CVaR is the minimum value of the convex function in α_j given by $F(\alpha_j \mid \beta, C^j, \mu_\gamma) := \alpha_j + \frac{1}{1-\beta}\mathbb{E}_{x\sim\mu_\gamma}[[C^j(x) - \alpha_j]_+]$ Rockafellar and Uryasev (2000), thus F provides an upper bound on CVaR. Some rearranging leads to the result. ■

Similar to the chance constrained case, Proposition 8 makes it easy to enforce CVaR constraints in the primal-dual algorithm. Here, the penalty term used is $[C^j(x) - \alpha_j]_+ - \eta_j$. Using this penalty, the primal-dual algorithm enforces $\mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty}[[C^j(x_t) - \alpha_j]_+]] - \eta_j \leq 0$ which by Proposition 8 implies $\text{CVaR}(\beta, C^j, \mu_\gamma) \leq \alpha_j + (1-\beta)^{-1}\eta_j$. By repeating for each $j \in \{1, \ldots, m\}$, we can bound the expected severity of constraint violations for each of the m constraints. Since the CVaR constraint is defined with x varying over μ_γ, the resulting guarantee is called a *near-term* or *discounted CVaR constraint.*

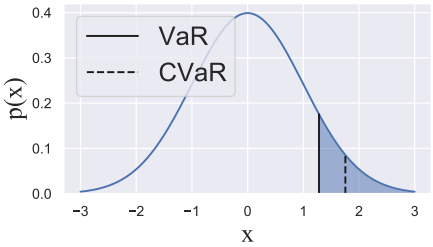

Figure 4: Example of VaR and CVaR at risk level $\beta = 0.9$.

In order to obtain a tight bound on the CVaR, α_j must be set to $\text{VaR}(\beta, C^j, \mu_\gamma)$ which minimizes the function F introduced in the proof of Proposition 8 Rockafellar and Uryasev (2000). Unfortunately, the VaR is not known ahead of time. Chow et al. (2018) includes α_j as an optimization variable in the learning procedure, but extending their technique to the multiagent setting is not straightforward. Our approach is to include it as a tunable hyperparameter. Simulation results in Section 6 show that it is easy to choose α_j to give a nearly tight bound.

5. Primal-dual value functions

In this section, we investigate challenges with value estimation in the primal-dual regime. The fact that the reward to each agent is constantly changing (due to dual variable updates) makes it difficult to estimate state values accurately. In order to quantify this decrease in accuracy, we introduce the value functions induced by the joint policy π, $\{V_\pi^i : \mathcal{X}\times\mathbb{R} \to \mathbb{R}\}_{i\in\mathcal{N}}$, $\{V_{R,\pi}^i : \mathcal{X} \to \mathbb{R}\}_{i\in\mathcal{N}}$, $V_{C,\pi} : \mathcal{X} \to \mathbb{R}^m$ where

$$V_\pi^i(x,\lambda) = \mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} r_t^i - \lambda^T c_t \mid x_0 = x], \tag{3}$$

$$V_{R,\pi}^i(x) = \mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} r_t^i \mid x_0 = x],\ V_{C,\pi}(x) = \mathbb{E}_{\tau\sim\mathcal{M}}[\prod_{t=0}^{\infty} \lambda^T c_t \mid x_0 = x]. \tag{4}$$

Note that c_t could be modified as indicated in Section 4 and the results below would hold for the modified penalty function.

Obviously, it is impossible to learn an accurate value function when λ is unknown and changing. However, simply making λ available to a value function approximator does not guarantee good generalization beyond previously seen values of λ. Having a good estimate of the *derivative* of the value function with respect to λ will ensure accuracy under small perturbations to the dual variables. Fortunately, this derivative is easy to obtain. Under Assumption 1 we can write $V^i_\pi(x,\lambda) = V^i_{R,\pi}(x) - \lambda^T V_{C,\pi}(x)$ Tessler et al. (2019). By learning $V^i_{R,\pi}$ and $V_{C,\pi}$ as separate functions and then combining them using the true value of λ, we can construct a value estimate whose derivative with respect to the dual variables is as accurate as $V_{C,\pi}$ itself. This estimate will be more robust to small changes in λ. We will refer to this type of value estimate as a *structured value function* or a *structured critic*.

Proposition 9 *Let* $\bar{c} = \mathbb{E}_{x\sim\mu_\gamma}[C(x)]$ *and* $\Sigma^2_C = \mathbb{E}_{x\sim\mu_\gamma}[(\bar{c} - C(x))(\bar{c} - C(x))^T]$. *Suppose* λ *is randomly varying with mean* $\bar{\lambda}$ *and variance* Σ^2_λ. *Using a structured value function approximator can reduce the mean square temporal difference error by up to* $\mathrm{Tr}[\Sigma^2_\lambda \cdot (\Sigma^2_C + \bar{c}\bar{c}^T)]$.

The proof of Proposition 9 is in Appendix A. Figure 5 illustrates Proposition 9 in a simple value estimation task with quadratic rewards, linear dynamics and policies, linear state constraints, and randomly varying λ. The *generic critic* (GC) is a value function modeled as a quadratic function of the state only. The *input-augmented critic* (IAC) is a value function modeled as an unknown quadratic function of the state and dual variables, while the *structured critic* (SC) is modeled using $\hat{V}^i_\pi = \hat{V}^i_{R,\pi} - \lambda^T \hat{V}_{C,\pi}$ with quadratic $\hat{V}^i_{R,\pi}$ and linear $\hat{V}_{C,\pi}$.

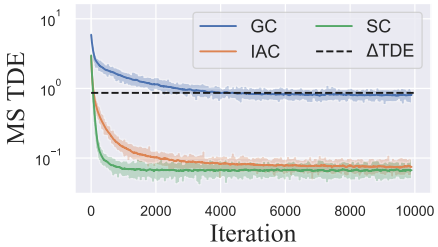

Figure 5: TD error training trajectories in a simple policy evaluation task.

The dashed line in Figure 5 is at the value $\mathrm{Tr}[\Sigma^2_\lambda \cdot (\Sigma^2_C + \bar{c}\bar{c}^T)]$ predicted in Proposition 9. In this simple value estimation task, high accuracy can be achieved when conditioning on the randomly varying λ. However, having an accurate estimate of $\nabla_\lambda V^i_\pi$ by using a structured critic is also shown to help. Although the assumptions of Proposition 9 do not hold in general, the results in Section 6 show that using the structured value function still yields improved results.

The loss function for value function approximation is therefore given by

$$TDE(x,x') = [R^i(x^i) + \gamma V^i_{R,\pi}((x^i)') - V^i_{R,\pi}(x^i)]^2 + \sum_{j=1}^{m}[C^j(x) + \gamma V_{C,\pi}(x') - V_{C,\pi}(x)]^2 \quad (5)$$

where $x \in \mathcal{X}$ and $x' \sim f^\pi(x)$. Equation (5) is simply a sum of squared temporal difference errors over the set of $m+1$ value functions.

6. Simulations

In our simulations, we sought to demonstrate the effectiveness of the changes proposed in Sections 4 and 5. We tested our findings in a modified Multiagent Particle Environment Lowe et al. (2017) with

two agents pursuing individual objectives subject to a constraint on the joint state. The objective of each agent i is to steer its state $x^i \in \mathbb{R}^2$ towards a target $x^{i*} \in \mathbb{R}^2$, while making sure that the agent ensemble satisfies the safety constraint. The reward and constraint functions are given by

$$R^1(x^1) = -\|x^1 - x^{1*}\|,\ R^2(x^2) = -2\|x^2 - x^{2*}\|,\ C(x) = \sum_{i=1}^{2} \mathbf{1}^T x^i. \tag{6}$$

The agents occupy a state space given by $\mathcal{X} = \{x \in \mathbb{R}^4 \mid \|x^1\|_\infty \leq 1, \|x^2\|_\infty \leq 1\}$. The target $x^* = \begin{bmatrix} x^{1*T} & x^{2*T} \end{bmatrix}^T$ is stationed outside of the safe region $\mathcal{S} = \{x \in \mathcal{X} \mid C(x) \leq 0\}$. Thus, the agents cannot both reach their goals while satisfying $C(x) \leq 0$. To train the agents to interact in this environment, we used a modified version of the EPyMARL codebase Papoudakis et al. (2020). All code for the algorithms is available at `github.com/dtabas/epymarl`, and the code for the environments is at `github.com/dtabas/multiagent-particle-envs`. We tested several MARL algorithms including MADDPG Lowe et al. (2017), COMA Foerster et al. (2018), and MAA2C Papoudakis et al. (2020). We chose to focus on using the MAA2C algorithm because it consistently produced the best results and because as a value function-based algorithm it provided the most straightforward route to implementing the changes proposed in Section 5.

For each risk metric described in Section 4, we tested the convergence of the agents to a safe policy with and without modifications to the penalty and value functions. Figure 6 shows the results when we make the substitution $C(x) \to I[C(x) \geq \delta] - \beta$ in the penalty function in order to enforce a chance constraint, $\Pr\{C(x) \geq \delta \mid x \sim \mu_\gamma\} \leq \beta$ with δ and β each set to 0.1. The modified penalty function serves as a superior chance constraint-enforcing signal. The structured critic architecture further improves the time it takes to converge to a policy satisfying the chance constraint.

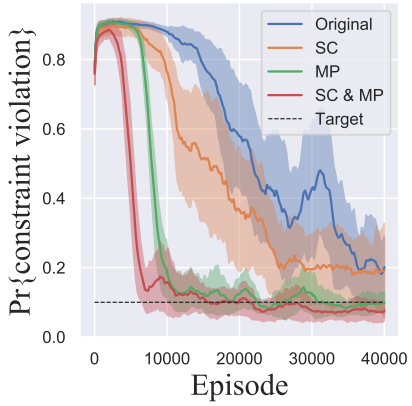

Figure 6: $Pr\{C(x) \geq 0.1 \mid x \sim \mu_\gamma\}$ measured throughout training. Key: SC = structured critic, MP = modified penalty (Prop. 6). Both modifications speed convergence to a safe policy.

Figure 7 shows the results when we make the substitution $C(x) \to [C(x) - \alpha]_+ - \eta$ in the penalty function in order to enforce the chance constraint $\text{CVaR}(\beta, C, \mu_\gamma) \leq \alpha + (1-\beta)^{-1}\eta$. Using the modified penalty drives the CVaR upper bound (drawn in dashed lines) directly to the target, and due to choice of α, this bound is nearly tight. On the other hand, using the original penalty results in an overly conservative policy that achieves low risk at the expense of rewards (right panel).

We chose α using the following heuristic, in order to make the bound on CVaR nearly tight. The "correct" value of α that would achieve a tight bound is $\text{VaR}(\beta, C, \mu_\gamma)$. Moreover, the upper bound that we used is convex and continuously differentiable in α Rockafellar and Uryasev (2000). Therefore, small errors in α will lead to small errors in the upper bound on CVaR, and any approximation of VaR will suffice. We obtained an approximation simply by running the simulation once with α set to zero and computing $\text{VaR}(\beta, C, \mu_\gamma)$ over some test trajectories. If necessary, the process could be

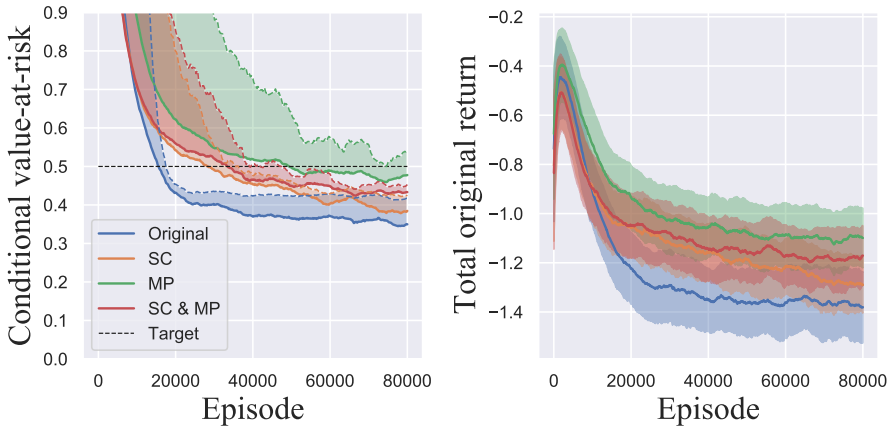

Figure 7: CVaR$(\beta = 0.9, C, \mu_\gamma)$ measured throughout training. Key: SC = structured critic, MP = modified penalty (Prop. 8). The dashed lines represent the CVaR upper bound used in Prop. 8. The unmodified penalty function enforces safety at the expense of rewards, whereas the modified penalty function affords precise control over the tail of the distribution of $C(x)$. The panel on the right shows progress towards the original objective through the total original returns, $\sum_{i=1}^{2} \prod_{t=0}^{T} r_t^i$, without any penalty terms.

repeated additional times. Alternatively, α could be tuned adaptively by computing VaR online, but the stability of this procedure needs further investigation.

7. Conclusion

In this paper, we studied the effect of primal-dual algorithms on the structure of constrained multiagent reinforcement learning problems. First, we used the occupation measure to study the effect of the penalty term on safety. We showed that the constraint evaluation itself enforces safety only in expectation. However, by making simple modifications to the penalty term we were able to enforce meaningful probabilistic safety guarantees, namely chance constraints and CVaR constraints. These risk metrics are defined over the occupation measure, leading to notions of safety in the near term. We used the concept of *effective horizon* to make the concept of "near term" concrete. Next, we studied the effect of the penalty term on the value function. We showed that when the dual variable and constraint evaluation signals are available, it is easy to model the relationship between the penalty term and the value function. By exploiting this structure, the accuracy of the value function can be improved. We demonstrated the usefulness of both of these insights in a constrained multiagent particle environment, showing that convergence to a low-risk policy is accelerated.

After studying the effect of primal-dual methods on the constraints and value functions, the next step is to study their effect on game outcomes. In general, some agents may pay a higher price than others for modifying their policies to satisfy the system-wide constraints. This phenomenon, and possible remediation, will be the focus of future work.

Acknowledgments

References

David Biagioni, Xiangyu Zhang, Dylan Wald, Deepthi Vaidhynathan, Rohit Chintala, Jennifer King, and Ahmed S. Zamzam. PowerGridworld: A Framework for Multi-Agent Reinforcement Learning in Power Systems. *e-Energy 2022 - Proceedings of the 2022 13th ACM International Conference on Future Energy Systems*, pages 565–570, 2022. doi: 10.1145/3538637.3539616.

Vivek S. Borkar and Abhay G. Bhatt. Occupation Measures for Controlled Markov Processes: Characterization and Optimality. *The Annals of Probability*, 24(3):1531–1562, 1996. ISSN 0091-1798. URL http://projecteuclid.org/euclid.aop/1065725192.

Steven Chen, Kelsey Saulnier, Nikolay Atanasov, Daniel D. Lee, Vijay Kumar, George J. Pappas, and Manfred Morari. Approximating Explicit Model Predictive Control Using Constrained Neural Networks. *Proceedings of the American Control Conference*, 2018-June:1520–1527, 2018. ISSN 07431619. doi: 10.23919/ACC.2018.8431275.

Yu Jia Chen, Deng Kai Chang, and Cheng Zhang. Autonomous Tracking Using a Swarm of UAVs: A Constrained Multi-Agent Reinforcement Learning Approach. *IEEE Transactions on Vehicular Technology*, 69(11):13702–13717, 2020. ISSN 19399359. doi: 10.1109/TVT.2020.3023733.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18: 1–51, 2018. ISSN 15337928.

Wenqi Cui, Jiayi Li, and Baosen Zhang. Decentralized safe reinforcement learning for inverter-based voltage control. *Electric Power Systems Research*, 211(October 2021):108609, 2022. ISSN 03787796. doi: 10.1016/j.epsr.2022.108609. URL https://doi.org/10.1016/j.epsr.2022.108609.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *32nd AAAI Conference on Artificial Intelligence*, pages 2974–2982, 2018.

Javier García and Fernando Fernández. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

Donghwan Lee, Hyungjin Yoon, and Naira Hovakimyan. Primal-Dual Algorithm for Distributed Reinforcement Learning: Distributed GTD. *Proceedings of the IEEE Conference on Decision and Control*, 2018-Decem(Cdc):1967–1972, 2018. ISSN 25762370. doi: 10.1109/CDC.2018.8619839.

Wenhao Li, Bo Jin, Xiangfeng Wang, Junchi Yan, and Hongyuan Zha. F2A2: Flexible Fully-decentralized Approximate Actor-critic for Cooperative Multi-agent Reinforcement Learning. *arXiv: 2004.11145*, pages 1–42, 2020. URL http://arxiv.org/abs/2004.11145.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *31st Conference on Neural Information Processing Systems*, 6 2017. URL http://arxiv.org/abs/1706.02275.

Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized Policy Gradient Descent Ascent for Safe Multi-Agent Reinforcement Learning. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 10A:8767–8775, 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i10.17062.

Haitong Ma, Jianyu Chen, Shengbo Eben, Ziyu Lin, Yang Guan, Yangang Ren, and Sifa Zheng. Model-based Constrained Reinforcement Learning using Generalized Control Barrier Function. *IEEE International Conference on Intelligent Robots and Systems*, pages 4552–4559, 2021. ISSN 21530866. doi: 10.1109/IROS51168.2021.9636468.

Ali Mesbah. Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine*, 36(6):30–44, 2016. ISSN 1066033X. doi: 10.1109/MCS.2016.2602087.

Miguel Molina-Solana, María Ros, M. Dolores Ruiz, Juan Gómez-Romero, and M. J. Martin-Bautista. Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*, 70(August 2015):598–609, 2017. ISSN 18790690. doi: 10.1016/j.rser.2016.11.132. URL http://dx.doi.org/10.1016/j.rser.2016.11.132.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In *35th Conference on Neural Information Processing Systems*, 2020. URL http://arxiv.org/abs/2006.07869.

P. Parnika, Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, and Shalabh Bhatnagar. Attention actor-critic algorithm for multi-agent constrained co-operative reinforcement learning. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 3(Aamas 2021):1604–1606, 2021. ISSN 15582914.

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

Santiago Paternain, Miguel Calvo-Fullana, Luiz F.O. Chamon, and Alejandro Ribeiro. Safe Policies for Reinforcement Learning via Primal-Dual Methods. *IEEE Transactions on Automatic Control*, 9286(c):1–16, 2022. ISSN 15582523. doi: 10.1109/TAC.2022.3152724.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–42, 2000. doi: 10.2307/1165345.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26(7):1443–1471, 2002. ISSN 03784266. doi: 10.1016/S0378-4266(02)00271-6.

Daniel Tabas and Baosen Zhang. Computationally Efficient Safe Reinforcement Learning for Power Systems. In *Proceedings of the American Control Conference*, pages 3303–3310. American Automatic Control Council, 2022. ISBN 9781665451963. doi: 10.23919/ACC53348.2022.9867652.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *7th International Conference on Learning Representations, ICLR 2019*, 5 2019. URL http://arxiv.org/abs/1805.11074.

Wei Zhou, Dong Chen, Jun Yan, Zhaojian Li, Huilin Yin, and Wanchen Ge. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. *Autonomous Intelligent Systems*, 2(1), 2022. doi: 10.1007/s43684-022-00023-5. URL http://dx.doi.org/10.1007/s43684-022-00023-5.

Appendix A. Theoretical results

A.1. Proof of Proposition 3

Applying the definition of μ_γ, we have $\int_{\mathcal{X}} \mu_\gamma(x)dx = \int_{\mathcal{X}} \prod_{t=0}^{\infty} p_t(x)dx$. Using the Dominated Convergence Theorem, we can exchange the order of the sum and integral. Each individual p_t integrates to 1. The geometric sum property ensures that the resulting expression evaluates to 1.

A.2. Proof of Proposition 4

1. By definition, we have $\lim_{\gamma\to 0^+} \mu_\gamma(x) = \lim_{\gamma\to 0^+} \prod_{t=0}^{\infty} p_t(x)$. Using Tannery's theorem, we can exchange the order of the limit and the infinite sum. The zeroth term in the sum evaluates to $p_0(x)$ and all other terms evaluate to 0.

2. Assume $\lim_{t\to\infty} p_t$ exists, and denote it p_∞. Using the triangle inequality, we have

$$|\mu_\gamma(x) - p_\infty(x)| \leq \sum_{t=0}^{\infty} |p_t^\theta(x) - p_\infty^\theta(x)| \tag{7}$$

$$= \sum_{t=0}^{N} |p_t^\theta(x) - p_\infty^\theta(x)| + \sum_{t=N+1}^{\infty} |p_t^\theta(x) - p_\infty^\theta(x)| \tag{8}$$

   for some $N \in \mathbb{N}$. Since $p_t(x) \to p_\infty(x)$, we can choose N large enough to make the second term in (8) arbitrarily small. Then, using boundedness of p_t for all t, we can take $\gamma \to 1^-$ to make the first term arbitrarily small.

A.3. Proof of Proposition 5

By the geometric sum property, we have $T_{wd}(\gamma,\varepsilon) = \min\{K \in \mathbb{N} : \prod_{t=0}^{K-1}[1] \geq 1-\varepsilon\} = \min\{K \in \mathbb{N} : 1-\gamma^K \geq 1-\varepsilon\} = \min\{K \in \mathbb{N} : K \geq \frac{\log\varepsilon}{\log\gamma}\} = \left\lceil \frac{\log\varepsilon}{\log\gamma} \right\rceil$. The termination time follows a geometric distribution with parameter $(1-\gamma)$, and thus has expected value $\frac{1}{1-\gamma}$. Setting $T_{wd}(\gamma,\varepsilon) = T_{tt}(\gamma)$ and solving for ε (ignoring the integer constraint) yields $\varepsilon = \gamma^{\frac{1}{1-\gamma}}$. Finally, taking $\lim_{\gamma\to 1} \gamma^{\frac{1}{1-\gamma}}$ yields $\frac{1}{e}$.

A.4. Proof of Proposition 9

Let $x \sim \mu_\gamma$, $x' \sim f^\pi(x)$, $\bar{c} = \mathbb{E}[C(x)]$, and $\Sigma_C^2 = \mathbb{E}[(\bar{c} - C(x))(\bar{c} - C(x))^T]$. Suppose λ is randomly distributed with mean $\bar{\lambda}$ and variance Σ_λ^2. For any value function approximator $\hat{V}_\pi^i$, assume λ and $\hat{V}_\pi^i$ are independent. Let $\eta = \begin{bmatrix}1 & \lambda^T\end{bmatrix}^T$, $d = \begin{bmatrix}R^i(x) & C(x)^T\end{bmatrix}^T$, $\hat{V}_\pi^i : \mathcal{X} \to \mathbb{R}$, $\hat{V}_{R,\pi}^i : \mathcal{X} \to \mathbb{R}$, and $\hat{V}_{C,\pi} : \mathcal{X} \to \mathbb{R}^m$. Let $\mathcal{D}$ be a dataset of trajectories sampled from $\mathcal{M}$ that is used to train $\hat{V}_\pi^i$, $\hat{V}_{R,\pi}^i$, and $\hat{V}_{C,\pi}$. The mean square temporal difference error achieved by using a generic value function is

$$MSTDE_1 = \mathbb{E}_{x,x',\lambda,\mathcal{D}}[(\eta^T d + \gamma\hat{V}_\pi^i(x') - \hat{V}_\pi^i(x))^2] \tag{9}$$

while the error achieved using the structured value function is

$$MSTDE_2 = \mathbb{E}_{x,x',\mathcal{D}}[(\eta^T d + \gamma[\hat{V}_{R,\pi}^i(x') - \lambda^T\hat{V}_{C,\pi}(x')] - [\hat{V}_{R,\pi}^i(x) - \lambda^T\hat{V}_{C,\pi}(x)])^2]. \tag{10}$$

Note that in (10) we do not take the expectation over λ since the dual variables are available to this function approximator.

Begin with the states and dual variables fixed at $(\bar{x}, \bar{x}', \bar{\lambda})$. Let $\hat{g}(\bar{x}, \bar{x}') = \begin{bmatrix} \hat{V}^i_{R,\pi}(\bar{x}) & \hat{V}_{C,\pi}(\bar{x})^T \end{bmatrix}^T - \gamma \begin{bmatrix} \hat{V}^i_{R,\pi}(\bar{x}') & \hat{V}_{C,\pi}(\bar{x}')^T \end{bmatrix}^T$ and $\hat{h}(\bar{x}, \bar{x}') = \hat{V}^i_\pi(\bar{x}) - \gamma \hat{V}^i_\pi(\bar{x}')$. Then, suppressing the arguments $(\bar{x}, \bar{x}')$ and setting $\bar{\eta} = \begin{bmatrix} 1 & \bar{\lambda}^T \end{bmatrix}^T$, we can write the squared temporal difference error at $(\bar{x}, \bar{x}', \bar{\lambda})$ as

$$STDE_1(\bar{\eta}) = \mathbb{E}_{\mathcal{D}}[(\bar{\eta}^T d - \hat{h})^2], \tag{11}$$

$$STDE_2(\bar{\eta}) = \mathbb{E}_{\mathcal{D}}[(\bar{\eta}^T d - \bar{\eta}^T \hat{g})^2]. \tag{12}$$

The loss function used to train $\hat{V}^i_{R,\pi}$ and $\hat{V}_{C,\pi}$ is

$$\mathbb{E}_{\mathcal{D}}[\|d - \hat{g}\|^2]. \tag{13}$$

Since d is a deterministic function of x, (13) can be decomposed into bias and variance terms:

$$\mathbb{E}_{\mathcal{D}}[\|d - \hat{g}\|^2] = \mathbb{E}_{\mathcal{D}}[\sum_{k=0}^{m} (d_k - \hat{g}_k)^2] \tag{14}$$

$$= \sum_{k=0}^{m} \mathbb{E}_{\mathcal{D}}[(d_k - \hat{g}_k)^2] \tag{15}$$

$$= \sum_{k=0}^{m} [(d_k - \mathbb{E}_{\mathcal{D}} \hat{g}_k)^2 + \mathbb{E}_{\mathcal{D}}[(\hat{g} - \mathbb{E}_{\mathcal{D}} \hat{g})^2]] \tag{16}$$

$$:= \sum_{k=0}^{m} [b_k^2 + \sigma_k^2] \tag{17}$$

$$:= \text{Tr}[bb^T + \Sigma^2] \tag{18}$$

where $k = 0$ corresponds to the reward signal and $k = 1, \ldots, m$ corresponds to the cost signals.

Following a similar line of reasoning, we can use (18) to rewrite (12) as

$$STDE_2(\bar{\eta}) = \text{Tr}[(bb^T + \Sigma^2)(\bar{\eta}\bar{\eta}^T)]. \tag{19}$$

For the sake of argument, we assume that $\hat{g}$ and $\hat{h}$ achieve the same performance at (x, x', λ), that is,

$$STDE_1(\bar{\eta}) = STDE_2(\bar{\eta}) = \text{Tr}[(bb^T + \Sigma^2)(\bar{\eta}\bar{\eta}^T)] \tag{20}$$

where $\text{Tr}[(bb^T)(\bar{\eta}\bar{\eta}^T)]$ and $\text{Tr}[\Sigma^2 \bar{\eta}\bar{\eta}^T]$ reflect the bias squared and variance terms, respectively. How do $STDE_1$ and $STDE_2$ change when λ is allowed to vary? Using the generic estimator, the noise in λ will introduce some amount of *irreducible error* into $STDE_1$. On the other hand, using $\lambda = \bar{\lambda} + \Delta\lambda$ in our proposed estimator will change the bias and variance terms in $STDE_2$ while the irreducible error remains at zero (since there is no uncertainty when $\Delta\lambda$ is known). Setting $\Delta\eta = \begin{bmatrix} 0 & \Delta\lambda^T \end{bmatrix}^T$, the temporal difference errors at $(\bar{x}, \bar{x}', \bar{\lambda} + \Delta\lambda)$ are

$$STDE_1(\bar{\eta} + \Delta\eta) = \text{Tr}[(bb^T + \Sigma^2)(\bar{\eta}\bar{\eta}^T)] + (\Delta\eta^T d)^2, \tag{21}$$

$$STDE_2(\bar{\eta} + \Delta\eta) = \text{Tr}[(bb^T + \Sigma^2)((\bar{\eta} + \Delta\eta)(\bar{\eta} + \Delta\eta)^T)]. \tag{22}$$

Taking the expectation over $\Delta\lambda$ which has a mean of zero and a variance of Σ_λ^2, and setting $\Sigma_\eta^2 = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_\lambda^2 \end{bmatrix}$, yields

$$\mathbb{E}_{\Delta\lambda}[STDE_1(\bar{\eta}+\Delta\eta) - STDE_2(\bar{\eta}+\Delta\eta)] = \text{Tr}[\Sigma_\eta^2(dd^T - bb^T - \Sigma^2)] \tag{23}$$

$$= \text{Tr}[\Sigma_\lambda^2(cc^T - \tilde{b}\tilde{b}^T - \tilde{\Sigma}^2)] \tag{24}$$

where $\tilde{b} = (c - \mathbb{E}_\mathcal{D}\hat{g}_C)$, $\tilde{\Sigma}^2 = \mathbb{E}_\mathcal{D}[(\hat{g}_C - \mathbb{E}_\mathcal{D}\hat{g}_C)^2]$, and $\hat{g}_C = \hat{V}_{C,\pi}(x) - \gamma\hat{V}_{C,\pi}(x')$. Note that $\mathbb{E}_\mathcal{D}[\|c - \hat{g}_C\|^2] = \text{Tr}[\tilde{b}\tilde{b}^T + \tilde{\Sigma}^2]$. Taking $\tilde{b}, \tilde{\Sigma}^2 \to 0$ as the accuracy of $\hat{g}_C$ improves, (24) can be estimated as

$$\text{Tr}[\Sigma_\lambda^2 cc^T]. \tag{25}$$

Taking the expectation over $c \sim C(x), x \sim \mu_\gamma$ yields the final result.

Appendix B. Simulation details